Inaugural-Dissertation

zur

Erlangung der Doktorwürde

der

Naturwissenschaftlich-Mathematischen Gesamtfakultät

der

Ruprecht-Karls-Universität Heidelberg

vorgelegt von

Diplom-Physiker Christian Hoffmann

aus Schwetzingen in Baden-Württemberg

Tag der mündlichen Prüfung

. . . .

Numerical Aspects of Uncertainty in the Design of Optimal Experiments for Model Discrimination

Gutachter

Professor Dr. Dr. h. c. mult. Hans Georg Bock

Zusammenfassung

Diese Arbeit betrachtet robuste Strategien der optimalen Versuchsplanung zur Diskriminierung zwischen mehreren nichtlinearen Regressionsmodellen. Für solche Strategien entwickelt sie neue Theorie, effiziente Algorithmen und Implementierungen. Darüber hinaus schlägt sie neue Techniken und Algorithmen zum Vergleich und zur Bewertung der praktischen Leistungsfähigkeit solcher Strategien vor, und setzt diese in mehreren umfangreichen Fallstudien ein. Die gewonnenen Ergebnisse zeigen den Erfolg der neuen Strategien.

Die Beiträge der Arbeit sind in verschiedenen Gebieten Fortschritte gegenüber existierenden Theorien und Methoden:

- Die Arbeit schlägt neuartige "modell-robuste" datenbasierte Approximationsformeln vor für die Kovarianzen von Maximum-Likelihood-Schätzern und von Bayesschen A-Posteriori Verteilungen von Parametern. Diese Formeln sind geeignet um Parameterunsicherheit zu quantifizieren, selbst wenn das zugrundeliegende Modell sowohl nichtlinear als auch systematisch falsch ist.

- Im Rahmen der Arbeit werden statistische Maße und angepasste effiziente Algorithmen entwickelt, mit denen Approximationen für die Kovarianz von Maximum-Likelihood-Schätzern für Parameter auf Basis von Simulationsstudien bewertet werden können. Die Algorithmen sind in vollständig parallelisierter Form im Programmpaket DoeSim implementiert.

- In einer umfangreichen numerischen Fallstudie wird mit Hilfe von DoeSim die modell-robuste Formel für die Kovarianz von Maximum-Likelihood-Schätzern für Parameter mit ihrem "klassischen" Gegenstück verglichen. Die Ergebnisse zeigen die klare Überlegenheit der modell-robusten Formel.

- Die Arbeit schlägt zwei neuartige sequenzielle Designkriterien zur Modelldiskriminierung vor. Diese berücksichtigen Parameterunsicherheit mit Hilfe der neuen modell-robusten Formel für die Kovarianz der A-Posteriori Verteilung der Parameter. Es wird gezeigt, dass beide Kriterien eine Verbesserung gegenüber einer häufig angewendeten Approximation des Box-Hill-Hunter-Kriteriums darstellen, da sie dessen Überbewertung der erwarteten Informationsmenge auf einem Experiment vermeiden.

- Die Arbeit stellt klar, dass das verbreitete Gauss-Newton-Verfahren im Allgemeinen ungeeignet ist, um im Kontext von Modelldiskriminierung Least-Squares-Schätzer für Parameter zu berechnen. Darüber hinaus zeigt sie, dass eine grosse Klasse von Optimierungsproblemen der Optimalen Versuchsplanung zur Modelldiskriminierung intrinsisch nicht-konvex ist, und dass dies sogar unter stark vereinfachenden Annahmen gilt. Nicht-konvexe Probleme sind $\mathcal{NP}$-schwer und damit besonders schwer mit numerischen Methoden zu lösen.

- Die Arbeit entwickelt ein Paket zur quantitativen Bewertung und zum Vergleich sequentieller Versuchsplanungsstrategien zur Modelldiskriminierung. Das Paket umfasst neue statistische Maße für deren praktische Effizienz und problemangepasste Algorithmen für die Berechnung dieser Maße. Eine moderne, modulare und parallelisierte Implementation wird im Programmpaket DoeSim realisiert. Damit ist es möglich, ein breites Spektrum der Eigenschaften von Designstrategien zu analysieren, welche diese unter den Schwankungen von Messdaten aufweisen.

- Die praktische Leistungsfähigkeit von vier etablierten und drei neuen sequenziellen Designkriterien zur Modelldiskriminierung wird in ein einer umfangreichen Simulationsstudie untersucht. Die Studie wurde mit DoeSim durchgeführt und umfasst eine grosse Zahl von Modelldiskriminierungs-Problemen. Sie untersucht unter anderem den Einfluss von verschiedenen Größenordnungen von Messunsicherheit und von der Anzahl der rivalisierenden Modelle.

  Zentrale Ergebnisse sind, dass eine häufig angewendete Approximation des Box-Hill-Hunter-Kriteriums bei Problemen mit mehr als zwei Modellen ineffzient ist, dass alle parameter-robusten Designkriterien tatsächlich die einfache Hunter-Reiner-Strategie übertreffen, und dass die neu vorgeschlagenen Designkriterien immer unter den effizientesten zu finden sind. Sie zeigen besondere deutliche Vorteile in anspruchsvollen Modelldiskriminierung-Problemen zwischen vielen Modellen und grosser Messunsicherheit.

Abstract

This thesis investigates robust strategies of optimal experimental design for discrimination between several nonlinear regression models. It develops novel theory, efficient algorithms, and implementations of such strategies, and provides a framework for assessing and comparing their practical performance. The framework is employed to perform extensive case studies. Their results demonstrate the success of the novel strategies.

The thesis contributes advances over existing theory and techniques in various fields as follows:

- The thesis proposes novel "misspecification-robust" data-based approximation formulas for the covariances of maximum-likelihood estimators and of Bayesian posterior distributions of parameters in nonlinear incorrect models. The formulas adequately quantify parameter uncertainty even if the model is both nonlinear and systematically incorrect.

- The thesis develops a framework of novel statistical measures and tailored efficient algorithms for the simulation-based assessment of covariance approximations for maximum-likelihood estimator for parameters. Fully parallelized variants of the algorithms are implemented in the software package DoeSim.

- Using DoeSim, the misspecification-robust covariance formula for maximum-likelihood estimators (MLEs) and its "classic" alternative are compared in an extensive numerical case study. The results demonstrate the superiority of the misspecification-robust formula.

- Two novel sequential design criteria for model discrimination are proposed. They take into account parameter uncertainty with the new misspecification-robust posterior covariance formula. It is shown that both design criteria constitute an improvement over a popular approximation of the Box-Hill-Hunter-criterion. In contrast to the latter, they avoid to overestimate the expected amount of information provided by an experiment.

- The thesis clarifies that the popular Gauss-Newton method is generally *not* appropriate for finding least-squares parameter estimates in the context of

model discrimination. Furthermore, it demonstrates that a large class of optimal experimental design optimization problems for model discrimination is intrinsically non-convex even under strong simplifying assumptions. Such problems are $\mathcal{NP}$-hard and particularly difficult to solve numerically.

- A framework is developed for the quantitative assessment and comparison of sequential optimal experimental design strategies for model discrimination. It consists of new statistical measures of their practical performance and problem-adapted algorithms to compute these measures. A state-of-the-art modular and parallelized implementation is provided in the software package DoeSim. The framework permits quantitative analyses of the broad range of behavior that a design strategy shows under fluctuating data.

- The practical performance of four established and three novel sequential design criteria for model discrimination is examined in an extensive simulation study. The study is performed with DoeSim and comprises a large number of model discrimination problems. The behavior of the design criteria is examined under different magnitudes of measurement error and for different number of rival models.

  Central results from the study are that a popular approximation of the Box-Hill-Hunter-criterion is surprisingly inefficient, particularly in problems with three or more models, that all parameter-robust design criteria in fact outperform the basic Hunter-Reiner-strategy, and that the newly proposed novel design criteria are among the most efficient ones. The latter show particularly strong advantages over their alternatives when facing demanding model discrimination problems with many rival model and large measurement errors.

Acknowledgements

My deep gratitude goes to my teachers and mentors *Hans Georg Bock* and *Johannes Schlöder.* This thesis would not have been possible without their profound knowledge, and their financial and professional support. They trusted in my abilities, inspired me in many conversations and by example, and encouraged me to follow my ideas. Doing research in their group with its warm, joyful, open-minded and productive spirit was an extraordinary pleasure.

I would like to thank the Heidelberg University for providing a scholarship for my research from 2007 to 2009 and for support from the Heidelberg Graduate School for Mathematical and Computational Methods in the Sciences.

I am an extraordinary lucky person to have *Tom Kraus* on my side as loyal friend, skilled researcher, and trustful advisor. I do not try to list the countless ways how he contributed to the success of this thesis. He has my deepest gratitude. Furthermore, I wish to thank *Christian Kirches* for being such a cheerful friend and colleague. He was a constant source of fruitful ideas and council over the years, and convinced me that my findings are valuable when I was in doubt. In a critical moment, his commitment ensured that I continued my work. *Leonard Wirsching* deserves a special mention: his door was always open to me when I had the urge to talk over an idea. He patiently discussed and answered all my questions, regardless how trivial they turned out to be. I wish to thank *Andreas Potschka* for inspiring and fruitful discussions, and also for supporting me in difficult decisions. Altogether: thank you, gentlemen!

It was always a pleasure to discuss science and not-so-science with my former colleagues *Simon Lenz, Katrin Hatz,* and *Dörte Beigel* over countless pieces of cake and cups of coffee. Speaking of coffee: I was always happy that we have *really* good coffee in the group, for which I want to thank *Andreas Schmidt, Felix Lenders,* and all the other *true coffee nerds.*

I an indebted to *Johannes Schlöder, Tom Kraus,* and *Christian Kirches,* who invested a lot of their time and expertise to support me in preparing the final version of this thesis. Furthermore, I wish to thank *Thomas Kloepfer, Margret Rothfuss, Abir Al-Laham,* and *Anja Vogel* of the group administration for keeping things running smoothly in background.

I feel grateful to *Sebastian Sager,* now running his own group at the Otto-von-

Guericke-University Magdeburg, for many fruitful discussions, for his personal council in many regards, and for showing me that scientific success and joy of life are not necessarily opposites. Just after I started my research, *Moritz Diehl* at the K. U. Leuven entrusted to me the organization of the OPTEX workshop. I am grateful to him for giving me this extraordinary opportunity. Moreover, I wish to thank my former mentor and colleague *Stefan Körkel,* now at the University of Mannheim, for introducing me to the field of optimal experimental design and inspiring my further research. He is possibly the didactically most skilled teacher I ever had.

In the first years of my research, I had the opportunity to cooperate with the GVS/C Scientific Computing group of the BASF SE in Ludwigshafen. Working together with *Hergen Schultze* was deeply inspiring. Our experiences with the real-world model discrimination problem that we studied convinced me that parameter-robustness is the key to efficient optimal experimental design here, and thus laid the foundation for this thesis. I wish to thank *Alexander Badinski* and *Anna Schreieck* from BASF SE to maintain our contact even after the cooperation ended.

A constant source of power for this project were my flatmates and friends, among them *Steffen K., Katrin P., Janina H., Cédric G., Frank E., Steffi Ö.,* and *Cristina P.* I would like to thank all of them for enriching my life. I am eagerly awaiting the Dance-Your-PhD performance promised by Katrin and Janina. This thesis might possibly not exist if my caring friends Katrin and Cristina Prat-Knoll would not have asked the right questions in the right moments. I owe you one.

My parents *Therese* and *Gerald* deserve my sincerest gratitude, for their enduring support for me and my project. Their love and trust gave me the confidence and endurance I required for this thesis. It is dedicated to them.

My heartfelt gratitude goes to *Anna M.,* for being the wonderful extraordinary person she is. She constantly trusted in my success and always found a way of giving me new power to carry on. Thank you. I am looking forward for realizing all the small and big dreams we have together.

Heidelberg, 23. Juli 2016 *Christian Hoffmann*

Contents

Acronyms

BF Buzzi-Ferraris.

BFGS Broyden-Fletcher-Goldfarb-Shanno.

BHH Box-Hill-Hunter.

DAE differential-algebraic equation.

HR Hunter-Reiner.

IID independently and identically distributed.

INID independently but not identically distributed.

KL Kullback-Leibler.

KLD Kullback-Leibler distance.

KLIC Kullback-Leibler information criterion.

LD low-discrepancy.

LSQ least-squares.

MC Monte Carlo.

MCD minimum covariance determinant.

MD model discrimination.

MLE maximum-likelihood estimator.

MLE maximum-likelihood estimate.

MMLE model maximum-likelihood estimator.

MMLE model maximum-likelihood estimate.

ODE ordinary differential equation.

OED optimal experimental design.

OED/MD optimal experimental design for model discrimination.

OED/PE optimal experimental design for parameter estimation.

PDF probability density function.

PE parameter estimation.

PMF probability mass function.

PMLE parameter maximum-likelihood estimator.

PMLE parameter maximum-likelihood estimate.

QP quadratic program.

SPD symmetric positive definite.

SPSD symmetric positive semi-definite.

SQP sequential quadratic programming.

SSR sum of squared residuals.

WGSR water-gas shift reaction.

WLOG without loss of generality.

Introduction

Model-based simulation and optimization and the related numerical techniques play a central role in science, industry, and economy. They are in fact often regarded as a "third pillar" of science besides theory and experiment. Their practical success crucially depends on the quality with which the underlying models reproduce those aspects of the considered processes that are of interest.

Building sufficiently good models is often a challenging task that requires substantial experimental effort, making it a potentially time-consuming, costly, and error-prone procedure. Any field that applies model-based methods can thus greatly benefit from techniques that help to reduce this effort.

Preface

On an abstract level, many processes can be described as follows: they are manipulable trough a number of independent variables (experimental conditions) and provide a number of observables quantities (observations, outcome, data) as output. An experiment is characterized by the condition under which it is performed and the resulting outcome.

A *particular* outcome is unpredictable: replicated experiments under the same condition do not necessarily provide the same results. This unpredictability is called "experimental uncertainty," and is often attributed to the presence of measurement errors. In the limit of many replications, however, the *frequency* of the outcomes follows a well-defined distribution that is determined by the nature of the considered process. A process of this type can be represented mathematically by a collection of random variables, called a "stochastic process."

In practice, the actual distribution of the data is unknown, a lack of knowledge called "structural uncertainty." To deal with this uncertainty, one might formulate a number of parametric regression models. Each model, and each parameter of each model, specifies for all considered experimental conditions a candidate for the unknown distribution of the corresponding outcomes. We refer to such a collection as a "model family."

Parameter Estimation and Model Discrimination

Models and model families aim to approximate the process. The practical success of any model-based method is limited by the related approximation quality. It is hence natural to ask for a parameter under which a given model describes the process "best." In a model family, one might be interested in a corresponding "best" model. What might be considered as "best" depends, of course, on the intended purpose of the model.

In practice, the "best" parameters or "best" models are unknown. When experimental data is available, the following two classes of empirical (=data-based) problems arise:

Parameter estimation (PE)[1]

Given a parametric model and experimental data, identify a "best" parameter.

Model discrimination (MD)[2]

Given several parametric models and experimental data, find the "best" model.

A parameter of a given model is said to be "correct[3]," if the associated model predictions are experimentally indistinguishable from the process. A model is called correct, if a correct parameter exists for it.

A correct parameter or a correct model are "best" in any practical sense. If they exist, identifying them is the natural aim of parameter estimation and model discrimination, respectively. Albeit correctness is a fairly strong assumption, it is commonly made to simplify the statistical and mathematical problems arising in various model-based methods.

Parameter estimation is a well-examined central problem of statistical inference, since it appears in different variants in almost all empirical sciences. The statistical background of parameter estimation can be found in the book by Lehmann and Casella [170], for example. The resulting optimization problems and corresponding numerical methods are examined, for instance, by Bard [23] and Walter and Pronzato [263].

A frequently considered special case is that of least-squares estimation, for which efficient numerical methods and highly developed implementations are available. A general overview of suitable methods is given by Nocedal and

[1] Sometimes called "parameter identification."

[2] Also referred to as "model selection" and "model identification."

[3] Alternative terms are "true" or "correctly specified."

Wright [194]. Contributions for problems involving different types of differential equations come for example from Bock [35], Schlöder [223], Bock, Kostina, and Schlöder [36], Hatz [114], Kostina [149], Kühl et al. [155], and Lenz [172].

Model discrimination problems are classically approached with statistical hypothesis testing, for example discussed in detail by Lehmann and Romano [171]. Ando [7] discusses the corresponding Bayesian methods.

Whatever method is applied to these problems, it can only solve the problem approximately, because the underlying data is subject to random fluctuations. For many methods one can show, fortunately, that the approximation quality increases with the number of available experiments, assuming that certain regularity conditions are met. One can therefore expect to obtain better approximations for a "best" parameter or "best" model by performing additional experiments.

Optimal Experimental Design

Performing experiments may be costly in terms of money, time, or other limited resources. The aim of reducing these costs leads to the following optimal experimental design (OED) problems:

Optimal experimental design for parameter estimation (OED/PE)

Given a parametric model, certain experimental capabilities, and a parameter estimation method, determine the experimental conditions which are most suitable for approximating a sought-after "best" parameter with that method.

Optimal experimental design for model discrimination (OED/MD)

Given several parametric models, certain experimental capabilities, and a model discrimination method, determine the experimental conditions which are most suitable for approximating a sought-after "best" model with that method.

These problems lead to constrained optimization problems, their objective functions are called "design criteria."

Which of the experimental conditions are *actually* most useful depends on the process, and possibly on the "best" parameter or "best" model. Since they are unknown, OEDs problem arising in practice are generally optimization problems under uncertainty. Their solutions will typically deviate from the *actually* most useful experimental conditions. The smaller these deviations, the more "robust" is the design criterion.

This thesis focuses on the practically important sequential approach, in which experiments are designed, performed, and analyzed after one another [10, 152]. In each step of such a procedure, the additional data tends to reduce the structural uncertainty. A "sequential" design criterion applied there can increase its robustness by properly taking into account suitable empirical quantifications of the current uncertainty.

Model discrimination problems typically arise in early stages of model building, when the uncertainties are particularly large. It is the aim of OED/MD to efficiently reduce the uncertainty about the sought-after "best" model, yet not the corresponding parameter uncertainty. In fact, optimal designs for model discrimination are typically inefficient for parameter estimation [13, 97]. Therefore, model discrimination problems often involve large parameter uncertainties, and the related optimal experimental design can benefit particularly from suitable robustification techniques.

Optimal Experimental Design for Parameter Estimation

The fundamental theory of OED for parameter estimation is well studied. Practically dominant are the so-called "alphabetic" design criteria going back to Kiefer and Wolfowitz [143], extensively discussed in the books by Fedorov [95], Atkinson and Donev [11], and Pukelsheim [206].

A focal point of research is currently the development of related efficient numerical methods. Optimal designs for dynamic or distributed models based on ordinary differential equations, differential-algebraic equations, and partial differential equations are treated by Asprey and Macchietto [9], Bock, Kostina, and Schlöder [36], Körkel [147], and Körkel et al. [148], to mention a few.

The so-called "Bayesian" design criteria have recently gained popularity in several applied fields. These design criteria can be regarded as generalizations of their non-Bayesian counterparts to cases in which prior knowledge is not negligible compared to the available experimental data. Details can be found in the reviews by Chaloner and Verdinelli [66] and von Toussaint [259] and references provided therein.

Design Criteria for Model Discrimination

Optimal experimental design for model discrimination is a far less homogeneous field than that for parameter estimation. A plethora of design criteria are available that rely on a wide range of statistical concepts, make varying assumptions about process and models, and apply different approximations. Most of them are connected to one of the three strategies presented in the following. Details can be found in the reviews by Burke [59], Franceschini and Macchietto [103], Hill [116], Kreutz and Timmer [152], and Steinberg and Hunter [238].

Hunter-Reiner Strategy

Hunter and Reiner [129] are possibly the first to suggest a sequential design criterion for model discrimination. It is based on the idea to perform that experiment under which the model predictions are maximally different. Atkinson and Fedorov [21, Sec. 3] provide a rigorous justification for it. The design criterion can be generalized straightforwardly to accept multivariate data and respect experimental uncertainty [88]. In this form – referred to as "Hunter-Reiner (HR)-criterion" – it is restricted to two rival models and neglects parameter uncertainty.

It is nevertheless still popular, presumably because it is easy to implement and cheap to compute. At the time of writing (May 19, 2016), Web of Science[4] lists 115 citations of [129], of with 20 are from the year 2010 or later. With minor modifications it is applicable to dynamic processes with time-dependent controls, and to models based on ordinary differential equations (ODEs) or differential-algebraic equations (DAEs) [10, 80, 120, 221].

Buzzi-Ferraris Strategy

A novel sequential design criterion for model discrimination was proposed by Buzzi-Ferraris and Forzatti [63], and was further extended by Buzzi-Ferraris et al. [61] and Buzzi-Ferraris, Forzatti, and Canu [64]. This "Buzzi-Ferraris (BF)-criterion" essentially generalizes the HR-criterion such that it incorporates parameter uncertainty.

The BF-criterion has received considerable attention: at the time of writing, Web of Science knows 89 citations of the three papers suggesting the BF-criterion, of which 34 are from the year 2010 or later.

The design criterion is popular among practitioners, but also stimulated further theoretical work: Schwaab et al. [225] proposes a data-adaptive multi-model generalization, Chen and Asprey [67] adopt it to dynamic processes with time-dependent controls, and Schwaab, Monteiro, and Pinto [224] and Donckels et al. [81] improve its parameter-robustness. The suggestions of the latter two were

[4] http://www.webofscience.com

applied just recently by Stamati et al. [237]. None of these modifications, however, changed the- underlying concept. The study by Donckels et al. [82] compares different variants of the BF-criterion.

Box-Hill-Hunter Strategy

Box and Hill [42] and Hill and Hunter [118] follow a conceptually different approach. They propose to measure the model uncertainty by the Shannon entropy of the posterior model probabilities, and to use the expected reduction of model uncertainty resulting from an additional experiment as design criterion for MD. This "Box-Hill-Hunter (BHH)-criterion" makes very few assumptions about the considered process and models, and incorporates experimental uncertainty, model uncertainty and parameter uncertainty in a natural way.

The BHH-strategy was an early and seminal approach for OED for MD. It gave rise to a huge body of follow-up works that both advanced the underlying theory and applied the strategy in practice. Currently, no less than 270 citations of [42] are known on Web of Science, with 42 of them coming from the last six years. In its general form, the BHH-criterion involves several integrals that typically lack a closed-form solution. Due to the curse of dimensionality, numerical approximations are computationally intractable for all but very simple model discrimination problems.

As remedy, Box, Hill and Hunter suggested an approximation that has a closed form under the common assumption that the data is normally distributed with known covariance. Early on, this approximation has been criticized as being an *upper bound* of a design criterion to be *maximized* [185]. Nevertheless, this "upper-bound approximation" has become and remains a popular design criterion. Just recently, Zhang et al. [269] proposed its usage in process engineering, and Pham and Tsai [202] adopted it to spatio-temporal models and applied it to design optimal observation networks in a real-world problem.

Aims and Contributions of this Thesis

This thesis aims to develop theory, algorithms, and actual implementations of new and practically applicable robust methods for designing optimal experiments that discriminate between several nonlinear multivariate parametric regression models. Furthermore, it strives to establish a framework of theoretical concepts, methods, and implementations that allows to numerically assess and compare the practical performance of different model discrimination methods and related robustification techniques.

To this end, this thesis contributes novel results and advances over existing techniques in various areas that are described in the following.

Empirical Quantification of Parameter Uncertainty

It lies in the very nature of optimal experimental design (OED) problems for model discrimination that they are often subject to substantial parameter uncertainty. Corresponding design criteria may hence benefit strongly from techniques that quantify this uncertainty empirically, that is, based on available experimental data. This thesis makes the following contributions to this field.

Approximations of PMLE Covariance

From the point of view of frequentist statistics, the uncertainty about the parameters of a model can be quantified based on the covariance matrix of the corresponding parameter maximum-likelihood estimator (PMLE). In the common case that this covariance is unknown, one reverts to empirical approximations.

This thesis proposes a novel "misspecification-robust" empirical approximation that is based on the first and second derivatives of the model responses with respect to the parameters. The approximation has a number of appealing properties: (a) it consistently generalizes the commonly applied "classic" alternative that is based on first-derivatives only, (b) but – much in contrast to the latter – does not assume that the model is locally affine-linear or correct, (c) it is applicable to the practically important class of models for normally distributed data with known covariances, and (d) it can be expected to be exact in the limit of infinitely many experiments. No comparable formula has been reported in literature.

The novel approximation constitutes an improvement over its classic counterpart, since it quantifies parameter uncertainty more adequately in model discrimination problems, which contain incorrect model by definition.

Posterior Parameter Covariance Approximation

From a Bayesian perspective, uncertainty in parameters can be expressed through the covariance of their posterior distribution, supposed that the model meets the classic assumptions of local linearity or correctness.

Central results concerning the posterior distribution in incorrect nonlinear models have become available only recently [144]. Based thereon, this thesis suggests a novel “misspecification-robust” formula for the empirical approximation of the posterior parameter covariance in nonlinear incorrect models for normally distributed observations with known covariances.

The novel formula overcomes the assumptions of local linearity or correctness that underlie the commonly applied alternative, yet is a consistent generalization of it. As such, it promises to be a more adequate quantification of parameter uncertainty in model discrimination problems.

Framework for Assessing PMLE Covariance Approximations

The thesis develops a framework that allows to assess and compare the quality of different approximations of the parameter maximum-likelihood estimate (pmle) covariance. It consists of statistically well-founded quality measures, numerical algorithms for their efficient computation, and a state-of-the-art implementation in the software package DoeSim. The framework is the first that allows the quantitative analysis of different empirical approximations for maximum-likelihood estimator (mle) covariance.

Case Studies: Quantification of Parameter Uncertainty

The quality of the classic and the misspecification-robust pmle covariance approximations are examined in an extensive numerical case study, using the DoeSim implementation of the previously proposed framework. The case study comprises twelve nonlinear models for the water-gas shift reaction (wgsr) reaction, which were collected by Schwaab, Monteiro, and Pinto [224] and Schwaab et al. [225]. The rival models show different magnitudes of structural “incorrectness.”

The results demonstrate that the misspecification-robust formula is asymptotically exact, and show that it is clearly superior to its classic counterpart in all considered cases, except if little data is available. These results seem to be the first published quantitative analysis of PMLE covariance approximations.

Design Criteria for Model Discrimination

Established Design Criteria with Enhanced Parameter-Robustness

The derived misspecification-robust formulas for quantifying parameter uncertainty can be applied to improve any parameter-robust design criterion that uses the classic formulas.

We demonstrate that by proposing a new and misspecification-robust variant of the Buzzi-Ferraris (BF)-criterion. It is the first design criterion proposed in literature that is parameter-robust and does not rely on the assumption that all underlying models are correct or locally affine-linear. We show that if they are, however, the new design criteria consistently reduces to the original BF-criterion.

Novel Design Criteria

This thesis contributes two novel design criteria for model discrimination under the assumption of normally distributed data with known covariance. They both possess the following promising properties.

- They are *lower* bounds of the Box-Hill-Hunter (BHH)-criterion under regularity conditions. The proof is given that is based on the recent information-theoretic inequalities by Hershey and Olsen [115] and Huber et al. [125]. Despite being lower bounds to the actual quantity of interest, they are statistically meaningful by themselves.

- They are parameter-robust, quantifying parameter uncertainty based on the newly proposed misspecification-robust formulas for the posterior parameter covariance.

- They are model-robust, using a novel formula for the posterior probability of a model which is applicable even if the models have different numbers of parameters.

- They are consistent generalizations of the simple and sound established multivariate Hunter-Reiner-criterion [129].

- They come with a natural support for discrimination among more than two models.

Currently, no other design criterion for model discrimination has been reported which one of the first three properties. These properties provide strong arguments to expect that both design criteria outperform the popular upper-bound approximation of the BHH-criterion and the multivariate Hunter-Reiner (HR)-criterion.

Optimization Problems in the Context of Model Discrimination

Parameter estimation is an integral part of most strategies for optimal experimental design for model discrimination. Newton's method and the Gauss-Newton method are particularly popular for computing unconstrained least-squares parameter estimates, see Nocedal and Wright [194, Sec. 10] and the references provided therein. In the context of model discrimination, the Gauss-Newton is applied, for example, by Schwaab, Monteiro, and Pinto [224].

This thesis shows that these methods are generally *not* appropriate in the context of model discrimination. The same is shown for sequential quadratic programming (SQP) methods with exact Hessians or Gauss-Newton Hessian approximations, which might be applied to compute such estimates under equality constraints.

Furthermore, it is demonstrated that a large class of optimal experimental design problems for model discrimination (MD) is *intrinsically non-convex* even under strong simplifying assumptions. Such problems are $\mathcal{NP}$-hard and particularly difficult to solve numerically. The computational complexity of optimal experimental design for model discrimination has so far not been discussed in literature in the detail given here.

Numerical Framework for Analyzing Model Discrimination Strategies

The thesis provides a numerical framework for comparing and assessing strategies for optimal experimental design for model discrimination (OED/MD). It consists of (a) statistical measures of the practical performance of a design criterion for solving a MD problem, (b) a set of problem-adapted algorithms for their computation, and (c) a state-of-the-art implementation in the software package DoeSim.

The architecture of DoeSim is completely modular and allows to exchange and recombine the algorithmic components of an MD strategy without effort. The package can autonomously simulate replicated runs of the sequence of designing, performing, and analyzing experiments in an MD problem specified by the user. The replicated runs can be performed fully in parallel to take full advantage of today's hardware. Based on replicated runs, the performance of the applied design criterion can be studied on a quantitative level. The implementation also contains tools for in-depth a priori and a posteriori analyses of the MD problem and offers rich visualization capabilities.

It is common practice in literature to examine the behavior of design criteria for MD on the basis of a single simulation of the considered sequential procedure, neglecting the random nature of the data. Contrary to this practice, the provided framework permits to analyze the full spectrum of the behavior that a design criterion shows under fluctuating data.

Case Studies: Efficiency of Design Criteria for Model

Simulation studies of sequential design criteria for MD typically rely one set of simulated data. Examples are the results of [81, 186, 224, 225, 234, 248], to mention a few. Since the data is inherently subject to random fluctuations, little general conclusions can be drawn from such results. Comparisons between different sequential criteria based on simulations of are rarely found in literature. Those available, for example [34, 185], compare two design criteria at most.

This thesis contains an extensive simulation study of different design criteria for model discrimination, including (a) a multivariate generalization of the Hunter-Reiner-criterion, (b) the Buzzi-Ferraris-criterion, (c) its newly proposed misspecification-robust counterpart, (d) a variant of the classic upper-bound approximation of the BHH-criterion, (e) the two novel misspecification-robust lower-bound approximations of the BHH-criterion, and (f) a model-independent and data-independent strategy for reference. They are examined on the basis of various MD problems among rival models for the water-gas shift reaction, which were collected by Schwaab, Monteiro, and Pinto [224] and Schwaab et al. [225] to study MD strategies. The results are obtained from the previously described software package DoeSim.

The considered case studies are novel in several aspects:

- They capture the statistical properties of design criteria for MD under the random fluctuations of the input data.
- They examine the influence of different magnitudes of measurement error onto the behavior of the design criteria.
- They systematically examine the influence of the number of rival models on their behavior.

- They comprise several of the fundamental design criteria for MD.
- The considered set of MD problems is much larger than in previously known comparisons.

The following previously unknown observations are made in the study:

- The popular upper-bound approximation of the BHH-criterion is surprisingly inefficient, particularly in model discrimination problems with three or more models.
- All of the considered parameter-robust design criteria outperforms the Hunter-Reiner-strategy in most cases.
- In all considered cases, the novel robust design criteria were among the most efficient ones for solving the MD problems. They showed particularly strong advantages over their alternatives in demanding MD problems with many rival model and large measurement errors.

Thesis Overview

This thesis is subdivided in four parts composed of nine chapters and three appendices as follows.

Part I concerns the theoretical foundations.

Chapter 1 defines the necessary fundamental concepts of "process" and "model family," and states the central questions of model discrimination (MD) and related optimal experimental design (OED). Furthermore, it introduces the Kullback-Leibler information criterion (KLIC), a measure for the process–model discrepancy. It discusses the properties of its minimizers, which formalize the notions of "best" parameters and "best" models.

Chapter 2 is concerned with statistical inference, which forms the basis of any OED strategy. It focuses on results that do not rely on the common but strong assumption that the underlying model (family) is correct, and points out in which areas such non-classic results are lacking.

To that end, it defines central concepts of statistical inference – likelihood function, information matrices, estimators, consistency and efficiency – such that they are applicable in possibly incorrect models. Based thereon, it surveys central results of maximum-likelihood estimators, particularly the conditions for consistency and asymptotic normality. Furthermore, it sets forth how the

Bayesian approach to inference can be applied to collections of several models, and surveys recent results concerning the consistency and asymptotic normality of the posterior distributions of parameters in incorrect models.

Chapter 3 focuses on statistical inference under the common assumptions of normally distributed data, known observation covariances, and locally affine-linear models. It is shown that under these assumptions, the central quantities of maximum-likelihood estimation and Bayesian inference reduce to conveniently simple forms, mostly sums-of-squares and matrices. These quantities permit the efficient numerical treatment of statistical inference. Furthermore, the assumption of normality and known covariances is used to derive a novel "misspecification-robust" formula that quantifies the parameter uncertainty even if the underlying model is incorrect and nonlinear.

Chapter 4 is the first of Part I which treats optimal experimental design. The chapter introduces the theoretical basics and considers strategies for MD that are based on frequentist inference, particularly on maximum-likelihood inference. It discusses in detail Kullback-Leibler (KL)-optimal designs and T-optimal designs, which are the theoretically ideal designs for MD. Although they depend on quantities that are unknown in practice, they define the aim that any practical approach for efficiently solving MD problems should strive for.

Then, two of the most popular sequential strategies for optimal experimental design for model discrimination (OED/MD) are reviewed: the Hunter-Reiner (HR)-strategy and the Buzzi-Ferraris (BF)-strategy. The latter is used as basis for proposing a novel design criterion that uses the misspecification-robust formula for quantifying parameter uncertainty.

Chapter 5 focuses on Bayesian approaches to optimal experimental design for model discrimination. It examines the de-facto standard Box-Hill-Hunter (BHH) strategy, which is based on the information-theoretic concept of entropy. It is clarified that this design criterion has no closed-form solution even under the comfortable assumptions of normally distributed data with known covariances. The popular closed-form upper-bound approximation is briefly reviewed.

The remaining chapter is dedicated to novel design criteria for MD. To that end, two information-theoretic inequalities are discussed that were discovered only recently. Based thereon, two new lower-bound approximations of the BHH-criterion are derived and discussed, with a focus under their robustness properties. It is shown that they are consistent with the HR-criterion.

The thesis continues in Part III with numerical methods and results.

Chapter 6 considers numerical methods required in the context of OED/MD. It discusses optimization techniques for least-squares problems that result from the aim of finding maximum-likelihood estimates in the context of MD. In particular, it shows that such problems have some intrinsic properties which make Newton's method and the popular Gauss-Newton method are inappropriate for solving them.

Furthermore, it examines optimization problems arising from OED/MD, with a focus on their computational complexity. Essentially, it clarifies that such problems are $\mathcal{NP}$-hard even under strong simplifying assumptions, which makes them difficult to solve numerically. A simple grid search is described as remedy for low-dimensional problems.

The chapter finishes with a short introduction to low-discrepancy sequences, which can be used to generate start values for local optimization techniques, and to generate space-filling experimental designs.

Chapter 7 compares the classic empirical approximation for the covariance of a parameter maximum-likelihood estimators (PMLEs) with the novel robust alternative proposed in Chap. 3. It derives suitable quantities for measuring the approximation quality between covariance matrices and develops efficient algorithms for their numerical computation. The algorithm is implemented in the software package DoeSim.

The chapter then presents and discusses numerical results of an extensive case study performed with DoeSim. The study assesses and compares the classic and the misspecification-robust formulas based on models for the water-gas shift reaction (WGSR) reaction [225]. The results demonstrate that the misspecification-robust formula is clearly superior to its classic counterpart.

Chapter 8 develops a framework for the numerical assessment of sequential design criteria for MD. It derives two statistical measures for the performance of design criteria with respect to solving MD problems. One is based on the concept of T-optimality introduced in Chap. 4, the other on Bayesian posterior probabilities discussed in Chap. 3. It then briefly reviews various sequential design criteria for MD and shows that they can be expressed in a uniform form.

Based on this representation, a Monte Carlo method is developed that allows to efficiently compute the introduced performance measures for a given design criterion in a user-specified MD problem. The algorithm is implemented in the software package DoeSim.

Chapter 9 uses the framework from the preceding chapter to examine established and newly proposed sequential design criteria for MD in an extensive simulation study. The considered MD problems are based on the models for the WGSR reaction that were introduced in Chap. 7. The design criteria are examined in MD problems with a different number of models and different magnitudes of measurement error. The results demonstrate that the newly proposed design criteria perform significantly better than established alternatives and thus have the potential to save considerable experimental effort.

The appendix collects various results for the convenience of the reader. Appendix A treats some basics concerning norms, matrices and derivatives. Appendix B contains some results from probability theory and statistics. Appendix C summarizes and interprets essential concepts from information theory: entropy, Kullback-Leibler distance, and mutual information. A bibliography it provided at the end of the thesis.

Part I.

Theoretical Foundations

They say that Understanding ought to work by the rules of right reason. These rules are, or ought to be, contained in Logic; but the actual science of Logic is conversant at present only with things either certain, impossible, or entirely *doubtful, none of which (fortunately) we have to reason on. Therefore the true Logic for this world is the Calculus of Probabilities, which takes account of the magnitude of the probability (which is, or which ought to be in a reasonable man's mind).*

James Clerk Maxwell in a letter to Lewis Campbell, circa July 1850, cited by Campbell and Garnett [65, p. 80]

1. Processes, Model Families and their Discrepancy

Contents

This chapter introduces the fundamental concepts of this thesis: a formalization of a real-world process providing intrinsically random (not necessarily normally distributed) data and "model families", collections of competing parametric regression models for such a process.

After introducing these concepts in Sec. 1.2, we use them in Sec. 1.3 to outline the central questions of this thesis: (a) what are the limits to what we can learn about a particular process using a given model family, (b) what can we learn about it in practice from available data, and (c) which experiments are best for collecting additional data for improving our knowledge?

Question (b) gives rise to parameter estimation (PE) and model discrimination (MD) problems. Selected methods of statistical inference for their solution are treated in Chaps. 2 and 3. Question (c) leads to optimal experimental design (OED) problems, which are the focus of this thesis. They are considered in detail in Chaps. 4 and 5.

As preparation, Sec. 1.4 focuses on (a). We avoid the common but strong assumption that the model family contains a "perfect" description of the process, that is, we allow the model family to be misspecified or incorrect. Then, the aim can only be to identify the model family member that "most adequately" (but possibly not perfectly) describes the process. We introduce a suitable discrepancy measure and discuss the properties of its minimizers, which state the limit of attainable knowledge in the sense of (a).

The concepts, terminology and notation provided in this chapter form a framework that allows us to express problems of PE, MD and OED and the involved uncertainties in incorrect model families in a unified fashion in the subsequent chapters.

1.1. Notation

We use the following notational conventions throughout this thesis.

Scalars and vectors, and scalar-values and vector-valued function are typeset in Latin or Greek letters like μ, c, θ, or Ψ. Matrices and matrix-valued functions are displayed in Latin or Greek uppercase boldface letters like $\boldsymbol{A}$, $\boldsymbol{C}$, or $\tilde{\boldsymbol{F}}_n$. The i-th scalar component of vector v is written as $[v]_i$ or v_i. Likewise, the scalar component in row i and column j of matrix $\boldsymbol{M}$ is written $[\boldsymbol{M}]_{ij}$ or m_{ij}. Scalar or vector-valued random variables are represented by calligraphic uppercase letters like $\mathcal{M}$, $\mathcal{Y}$, or $\mathcal{Q}$. Sets and ordered lists use an alternative calligraphic font, for example $\mathscr{I}$ or $\mathscr{M}$.

A blackboard-bold style is used for different classes of objects, without risk of ambiguity: $\mathbb{N}$, $\mathbb{N}_0$, $\mathbb{R}$, $\mathbb{R}^+$, and $\mathbb{R}_0^+$ represent the natural numbers, the natural numbers including zero, the real numbers, the positive real numbers, and the non-negative real numbers, respectively. The symbols $\mathbb{P}[\cdot]$, $\mathbb{E}[\cdot]$, and $\mathbb{C}[\cdot]$ stand for the operators of probability, expectation, covariance from probability theory, respectively. Finally, $\mathbb{H}[\cdot]$, $\mathbb{D}[\cdot\|\cdot]$ and $\mathbb{I}[\cdot\|\cdot]$ are the information-theoretic operators of entropy, Kullback-Leibler distance, and mutual information, respectively. The latter are defined in Appendix C.

Let $\mathscr{X}$ and $\mathscr{Y}$ be arbitrary sets and let $f: \mathscr{X} \mapsto \mathscr{Y}$. The expressions

$$\{f(x) : x \in \mathscr{X}\} \text{ and } (f(x) : x \in \mathscr{X}) \tag{1.1}$$

denote the SET and the FAMILY, respectively, of elements $f(x)$ from $\mathscr{Y}$ with index x from the index set $\mathscr{X}$. In contrast to an indexed set, an indexed family may have several identical members.

Let $(x, y) \mapsto f(x, y)$ be an arbitrary two-argument function. The expression $f(\cdot, y)$ refers to the function $x \mapsto f(x, y)$, that is, to the one-argument function obtained by fixing the second argument of f to the value y. Therefore, $f(\cdot, y_1)$ and $f(\cdot, y_2)$ are generally different functions if $y_1 \neq y_2$. The notation is used likewise for functions with more than two arguments.

Different probability density functions (PDFs) share the same symbol – usually $p(\cdot)$ – and are distinguished by their arguments only. That is, if $\mathcal{U}$ and $\mathcal{V}$ are continuous (as opposed to discrete) random variables, their PDFs are denoted $p(u)$ and $p(v)$, respectively. Despite they both use the same symbol $p(\cdot)$, the expressions $p(u)$ and $p(v)$ refer to two different functions. The same convention is used for the probability mass function (PMF) of discrete random variables.

1.2. Data-Generating Processes and Model Families

This section introduces the fundamental concepts of "process" and "model family" that are used throughout this thesis.

1.2.1. Multivariate Nonlinear Data-Generating Processes

The center of interest in this thesis is a type of process that is manipulable trough a certain number of independent variables and yields a fixed number of observable quantities as output.

Definition 1.1 (Observation Domain, Control Domain)

Let $n_x, n_y \in \mathbb{N}$. The EXPERIMENTAL DOMAIN $\mathcal{X}$ is a non-empty Lebesgue-measurable subset of $\mathbb{R}^{n_x}$. The OBSERVATION DOMAIN $\mathcal{Y}$ is a non-empty Lebesgue-measurable subset of $\mathbb{R}^{n_y}$.

Elements from $\mathcal{X}$ represent the available means and methods of manipulating the process of interest. A point $x \in \mathcal{X}$ is hence called EXPERIMENTAL CONDITION. Element from $\mathcal{Y}$ represents observed outputs of the process. A point $y \in \mathcal{Y}$ is accordingly referred to as OBSERVATION, EXPERIMENTAL OUTCOME, or EXPERIMENTAL RESULT.

The considered type of process is intrinsically random (=ALEATORY) in the sense that the observations obtained from replicated experiments under the same experimental condition exhibit random fluctuations. The distribution characterizing these fluctuations, however, depends deterministically on the experimental condition. Such a process can be characterized as follows.

Definition 1.2 (Process, Observable)

A process is a function $q: \mathscr{Y} \times \mathscr{X} \mapsto \mathbb{R}_0^+$ for which $p(y|x)$ is a probability density function (pdf) in y under all experimental conditions $x \in \mathscr{X}$. An observable $\mathcal{Y}_x$ is a continuous (as opposed to discrete) $\mathscr{Y}$-valued random vector distributed according to pdf $q(\cdot|x)$.

The vertical bar in $q(y|x)$ indicates q is a pdf in the quantity y on the left, and that the definition of this pdf depends parametrically on the quantity x on the right.

Strictly speaking, q should be called "probabilistic description of the process", since the term "process" typically refers to an entity of the real-world and not to a mathematical concept. Since this thesis focuses, with very few exceptions, only on this probabilistic description, the chosen terminology is favored for its brevity.

The distributions specified by the process under different experimental conditions are not necessarily unique, it is possible that $q(\cdot|x) = q(\cdot|x')$ under different experimental conditions $x \neq x'$. The process may be nonlinear in the experimental condition in the sense that $q(y|x)$ for a given y is a nonlinear function of x.

The function q describes the probabilistic properties of observations. Given the process, the probability of observing a value in a measurable subset $\mathscr{A}$ of $\mathscr{Y}$ under experimental condition x is

$$\mathbb{P}[\mathcal{Y}_x \in \mathscr{A}] \equiv \int_{\mathscr{A}} q(y|x)\,\mathrm{d}y. \tag{1.2}$$

If the expectation of the observable $\mathcal{Y}_x$ exists, it can be written as

$$\mathcal{Y}_x \equiv \bar{\eta}(x) + \bar{\mathcal{E}}(x), \tag{1.3a}$$

consisting of the non-random observation mean $\bar{\eta}(x) := \mathbb{E}[\mathcal{Y}_x]$ and the random contribution $\bar{\mathcal{E}}(x) := \mathcal{Y}_x - \bar{\eta}(x)$, whose pdf is related to that of $\mathcal{Y}_x$ by a simple translation,

$$\bar{\mathcal{E}}(x) \sim q(y - \bar{\eta}(x)|x), \tag{1.3b}$$

which implies the identity $\mathbb{E}\left[\bar{\mathcal{E}}(x)\right] \equiv 0$. A particular observation $y \in \mathcal{Y}$ obtained from the process under experimental condition x can then be written as

$$y \equiv \bar{\eta}(x) + \epsilon, \text{ where } \epsilon \text{ is a realization of } \bar{\mathcal{E}}(x). \tag{1.4}$$

This representation suggests to identify $\bar{\eta}(x)$ with the observable part of a purely *deterministic* process and $\bar{\mathcal{E}}(x)$ with random fluctuations, for example caused by measurement errors or uncontrolled influences on the process. *Our notion of process may therefore comprise a description of the data acquisition methods or the measurement apparatus.*

1.2.2. Families of Multivariate Nonlinear Regression Models

This thesis considers the case that several competing parametric regression models, subsumed in what we call a MODEL FAMILY, are available for a given process.

Definition 1.3 (Regression Model, Model Family)

Let $\mathcal{M} := \{1, \ldots, n_{\mathcal{M}}\}$ be the MODEL INDEX SET. For all $\mu \in \mathcal{M}$, let the PARAMETER DOMAIN $\mathcal{Q}^{\mu}$ be a possibly empty subset of $\mathbb{R}^{n_{\theta^\mu}}$, with $n_{\theta^\mu} \in \mathbb{N}$. For all x from the experimental domain $\mathcal{X}$, all $\mu \in \mathcal{M}$ and all $\theta^\mu \in \mathcal{Q}^\mu$, let $p(y|x, \mu, \theta^\mu)$ be a PDF in the argument y from the observation domain $\mathcal{Y}$. The REGRESSION MODEL $\mu \in \mathcal{M}$ or simply a MODEL μ is the indexed family

$$\left(p(\cdot|\cdot, \mu, \theta^\mu) : \theta^\mu \in \mathcal{Q}^\mu\right), \tag{1.5}$$

and a MODEL FAMILY is the indexed family

$$\left(p(\cdot|\cdot, \mu, \theta^\mu) : \mu \in \mathcal{M}, \theta^\mu \in \mathcal{Q}^\mu\right). \tag{1.6}$$

We refer to a function $p(\cdot|\cdot, \mu, \theta^\mu)$ as MODEL (FAMILY) MEMBER.

The vertical bar in $p(y|x, \mu, \theta^\mu)$ indicates p is a PDF in the quantity y on the left, and that the definition of this PDF depends parametrically on the quantities on the right.

This definition implies that the parameters in a model family are MODEL-LOCAL, that is, each model $\mu \in \mathcal{M}$ depends only on the parameter $\theta^\mu \in \mathcal{Q}^\mu$ and not on any of the parameters $(\theta^\nu \in \mathcal{Q}^\nu : \nu \neq \mu)$. This property gives us the freedom to combine regression models with completely unrelated internal formulations within one a model family.

The PDF $p(\cdot\,|\,x,\mu,\theta^\mu)$ may depend nonlinearly on both the experimental condition x and the parameter θ^μ.

Let $\tilde{\mathcal{Y}}(x,\mu,\theta^\mu)$ be a continuous (as opposed to discrete) $\mathcal{Y}$-valued random vector distributed according to PDF $p(\cdot\,|\,x,\mu,\theta^\mu)$ and suppose that its expectation exists. Analogously to (1.3), $\tilde{\mathcal{Y}}(x,\mu,\theta^\mu)$ can then be written as

$$\tilde{\mathcal{Y}}(x,\mu,\theta^\mu) \equiv \eta^\mu(x,\theta^\mu) + \mathcal{E}(x,\mu,\theta^\mu), \tag{1.7}$$

with the non-random (MODEL) RESPONSE $\eta^\mu(x,\theta^\mu) := \mathbb{E}\left[\tilde{\mathcal{Y}}(x,\mu,\theta^\mu)\right]$ and the random contribution $\mathcal{E}(x,\mu,\theta^\mu) := \tilde{\mathcal{Y}}(x,\mu,\theta^\mu) - \eta^\mu(x,\theta^\mu)$ distributed according to

$$\mathcal{E}(x,\mu,\theta^\mu) \sim p(y - \eta^\mu(x,\theta^\mu)\,|\,x,\mu,\theta^\mu), \tag{1.8}$$

which implies $\mathbb{E}\left[\mathcal{E}(x,\mu,\theta^\mu)\right] = 0$. This representation suggests to interpret $\eta^\mu(x,\theta^\mu)$ as a description of the deterministic part $\bar{\eta}(x)$ of the process and $\mathcal{E}(x,\mu,\theta^\mu)$ as a description of its additive random contribution $\bar{\mathcal{E}}(x)$. In practice, regression models are in fact often specified in the form of (1.7) and (1.8) in the first place.

The responses η^μ might be determined implicitly. For example, they might be composed of the values that a solution of a system of ordinary differential equations or partial differential equations takes at the points in time or space where measurements are made.

The considered concepts of process and model family are rather general. They may be UNIVARIATE, $n_y = 1$, or MULTIVARIATE, $n_y > 1$. They may have *non-normal* distributions, with covariances that (if they exist) *depend on the experimental condition* (=heteroscedasticity) and non-diagonal, representing *correlations*. Furthermore, the PDFs specified by a model family may be OVERLAPPING, meaning that there may be values $(x,\mu,\theta^\mu) \neq (x',\nu,\theta^\nu)$ for which $p(\cdot\,|\,x,\mu,\theta^\mu) = p(\cdot\,|\,x',\nu,\theta^\nu)$. This comprises the important special case that a model is NESTED within another model, that is, is a special case of it.

1.2.3. Experimental Designs

The following definition introduces a convenient way of representing collections of several experimental conditions.

Definition 1.4 (Design, Exact Design)

A DESIGN is a function

$$\xi\colon \mathcal{X} \mapsto [0,1] \text{ with finite support and } \sum_{x \in \operatorname{supp}(\xi)} \xi(x) = 1. \tag{1.9}$$

The value $\xi(x)$ is the WEIGHT assigned to experimental condition x. The set of designs is denoted Ξ. A design $\xi \in \Xi$ is EXACT, iff for all $x \in \mathcal{X}$ it holds that

$$\xi(x) = \frac{r(x)}{n}, \text{ where } r\colon \mathcal{X} \mapsto \mathbb{N} \text{ and } \sum_{x \in \operatorname{supp}(\xi)} r(x) = n. \tag{1.10}$$

The value $n\xi(x) = r(x)$ is the NUMBER OF REPLICATIONS assigned to experimental condition x. The set of all n-experiment exact designs is denoted $\Xi_n \subseteq \Xi$.

Formally, a design is a normed measure or probability measure over the experimental domain $\mathcal{X}$ with finite support. Where necessary, we write ξ_n to emphasize that a design is a n-experiment exact design.

A n-experiment exact design specifies a set of *mutually distinctive* experimental conditions $\operatorname{supp}(\xi) = \{x_1, \ldots, x_s\}$ and associated *integer* replication numbers $n\xi(x_i) = n\xi(x_i)$, for all $i \in \{1, \ldots, s\}$. The conditions of any finite number of experiments can thus be represented by an exact design and vice versa.

For a non-exact design ξ there exists no number of experiments $n \in \mathbb{N}$ such that $n\xi(x)$ is integer. In general, a (possibly non-exact) design is therefore not uniquely associated with the conditions of a finite number of experiments. It can, however, be approximated arbitrarily well by an n-experiment exact design with large n, since the set of rational numbers (containing the weights of exact designs) is dense in the set of reals (containing the weights of exact and non-exact designs). The other way round, there are efficient strategies for rounding non-exact designs to exact ones, for example described by Pukelsheim and Rieder [207] and the references given therein.

Let be two designs and let $c \in (0,1)$. Any convex combination $c\xi + (1-c)\tilde{\xi}$

defined by

$$c\xi(x) + (1-c)\tilde{\xi}(x), \text{ for all } x \in \operatorname{supp}(\xi) \cup \operatorname{supp}(\tilde{\xi}), \tag{1.11}$$

is again a design, with $\operatorname{supp}\left(c\xi + (1-c)\tilde{\xi}\right) = \operatorname{supp}(\xi) \cup \operatorname{supp}\left(\tilde{\xi}\right)$. If the exact designs ξ_n and ξ_l describe the conditions of $n \in \mathbb{N}$ and $l \in \mathbb{N}$ different experiments, then the $n + l$-experiment exact design $\frac{n}{n+l}\xi_n + \frac{l}{n+l}\xi_l$ describes the conditions of the joint collection of all $n + l$ experiments.

1.3. Outline: Data-Related Problems

Suppose we are interested in a particular real-world process. We do not know the process behavior, but have assembled one or several tentative regression models for it and have performed some experiments. This typical practical situation might be formalized as follows.

Scenario 1.5 (Fundamental Setting)

(i) A process according to Def. 1.2 is given.

(ii) The function q characterizing the process is unknown.

(iii) The data $d_n \in \mathcal{Y}^n$ is available from the process, consisting of n observations from the observation domain $\mathcal{Y}$, obtained in $n \in \mathbb{N}$ statistically independent experiments performed under known conditions described by the n-experiment exact design ξ_n.

(iv) A model family from Def. 1.3 is given for describing the process.

For each model $\mu \in \mathcal{M}$, the parameter domain $\mathcal{Q}^\mu$ is compact and $p(y|x, \mu, \theta^\mu)$ is continuous with respect to θ^μ for all $y \in \mathcal{Y}$ and all $x \in \operatorname{supp}(\xi_n)$ for all $n \in \mathbb{N}$.

(v) Additional experiments can be performed under any condition from the experimental domain $\mathcal{X}$. Under given conditions, the additional experiments are statistically independent of each other and from the previous ones.

Here and in the following a function is unknown means that it is not possible to evaluate it for any argument from its domain. This scenario is starting point for the remaining thesis. Various variants and special cases of it appear in the subsequent chapters. The continuity assumed in assumption (iv) is required to ensure that the Kullback-Leibler information criterion (KLIC) that we introduce in Sec. 1.4.2 is well defined.

1.3.1. Extended Notation

In the context of scenario 1.5, the following extended notation is used for all $n \in \mathbb{N}$.

According to Def. 1.4, the number of replications of the experiment under condition $x \in \operatorname{supp}(\xi_n)$ is $r_n(x) := n\xi_n(x)$. The observation resulting from replication no. $j \in \{1, \ldots, r_n(x)\}$ of the experiment under $x \in \operatorname{supp}(\xi_n)$ is denoted $y_j(x) \in \mathcal{Y}$.

An observation $y_j(x)$ is considered as a realization of the observable $\mathcal{Y}_x$. Likewise, the vector of data $d_n \in \mathcal{Y}^n$ which summarizes the observations from the n experiments is a realization of the SAMPLE $\mathcal{D}_n$, a random variable taking values in $\mathcal{Y}^n$. The sample $\mathcal{D}_n$ is composed of the observables $\mathcal{Y}_x$ with $x \in \operatorname{supp}(\xi_n)$, replicated according to $r_n(x)$.

The probability density function (PDF) of $\mathcal{D}_n$ is denoted $q(d_n \,|\, \xi_n)$, and the corresponding PDF specified by model μ with parameter θ^μ is denoted $p(d_n \,|\, \xi_n, \mu, \theta^\mu)$. Since the experiments are independent by assumption, the density assigned by the process to the data d_n obtained under ξ_n is

$$q(d_n \,|\, \xi_n) = \prod_{x \in \operatorname{supp}(\xi_n)} \prod_{j=1}^{r_n(x)} q\big(y_j(x) \,\big|\, x\big) \tag{1.12}$$

and the corresponding density of model μ with parameter θ^μ is

$$p(d_n \,|\, \xi_n, \mu, \theta^\mu) = \prod_{x \in \operatorname{supp}(\xi_n)} \prod_{j=1}^{r_n(x)} p\big(y_j(x) \,\big|\, x, \mu, \theta^\mu\big). \tag{1.13}$$

Analogously to Defs. 1.2 and 1.3, the vertical bar in $q(\cdot \,|\, \cdot)$ and $p(\cdot \,|\, \cdots)$ indicates that the functions are PDFs in the quantity on the left, and that the definitions of these PDFs depend parametrically on the quantities on the right.

1.3.2. Experimental and Structural Uncertainty

Scenario 1.5 involves two types of uncertainty that need to be clearly distinguished.

Repeated observations obtained under the same experimental condition $x \in \mathcal{X}$ exhibit random fluctuations. The ensuing uncertainty about the particular outcome of a not yet performed experiment is called EXPERIMENTAL UNCERTAINTY[1]. It complicates inferences from the data about the process.

The probabilities of the possible observations under experimental condition $x \in \mathcal{X}$ are determined by the PDF $q(\cdot\,|\,x)$. The lack of knowledge about the function q is called STRUCTURAL UNCERTAINTY.

We shall see in Chaps. 2 and 3, that under certain regularity conditions, the random fluctuations of the data tend to cancel out in the long run, so that structural uncertainty can be reduced by extending the data base. To what extent it can be reduced is determined by the model family, in particular by the bias or mismatch between the "best" model family member and the process.

1.3.3. Central Questions

We can now give a first outline of the central problems considered in this thesis. They arise from scenario 1.5 and can in general terms be stated as follows:

(Q1.1) *Given a model family, what is theoretical limit of what we can learn about the process q, that is, to which theoretical limit can we reduce the structural uncertainty?*

(Q1.2) *Given the model family and the available experiments, what can we learn about the unknown process q, that is, how far can we reduce the structural uncertainty? (statistical inference)*

(Q1.3) *Given the model family and the available experiments, under which conditions shall we perform additional experiments to improve our knowledge about the unknown process q, that is, to reduce the structural uncertainty? (optimal experimental design)*

The questions build on each other and must be considered consecutively. Question (Q1.1) can be interpreted as a question of finding the model family member with the lowest mismatch to the process. In this sense, (Q1.1) is answered in the next section. Question (Q1.2) is a question from the field of statistical inference, which comprises, among others, the problem classes of parameter

[1] Alternative terms are "observation error" or "observational variability".

estimation and model discrimination. Methods for statistical inference are considered in Chaps. 2 and 3. Question (Q1.3) leads to problems of optimal experimental design (OED), which are the focus of this thesis. They are treated in Chaps. 4 to 5.

1.4. Measuring the Process–Model Discrepancy

Essentially, all models are wrong, but some are useful.

Box and Draper [44, p. 424]

A model family can be considered as a collection of attempts to mimic or describe the unknown process. In particular, model $\mu \in \mathcal{M}$ with parameter $\theta^\mu \in \mathcal{Q}^\mu$ gives rise to the approximation

$$p(\cdot \mid \xi, \mu, \theta^\mu) \approx q(\cdot \mid \xi) \tag{1.14}$$

for the process under the exact design ξ. We refer to the error of this approximation (in an as yet unspecified measure) as DISCREPANCY. Among the members of the model family, usually some approximate the process more adequately than others. Typically, the discrepancy is design-dependent: a model family member that adequately describes the process well under a particular design might perform badly under a different design.

Suppose we are interested in the process behavior under the exact design ξ. Without additional assumptions, our aim in scenario 1.5 can at most be to identify that model family member with exhibits the lowest mismatch under ξ. Question (Q1.1) can thus be split up into the following two questions.

(Q1.4) How can we measure the discrepancy[2] of a model $\mu \in \mathcal{M}$ with parameter $\theta^\mu \in \mathcal{Q}^\mu$ under a given design, that is, how can we measure the error of approximation (1.14)?

(Q1.5) For which model $\bar{\mu}(\xi) \in \mathcal{M}$ and which corresponding parameter $\bar{\theta}(\xi) \in \mathcal{Q}^\mu$ is the mismatch minimal in terms of the measure from (Q1.4), and how low is it?

To be well-posed, (Q1.5) requires the sought-after best quantities to be unique, or, in the language of statistical inference, to be IDENTIFIABLE. Depending on our

[2] Other common terms are "mismatch" or "bias".

interests, we may in (Q1.5) be interested in the values $\bar{\mu}(\xi)$ and $\bar{\theta}(\xi)$ themselves, as they might give us insights about the mechanism governing the process under ξ, or in the corresponding probability density function (PDF) $p\big(\cdot \,\big|\, \xi, \bar{\mu}(\xi), \bar{\theta}(\xi)\big)$ for predicting the process behavior under ξ.

Any of these "minimal-mismatch" quantities depend on the design of interest ξ. Without additional assumption, we can generally not conclude from having minimal mismatch under ξ to having minimal mismatch under any other design $\xi' \neq \xi$.

When we answer (Q1.4) and (Q1.5), we *define* which quantities are of interest for us in the given model family. The aim of learning something from data about the unknown process, as stated in (Q1.2), can then be expressed as learning something about these unknown "best" quantities.

1.4.1. Correct Parameters, Models and Model Families

Let us first consider the special case that the model family is capable of "perfectly" describing the process in the sense that approximation (1.14) is exact.

Definition 1.6 (Correct Parameter, Model and Model Family)

Let $\xi \in \Xi$ be a possibly non-exact design.

(i) Parameter $\theta^{\mu} \in \mathcal{Q}^{\mu}$ of model $\mu \in \mathcal{M}$ is CORRECT UNDER ξ, iff

$$q(y \mid x) = p(y \mid x, \mu, \theta^{\mu}) \text{ for all observations } y \in \mathcal{Y} \qquad (1.15)$$

under all experimental conditions $x \in \operatorname{supp}(\xi)$. The parameter is CORRECT, iff (1.15) holds under all $x \in \mathcal{X}$. The parameter is INCORRECT (UNDER ξ), iff it is not correct (under ξ).

(ii) Model $\mu \in \mathcal{M}$ is CORRECT (UNDER ξ), iff there exists a correct parameter (under ξ) in $\mathcal{Q}^{\mu}$. The model is INCORRECT (UNDER ξ), iff it is not correct (under ξ).

(iii) The model family is CORRECT (UNDER ξ), iff there exists a correct model (under ξ) in $\mathcal{M}$. The model family is INCORRECT (UNDER ξ), iff it is not correct (under ξ).

To emphasize that we do *not* assume that a particular parameter, model or model family is correct (for a design), nor assume that it is incorrect, we say that it is possibly incorrect. A parameter or a model that are correct (under a design) are not necessarily unique. Parameters and models that are correct do not depend on any particular design. Accordingly, a correct parameter and a correct model are in fact correct for any design $\xi \in \Xi$.

Under a model $\mu \in \mathcal{M}$ and a parameter $\theta^\mu \in \mathcal{Q}^\mu$ that are correct for design ξ, approximation (1.14) is *exact*. Then, it is theoretically possibly to *identify* the unknown process under ξ using the given model family, that is, it is possible to completely dispose of the structural uncertainty under ξ in terms of (Q1.1) by identifying the model and the parameter that are correct parameter for ξ.

Many important results from statistical inference are derived under correctness assumptions. Under real-world conditions, however, they rarely hold: looking close enough typically reveals a mismatch between model family and process.

Correctness can be considered as a binary measure for the mismatch in the sense of (Q1.4). A correct model family member has mismatch zero, non-correct ones have non-zero mismatch. Among the latter, some will describe the process better than others. We now introduce a continuous measure for their mismatch.

1.4.2. Kullback-Leibler Information Criterion (KLIC)

In principle, any measure for the dissimilarity of pdfs is a candidate for answering (Q1.4). A comprehensive class of such measures are the so-called f-divergences, independently introduced by Csiszár [76] and Ali and Silvey [6]. A summary of their properties plus some novel results are given by Liese and Vajda [175].

A popular member of this class is the Kullback-Leibler distance (kld). In its terms, the discrepancy of model family member $p(\cdot\,|\,x,\mu,\theta^\mu)$ for describing the process $q(\cdot\,|\,x)$ under the experimental condition $x \in \mathcal{X}$ is

$$\int_{\mathcal{Y}} q(y\,|\,x) \ln \frac{q(y\,|\,x)}{p(y\,|\,x,\mu,\theta^\mu)}\, \mathrm{d}y. \tag{1.16}$$

An overview of the kld and its properties is given in Appendix C. The additivity of the kld for independent random variables, see Prop. C.4property (v), suggests the following measure for the discrepancy under a given design.

Definition 1.7 (Kullback-Leibler Information Criterion (KLIC))

The Kullback-Leibler information criterion (klic) of model $\mu \in \mathcal{M}$

with parameter $\theta^{\mu} \in \mathcal{Q}^{\mu}$ under design ξ is

$$\delta(\mu, \theta^{\mu}, \xi) := \sum_{x \in \operatorname{supp}(\xi)} \xi(x) \int_{\mathcal{Y}} q(y\,|\,x) \ln \frac{q(y\,|\,x)}{p(y\,|\,x, \mu, \theta^{\mu})} \,\mathrm{d}y, \tag{1.17}$$

supposed the right-hand side exists.

Under mild regularity conditions on the process and the model, notably parameter continuity as assumed in assumption (iv) of scenario 1.5, the KLIC exists and is continuous with respect to θ^{μ}. They are satisfied, for example, if both process and model are normal distributions with unit covariance and if the model responses are continuous in the parameter. Details on the regularity conditions can be found in the works of White [267, Asmps. A1–A3(a) and subsequent comments].

The KLIC assigns a real number to each pair of a model and a parameter: the smaller its value, the more does the corresponding model family member adequately describe the process. Note that the KLIC is also defined for non-exact designs. The KLIC is widely used for measuring the discrepancy of a hypothesized distribution with respect to the actual one. Several authors use it as a starting point for deriving empirical criteria for the selection of the most adequate model for a given process. The probably most prominent representative here is the information criterion of Akaike [2, 3], other examples are the criteria of Bozdogan [47, 48] and Sawa [220] and Sin and White [233]. Vuong [261] starts from the KLIC to develop generalized likelihood-ratio tests for falsifying incorrect models. The KLIC also plays a key role in the field of estimation, particularly in the extension of the maximum-likelihood method for possibly incorrect models developed by White [267] which are treated in Sec. 2.4.

The following properties make the KLIC an attractive candidate for measuring the discrepancy in the sense of (Q1.1).

Proposition 1.8 (Fundamental Properties of the KLIC)

Under the previously mentioned regularity conditions, the KLIC exists and has the following properties.

(i) $\delta(\mu, \theta^{\mu}, \xi) \geqslant 0$ for all $\theta^{\mu} \in \mathcal{Q}^{\mu}$. (non-negativity)

(ii) $\delta(\mu, \theta^{\mu}, \xi) = 0$ if and only if $q(\cdot\,|\,x) = p(\cdot\,|\,x, \mu, \theta^{\mu})$ for all $x \in \operatorname{supp}(\xi)$, which is equivalent to the statement that μ is a correct model under ξ and θ^{μ} is a correct parameter under ξ. (consistency with correctness)

(iii) If $p(\cdot\,|\,x,\mu,\theta^\mu) = p(\cdot\,|\,x,\nu,\theta^\nu)$ for all $x \in \operatorname{supp}(\xi)$, then $\delta(\mu,\theta^\mu,\xi) = \delta(\nu,\theta^\nu,\xi)$. (invariance)

(iv) $\delta(\mu,\theta^\mu,\xi)$ is continuous with respect to θ^μ. (continuity)

Proof Items (i) to (iii) are direct carry-overs from the KLD properties properties (i)–(iii) of Prop. C.4, respectively. In particular, (i) is a corollary of Prop. C.4property (i), (ii) follows from Prop. C.4property (ii) and Def. 1.6, and (iii) results from Prop. C.4property (iii). A proof of (iv) is given by White [267]. □

The KLIC is *not* a metric, since it is not symmetric and does not satisfy the triangle inequality. It is, however, a premetric which implies a concept of "closeness" between the process and the model family member specified by μ and θ^μ. Furthermore, property (ii) tells us that the KLIC consistently extends the concept of correctness under a design.

As per (iii), the KLIC is invariant to distribution-preserving transformations and as such independent of formulation details of the model family. Using a different ordering, scaling or physical interpretation of the parameter vector, permuting model indices, or reformulating the underlying equations leaves the KLIC unchanged as long as the resulting distribution is the same.

From the experimenter point of view, errors in approximation (1.14) for observations that are frequent in practice (areas where $q(\cdot\,|\,x)$ is large) are practically more severe than approximation errors for rarely encountered observations (areas where $q(\cdot\,|\,x)$ is small). The more likely an observation, the more should an approximation error for this observation contribute to the overall approximation error. The KLIC fulfills this requirement, since the deviations between process and its approximation for given y and x, as measured by the log-term in the integrand of (1.17), are weighted with the probability density $q(y\,|\,x)$ of actually observing y under x.

1.4.3. KLIC-Best Parameters and Models

The properties of the KLIC suggest the following rules:

- Given a model $\mu \in \mathcal{M}$, consider parameter $\theta^\mu \in \mathcal{Q}^\mu$ as better under design ξ than parameter $\tilde{\theta}^\mu \in \mathcal{Q}^\mu$, iff $\delta(\mu,\theta^\mu,\xi) < \delta\big(\mu,\tilde{\theta}^\mu,\xi\big)$.
- Given a model family, consider model $\mu \in \mathcal{M}$ as better under design ξ than model $\nu \in \mathcal{M}$, iff $\min_{\theta^\mu} \delta(\mu,\theta^\mu,\xi) < \min_{\theta^\nu} \delta(\nu,\theta^\nu,\xi)$.

The second rule means that models are compared based on the minimal KLIC they can achieve by varying their parameter. This rule is consistent with the proposal that "[...] the adequacy of a postulated model is measured by the minimum possible KLIC distance between the model and the true distribution" given by Sawa [220, Rule 2.1(i)]. Equipped with these rules we can now rigorously define minimum-mismatch models and parameters in the sense of (Q1.5) on p. 29.

Definition 1.9 (KLIC-Best Parameter, KLIC-Best Model)

(i) Parameter $\theta^{\mu}(\xi)$ of model $\mu \in \mathcal{M}$ is KLIC-BEST UNDER DESIGN ξ, iff

$$\theta^{\mu}(\xi) \in \operatorname*{argmin}_{\theta^{\mu} \in \mathcal{Q}^{\mu}} \delta(\mu, \theta^{\mu}, \xi). \tag{1.18}$$

(ii) Model $\mu(\xi)$ is KLIC-BEST UNDER DESIGN ξ, iff

$$\mu(\xi) \in \operatorname*{argmin}_{\mu \in \mathcal{M}} \min_{\theta^{\mu} \in \mathcal{Q}^{\mu}} \delta(\mu, \theta^{\mu}, \xi), \tag{1.19}$$

or equivalently, iff

$$\mu(\xi) \in \operatorname*{argmin}_{\mu \in \mathcal{M}} \delta(\mu, \theta^{\mu}(\xi), \xi). \tag{1.20}$$

These definitions are consistent generalizations of those given by Akaike [2], Sawa [220], and White [267] and by Vuong [261]. Parameters satisfying (i) are occasionally referred to as "pseudo-correct" or "pseudo-true" in literature. We prefer the term "KLIC-best" due to its suggestiveness. In situations where it introduces no ambiguity, we sometimes simply speak of "best" instead of KLIC-best models and parameters.

Corollary 1.10 (Consistency of KLIC-Best with Correct Models and Parameters)

(i) Suppose the model family is correct under design ξ. If model $\mu \in \mathcal{M}$ is KLIC-best under ξ, then it is also correct under ξ, and vice versa.

(ii) Suppose model $\mu \in \mathcal{M}$ is correct under design ξ. If parameter $\theta^{\mu} \in \mathcal{Q}^{\mu}$ is KLIC-best under ξ, then it is also a correct under ξ, and vice versa.

Proof Follows immediately from Prop. 1.8(ii) and Def. 1.9. □

The set of parameters (models) that are KLIC-best for a particular design are hence a superset of the set of parameters (models) that are correct under that design. Being KLIC-best under a design is thus a consistent generalization of being correct under a design. Under a model and a parameter that are KLIC-best under a design, the error in approximation (1.14) is minimal in the KLIC sense; if the parameter is correct under the design, the approximation is exact.

Existence and Uniqueness

The parameter domain $\mathcal{Q}^\mu$ is compact and under certain regularity conditions, the KLIC has a lower bound according to Prop. 1.8(i) and is continuous in θ^μ according to Prop. 1.8(iv). Thus, a best parameter for model μ follows from the extreme value theorem. If a best parameter exist for each model of the family, a best model exists since $\mathcal{M}$ is finite.

Recall from Def. 1.6 that correct models are defined only within correct model families, and that correct parameters are defined only within correct models. In contrast, *every model family contains a KLIC-best model, and every model has a KLIC-best parameter,* much in accordance with the intuitive understanding of the word “best”.

Additional conditions are required to ensure that they are unique, or, in the language of statistical inference, IDENTIFIABLE.

Definition 1.11 (Identifiability)

(i) A KLIC-best parameter of model $\mu \in \mathcal{M}$ is IDENTIFIABLE UNDER DESIGN ξ, iff it is *unique,* that is, iff there is exactly one minimizer of $\delta(\mu, \theta^\mu, \xi)$ with respect to $\theta^\mu \in \mathcal{Q}^\mu$.

(ii) The KLIC-best model is IDENTIFIABLE UNDER DESIGN ξ, iff it is *unique,* that is, iff there is exactly one minimizer of $\min_{\theta^\mu \in \mathcal{Q}^\mu} \delta(\mu, \theta^\mu, \xi)$ with respect to $\mu \in \mathcal{M}$.

Identifiability, particularly for KLIC-best parameters, is crucial for several results from statistical inference (discussed in detail in Chap. 2) that are central for this thesis. In practice, it is unknown whether identifiability holds, since the process and thus the KLIC $\delta(\mu, \theta^\mu, \xi)$ and its minimizers are unknown.

Techniques for empirically detecting non-identifiability are a well-established part of statistical inference, which we shall not consider here. Instead, we shall often assume that identifiability holds, silently implying that the adequate statistical methods are applied to detect violations of this assumption. A fairly general result concerning identifiability of KLIC-best parameters is White [267, Thm. 3.1], which comprises the “classic” results of Rothenberg [214, Thm. 1] and Bowden [38] as special cases.

1.4.4. KLIC-Based Assessment of Model Families

Summarizing the results of this section, (Q1.4) and (Q1.5) on p. 29 can be answered as follows:

(A1.4) The mismatch of model $\mu \in \mathcal{M}$ with parameter $\theta^\mu \in \mathcal{Q}^\mu$ under design ξ, that is, the error of approximation (1.14), is given by the KLIC $\delta(\mu, \theta^\mu, \xi)$.

(A1.5) The mismatch of model $\mu \in \mathcal{M}$ is minimal under parameters minimizing $\delta(\mu, \theta^\mu, \xi)$ with respect to $\theta^\mu \in \mathcal{Q}^\mu$, that is, under KLIC-best parameters.

The mismatch of a model family is minimal under models and corresponding parameters minimizing $\delta(\mu, \theta^\mu, \xi)$ with respect to $\mu \in \mathcal{M}$ and $\theta^\mu \in \mathcal{Q}^\mu$, that is, under KLIC-best models and corresponding KLIC-best parameters.

KLIC-best parameters (models) are theoretical concepts, characterizing the best approximations of the process available within a given model (family). In practice, the process and thus the KLIC-best parameters and models are *unknown*. We refer to our lack of knowledge about them as MODEL UNCERTAINTY and PARAMETER UNCERTAINTY, respectively. In principle, they can be reduced empirically. Suitable methods are discussed in the next chapter. By reducing the parameter and model uncertainty we *indirectly* reduce the structural uncertainty, down to the (possibly non-zero) limit defined by the KLIC-best model and parameter.

Alternatives to the KLIC

The KLIC is a well motivated and broadly accepted measure of discrepancy. Nevertheless, any member of the large class of so-called f-divergences might be used alternatively, as mentioned in the introduction of Sec. 1.4.2. One might ask for the reason of choosing the KLIC.

This thesis aims to improve certain optimal experimental design (OED) strategies which are based on (a) likelihood-based inference and (b) Bayesian inference. Both approaches are broadly accepted and have a well elaborated body of theory and widely available technically mature software. With their help, one can empirically identify KLIC-best parameters and models – but *only* them. If one chooses a different discrepancy measure than the KLIC, the established methods and implementations from (a) and (b) cannot longer be used.

The results of Liese and Vajda [175, Sec. VIII] indicate that it is possible to generalize methods from likelihood-based inference to general discrepancy measures based on f-divergences. These methods are, however, still subject to ongoing research and do not have gained the maturity of (a) and (b). Since the aim of this thesis is to improve OED strategies, and not to develop novel inference methods, we use the KLIC.

2. Statistical Inference in Families of Possibly Incorrect Models

> *Standard statistical practice ignores model uncertainty. Data analysts typically select a model from some class of models and then proceed as if the selected model had generated the data. This approach ignores the uncertainty in model selection, leading to over-confident inferences and decisions that are more risky than one thinks they are.*
>
> Hoeting et al. [119, Abstract]

Contents

This chapter is concerned with the question of statistical inference raised in (Q1.2) on p. 28: "Given the model family and the available data, what can we learn about the unknown process?" The theory provided by this chapter (and its special cases considered in the next chapter) form the basis for the optimal experimental design (OED) strategies considered in the second part of this thesis.

Statistical inference under the "classic" assumption that the model is correct has gained a certain maturity and is treated in many textbooks. This chapter, however, takes the view that "essentially, all models are wrong," (Box and Draper [44, p. 424]) and tries to avoid this common, but strong assumption.

Certain "classic" results of statistical inference can and have been generalized to possibly incorrect models, while for other results it remains unclear whether such generalizations are possible at all, or the necessary steps have not been taken yet. Some areas of inference in possibly incorrect models are still subject to active research. Essential results have emerged only in the last years and cannot be assumed to be commonly known.

This chapter focuses on "non-classic" results from statistical inference which do *not* rely on correctness assumptions, and points out in which areas such non-classic results are lacking. The discussions of this chapter help to clarify the explicit and implicit assumptions made by the OED strategies considered later.

Section 2.1 defines the considered scenario and formalizes the central questions of statistical inference. Section 2.2 introduces the likelihood and the related information matrices, central quantities for statistical inference, and Sec. 2.3 summarizes required essential concepts of estimation theory. After these preparations, we survey the major frequentist approach of maximum-likelihood estimation in Sec. 2.4. The alternative Bayesian approach to inference is examined in Sec. 2.5. Both are summarized and compared in Sec. 2.5.4.

Besides certain regularity conditions, this chapter makes no assumptions about the distributions specified by the process and the model family. In the next chapter, we focus on the special cases of normal distributions and (local) linearity, under which many of the complex expressions introduced here simplify to both intuitively appealing and computationally tractable forms. Much of the widely-used classic formulas and results can be found there.

2.1. Fundamental Assumptions and Questions

The following scenario summarizes the central assumptions of this chapter.

Scenario 2.1 (Statistical Inference)

(i) A process according to Def. 1.2 is given.

(ii) The function q characterizing the process is unknown.

(iii) Data is available from the process, consisting of observations from the observation domain $\mathcal{Y}$, obtained from a sequence of statistically independent experiments numbered $1, 2, \ldots$ performed under known conditions from the experimental domain $\mathcal{X}$.

For all $n \in \mathbb{N}$, the n-experiment exact design describing experiments 1 to n is denoted ξ_n, the data resulting from these experiments is denoted $d_n \in \mathcal{Y}^n$, and the corresponding sample is denoted $\mathcal{D}_n$.

The total number of replications of experiments under condition $x \in \operatorname{supp}(\xi_n)$ is $n\xi_n(x)$. The observation resulting from the j-th replicated experiment under x is denoted $y_j(x) \in \mathcal{Y}$.

(iv) A model family from Def. 1.3 is available for describing the process.

For each model $\mu \in \mathcal{M}$, the parameter domain $\mathcal{Q}^\mu$ is compact (and thus Lebesgue-measurable), and $p(y \mid x, \mu, \theta^\mu)$ is twice continuously differentiable and Lebesgue-measurable with respect to θ^μ for all $y \in \mathcal{Y}$ and all $x \in \operatorname{supp}(\xi_n)$ for all $n \in \mathbb{N}$.

(v) As a consequence of assumption (ii), it is not known whether the model family is correct for any of the designs $(\xi_n : n \in \mathbb{N})$ and the Kullback-Leibler information criterion (KLIC)-best models and KLIC-best parameters for these designs are unknown.

The remaining chapter considers this scenario without explicitly referring to it. This scenario is closely related to scenario 1.5. In contrast to the latter, it does not allow performing new experiments, but instead assumes in assumption (iii) that a *sequence* of experiments has already been performed. In practice, this sequence will be finite, yet for the examination of asymptotic behavior, we allow it to be infinite.

The additional assumptions of compactness, differentiability and measurability in assumption (iv) – quite common in statistical inference – permit using differential calculus and the extreme value theorem for parameter inference, and to define probability density functions (PDFs) over the parameter domain, which is required for Bayesian inference.

The sequence of exact designs $\xi_1, \xi_2, \dots$ defined in assumption (iii) are built upon each other by successively adding the condition of the next experiment. The index n of the design ξ_n thus serves two purposes: it indicates its position in the design sequence and specifies the number of experiments that is describes.

We continue to use the notation introduced in Sec. 1.3.1. In particular, $r_n(x) := n\xi_n(x)$ denotes the number of replications of the experiment under condition $x \in \operatorname{supp}(\xi_n)$, and $y_j(x) \in \mathcal{Y}$ denotes the observation resulting from replication no. $j \in \{1, \dots, r_n(x)\}$ of the experiment under $x \in \operatorname{supp}(\xi_n)$, for all $n \in \mathbb{N}$. Furthermore, $q(d_n \mid \xi_n)$ denotes the PDF of the sample $\mathcal{D}_n$, and $p(d_n \mid \xi_n, \mu, \theta^\mu)$ denotes the corresponding PDF specified by model μ with parameter θ^μ, for all $n \in \mathbb{N}$.

2.1.1. Central Questions

Suppose we agree to measure the discrepancy between model family members and the process with the KLIC. Then, the aim of learning something about the unknown process, informally stated in (Q1.2) on p. 28, can be stated as follows.

(Q2.1) *What can we learn about the unknown KLIC-best parameter(s) of a given model EMPIRICALLY, that is, based on the data d_n obtained under design ξ_n? How can we quantify the corresponding parameter uncertainty empirically?*

(Q2.2) *Based on (Q2.1), what can we learn empirically about the unknown PDF(s) of a given model corresponding to the KLIC-best parameter(s)?*

(Q2.3) *Based on (Q2.1), what can we learn empirically about the unknown KLIC-best model(s) of the family? How can we quantify the model uncertainty empirically?*

(Q2.4) *Based on (Q2.1) and (Q2.3), what can we learn empirically about the unknown PDF(s) of the model family corresponding to the KLIC-best model(s) and the KLIC-best parameter(s)?*

These rather general questions can be stated more precisely once one decides for a certain approach for statistical inference. Question (Q2.1) leads to problems of PARAMETER ESTIMATION and (Q2.3) to problems of MODEL SELECTION. If it is not possible in (Q2.3) to determine a unique best approximation for the KLIC-best model from the available data, one speaks of a MODEL DISCRIMINATION (MD) problem. Questions (Q2.2) and (Q2.4) are the basis for making predictions of future observations and derived quantities.

2.1.2. Assumptions on the Experimental Conditions

To answer these questions, the methods considered here impose additional assumptions on the experimental conditions. Consider the sequence of experiments from assumption (iii) of scenario 2.1 and let $x_{(i)} \in \mathcal{X}$ denote the condition of the i-the experiment, for all $i \in \mathbb{N}$.

Sampling Experiments from a Design

We are particularly interested in experiments that aim to approximate a certain design, for example one of the optimal designs introduced later. The sequence of experiments is sampled from design ξ, iff their conditions $x_{(1)}, x_{(2)}, \ldots$ are chosen such that the corresponding design sequence $\xi_1, \xi_2, \ldots$ converges to a limit design ξ, which may be non-exact. When speaking of the convergence of designs we speak of the convergence of normed (=probability) measures, considered in detail by Bilingsley [31].

Independently and Identically Distributed (IID) Observables

An important special case of experiments sampled from an design are experiments with independently and identically distributed (iid) observables. The observables of the experiments under conditions $x_{(1)}, x_{(2)}, \ldots$ are iid, iff they have the sample distribution for all $i, j \in \mathbb{N}$, that is, iff

$$q\left(y \,\middle|\, x_{(i)}\right) = q\left(y \,\middle|\, x_{(j)}\right) \forall y \in \mathcal{Y}. \tag{2.1}$$

Under iid observables, any of the designs ξ_n from assumption (iii) can without loss of generality (wlog) be considered as a one-point design putting full weight at the fixed experimental condition $x \in \left\{x_{(i)} : i \in \mathbb{N}\right\}$. Consequentially, parameters that are best or correct for ξ_n are independent of n, and the corresponding sample $\mathcal{D}_n$ is composed of n repetitions of the observable $\mathcal{Y}_x$.

Several important results from statistical inference are based on the assumption of iid observables. It is, however, usually not satisfied in scenarios 1.5 and 2.1, where experiments have been and can be performed under *arbitrary* conditions.

Experiments Sampled from a Design have Asymptotically IID Observables

The following argumentation shows that the assumption of IID observables is less restrictive than it seems.

Consider the experimental setting described by the n-experiment exact design ξ_n with $\operatorname{supp}(\xi_n) = \{x_1, \dots, x_s\}$. The number of replications of the experiment under condition $x \in \operatorname{supp}(\xi_n)$ is $r_n(x) := n\xi_n(x)$, see Def. 1.4. Without additional assumptions about the distributions of the corresponding observables $\mathcal{Y}_{x_1}, \dots, \mathcal{Y}_{x_s}$, the experiments are independently but not identically distributed (INID). Now summarize the experimental conditions of all experiments in the "extended experimental condition"

$$x' := \big[\underbrace{x_1^\top \,\dots\, x_1^\top}_{r_n(x_1)\text{ reps.}} \;\dots\; \underbrace{x_s^\top \,\dots\, x_s^\top}_{r_n(x_s)\text{ reps.}}\big]^\top, \tag{2.2}$$

and merge the corresponding observables into the "extended" observable

$$\mathcal{Y}'_{x'} := \big[\underbrace{\mathcal{Y}_{x_1}^\top \,\dots\, \mathcal{Y}_{x_1}^\top}_{r_n(x_1)\text{ reps.}} \;\dots\; \underbrace{\mathcal{Y}_{x_s}^\top \,\dots\, \mathcal{Y}_{x_s}^\top}_{r_n(x_s)\text{ reps.}}\big]^\top. \tag{2.3}$$

If the experiments under each condition x_i, $i \in \{1, \dots, s\}$ is replicated $c \cdot r_n(x_i)$ times, with $c \in \mathbb{N}$, the resulting sample consists of c *identically distributed* replications of the "extended" observable $\mathcal{Y}'_{x'}$. In other words, when sampling from an n-experiment exact design, the assumption of IID observables is automatically satisfied on a "higher level," whenever the total number of experiments is an integer multiple of n.

This argument can be generalized to experiments that are sampled from a particular design. If the design sequence $\xi_1, \xi_2, \dots$ converges to the *exact* design ξ with $s \in \mathbb{N}$ support points, the observables resulting from ξ_i are approximately IID (in the previously described generalized sense) if $i \gg s$. This asymptotic result remains even true if ξ is *non-exact,* since any non-exact design with s support points can be approximated arbitrary well by an exact i-experiment design if $i \gg s$.

In summary, if we sample statistically independent experiments from a (possibly non-exact) design, the resulting sequence of appropriately summarized observables is approximately IID in the large-sample limit or ASYMPTOTICALLY IID.

2.2. Likelihood and Information Matrices

The likelihood function and derived quantities based on the likelihood function are the basis for all statistical inference based on mathematical modeling.

Reid [211, Conclusion]

This section introduces and discusses central quantities of statistical inference: the likelihood function and so-called information matrices derived from it. They are used extensively in this and the next chapter dealing with statistical inference, and in Chaps. 4 to 5 dealing with optimal experimental design (OED). Further details can be found in the overview article of Reid [211] or in the standard works of Edwards [86] and Pawitan [200].

2.2.1. Likelihood

The likelihood is proportional to the joint probability density that a model and a parameter assign to a sample, evaluated at *given* data and a *given* design, considered as function of the model index and the parameter. In our case, the joint probability density is given by (1.13), leading to the following definition.

Definition 2.2 (Likelihood)

Given the data $d_n \in \mathcal{Y}^n$ obtained from $n \in \mathbb{N}$ experiments performed under the n-experiment exact design ξ_n, the LIKELIHOOD is the function $p(d_n \mid \xi_n, \mu, \theta^\mu)$ defined in (1.13) considered as function of $\mu \in \mathcal{M}$ and $\theta^\mu \in \mathcal{Q}^\mu$, with d_n and ξ_n being *fixed*.

By definition, the likelihood is a non-negative scalar function. Recall that $r_n(x) = n\xi_n(x)$ denotes total number of replications of experiments under condition $x \in \operatorname{supp}(\xi_n)$, and that $y_j(x) \in \mathcal{Y}$ denotes the observation resulting from corresponding j-th replication. Since the experiments are statistically independent, the likelihood has the representation

$$p(d_n \mid \xi_n, \mu, \theta^\mu) = \prod_{x \in \operatorname{supp}(\xi_n)} \prod_{j=1}^{r_n(x)} p\big(y_j(x) \,\big|\, x, \mu, \theta^\mu\big), \tag{2.4}$$

see (1.13). For the corresponding LOG-LIKELIHOOD

$$\ln p(d_n \mid \xi_n, \mu, \theta^\mu) = \sum_{x \in \operatorname{supp}(\xi_n)} \sum_{j=1}^{r_n(x)} \ln p\big(y_j(x) \,\big|\, x, \mu, \theta^\mu\big) \tag{2.5}$$

we use the convention $\ln 0 := -\infty$ so that it takes values in $[-\infty, \infty)$.

Some authors reserve the term "likelihood" for the case that the model (family) is correct for ξ_n, and use the term "pseudo-likelihood" if the model (family) is incorrect. For convenience, we do not make this distinction, but explicitly mention whenever we assume correctness.

The notation demands some explanation. The likelihood $p(d_n \mid \xi_n, \mu, \theta^\mu)$ is a function of μ and θ^μ, while the definition of that function depends parametrically on the data d_n and the design ξ_n. The (seemingly unusual) order of arguments was chosen to be compliant with the usual notation for conditional probabilities, which we require later in the context of Bayesian inference. At this point, neither of the quantities ξ_n, μ or θ^μ has a probabilistic interpretation, so that the vertical bar (|) is semantically equivalent to a comma.

The likelihood is meaningful only in relative terms. In general, it can be defined to be any function *proportional* to $p(d_n \mid \xi_n, \mu, \theta^\mu)$, with any non-negative factor of proportionality $c(d_n, \xi_n)$ that is independent of the model family. Inferences drawn from the likelihood are invariant to the choice of $c(\cdot)$. For more details we refer to the paper of Reid [211]. Definition 2.2 represents the common and convenient choice $c(d_n, \xi_n) \equiv 1$.

The likelihood of model μ and parameter θ^μ equals the probability density of obtaining the data d_n under design ξ_n, *if* the data follows the distribution specified by model μ with parameter θ^μ. In this sense, *the likelihood $p(d_n \mid \xi_n, \mu, \theta^\mu)$ can be regarded as the "plausibility" that the data d_n obtained under ξ_n originates from the distribution specified by model μ with parameter θ^μ.*

The likelihood has several attractive properties: it is *universal* in the sense that it is defined for any model, it is *simple,* its definition requiring nothing more than a model and data, and is *invariant under bijective reparameterizations*. Furthermore, one can show that the likelihood is a *maximally condensed representation of the information* in the data from which nothing can be omitted without loss, a property called "minimal sufficiency." These properties are main reasons of the outstanding role of the likelihood for statistical inference. Details are available in several textbooks, for example in that of Pawitan [200, Chaps. 2 and 3].

2.2.2. Information Matrices

For models that are differentiable with respect to the parameter one can define so-called information matrices, which are matrix-valued functions capturing some kind of "information" contained in the gradient and the curvature of the likelihood function. Information matrices play a central role for quantifying the parameter uncertainty in both classic maximum-likelihood estimation and Bayesian inference.

In the following, let ∇ and ∇^2 denote the gradient and the Hessian differential operator, respectively, with respect to the parameter vector θ^μ, and let $\mathbb{C}\left[\cdot\right]$ denote the covariance of a random vector.

Expected Information Matrices

Definition 2.3 (Expected Information Matrices)

Let ξ be a (possibly non-exact) design. The expected Hessian-based Fisher information of model $\mu \in \mathcal{M}$ is

$$\tilde{\boldsymbol{F}}^\mu(\theta^\mu, \xi) := -\sum_{x \in \operatorname{supp}(\xi)} \xi(x)\, \mathbb{E}\left[\nabla^2 \ln p(\mathcal{Y}_x \,|\, x, \mu, \theta^\mu)\right], \tag{2.6}$$

its expected gradient-based Fisher information is

$$\tilde{\boldsymbol{G}}^\mu(\theta^\mu, \xi) := \sum_{x \in \operatorname{supp}(\xi)} \xi(x)\, \mathbb{C}\left[\nabla \ln p(\mathcal{Y}_x \,|\, x, \mu, \theta^\mu)\right], \tag{2.7}$$

and its expected sandwich information is

$$\tilde{\boldsymbol{S}}^\mu(\theta^\mu, \xi) := \left(\tilde{\boldsymbol{F}}^\mu(\theta^\mu, \xi)\right)^{-1} \tilde{\boldsymbol{G}}(\theta^\mu, \xi) \left(\tilde{\boldsymbol{F}}^\mu(\theta^\mu, \xi)\right)^{-1}, \tag{2.8}$$

supposed the right-hand sides exists.

By definition, all introduced matrices are $n_{\theta^\mu} \times n_{\theta^\mu}$ and symmetric positive semi-definite (spsd). The involved expectations and covariances are calculated using the process-specified probability density functions (pdfs) $q(\cdot \,|\, x)$ for the observables $\mathcal{Y}_x$.

If evaluated at an n-experiment *exact* design ξ_n, comparison with Def. 2.2

reveals that they simplify according to

$$\tilde{\boldsymbol{F}}^{\mu}(\theta^{\mu}, \xi_n) = -\tfrac{1}{n}\,\mathbb{E}\left[\nabla^2 \ln p(\mathcal{D}_n \mid \xi_n, \mu, \theta^{\mu})\right], \text{ and} \tag{2.9}$$

$$\tilde{\boldsymbol{G}}^{\mu}(\theta^{\mu}, \xi_n) = \tfrac{1}{n}\,\mathbb{C}\left[\nabla \ln p(\mathcal{D}_n \mid \xi_n, \mu, \theta^{\mu})\right]. \tag{2.10}$$

Since the log-likelihood is of $\mathcal{O}(n)$, the expected information matrices are of $\mathcal{O}(1)$.

Relation to Identifiability of KLIC-Best Parameters

Kullback-Leibler information criterion (KLIC)-KLIC-best parameters are minimizers of the KLIC. Under certain regularity conditions, their identifiability can be related to the curvature of the KLIC in their vicinity.

Theorem 2.4 (Identifiability of KLIC-Best Parameters, White [267, Thm. 3.1])

Suppose that for model $\mu \in \mathcal{M}$ and some design ξ, the KLIC $\delta(\mu, \theta^{\mu}, \xi)$ and $\tilde{\boldsymbol{F}}^{\mu}(\theta^{\mu}, \xi)$ exist and are continuous in θ^{μ}, for all $\theta^{\mu} \in \mathcal{Q}^{\mu}$. Let $\bar{\theta}^{\mu}$ be an interior point of $\mathcal{Q}^{\mu}$. Then, under regularity conditions, the following statements hold.

(i) If $\bar{\theta}^{\mu}$ is an identifiable KLIC-best parameter under ξ and $\tilde{\boldsymbol{F}}^{\mu}(\theta^{\mu}, \xi)$ has constant rank for all θ^{μ} from an open neighborhood of $\bar{\theta}^{\mu}$, then $\tilde{\boldsymbol{F}}^{\mu}\left(\bar{\theta}^{\mu}, \xi\right)$ has full rank (and is thus invertible).

(ii) If $\bar{\theta}^{\mu}$ is a KLIC-best parameter under ξ and $\tilde{\boldsymbol{F}}^{\mu}\left(\bar{\theta}^{\mu}, \xi\right)$ has full rank (and is thus invertible), then $\bar{\theta}^{\mu}$ is identifiable.

This is a straightforward generalization of a result of White [267, Thm. 3.1] to experimental settings described by design ξ. The mentioned regularity conditions are listed and discussed there in detail. It comprises as special cases the “classic” results of Rothenberg [214, Thm. 1] and Bowden [38], which assume a correct model.

Item (ii) clarifies that a full-rank matrix $\tilde{\boldsymbol{F}}^{\mu}\left(\bar{\theta}^{\mu}, \xi\right)$ is a necessary condition for identifiability of $\bar{\theta}^{\mu}$. In several of the subsequent theorems, identifiability is assumed anyhow. To ensure invertibility of $\tilde{\boldsymbol{F}}^{\mu}\left(\bar{\theta}^{\mu}, \xi\right)$, they hence only have to add the assumption that $\tilde{\boldsymbol{F}}^{\mu}(\theta^{\mu}, \xi)$ has constant rank in vicinity of $\bar{\theta}^{\mu}$.

Expected Information Matrices under Correctness

In classic likelihood theory, the term "expected Fisher information" sometimes refers to $\tilde{\boldsymbol{F}}^{\mu}$ and sometimes to $\tilde{\boldsymbol{G}}^{\mu}$. The next theorem states that both matrices are in fact equal, if the classic correctness assumption is met.

Theorem 2.5 (Information Matrix Equality)

Suppose that model $\mu \in \mathcal{M}$ is correct for the exact design ξ and that

(i) the model has an identifiable correct parameter $\bar{\theta}^{\mu}$ under ξ,

(ii) $\bar{\theta}^{\mu}$ is an interior point of the parameter domain $\mathcal{Q}^{\mu}$, and

(iii) $\tilde{\boldsymbol{F}}^{\mu}(\theta^{\mu}, \xi)$ has constant rank for all θ^{μ} in an open neighborhood of $\bar{\theta}^{\mu}$.

Then, under some regularity conditions,

$$\tilde{\boldsymbol{S}}^{\mu}(\bar{\theta}^{\mu}, \xi) = \left(\tilde{\boldsymbol{F}}^{\mu}(\bar{\theta}^{\mu}, \xi)\right)^{-1} = \left(\tilde{\boldsymbol{G}}^{\mu}(\bar{\theta}^{\mu}, \xi)\right)^{-1}. \tag{2.11}$$

Proof Given, for example, by Vuong [260, Lem. 2.1(ii)]. □

The regularity conditions are listed and discussed in detail by Vuong [260, Asmps. 1–4]. They ensure, among others, that the KLIC and the expected information matrices exists and are continuous with respect to the parameter, that the KLIC-best parameter exists, and that differentiation under the integral sign of the expectation is possible. According to Vuong [260, Lem. 5.1], they are met under the common assumptions of normality and (local) linearity which we consider in Chap. 3 and for all our numerical results in Chap. 9.

The matrix $\tilde{\boldsymbol{F}}^{\mu}$ is based on the Hessian of the log-likelihood, while the matrix $\tilde{\boldsymbol{G}}^{\mu}$ is defined in terms of its gradient. The second equality in (2.11) thus allows to conveniently replace second derivatives by first derivatives. For this reason, the simpler form $\tilde{\boldsymbol{G}}^{\mu}$ is often preferred in classic likelihood theory, where the model is assumed to be correct for ξ_n. Since we explicitly do *not* want to make this assumption in general, we have to distinguish between $\tilde{\boldsymbol{F}}^{\mu}$ and $\tilde{\boldsymbol{G}}^{\mu}$.

Empirical Hessian-Based Fisher Information

The expectation and covariance appearing in the expected information matrices are calculated over the PDFs $q(\cdot \mid x)$ specified by the process. Therefore, *the* *expected information matrices are unknown in practice (scenario 2.1).* In contrast, the *empirical* Hessian-based Fisher information defined in the following are independent of the unknown process, but in turn depend on the observed data.

Definition 2.6 (Empirical Hessian-Based Fisher Information Matrix)

Let d_n be the data obtained under n-experiment exact design ξ_n, and let $r_n(x) := n\xi_n(x)$ denote the number of replications of the experiment under $x \in \operatorname{supp}(\xi_n)$. The empirical Hessian-based Fisher information matrix of model $\mu \in \mathcal{M}$ is

$$\begin{aligned} \boldsymbol{F}_n^{\mu}(\theta^{\mu}, d_n, \xi_n) &:= -\tfrac{1}{n}\nabla^2 \ln p(d_n \,|\, \xi_n, \mu, \theta^{\mu}) \\ &= -\tfrac{1}{n}\sum_{x \in \operatorname{supp}(\xi_n)} \sum_{j=1}^{r_n(x)} \nabla^2 \ln p\big(y_j(x) \,\big|\, x, \mu, \theta^{\mu}\big). \end{aligned} \tag{2.12}$$

The definition implies that $\boldsymbol{F}_n^{\mu}$ is a SPSD $n_{\theta^{\mu}} \times n_{\theta^{\mu}}$ matrix of $\mathcal{O}(1)$ with respect to n. It is defined as empirical counterpart of $\tilde{\boldsymbol{F}}^{\mu}$ from (2.6) with a law of large numbers in mind, which states that the sum $\frac{1}{n}\sum_{i=1}^{n}(\mathcal{U}_i - \mathbb{E}[\mathcal{U}_i])$ converges to zero in some sense as n goes to infinity, supposed that the random vectors $\mathcal{U}_i$ satisfy various regularity conditions. Note that there is no direct empirical counterpart of $\tilde{\boldsymbol{G}}^{\mu}$ in this sense, and thus also not of $\tilde{\boldsymbol{S}}^{\mu}$.

2.3. Basics of Estimation Theory

This section examines properties of estimators, that is, data-based approximations for unknown quantities, with a focus on the assessment of their quality.

2.3.1. Estimation Problems

For the sake of generality, let us consider a generalized type of model: Let $\mathcal{W}$ be an non-empty subset of $\mathbb{R}^u$ and $p(y\,|\,x, w)$ be a probability density function (PDF) in the argument y from the observation domain $\mathcal{Y}$ for all experimental conditions x from the experimental domain $\mathcal{X}$ and all $w \in \mathcal{W}$. We shall consider the indexed family $(p(\cdot\,|\,\cdot, w) : w \in \mathcal{W})$ as model in this section. No confusion with the "model" from Def. 1.3 shall occur.

An estimator is a function of the data that aims to approximate a (typically unknown) quantity, the estimand.

Definition 2.7 (Estimand, Estimator, Estimate)

Let $\mathcal{T}$ be a non-empty subset of $\mathbb{R}^{\nu}$.

(i) An ESTIMAND is a value $\bar{t} := f(\bar{w}) \in \mathcal{T}$, where $f\colon \mathcal{W} \mapsto \mathcal{T}$ is a function that is independent of any particular data and $\bar{w}$ is a fixed, non-random element from $\mathcal{W}$.

Let $\mathcal{D}_n$ be the sample corresponding to n experiments under the exact design ξ_n, and let the data d_n be a realization of $\mathcal{D}_n$.

(ii) An ESTIMATOR for an estimand $\bar{t}$ is a $\mathcal{T}$-valued function t defined over the domains of the sample and the corresponding design. Given $\mathcal{D}_n$ and ξ_n, the estimator is a $\mathcal{T}$-valued random variable written $\hat{\mathcal{T}}_n := t(\mathcal{D}_n, \xi_n)$.

(iii) An ESTIMATE $\hat{t}_n := t(d_n, \xi_n)$ is a realization of an estimator $\hat{\mathcal{T}}_n$, that is, a value that it takes under the particular data d_n.

An estimand is some quantity of interest which is unknown – otherwise there would be no need to approximate it from data. The function f introduces some flexibility in the choice of the quantity of interest. The case that we are interested in a particular value in $\mathcal{W}$ corresponds to the special case $f = id_{\mathcal{W}}$.

An estimate is considered as an approximation for the estimand, $\hat{t}_n \approx \bar{t}$. The problem of finding an estimate from a given design and corresponding data is called an ESTIMATION PROBLEM. Any $\mathcal{T}$-valued function of data and design, whatever simple or crude, might be considered as an estimator. A reasonable measure for the quality of this approximation has to take into account the distribution of the estimator $\hat{\mathcal{T}}_n$. It is convenient to distinguish between the ACCURACY and the PRECISION of the approximation: the smaller the discrepancy between estimator and estimand "in the average", the more accurate is the approximation; the smaller the amount of random fluctuations around this mean, the more precise is it.

If the domain $\mathcal{W}$ can be equipped with the Euclidean norm $\|\cdot\|_2$, a popular choice for the estimator quality is the MEAN-SQUARED ERROR

$$\mathbb{E}\left[\left\|\hat{\mathcal{T}}_n - \bar{t}\right\|_2^2\right] = \left\|\mathbb{E}\left[\hat{\mathcal{T}}_n\right] - \bar{t}\right\|_2^2 + \operatorname{tr}\mathbb{C}\left[\hat{\mathcal{T}}_n\right], \tag{2.13}$$

where the equality is due to (B.1). The smaller the first term on the right-hand side, a l_2 norm on the BIAS $\mathbb{E}\left[\hat{\mathcal{T}}_n\right] - \bar{t}$, the higher the ACCURACY of the estimator.

The lower the second term on the right-hand side, the trace of the estimator covariance, the higher its PRECISION.

Other measures for estimator quality than the mean-squared error lead to other measures of accuracy and precision. The expected *absolute* loss, for example, leads to median-unbiasedness estimators and median absolute deviation. For the considerations in this thesis, however, the mean-squared error suffices.

2.3.2. Estimator Accuracy: Unbiasedness and Consistency

A common notion, motivated by (2.13), is to consider an estimator as "perfectly accurate" if it coincides with the estimand in the (arithmetical) average.

Definition 2.8 (Unbiased Estimator)

The estimator $\hat{\mathcal{T}}_n$ is UNBIASED FOR ESTIMAND $\bar{t}$, iff its expectation exists and $\mathbb{E}\left[\hat{\mathcal{T}}_n\right] = \bar{t}$.

Unbiasedness is a desirable property for an estimator, but is, however, not invariant to nonlinear reparameterizations. An unbiased estimator for the variance, for example, is biased for the standard deviation and vice versa. Furthermore, unbiased estimators often perform badly in other terms of estimator quality. In many cases, it suffices to assess an estimator based on its accuracy in the large-sample limit only.

Definition 2.9 (Asymptotically Unbiased and Consistent Estimators)

The sequence of estimators $\hat{\mathcal{T}}_1, \hat{\mathcal{T}}_2, \ldots$ is

(i) ASYMPTOTICALLY UNBIASED FOR $\bar{t}$, iff their expectations exists for all $n \geqslant s \in \mathbb{N}$ and $\lim_{n\to\infty} \mathbb{E}\left[\hat{\mathcal{T}}_n\right] = \bar{t}$,

(ii) WEAKLY CONSISTENT FOR $\bar{t}$, iff $\hat{\mathcal{T}}_n \xrightarrow{\text{p}} \bar{t}$, for $n \to \infty$, and

(iii) STRONGLY CONSISTENT FOR $\bar{t}$, iff $\hat{\mathcal{T}}_n \xrightarrow{\text{a.s.}} \bar{t}$, for $n \to \infty$.

The symbols $\xrightarrow{\text{p}}$ and $\xrightarrow{\text{a.s.}}$ denote convergence in probability and almost sure convergence, respectively, see Def. B.2 on p. 293. In cases where it introduces no ambiguity, we adopt the usual terminology and speak of "the estimator $\hat{\mathcal{T}}_n$"

instead of "the sequence of estimators $\hat{\mathcal{T}}_1, \hat{\mathcal{T}}_2, \ldots$" As laid out in Appendix B, strong consistency implies weak consistency and asymptotic unbiasedness, but weak consistency does not imply asymptotic unbiasedness.

If the domain $\mathcal{T}$ is non-convex, the expectation of the estimator $\mathbb{E}\left[\hat{\mathcal{T}}_n\right]$ may take a value *outside* of $\mathcal{T}$ in $\mathbb{R}^\nu$. Such a value might not have an interpretation in terms of the underlying model and might hence not be a reasonable approximation for the estimand. For non-convex $\mathcal{T}$, the concepts of (asymptotic) unbiasedness might hence be meaningless. The notions of weak and strong consistency, however, remain sensible.

This limitation particularly applies to all *discrete* estimators and estimands, since all discrete subsets (e.g. the model index set $\mathcal{M}$) of $\mathbb{R}^\nu$ are non-convex. Furthermore, weak and strong consistency are identical for discrete $\mathcal{T}$, since almost sure convergence and convergence in probability are identical for discrete random variables, see Def. B.2 and the subsequent comments.

2.3.3. Precision of Parameter Estimators: Efficiency

Suppose the aim is to estimate the Kullback-Leibler information criterion (KLIC)-best parameter of model $\mu \in \mathcal{M}$ under a given design. According to (2.13), we can in this case (the parameter domain is Euclidean) measure the precision of an estimator based on its covariance. For this important case one can state a general lower bound for the covariance of a large class of estimators in terms of the expected sandwich information (2.8).

Theorem 2.10 (Cramér-Rao Inequality for Possibly Incorrect Models)

Let ξ_n be a n-experiment exact design and assume that

(i) the model has an identifiable KLIC-best parameter $\bar{\theta}_n^\mu$ under ξ_n,

(ii) $\bar{\theta}_n^\mu$ is an interior point of the parameter domain $\mathcal{Q}^\mu$,

(iii) the expected Hessian-based Fisher information $\tilde{\boldsymbol{F}}^\mu(\theta^\mu, \xi_n)$ has constant rank for all θ^μ in an open neighborhood of $\bar{\theta}_n^\mu$, and

(iv) $\hat{\mathcal{Q}}_n^\mu$ is an unbiased estimator of $\bar{\theta}_n^\mu$ whose covariance matrix $\mathbb{C}\left[\hat{\mathcal{Q}}_n^\mu\right]$ exists and has full rank.

Then, under some regularity conditions,

$$\mathbb{C}\left[\hat{\mathcal{Q}}_n^\mu\right] \geqslant \tfrac{1}{n}\tilde{\mathbf{S}}^\mu\left(\bar{\theta}_n^\mu, \xi_n\right), \qquad (2.14)$$

where the matrix inequality is meant in the sense that the left-hand side minus the right-hand side is a positive semi-definite matrix. If the assumptions hold in the limit $n \to \infty$, then also the inequality holds asymptotically.

This inequality is a generalization of the eponymous theorem of Cramér [75] and Rao [210] (explicitly stated in the next section) for possibly incorrect models. The right-hand side of (2.14) is also referred to as the Cramér-Rao lower bound. This theorem is a simplified variant of that given by Vuong [260, Thm. 4.1]. The regularity conditions are the same required for Thm. 2.5.

This theorem establishes a lower bound to the unknown covariance of any unbiased estimator for the KLIC-best parameter in terms model-based quantities. *Remarkably, this bound also holds for incorrect models.* An estimator which always meets this bound is termed EFFICIENT.

Definition 2.11 (Efficient Estimator)

An estimator $\hat{\mathcal{Q}}_n^\mu$ is EFFICIENT, iff equality in (2.14) holds regardless which value $\bar{\theta}_n^\mu$ takes in $\mathcal{Q}^\mu$.

It is practically most crucial that equality in (2.14) holds *regardless* of the value of KLIC-best parameter. Since the latter is unknown in practice, it would be of little practical help to know that an estimator is efficient for a particular value only.

The mean-squared error (2.13) of an unbiased estimator is

$$\underbrace{\left\| \mathbb{E}\left[\hat{\mathcal{Q}}_n^\mu\right] - \bar{\theta}_n^\mu \right\|_2^2}_{=0 \text{ due to unbiasedness}} + \operatorname{tr}\left(\mathbb{C}\left[\hat{\mathcal{Q}}_n^\mu\right]\right) = \tfrac{1}{n} \operatorname{tr}\left(\tilde{\mathbf{S}}^\mu\left(\bar{\theta}_n^\mu, \xi_n\right)\right). \tag{2.15}$$

Efficient estimators are unbiased by definition. Taking this equality together with inequality (2.14) tells us that *efficient estimators are the best possible estimators in the sense of the mean-squared error.*

In practice (scenario 2.1), the Cramér-Rao lower bound is unknown, since the matrix $\tilde{\mathbf{S}}^\mu$ as well as the KLIC-best parameter $\bar{\theta}_n^\mu$ depend on the unknown process. That is, if we can show that an estimator is efficient, we know that it is best possible estimator in terms of the mean-squared error, but we still cannot quantify its precision.

The Special Case of a Correct Model

If the model is correct under the considered design, Thm. 2.10 reduces to the well-known classic Cramér-Rao inequality which states that the estimator covariance is bounded from below by the inverse of the expected Fisher information at the correct parameter for that design.

Theorem 2.12 (Classic Cramér-Rao Inequality)

Consider the same setting as in Thm. 2.10. In addition, assume the model is correct for design ξ_n and let $\bar{\theta}_n^\mu$ denote the correct parameter. Then,

$$\mathbb{C}\left[\hat{Q}_n^\mu\right] \geqslant \tfrac{1}{n}\tilde{\boldsymbol{F}}^{-1}\left(\bar{\theta}_n^\mu, \xi_n\right) \overset{(2.11)}{=} \tfrac{1}{n}\tilde{\boldsymbol{G}}^{-1}\left(\bar{\theta}_n^\mu, \xi_n\right) \tag{2.16}$$

where the matrix inequality is meant as previously.

Proof Originally given by Cramér [75] and Rao [210]. Is a corollary of Thms. 2.5 and 2.10.□

2.4. Maximum-Likelihood Estimation

> *Sampling experiments [...] have shown, however, that the maximum likelihood method produces acceptable estimates in many situations. Whereas better methods may be available for specific cases, a powerful argument for the use the maximum likelihood method is the generality and relative ease of application.*
>
> Bard [23]

Maximum-likelihood estimation is a major branch of frequentist inference and a popular choice of practitioners thanks to its generality and simplicity. In particular, it is an essential ingredient for the frequentist optimal experimental design (OED) strategies considered in Chap. 4.

A classic assumption of maximum-likelihood theory, made in many publications and textbooks, is that the considered model is correct. We want to avoid this comparably strong assumption as far as possibly. We hence survey less familiar non-classic results that allow incorrect models, but contain the classic results as special case.

The corresponding branch of non-classic maximum-likelihood theory – called "quasi-likelihood" by some – goes back to the works of Huber [126] and White [267] and Burguete, Gallant, and Souza [58]. An extensive source about likelihood theory is the book of Pawitan [200].

2.4.1. Basic Approach

Recall from Sec. 2.2 that the likelihood of a given parameter can be interpreted as the plausibility that the data originates from the distribution specified by the model with that parameter. This interpretation suggests to use a maximizer of the likelihood – a maximum-likelihood estimate (MLE) – to approximate the Kullback-Leibler information criterion (KLIC)-best parameter.

Definition 2.13 (Maximum Likelihood Estimate/Estimator (MLE))

(i) A parameter $\hat{\theta}^{\mu}_{n} := \hat{\theta}^{\mu}_{n}(d_n, \xi_n)$ is a PARAMETER MAXIMUM-LIKELIHOOD ESTIMATE (PMLE), iff

$$\hat{\theta}^{\mu}_{n}(d_n, \xi_n) \in \operatorname*{argmin}_{\theta^{\mu} \in \mathfrak{Q}^{\mu}} p(d_n \mid \xi_n, \mu, \theta^{\mu}). \tag{2.17}$$

The corresponding estimator $\hat{\mathcal{Q}}^{\mu} := \hat{\theta}^{\mu}_{n}(\mathcal{D}_n, \xi_n)$ is a $\mathcal{Q}^{\mu}$-valued random variable, its probability density function (PDF) is denoted $p\big(\hat{\theta}^{\mu} \,\big|\, \mu, \xi_n\big)$.

(ii) A model $\hat{\mu}_n := \hat{\mu}_n(d_n, \xi_n)$ is a MODEL MAXIMUM-LIKELIHOOD ESTIMATE (MMLE), iff

$$\hat{\mu}_n(d_n, \xi_n) \in \operatorname*{argmin}_{\mu \in \mathcal{M}} \max_{\theta^{\mu} \in \mathcal{Q}^{\mu}} p(d_n \,|\, \xi_n, \mu, \theta^{\mu}), \tag{2.18}$$

or equivalently, iff

$$\hat{\mu}_n(d_n, \xi_n) \in \operatorname*{argmin}_{\mu \in \mathcal{M}} p\big(d_n \,\big|\, \xi_n, \mu, \hat{\theta}^{\mu}_{n}(d_n, \xi_n)\big). \tag{2.19}$$

The corresponding estimator $\hat{\mathcal{M}}_n := \hat{\mu}_n(\mathcal{D}_n, \xi_n)$ is a $\mathcal{M}$-valued random variable, its probability mass function (PMF) is denoted $p(\hat{\mu} \,|\, \xi_n)$.

It follows from assumption (iv) of scenario 2.1 that the likelihood (2.4) is continuous with respect to the parameter. Since the parameter domain is compact, the existence of PMLEs follows from the extreme value theorem. If PMLEs exists for all models $\mu \in \mathcal{M}$, a MMLE exists since $\mathcal{M}$ is finite. In general, parameter and MMLEs are not unique.

A PMLE is considered as a point approximation for the corresponding unknown KLIC-best parameter, $\bar{\theta}^{\mu}_{n} \approx \hat{\theta}^{\mu}_{n}$. The uncertainty about the unknown KLIC-best parameter is determined by the density $p\big(\hat{\theta}^{\mu} \,\big|\, \mu, \xi_n\big)$ of the PMLE. The more it "accumulates" in vicinity of $\bar{\theta}^{\mu}_{n}$, the smaller the parameter uncertainty. Various uncertainty quantifications can be derived from $p\big(\hat{\theta}^{\mu} \,\big|\, \mu, \xi_n\big)$, for example confidence intervals, the mean-squared error (2.13), or similar measures for the quality of the approximation $\bar{\theta}^{\mu}_{n} \approx \hat{\theta}^{\mu}_{n}$.

Likewise, a MMLE is considered as a point approximation for the corresponding unknown best model, $\bar{\mu}_n \approx \hat{\mu}_n$. Being a discrete index set, the domain $\mathcal{M}$ of a model maximum-likelihood estimator (MMLE) has no associated concept of "closeness" whatsoever, which complicates the quantification of model uncertainty. Appealing concepts like confidence regions or mean-squared error, for example, cannot be reasonably defined for MMLEs. Consequentially, model uncertainty is usually not expressed in terms of the density $p(\hat{\mu} \,|\, \xi_n)$ of the MMLE. In fact, there is no single commonly agreed "standard" approach in frequentist inference for quantifying model uncertainty.

An important argument for maximum-likelihood estimators (MLEs) is their asymptotic behavior. It justifies to use them as approximation for the unknown KLIC-best quantities in large samples, allows to empirically quantify the associated uncertainty, and to make robust predictions. The central asymptotic results are considered in the subsequent Secs. 2.4.2 to 2.4.4, their practical application for answering (Q2.1)–(Q2.4) is discussed in Sec. 2.4.5.

2.4.2. Consistency and Asymptotic Normality of Parameter MLEs

Suppose the experiments are sampled from design ξ, such that $\xi_1, \xi_2, \dots$ converges to ξ. Let $\bar{\theta}^\mu$ be a KLIC-best parameter of model $\mu \in \mathcal{M}$ under ξ. Under regularity conditions, parameter maximum-likelihood estimators (PMLEs) are strongly consistent estimators of the KLIC-best parameter under the limit design,

$$\hat{\mathcal{Q}}_n^\mu \xrightarrow{\text{a.s.}} \bar{\theta}^\mu, \text{ for } n \to \infty, \tag{2.20}$$

Furthermore, they are under additional regularity assumptions normally distributed in the large-sample limit,

$$\hat{\mathcal{Q}}_n^\mu \overset{\infty}{\sim} \mathcal{N}\big(\bar{\theta}^\mu, n^{-1} \tilde{\mathbf{S}}^\mu\big(\bar{\theta}^\mu, \xi\big)\big), \tag{2.21}$$

with a mean given by the KLIC-best parameter of the limit design, and a covariance given by the corresponding expected sandwich information divided by the sample size. The mentioned regularity conditions are discussed later.

Since ξ_n converges to ξ, also $\bar{\theta}_n^\mu$ converges to $\bar{\theta}^\mu$, so that (2.21) can also be stated as

$$\hat{\mathcal{Q}}_n^\mu \overset{\infty}{\sim} \mathcal{N}\big(\bar{\theta}_n^\mu, n^{-1} \tilde{\mathbf{S}}^\mu\big(\bar{\theta}_n^\mu, \xi_n\big)\big). \tag{2.22}$$

Relation (2.20) implies that PMLEs are asymptotically unbiased, so that the generalized Cramér-Rao inequality (Thm. 2.10) applies. Comparing the asymptotic covariance from (2.22) with the Cramér-Rao lower bound from (2.14) reveals that *PMLEs are asymptotically efficient.* Hence, *PMLEs are in the large-sample limit the most accurate and most precise estimators of the KLIC-best parameter in terms of the mean-squared error.*

References and Historical Remarks

The first proof of consistency and asymptotic normality of what we call PMLEs is given by Doob [84]. A frequently cited variant is the proof given by Wald [262]. They both make the classic assumptions of a correct model and independently and identically distributed (IID) observables. Proofs under these assumptions can nowadays be found in most relevant textbooks, for example in that of Lehmann and Casella [170, Sec. 6.3].

A series of publications examines PMLEs in possibly incorrect models, but still assume IID observables. In that setting, Huber [126] seems to be the first to state sufficient conditions for consistency and asymptotic normality. While being very general, his conditions are practically difficult to verify. The seminal work of White [267] provides comparably simpler conditions. They can nowadays also be found in some textbooks, for example in that of Pawitan [200, Sec. 13.4]. White's paper is likely to be one of the most influential publications concerning maximum-likelihood estimation in possibly incorrect models. Some of the numerous follow-up publications are compiled in the book of Fomby and Hill [101]. White [265] himself elaborated his ideas further in his book.

These publications limit their considerations to IID observables. As argued in Sec. 2.1.2, one can hope that their results can be generalized to independently but not identically distributed (INID) experiments that are sampled from an experimental design, as assumed here. Such generalizations in fact exist, yet are little known. Based on the work of Souza [236] and Gallant and Holly [104], Burguete, Gallant, and Souza [58, Thms. 2 and 4] prove strong consistency and asymptotic normality for a broad class of estimators under a rather general set of sufficient conditions. Their results comprise (2.20) and (2.21) in possibly incorrect models and INID experiments sampled from a design as special case.

Regularity Conditions

The full list of conditions sufficient for (2.20) and (2.21) in the considered scenario is rather long. They are generalizations of those listed by White [267, A1–A6], and special cases of those given by Burguete, Gallant, and Souza [58, Asmps. 1–6]. Besides certain technicalities, the strong consistency (2.20) requires that

(a) the experiments are sampled from design ξ, as already mentioned,

(b) the KLIC $\delta(\mu, \theta^{\mu}, \xi)$ exists and is continuous in θ^{μ}, for all $\theta^{\mu} \in \mathcal{Q}^{\mu}$, and

(c) the model has an identifiable KLIC-best parameter $\bar{\theta}^{\mu} \in \mathcal{Q}^{\mu}$ under ξ.

In addition, the asymptotic normality (2.21) requires – again omitting certain technicalities – that

(d) $\bar{\theta}^\mu$ is an interior point of $\mathcal{Q}^\mu$,

(e) $\tilde{\boldsymbol{F}}^\mu(\theta^\mu, \xi)$ exist and is continuous in θ^μ, for all $\theta^\mu \in \mathcal{Q}^\mu$,

(f) $\tilde{\boldsymbol{F}}^\mu(\theta^\mu, \xi)$ has constant rank for all θ^μ in an open neighborhood of $\bar{\theta}^\mu$,

(g) $\tilde{\boldsymbol{G}}^\mu(\theta^\mu, \xi)$ exist and is continuous in θ^μ, for all $\theta^\mu \in \mathcal{Q}^\mu$, and

(h) $\tilde{\boldsymbol{G}}^\mu\left(\bar{\theta}^\mu, \xi\right)$ has full rank.

This list is not meant to be exhaustive. For details, we refer to the previously cited original works.

Condition (a) ensures that there is asymptotically enough "repetition" in the sample, which is required for applying central limit theorems and laws of large numbers. Together with the assumed compactness of $\mathcal{Q}^\mu$, condition (b) ensures the existence of a KLIC-best parameter. Condition (c) makes the inference problem well posed. Condition (d) ensures the existence of an open neighborhood of $\bar{\theta}^\mu$, required for local Taylor series approximations and certain convergence theorems. Condition (e) allows to apply certain mean value theorems and uniform laws of large numbers. Together with condition (c), condition (f) guarantees that $\tilde{\boldsymbol{F}}^\mu\left(\bar{\theta}^\mu, \xi\right)$ is invertible according to Thm. 2.4(i). Condition (g) is analog to condition (e). Finally, condition (h) is necessary to guarantee that $\tilde{\boldsymbol{S}}^\mu\left(\bar{\theta}^\mu, \xi\right)$ has full rank and is thus a proper covariance matrix.

As pointed out by Huber [126, Sec. 2] and Bunke and Milhaus [55, Rem. 4], a generalized type of strong consistency holds even if the KLIC-best parameter is not identifiable. To the best of our knowledge, there is not proof of asymptotic normality not requiring identifiability.

The Classic Special Case of a Correct Model

Suppose the model is correct for ξ and let $\bar{\theta}^\mu$ be a corresponding correct parameter. Under regularity conditions, a PMLE is a strongly consistent estimator of the correct parameter,

$$\hat{\mathcal{Q}}^\mu_n \xrightarrow{\text{a.s.}} \bar{\theta}^\mu, \text{ for } n \to \infty, \tag{2.23}$$

and is asymptotically normally distributed,

$$\hat{\mathcal{Q}}_n^\mu \overset{\infty}{\sim} \mathcal{N}\Big(\bar{\theta}^\mu, \big(n\tilde{\boldsymbol{F}}^\mu\big(\bar{\theta}^\mu, \xi\big)\big)^{-1}\Big). \tag{2.24}$$

According to (2.11), the matrix $\tilde{\boldsymbol{F}}^\mu$ may be replaced by $\tilde{\boldsymbol{G}}^\mu$. Recalling Cor. 1.10 and the information matrix equality (Thm. 2.5), one can easily see that relations (2.23) and (2.24) are special cases of (2.20) and (2.21), respectively.

Relations (2.23) and (2.24) are well-known classic results of maximum-likelihood theory. For the case of IID observables, they go back to Le Cam [166], and can nowadays be found in most textbooks, for example in those of Lehmann and Casella [170, Chap. 6] or Pawitan [200, Chap. 9]. The IID assumption is not essential and can be replaced by weaker requirements. Philippou and Roussas [203] and Sweeting [242] prove consistency and asymptotic normality of PMLEs in correct models under very general conditions. Their results imply that (2.23) and (2.24) hold under conditions (a)–(f) on p. 59 and on the facing page.

2.4.3. Consistent Estimation of Parameter MLE Covariance

In practice, the asymptotic covariances given in (2.21) and (2.24) are unknown, since they directly depend on the unknown process via the involved expected information matrices, and indirectly via the best or correct parameter. Under certain conditions examined in this section, these covariance can be estimated consistently.

This section uses the same setting and notation as the previous one, with the generalization that $\hat{\mathcal{Q}}_n^\mu$ is not necessarily a PMLE, but may be any strongly consistent estimator of $\bar{\theta}^\mu$.

Correct Models

Suppose that model $\mu \in \mathcal{M}$ is correct under ξ and let $\bar{\theta}^\mu$ be a corresponding correct parameter. The PMLE covariance is then asymptotically the inverse of $n\tilde{\boldsymbol{F}}^\mu\big(\bar{\theta}^\mu, \xi\big)$. Under regularity conditions, notably conditions (a)–(e) on p. 59 and on the facing page,

$$\boldsymbol{F}_n^\mu\big(\hat{\mathcal{Q}}_n^\mu, \mathcal{D}_n, \xi_n\big) \xrightarrow{\text{a.s.}} \tilde{\boldsymbol{F}}^\mu\big(\bar{\theta}^\mu, \xi\big) \text{ element-wise, for } n \to \infty. \tag{2.25}$$

Due to the information matrix equality (Thm. 2.5), the relation remains valid if $\tilde{\boldsymbol{F}}^\mu\big(\bar{\theta}^\mu, \xi\big)$ is replaced by $\tilde{\boldsymbol{G}}^\mu\big(\bar{\theta}^\mu, \xi\big)$.

According to Thm. 2.4(ii), conditions (c) and (f) imply that $\tilde{\boldsymbol{F}}^{\mu}(\theta^{\mu}, \xi)$ is invertible in vicinity $\bar{\theta}^{\mu}$. Consequentially, the inverse of $\boldsymbol{F}_n^{\mu}\big(\hat{\mathcal{Q}}_n^{\mu}, \mathcal{D}_n, \xi_n\big)$ under conditions (a)–(f) on p. 59 and on p. 60 exists asymptotically almost surely, allowing the strongly consistent estimation of the asymptotic PMLE covariance.

These are classic results of maximum-likelihood theory. Proofs and more details can be found in most relevant textbooks, for example in that of Pawitan [200, Sec. 9.9]. They are also contained as special cases in the more general results of Burguete, Gallant, and Souza [58, Thm. 4].

Possibly Incorrect Models

Now drop the assumption that model μ is correct under ξ and let $\bar{\theta}^{\mu}$ be a corresponding KLIC-best parameter. The PMLE covariance is then asymptotically

$$n^{-1}\tilde{\boldsymbol{S}}^{\mu}\big(\bar{\theta}^{\mu}, \xi\big) = \big(n\tilde{\boldsymbol{F}}^{\mu}\big(\bar{\theta}^{\mu}, \xi\big)\big)^{-1}\tilde{\boldsymbol{G}}^{\mu}\big(\bar{\theta}^{\mu}, \xi\big)\big(\tilde{\boldsymbol{F}}^{\mu}\big(\bar{\theta}^{\mu}, \xi\big)\big)^{-1}. \tag{2.26}$$

Relation (2.25) can straightforwardly be generalized to this case. Under regularity conditions,

$$\boldsymbol{F}_n^{\mu}\big(\hat{\mathcal{Q}}_n^{\mu}, \mathcal{D}_n, \xi_n\big) \xrightarrow{\text{a.s.}} \tilde{\boldsymbol{F}}^{\mu}\big(\bar{\theta}^{\mu}, \xi\big) \text{ element-wise, for } n \to \infty. \tag{2.27}$$

The results of White [267, Thm. 3.2] imply that (2.27) holds under the assumption of IID observables and under certain regularity conditions including conditions (b)–(e) on p. 59 and on p. 60. The more general results of Burguete, Gallant, and Souza [58, Thm. 4] imply that it remains valid if the IID assumption is replaced by condition (a) on p. 59. As in the correct case, adding condition (f) ensures that the inverse of $\boldsymbol{F}_n^{\mu}\big(\hat{\mathcal{Q}}_n^{\mu}, \mathcal{D}_n, \xi_n\big)$ exists asymptotically almost surely and is a strongly consistent estimator of the inverse of $\tilde{\boldsymbol{F}}^{\mu}\big(\bar{\theta}^{\mu}, \xi_n\big)$.

It remains to derive a strongly consistent estimator of $\tilde{\boldsymbol{G}}^{\mu}\big(\bar{\theta}^{\mu}, \xi\big)$. Since the information matrix equality does not apply here, (2.27) does not automatically provide a consistent estimator for $\tilde{\boldsymbol{G}}^{\mu}\big(\bar{\theta}^{\mu}, \xi\big)$, as it does in the case of a correct model.

Assuming IID observables, White [267, Thm. 3.2] derives an estimator for $\tilde{\boldsymbol{G}}^{\mu}\big(\bar{\theta}^{\mu}, \xi\big)$ and thus for $\tilde{\boldsymbol{S}}^{\mu}\big(\bar{\theta}^{\mu}, \xi\big)$ that are strongly consistent under regularity conditions including conditions (b)–(h). White [267, Footnote 3] believed that the assumption of IID observables is not crucial and conjectured that the straightforward generalizations of his estimators to INID experiments remain strongly consistent. *Chow [70] shows, however, that this is not the case.* In a reply to Chow, White [266] admits that in possibly incorrect models

> with observables not identically distributed [...] a consistent estimator [of $\tilde{\boldsymbol{G}}^{\mu}\big(\bar{\theta}^{\mu}, \xi\big)$ and thus of $\tilde{\boldsymbol{S}}^{\mu}\big(\bar{\theta}^{\mu}, \xi\big)$] is not generally available unless the true model [= the process in our terminology] is known.

Ten years later, he repeated this statement [265, Sec. 8.3]. The results of Burguete, Gallant, and Souza [58, Thm. 4] support this conclusion.

To the best of our knowledge, no *generally valid* consistent estimators of $\tilde{\boldsymbol{G}}^{\mu}\big(\bar{\theta}^{\mu}, \xi\big)$ and thus of $\tilde{\boldsymbol{S}}^{\mu}\big(\bar{\theta}^{\mu}, \xi\big)$ are known at present. The practical implications of this lack are discussed in Sec. 2.4.5. In Sec. 3.4, we propose a novel estimator for $\tilde{\boldsymbol{S}}^{\mu}\big(\bar{\theta}^{\mu}, \xi\big)$ that is strongly consistent under certain *additional assumptions* which are frequently made in practice.

2.4.4. Consistency of Model MLEs

Suppose experiments are sampled from design ξ, so that the sequence of designs $\xi_1, \xi_2, \ldots$ converges to the ξ. Let $\bar{\mu}$ be a KLIC-best model under ξ.

Under regularity conditions, MMLEs are consistent estimators of the best model,

$$\hat{\mathcal{M}}_n \xrightarrow{\text{a.s.}} \bar{\mu}, \text{ for } n \to \infty. \tag{2.28}$$

Since MMLEs take values in the finite set $\mathscr{M}$ they are *discrete* estimators and the concepts of strong and weak consistency coincide, see (B.3). We hence simply speak of "consistency."

Discrete MLEs like $\hat{\mathcal{M}}_n$ have received little explicit attention in literature so far, particularly in possibly incorrect models. Recently, Choirat and Seri [69, Prop. 1] showed that discrete MLEs in possibly incorrect models are consistent estimators of the corresponding KLIC-minimal value. Their proof requires only mild regularity conditions, but is formulated for IID observables only. As discussed in Sec. 2.1.2, one can expect that their result remains valid for INID experiments as long as the corresponding designs describing them converges to some limit design as the sample size increases.

Suppose that for each model $\mu \in \mathscr{M}$, the conditions for the strong consistency of PMLEs are satisfied, in particular conditions (a)–(c) on p. 59, and $\bar{\mu}$ is identifiable. Then, MMLEs meet the prerequisites of Choirat and Seri in the large-sample limit, leading to (2.28).

2.4.5. Practical Application

How can maximum-likelihood estimation be applied in scenario 2.1 to answer the central questions (Q2.1)–(Q2.4)?

To ease the following discussion, presume that all regularity conditions required for consistency and asymptotic normality of parameter and MMLEs are satisfied. In particular, suppose that experiments are sampled from design ξ, that the KLIC-best parameter $\bar{\theta}^{\mu}$ under ξ is identifiable in each model $\mu \in \mathcal{M}$, and that the KLIC-best model $\bar{\mu}$ under ξ is identifiable.

Basic Inferences not Taking into Account Parameter Uncertainty

In practice, a PMLE $\hat{\theta}^{\mu}_{n}$ for model $\mu \in \mathcal{M}$ can be determined from the data d_n obtained under design ξ_n by maximizing the likelihood $p(d_n \,|\, \xi_n, \mu, \theta^{\mu})$ with respect to $\theta^{\mu} \in \mathcal{Q}^{\mu}$. It us an empirical approximation of the corresponding unknown KLIC-best parameter,

$$\bar{\theta}^{\mu} \overset{\infty}{\approx} \hat{\theta}^{\mu}_{n}. \tag{2.29}$$

Given PMLEs $\hat{\theta}^{\mu}_{n}$ for all models $\mu \in \mathcal{M}$, a MMLE $\hat{\mu}_n$ can be determined in practice by maximizing $p\big(d_n \,\big|\, \xi_n, \mu, \hat{\theta}^{\mu}_{n}\big)$ with respect to $\mu \in \mathcal{M}$. It is an empirical approximation of the unknown best model,

$$\bar{\mu} \overset{\infty}{\approx} \hat{\mu}_n. \tag{2.30}$$

Approximation (2.29) suggests the empirical approximation

$$p\big(y \,\big|\, x, \mu, \bar{\theta}^{\mu}\big) \overset{\infty}{\approx} p\big(y \,\big|\, x, \mu, \hat{\theta}^{\mu}_{n}\big) \tag{2.31}$$

for the KLIC-best PDF of model μ under experimental condition $x \in \mathcal{X}$. A corresponding approximation for the unknown KLIC-best PDF $p\big(y \,\big|\, x, \bar{\mu}, \bar{\theta}\big)$ of the model family under x is obtained by evaluating the right-hand side of (2.31) at the MMLE, that is, for $\mu = \hat{\mu}_n$.

All these approximations are EMPIRICAL, meaning that they depend *only* on known quantities and can thus be evaluated in practice. The consistency of PMLEs (2.20) and MMLEs (2.28) tells us that they improve with the sample size and are asymptotically exact. They are hence justified in sufficiently large samples, as indicated ($\overset{\infty}{\approx}$) in the formulas.

Taking into Account Parameter Uncertainty in Correct Models

If model μ is correct, then the PMLE is asymptotically normal with mean $\bar{\theta}^{\mu}$ and covariance $\left(n\tilde{\boldsymbol{F}}^{\mu}\left(\bar{\theta}^{\mu}, \xi\right)\right)^{-1}$, see (2.24). The matrix $\tilde{\boldsymbol{F}}^{\mu}\left(\bar{\theta}^{\mu}, \xi\right)$ is unknown, but can be approximated by its empirical counterpart,

$$\tilde{\boldsymbol{F}}^{\mu}\left(\bar{\theta}^{\mu}, \xi\right) \overset{\infty}{\approx} \hat{\boldsymbol{F}}_{n}^{\mu} := \boldsymbol{F}_{n}^{\mu}\left(\hat{\theta}_{n}^{\mu}, d_{n}, \xi_{n}\right), \tag{2.32}$$

Combined with (2.29) these relations suggest the empirical approximation

$$p\left(\hat{\theta}^{\mu} \,\middle|\, \mu, \xi_{n}\right) \overset{\infty}{\approx} \phi\left(\hat{\theta}^{\mu} \,\middle|\, \hat{\theta}_{n}^{\mu}, \left(n\hat{\boldsymbol{F}}_{n}^{\mu}\right)^{-1}\right) \tag{2.33}$$

for the unknown distribution of the PMLE, where $\phi(\cdot)$ denotes the PDF of a normal distribution, see Def. B.8 on p. 296.

This approximation allows to empirically quantify the parameter uncertainty. Common characterizations of uncertainty like confidence regions can be derived from it. Furthermore, it allows to empirically approximate functions of the unknown KLIC-best parameter in a PARAMETER-ROBUST way, taking into account the variability of the estimate. A popular example using an expected value approach is the empirical approximation

$$\begin{aligned} p\left(y \,\middle|\, x, \mu, \bar{\theta}^{\mu}\right) &\approx \int_{\mathfrak{Q}^{\mu}} p\left(y \,\middle|\, x, \mu, \hat{\theta}^{\mu}\right) p\left(\hat{\theta}^{\mu} \,\middle|\, \mu, \xi_{n}\right) \mathrm{d}\hat{\theta}^{\mu} \\ &\overset{\infty}{\approx} \int_{\mathfrak{Q}^{\mu}} p\left(y \,\middle|\, x, \mu, \hat{\theta}^{\mu}\right) \phi\left(\hat{\theta}^{\mu} \,\middle|\, \hat{\theta}_{n}^{\mu}, \left(n\hat{\boldsymbol{F}}_{n}^{\mu}\right)^{-1}\right) \mathrm{d}\hat{\theta}^{\mu}, \end{aligned} \tag{2.34}$$

a parameter-robust counterpart of (2.31).

The strong consistency and the asymptotic normality of PMLEs and of the empirical Hessian-based Fisher information matrix justifies to use (2.33) and (2.34) in sufficiently large samples, as indicated ($\overset{\infty}{\approx}$) in the formulas.

Taking into Account Parameter Uncertainty, Incorrect Models

If the model is possibly incorrect, then the PMLE is still asymptotically normal around $\bar{\theta}^{\mu}$, yet with covariance $n^{-1}\tilde{\boldsymbol{S}}^{\mu}\left(\bar{\theta}^{\mu}, \xi\right)$, see (2.21). Unfortunately, as discussed in Sec. 2.4.3, no generally valid consistent estimator is available for $\tilde{\boldsymbol{S}}^{\mu}\left(\bar{\theta}^{\mu}, \xi\right)$, except for the special case of IID observables. For possibly incorrect models, there is hence no analog to approximations (2.33) and (2.34)

Consequentially, *if models that might be incorrect, it is not possible using PMLEs to empirically quantify the parameter uncertainty and to make parameter-robust approximations.* The empirical parameter-unrobust point approximations (2.29)–(2.31) remain untouched from this problem.

A nearby remedy is to use (2.33) nevertheless even if the model might be incorrect, tolerating that a central underlying assumption is violated. This is the choice made – knowingly or not – by most of the established "parameter-robust" OED strategies, particularly by those considered in Chaps. 4 and 5.

As a main result of this thesis, we show in Sec. 3.4 that *the asymptotic covariance of* PMLEs can *be estimated consistently under the common additional assumptions that the observation covariances are known, the process is normal and the rival models are normal.* Based on that novel result, we determine the error introduced when applying (2.33) to incorrect models, formulate a counterpart of this approximation that is applicable also in incorrect models, and propose new parameter-robust OED strategies.

Summary

It is their generality which gives the discussed MLE-based empirical approximations their great practical significance. Since they hold for a broad class of model families, they constitute a unified set of methods with a wide range of applications. And because they hold regardless of the actual distribution of the sample (except for regularity conditions), they permit to use these methods in practice where this distribution is unknown.

Using maximum-likelihood estimation one can answer the central questions (Q2.1)–(Q2.4) as follows: using PMLEs and MMLEs, the unknown KLIC-best parameters of each model, the unknown best model and the associated best PDFs can be approximated arbitrarily well as the amount of available data increases. The relevant parameter-unrobust empirical approximations are applicable regardless if the model is correct or not.

In a correct model, the parameter uncertainty can be quantified via an empirical large-sample approximation of the density of the PMLE. Based thereon one can evaluate the parameter-robust approximations for the unknown KLIC-best PDF of the model, which can be used for predicting the process behavior. In incorrect models, however, the parameter uncertainty cannot be quantified empirically in a reliable way, corresponding parameter-robust approximations are not available.

To quantify the model uncertainty and to make corresponding model-robust predictions, one typically resorts to different techniques than maximum-likelihood estimation.

2.5. Bayesian Inference

> *[...] a* probability *p is an abstract concept, a quantity that we* assign *theoretically, for the purpose of representing a state of knowledge [...] or that we* calculate *from previously assigned probabilities using the rules [...] of probability theory.*
>
> Jaynes [130, p. 8] on probability in Bayesian statistics.

This section deals with Bayesian inference, that is, methods of inference based on Bayesian statistics. These methods form the basis of the optimal experimental design (OED) strategies considered in Chap. 5.

Bayesian inference under the "classic" assumption that the underlying model is correct has gained a certain maturity. As in the previous sections, we try to avoid this assumption, since it is usually violated in practical problems of model discrimination (MD). In the "non-classic" setting of possibly incorrect models, central results were obtained rather recently. This section gives an overview of these little-known results.

Lee [169] provides an introduction to Bayesian statistics, Robert [213] gives a decision-theoretic motivation. The extensive book of Jaynes [134] even advocates Bayesian statistics as "the logic of science." Comprehensive references for Bayesian inference are the books of Bernardo and Smith [28] and Box and Tiao [45] and O'Hagan and Forster [195].

2.5.1. Outline of the General Approach

In frequentist inference, like in the previously discussed maximum-likelihood estimation, the only genuine source of "randomness" are the fluctuations of the data. This EXPERIMENTAL UNCERTAINTY is described by the distribution of the sample. EPISTEMIC UNCERTAINTY, that is, uncertainty due to a lack of knowledge (e.g. about the Kullback-Leibler information criterion (KLIC)-best parameters and models) is quantified indirectly based on STATISTICS, that is, functions of the data, for example maximum-likelihood estimates (MLES). These statistics are only random inasmuch as they are functions of the sample. Without experimental uncertainty, there is no natural way of expressing epistemic uncertainty in frequentist inference.

This is different in Bayesian inference: there, probability distributions are directly used to represent "states of knowledge" or "beliefs" (see introductory quote of the section). That is, *both* experimental uncertainty *and* epistemic uncertainty are represented by probability distributions.

Let us outline some key ingredients of Bayesian inference. Suppose the hypotheses $\mathrm{H}_1, \ldots, \mathrm{H}_m$ shall be assessed in the light of the evidence E. The probability of H_i under E, denoted $\mathbb{P}[\mathrm{H}_i \mid \mathrm{E}]$, can be calculated via Bayes' theorem,

$$\mathbb{P}[\mathrm{H}_i \mid \mathrm{E}] = \mathbb{P}[\mathrm{H}_i] \frac{\mathbb{P}[\mathrm{E} \mid \mathrm{H}_i]}{\mathbb{P}[\mathrm{E}]}, \tag{2.35}$$

which is actually a corollary from the definition of conditional probability. Bayes [25] was the first to propose it in a non-trivial case for statistical inference, but the formula did not gain much attention until it was "rediscovered" by Laplace [163].

In Bayesian terminology, $\mathbb{P}[\mathrm{H}_i]$ is the prior (probability), $\mathbb{P}[\mathrm{H}_i \mid \mathrm{E}]$ is the posterior (probability), $\mathbb{P}[\mathrm{E} \mid \mathrm{H}_i]$ is the likelihood, and $\mathbb{P}[\mathrm{E}] := \sum_{i=1}^{m} \mathbb{P}[\mathrm{H}_i] \mathbb{P}[\mathrm{E} \mid \mathrm{H}_i]$ is the marginal likelihood. The latter is a normalizing factor ensuring that the posterior probabilities sum up to 1. It is often omitted and Bayes' theorem is simple written as $\mathbb{P}[\mathrm{H}_i \mid \mathrm{E}] \propto \mathbb{P}[\mathrm{H}_i] \mathbb{P}[\mathrm{E} \mid \mathrm{H}_i]$.

In Bayesian inference, this theorem is applied as follows to assess the validity of the hypotheses. The prior and posterior probabilities are interpreted as the belief in H_i *before* and *after* taking into account E, respectively. The likelihood corresponds to a model that specifies probabilities for obtaining certain evidence if the hypothesis was true. Given a model and a prior, Bayes' theorem allows to calculate the posterior once the evidence is available. The posterior is considered as an improvement over the prior and is used for all further calculations, possibly as new prior in the next step of inference when additional evidence gets available. This procedure is called Bayesian updating.

Predictions or approximations for unknowns are obtained through averaging over the available priors or posteriors. If X is some unknown quantity of interest (for example an unobserved experimental result), and $\mathbb{P}[\mathrm{X} \mid \mathrm{H}_i]$ it its probability under the assumption that hypothesis H_i is true, then

$$\mathbb{P}[\mathrm{X} \mid \mathrm{E}] = \sum_{i=1}^{m} \mathbb{P}[\mathrm{X} \mid \mathrm{H}_i] \mathbb{P}[\mathrm{H}_i \mid \mathrm{E}] \tag{2.36}$$

is the Bayesian prediction for X given the evidence E, or simply the POSTERIOR PREDICTION. Bayesian predictions are thus intrinsically probabilistic.

The prior offers a natural and consistent way to regard previous knowledge both from preceding experiments as well as from other sources like literature or a priori considerations. This property is often considered as a main strength of the Bayesian approach. It is, however, also a primary point of critique, since it forces the Bayesian to formulate a prior even if no previous information is available, which might introduce a certain arbitrariness into the inferences.

2.5.2. Application to Model Families

The Bayesian approach applies to scenario 2.1 as follows. For clarity, we use a convenient "overloaded" notation in which different densities share the same symbol $p(\cdot)$ and are distinguished solely by their arguments, as noted in the introduction of Sec. 1.3.1 on p. 27.

Inference in Single Regression Models

Previous knowledge (or previous uncertainty, we use these terms interchangeably here) associated with the parameter of model $\mu \in \mathcal{M}$ is represented by the PARAMETER PRIOR $p(\theta^\mu)$, a probability density function (PDF) over the parameter domain $\mathcal{Q}^\mu$. Knowledge or uncertainty *after* taking into account the data d_n obtained from $n \in \mathbb{N}$ experiments described by the exact design ξ_n is represented by the PARAMETER POSTERIOR $p(\theta^\mu \,|\, d_n, \xi_n)$, also a PDF over $\mathcal{Q}^\mu$. The parameter posterior can for all $\theta^\mu \in \mathcal{Q}^\mu$ be determined via Bayes' theorem

$$p(\theta^\mu \,|\, d_n, \xi_n) \propto p(\theta^\mu) p(d_n \,|\, \xi_n, \mu, \theta^\mu) \tag{2.37}$$

from the corresponding prior and from the likelihood $p(d_n \,|\, \xi_n, \mu, \theta^\mu)$ of model μ discussed in Sec. 2.2. The factor of proportionality in (2.37) is determined by the requirement that the PDF integrates up to 1 over $\mathcal{Q}^\mu$.

The parameter posterior is the pivot for any further inferences in the model. It can be used to derive point estimators, to quantify uncertainty, or to make predictions. We shall see in Sec. 2.5.3 that it does under certain assumptions actually represent knowledge *about the KLIC-best parameter of the model.*

Predictions based on Single Regression Models

The *actual* distribution of experimental outcomes under condition $x \in \mathcal{X}$ is described by the unknown PDF $q(y \mid x)$. For all $x \in \mathcal{X}$, the PDF

$$p(y \mid x, \mu, d_n, \xi_n) := \int_{\mathcal{Q}^\mu} p(y \mid x, \mu, \theta^\mu) p(\theta^\mu \mid d_n, \xi_n) \, \mathrm{d}\theta^\mu \tag{2.38}$$

is the POSTERIOR PREDICTION OF MODEL μ for the experimental outcome under x, that is,

$$q(y \mid x) \approx p(y \mid x, \mu, d_n, \xi_n). \tag{2.39}$$

This prediction takes into account all available knowledge about the parameter in terms of the posterior distribution (as opposed to a point estimator) and can thus be considered as "robust" with respect to the parameter uncertainty, or simply as "parameter-robust."

Inference in a Family of Regression Models

When the parameters are model-local, as we assume here, their priors and posteriors can be specified and treated independently for each model $\mu \in \mathcal{M}$. In a given family of models distinguished with a finite model index set $\mathcal{M}$, it then remains to represent the knowledge (or uncertainty) about the models themselves. This is done via the MODEL PRIOR $p(\mu)$, a probability mass function (PMF) over the model index set $\mathcal{M}$. After taking into account the data d_n obtained under design ξ_n, the knowledge is represented by the MODEL POSTERIOR $p(\mu \mid d_n, \xi_n)$, also a PMF over the model index set $\mathcal{M}$. For all $\mu \in \mathcal{M}$, it can be determined via Bayes' theorem

$$p(\mu \mid d_n, \xi_n) \propto p(\mu) p(d_n \mid \xi_n, \mu) \tag{2.40}$$

from the model prior and from the MARGINAL LIKELIHOOD

$$p(d_n \mid \xi_n, \mu) := \int_{\mathcal{Q}^\mu} p(d_n \mid \xi_n, \mu, \theta^\mu) p(\theta^\mu) \, \mathrm{d}\theta^\mu. \tag{2.41}$$

The factor of proportionality in (2.40) is determined by the condition that the probabilities add up to 1 over $\mathcal{M}$.

We shall see in Sec. 2.5.4 that the model posterior does actually represents knowledge *about the KLIC-best model in the family.* Based on the model posterior one can derive point estimators for the latter or quantify the associated model uncertainty. Together with parameter posteriors for each model, one can also make predictions.

Predictions based on a Family of Regression Models

For all $x \in \mathcal{X}$, the PDF

$$p(y\,|\,x, d_n, \xi_n) := \sum_{\mu \in \mathcal{M}} p(y\,|\,x, \mu, d_n, \xi_n)\, p(\mu\,|\,d_n, \xi_n) \tag{2.42}$$

is the POSTERIOR PREDICTION OF THE MODEL FAMILY for the experimental outcome under x, that is,

$$q(y\,|\,x) \approx p(y\,|\,x, d_n, \xi_n). \tag{2.43}$$

Since it incorporates all parameter posteriors and the model posterior, this prediction can be considered as both "model-robust" and "parameter-robust."

The technique of forecasting the process behavior based on the weighted predictions of several individual models is known as "Bayesian model averaging." *It provides generally better average predictive accuracy than using a single model.* For details we refer to the works of Dawid [77], Draper [85], and Madigan and Raftery [179] and to the overview given by Hoeting et al. [119] and the references provided therein.

2.5.3. Large-Sample Behavior in Single Regression Models

Consistency of the Parameter Posterior

Under regularity conditions, the parameter posterior is consistent in the sense that it accumulates arbitrarily close to the KLIC-best parameter with probability 1 in the large-sample limit:

$$\int_{\mathcal{B}^\mu} p(\theta^\mu\,|\,\mathcal{D}_n, \xi_n)\, \mathrm{d}\theta^\mu \xrightarrow{\text{P}} \begin{cases} 1 & \text{if } \bar{\theta}^\mu \in \mathcal{B}^\mu \\ 0 & \text{otherwise} \end{cases}, \text{ as } n \to \infty, \tag{2.44}$$

for any open subset $\mathcal{B}^\mu$ of $\mathcal{Q}^\mu$. Furthermore, its maximizer or MODE is under regularity conditions a consistent estimator for the KLIC-best parameter:

$$\operatorname*{argmax}_{\theta^\mu \in \mathcal{Q}^\mu} p(\theta^\mu \,|\, \mathcal{D}_n, \xi_n) \xrightarrow{\text{p}} \bar{\theta}^\mu, \text{ for } n \to \infty. \tag{2.45}$$

The regularity conditions are discussed later.

These relations suggest the following conclusions with regard to (Q2.1) on p. 42: by increasing the amount of available data, the Bayesian approach permits to *identify* the unknown KLIC-best parameter empirically with arbitrary precision. The knowledge which the parameter posterior represents is under the given assumptions in fact knowledge *about the KLIC-best parameter.*

Asymptotic Normality of the Parameter Posterior

Under certain regularity conditions, including those required for consistency, the parameter posterior is in the large-sample limit normally distributed around KLIC-best parameter,

$$p(\theta^\mu \,|\, d_n, \xi_n) \overset{\infty}{\approx} \phi\Big(\theta^\mu \,\Big|\, \bar{\theta}^\mu, \big(\boldsymbol{P}^\mu(\bar{\theta}^\mu) + n\tilde{\boldsymbol{F}}^\mu(\bar{\theta}^\mu, \xi)\big)^{-1}\Big). \tag{2.46}$$

Its covariance is determined by the expected Hessian-based Fisher information matrix $\tilde{\boldsymbol{F}}^\mu$ from (2.6) and the PRIOR INFORMATION MATRIX, defined as

$$\boldsymbol{P}^\mu(\theta^\mu) := -\nabla^2 \ln p(\theta^\mu) \tag{2.47}$$

for all $\theta^\mu \in \mathcal{Q}^\mu$, where ∇^2 denotes the Hessian differential operator with respect to θ^μ.

In nonlinear models, the parameter posterior (2.37) can typically not be expressed in a closed form. Numerical approximations are possible, but often computationally expensive. In practice, it often suffices to use the following easier-to-compute large-sample normal approximation that can be derived from (2.46). If the sample size n is large, the KLIC-best parameter $\bar{\theta}^\mu$ can be well approximated empirically by the parameter maximum-likelihood estimate (PMLE) $\hat{\theta}^\mu_n = \hat{\theta}^\mu_n(d_n, \xi_n)$, see (2.20), and the expected Fisher information $\tilde{\boldsymbol{F}}^\mu(\bar{\theta}^\mu, \xi)$ can be well approximated by its empirical counterpart $\hat{\boldsymbol{F}}^\mu_n := \boldsymbol{F}^\mu_n(\hat{\theta}^\mu_n, d_n, \xi_n)$, see (2.25). Applying these substitutions to (2.46) yields the empirical large-sample approximation

$$p(\theta^\mu \,|\, d_n, \xi_n) \overset{\infty}{\approx} \phi\Big(\theta^\mu \,\Big|\, \hat{\theta}^\mu_n, \big(\hat{\boldsymbol{P}}^\mu_n + n\hat{\boldsymbol{F}}^\mu_n\big)^{-1}\Big), \tag{2.48}$$

where $\hat{\boldsymbol{P}}^\mu_n := \boldsymbol{P}^\mu\big(\hat{\theta}^\mu_n\big)$. The right-hand side of (2.48) depends only on known quantities and on empirical data (unlike (2.46)), and can be evaluated in practice once the experiments have been performed.

Consistency of the Posterior Prediction

If the parameter posterior is consistent, it follows from a generalization of Slutsky's Theorem (Thm. B.4) that under mild regularity conditions

$$p(y\,|\,x, \mu, \mathcal{D}_n, \xi_n) \xrightarrow{\text{p}} p\big(y\,\big|\,x, \mu, \bar{\theta}^\mu\big), \text{ for } n \to \infty, \tag{2.49}$$

for all $y \in \mathcal{Y}$ under all $x \in \mathcal{X}$, so that

$$p\big(y\,\big|\,x, \mu, \bar{\theta}^\mu\big) \overset{\infty}{\approx} p(y\,|\,x, \mu, d_n, \xi_n). \tag{2.50}$$

These relation are the Bayesian answer to (Q2.2) on p. 42: by increasing the amount of available data, Bayesian inference permits to identify the KLIC-best PDF of a model with arbitrary precision.

Of all the PDFs $\{p(y\,|\,x, \mu, \theta^\mu) : \theta^\mu \in \mathcal{Q}^\mu\}$ specified by the model, that one associated with the KLIC-best parameter $\theta^\mu = \bar{\theta}^\mu$ exhibits the lowest discrepancy to the process under design ξ in terms of the KLIC (see Sec. 1.4), which suggests the approximation

$$q(y\,|\,x) \approx p\big(y\,\big|\,x, \mu, \bar{\theta}^\mu\big) \tag{2.51}$$

for experiments performed under $x \in \operatorname{supp}(\xi)$. Taking together (2.50) and (2.51) and regarding that by assumption $\xi_n \approx \xi$ for large n thus justifies (2.43) in large samples. In other words, the posterior prediction $p(y\,|\,x, \mu, d_n, \xi_n)$ is in large samples a "best guess" for the experimental outcome under x, given model μ and the data d_n obtained under ξ. The quality of this guess depends, of course, on the details of the model formulation.

If model μ is correct (under ξ), then (2.51) and thus (2.43) are in exact in the large-sample limit for all $x \in \mathcal{X}$ (for all $x \in \operatorname{supp}(\xi)$).

References and Historical Remarks

Results stating asymptotic normality of parameter posteriors in our terminology are also referred to as "Bernstein-Von-Mises theorems," in honor of Richard von Mises and Sergei Natanowitsch Bernstein, even if the earliest proof was given by Doob [83]. Under the assumptions of a correct model he proves that the posterior concentrates under mild conditions in an arbitrarily small neighborhood of the correct parameter in the sense of (2.44). Based thereon he proves consistency of Bayes estimators as in (2.45) for *almost all* values that are possible for the correct parameter. Under stronger assumptions, Le Cam [165, 166] proves consistency for *all* values of the correct parameter and also normality of the posterior in the sense of (2.46). This line of argumentation culminated in the work of Schwartz [226, 227], whose central result can be paraphrased as "a Bayes estimator is consistent if a consistent estimator exists."

It took some time until these results were generalized to possibly incorrect models. Berk [26, 27] shows that under some regularity conditions the posterior converges in a weak sense to a degenerate distribution over the set of KLIC-best parameters – similar to (2.44) – even if the model is incorrect. This property does not suffice, however, to ensure consistency of Bayes estimators. A big step was taken by Bunke and Milhaus [55], who state sufficient conditions for the consistency of Bayes estimators like (2.45) and for their large-sample normality. Some of their ideas were developed further recently by Lee and MacEachern [168] for the special case of models from the minimal standard exponential family, which also includes normal models that we consider in the next chapter.

Bunke and Milhaus and Lee and MacEachern consider the distribution of Bayesian (point) estimators which are derived from the posterior, but did not consider the distribution of the posterior itself. This gap was recently closed by Kleijn [145] and Kleijn and van der Vaart [144], who show that *for the large class of so-called "local asymptotic normality" models, the posterior is asymptotically normal as in (2.46), even if the model is incorrect.*

Regularity Conditions

Kleijn and van der Vaart [144] show that parameter posteriors are asymptotically normal under a set of fairly general conditions. In the setting considered here, they reduce to conditions similar to those required for asymptotic normality of parameter maximum-likelihood estimators (PMLEs). In addition, it is required that the parameter prior $p(\theta^{\mu})$ is positive in a neighborhood of $\bar{\theta}^{\mu}$. Kleijn and van der Vaart [144, Sec. 2.2] focus on independently and identically distributed

(IID) observables. This limitation can be relaxed to the assumption that the experiments sampled from a design, as discussed in Sec. 2.1.2. In the words of Gelman et al. [106, Appendix B.1], “the key condition [for asymptotic normality] is that there be ‘replication’ at some level [...]”

Summed up, asymptotic normality of parameter posteriors requires, besides certain technicalities, that conditions conditions (a)–(f) on p. 59 and on p. 60 are met and that the parameter prior does not vanish in vicinity of the KLIC-best parameter.

2.5.4. Large-Sample Behavior in Families of Regression Models

As previously, assume that the experiments are sampled from a design ξ, such that the design sequence $\xi_1, \xi_2, \ldots$ converges to ξ. Consider a family of regression models distinguished by indices from the finite set $\mathcal{M}$, and suppose that the KLIC-best model $\bar{\mu} \in \mathcal{M}$ under ξ is identifiable and that each model $\mu \in \mathcal{M}$ has an identifiable KLIC-best parameter $\bar{\theta}^{\mu} \in \mathcal{Q}^{\mu}$ under ξ.

Consistency of the Model Posterior

Under regularity conditions, the model posterior converges with increasing sample size to a degenerate distribution putting full mass at the best model $\bar{\mu}$ with probability 1, that is,

$$p(\mu \,|\, \mathcal{D}_n, \xi_n) \xrightarrow{\text{P}} \begin{cases} 1 & \text{if } \mu = \bar{\mu} \\ 0 & \text{otherwise} \end{cases}, \quad \text{for } n \to \infty. \tag{2.52}$$

which implies that its maximizer is a consistent estimator of the KLIC-best model,

$$\operatorname*{argmax}_{\mu \in \mathcal{M}} p(\mu \,|\, \mathcal{D}_n, \xi_n) \xrightarrow{\text{a.s.}} \bar{\mu}, \quad \text{for } n \to \infty. \tag{2.53}$$

Note that strong and weak consistency coincide in this case, since the model index set is discrete, see (B.3). These relations show that the Bayesian approach thus permits to identify the unknown KLIC-best model arbitrary well if enough experimental data is available, which answers (Q2.3) on p. 42. The model posterior can hence be interpreted as knowledge *about the KLIC-best model.*

The Marginal Likelihood in Large Samples

According to Bayes' theorem (2.40), the model posterior is proportional to the product of the model prior and the marginal likelihood (2.41). In general, the integral in the latter has no closed-form representation. Approximating it numerically is possible, yet often too expensive computationally. In the following we introduce popular closed-form approximations of the marginal likelihood. The resulting approximate formulas for the model posteriors are discussed in the next section.

The method of Laplace [162] allows to approximate indefinite integrals of the form $\int_{\mathcal{A}} g(x) \exp(-nf(x))\,\mathrm{d}x$, if they exist, where n is a large natural number, $\mathcal{A} \subseteq \mathbb{R}^m$, f is twice differentiable and has a unique maximum on $\mathcal{A}$, and g is differentiable and non-zero at the maximizer of f. The underlying idea is to use a second-order Taylor approximation of f around its maximum. The resulting integral is then a Gaussian integral with a known closed-form solution. Azevedo-Filho and Shachter [22] provide a detailed discussion of Laplace's method in the context of Bayesian inference.

Applying Laplace's method to the logarithm of the marginal likelihood (2.41) yields the large-sample approximation

$$\begin{aligned}\ln p(d_n \mid \xi_n, \mu) &\overset{\infty}{\approx} \ln p\left(d_n \,\middle|\, \xi_n, \mu, \check{\theta}_n^\mu\right)\\ &\quad - \tfrac{1}{2} \ln \det\left(\check{\boldsymbol{P}}_n^\mu + n\check{\boldsymbol{F}}_n^\mu\right) + \ln p\left(\check{\theta}_n^\mu\right) + \tfrac{1}{2} n_{\theta^\mu} \ln(2\pi) + \mathcal{O}\left(n^{-1}\right), \qquad (2.54)\end{aligned}$$

where $\check{\theta}_n^\mu$ denotes the maximizer (or "mode") of the parameter posterior. If the sample size n is large, the $\mathcal{O}(1)$ term $\check{\boldsymbol{P}}_n^\mu := \boldsymbol{P}^\mu\left(\check{\theta}_n^\mu\right)$ is negligible compared to the $\mathcal{O}(n)$ term $n\check{\boldsymbol{F}}_n^\mu := n\boldsymbol{F}_n^\mu\left(\check{\theta}_n^\mu, d_n, \xi_n\right)$, and the posterior mode $\check{\theta}_n^\mu$ approximately equals the PMLE $\hat{\theta}_n^\mu$ based on the n experiments. Furthermore, the empirical Fisher information $\hat{\boldsymbol{F}}_n^\mu := \boldsymbol{F}_n^\mu\left(\hat{\theta}_n^\mu, d_n, \xi_n\right)$ is under the regularity conditions discussed in Sec. 2.4.3 a consistent estimator of its expected counterpart $\hat{\tilde{\boldsymbol{F}}}_n^\mu := \tilde{\boldsymbol{F}}^\mu\left(\hat{\theta}_n^\mu, \xi_n\right)$. The relative error of these approximations is $\mathcal{O}\left(n^{-½}\right)$, so that

$$\begin{aligned}\ln p(d_n \mid \xi_n, \mu) &\overset{\infty}{\approx} \ln p\left(d_n \,\middle|\, \xi_n, \mu, \hat{\theta}_n^\mu\right)\\ &\quad - \tfrac{1}{2} \ln \det\left(n\hat{\tilde{\boldsymbol{F}}}_n^\mu\right) + \ln p\left(\hat{\theta}_n^\mu\right) + \tfrac{1}{2} n_{\theta^\mu} \ln(2\pi) + \mathcal{O}\left(n^{-½}\right). \qquad (2.55)\end{aligned}$$

Substituting $\ln \det\left(n\hat{\tilde{\boldsymbol{F}}}_n^\mu\right) = \ln \det\left(\hat{\tilde{\boldsymbol{F}}}_n^\mu\right) + n_{\theta^\mu} \ln n$ and omitting all $\mathcal{O}(1)$ terms yields

$$\ln p(d_n \mid \xi_n, \mu) \overset{\infty}{\approx} \ln p\left(d_n \mid \xi_n, \mu, \hat{\theta}_n^\mu\right) - \tfrac{1}{2} n_{\theta^\mu} \ln n + \mathcal{O}(1). \tag{2.56}$$

The prior

$$p(\theta^\mu) = \phi\left(\theta^\mu \,\middle|\, \hat{\theta}_n^\mu, \left(\hat{\bar{F}}_n^\mu\right)^{-1}\right), \tag{2.57}$$

may be regarded as representing "as much information as a single average experiment." Strictly speaking, it is not a proper parameter *prior,* though, since it depends on the design and the data. It is, however, a reasonable representation of the common situation that little, but not much prior information is available. Under this prior

$$\ln p\left(\hat{\theta}_n^\mu\right) \overset{\text{(B.12b)}}{=} \tfrac{1}{2} \ln \det\left(\hat{\bar{F}}_n^\mu\right) - \tfrac{1}{2} n_{\theta^\mu} \ln(2\pi), \tag{2.58}$$

so that (2.55) simplifies to the same form as (2.56), except that the error is then only of $\mathcal{O}\left(n^{-1/2}\right)$.

The Model Posterior in Large Samples

Applying approximation (2.56) for all models $\mu \in \mathcal{M}$ in Bayes' theorem (2.40) leads to the large-sample approximation

$$\ln p(\mu \mid d_n, \xi_n) \overset{\infty}{\approx} \ln p(\mu) + \ln p\left(d_n \mid \xi_n, \mu, \hat{\theta}_n^\mu\right) - \tfrac{1}{2} n_{\theta^\mu} \ln n + c_n. \tag{2.59}$$

for the model posterior. The constant $c_n \in \mathbb{R}_0^+$ is determined by the requirement that the posterior model probabilities sum up to 1. Note that both c_n and the PMLEs $\hat{\theta}_n^\mu$ in the right-hand side of (2.59) may depend on the data d_n and the design ξ_n.

In general, approximation (2.59) has an absolute error of $\mathcal{O}(1)$ like (2.56). The *relative* error, however, typically vanishes asymptotically with n^{-1} because the log-likelihood term $\ln p\left(d_n \mid \xi_n, \mu, \hat{\theta}_n^\mu\right)$ is of $\mathcal{O}(n)$, see (2.5). If one assumes that all parameter priors are "little informative" in the sense of (2.57), then (2.59) has even an asymptotically vanishing absolute error of only $\mathcal{O}\left(n^{-1/2}\right)$.

Consistency of the Posterior Prediction

Suppose that the parameter posterior of model $\bar{\mu}$ and the model posterior are consistent, and that (2.49) holds for the KLIC-best model $\bar{\mu}$. Then it follows from a generalization of Slutsky's Theorem (Thm. B.4) that under mild regularity conditions

$$p(y \,|\, x, \mathcal{D}_n, \xi_n) \xrightarrow{\mathrm{p}} p\big(y \,\big|\, x, \bar{\mu}, \bar{\theta}\big), \text{ as } n \to \infty, \tag{2.60}$$

for all $y \in \mathcal{Y}$ and under all $x \in \mathcal{X}$. In other words, the posterior prediction (2.42) converges in probability to the KLIC-best PDF of the model family.

This relation can be interpreted in a similar fashion as (2.49). The posterior prediction $p(y \,|\, x, d_n, \xi_n)$ of the model family is in large samples a "best guess" for the experimental outcome under x, given the data d_n obtained under ξ, justifying (2.43). If model $\bar{\mu}$ is correct (under ξ), then this approximation is also exact in the large-sample limit for all $x \in \mathcal{X}$ (for all $x \in \operatorname{supp}(\xi)$).

Discussion and References

Proofs for the consistency of posteriors over discrete sets under IID observables can be found in textbooks on Bayesian inference, for example in that of Gelman et al. [106, Appendix B]. As discussed in Sec. 2.1.2, one can expect that those results remain valid for independently but not identically distributed (INID) experiments, as long as they are sampled from an experimental design, as we assume here.

Suppose that for each model $\mu \in \mathcal{M}$, the conditions for the asymptotic normality of the parameter posterior are satisfied, in particular conditions (a)–(f) on p. 59 and on p. 60. Furthermore, assume that the model family has an identifiable KLIC-best model $\bar{\mu}$ under ξ and that the model prior does not vanish there, $p(\bar{\mu}) > 0$. Then, model posteriors asymptotically meet the prerequisites for consistency as described by Gelman et al., Appendix B, leading to (2.52) and (2.53).

Laplace-based approximations for the marginal likelihood like (2.54)–(2.56) are common in Bayesian inference. They were proposed and used, for example, by Kass and Raftery [139], Raftery [208], and Raftery, Madigan, and Volinsky [209] and Draper [85]. More applications are listed by the references provided therein.

Such Laplace-approximations are typically motivated heuristically. Kass, Tierney, and Kadane [140] seem to be the only ones providing a rigorous treatment, which, however, seems to apply to correct models only. They provide regularity conditions [140, Sec. 3, Items (i)–(iv)] ensuring that Laplace approximations of the marginal likelihood (and similar quantities) are asymptotically exact. Their conditions resemble those required for asymptotic normality of parameter posteriors discussed in Sec. 2.5.3. In addition, they comprise the requirement that the log-likelihood is six times continuously differentiable with respect to the parameter, and that these derivatives are asymptotically bounded in a neighborhood of the correct parameter.

To the best of our knowledge, Laplace-based approximations of the marginal likelihood in possibly incorrect models have not been treated rigorously so far. The similarity of the regularity conditions provided by Kass, Tierney, and Kadane [140] to the regularity conditions for parameter posteriors, however, suggests that a generalization to incorrect models is possible. From a strict point of view, the Laplace-based approximation (2.59) for the model posterior has to be considered as a heuristics in incorrect models.

Summary

This chapter dealt with statistical inference in families of parametric regression models. It focused on the real-world situation that the considered model (family) may be incorrect, and stated the empirical questions that arise if one agrees to measure its discrepancy to the process with the Kullback-Leibler information criterion (KLIC).

After the necessary preparatory steps, the major frequentist approach of maximum-likelihood estimation and the alternative Bayesian approach were surveyed. In both approaches there exist asymptotic results that allow to apply a small unified set of formulas for inferences in a wide range of processes and models, only restricted by certain regularity conditions.

In short, *given enough data, maximum-likelihood estimation as well as Bayesian inference allow to approximate unknown KLIC-best parameters, KLIC-best model and the associated model family members arbitrarily well. Remarkably, this is true regardless if the model family is correct or incorrect.* There are, however, subtle yet important differences between both approaches concerning the ability to empirically quantify the related uncertainties.

If experiments are sampled from design ξ, parameter maximum-likelihood estimators (PMLEs) and parameter posteriors of a model $\mu \in \mathcal{M}$ exhibit remarkably similar behavior in the large-sample limit: both are described by a normal distribution around the KLIC-best parameter $\bar{\theta}^{\mu}$ under the limit design ξ.

If the model is *correct* under ξ, the covariance of both is asymptotically equal to the inverse of $n\tilde{\boldsymbol{F}}^{\mu}\left(\bar{\theta}^{\mu}, \xi\right)$. For parameter posteriors, this asymptotic covariance formula remains valid even without the assumption of a correct model. In contrast, parameter maximum-likelihood estimates (PMLEs) in *possibly incorrect* models have an asymptotic covariance of $n^{-1}\tilde{\boldsymbol{S}}^{\mu}\left(\bar{\theta}, \xi\right)$. Kleijn and van der Vaart [144] discuss this discrepancy and show that the difference between the asymptotic covariances can be substantial. In large samples, the matrix $\tilde{\boldsymbol{F}}^{\mu}\left(\bar{\theta}^{\mu}, \xi\right)$ can be well approximated by its empirical counterpart $\hat{\boldsymbol{F}}_{n}^{\mu} := \boldsymbol{F}_{n}^{\mu}\left(\hat{\theta}^{\mu}, d_{n}, \xi_{n}\right)$. *For the matrix* $\tilde{\boldsymbol{S}}^{\mu}\left(\bar{\theta}, \xi\right)$*, however, such an empirical approximation is not available.*

In maximum-likelihood estimation and Bayesian inference, the distributions of a PMLE and a parameter posterior, respectively, describe the uncertainty about the unknown KLIC-best parameter. In correct models, both approaches lead to asymptotically compliant descriptions, which can be approximated empirically based on $\hat{\boldsymbol{F}}_{n}^{\mu}$. In possibly incorrect models, however, both approaches provide asymptotically different descriptions of the parameter uncertainty. Furthermore, empirically quantifying the parameter uncertainty is generally possible only in the Bayesian approach, but not using maximum-likelihood estimation.

The lack of a consistent estimator for $\tilde{\boldsymbol{S}}^{\mu}\left(\bar{\theta}^{\mu}, \xi\right)$ *in general* does not rule out the possibility that one exists under *additional* assumptions. In fact, as a main result of our thesis, we propose in Sec. 3.4.1 a novel estimator for it under the common assumptions of a normal model and known observation covariance. For all models meeting these assumptions, this estimator allows us to empirically quantify the parameter uncertainty in a maximum-likelihood framework even in possibly incorrect models. This estimator allows us to propose novel parameter-robust design criteria in Chap. 5.

Discrete maximum-likelihood estimators (MLEs) like model maximum-likelihood estimator (MMLE) are considered rather infrequently in frequentist inference, and less even their distribution. In particular, there is no equally general result about the asymptotic distribution of MMLEs as there is for PMLEs, to the best of our knowledge. Typically, a frequentist resorts to other techniques than maximum-likelihood estimation, like statistical hypothesis testing, if confronted with a finite family of rival models.

In Bayesian inference, however, asymptotic approximations for model posterior are available through the method of Laplace. They can be evaluated from given data and thus allow to quantify model uncertainty empirically. So far, however, strict regularity conditions for the asymptotic validity of those approximations are available only for correct models. There are reasons, however, to hope that they can be generalized to possibly incorrect models, as in the case of parameter posteriors.

In the next chapter, we consider the results provided here under the common assumptions of normality and (local) linearity. Many well-known formulas can be found there. The results of this and the next chapter form the basis for optimal experimental design (OED) strategies considered in the second part of this thesis.

3. Inference under Normality and Linearity for Practical Computation

> *[…] the statistician knows […] that in nature there never was a normal distribution, there never was a straight line, yet with normal and linear assumptions, known to be false, he can often derive results which match, to a useful approximation, those found in the real world.*
>
> Box [43, Sec. 2.5]

Contents

So far, few assumptions have been made about the distributions of the observables and the corresponding predictions specified by the model family. This chapter considers selected results of statistical inference under certain commonly considered additional assumptions: known observation covariances, normally distributed observables and models, and locally affine-linear models. While these assumptions are typically not exactly met in practice, they often lead to conveniently simple results which are often useful approximations and are the basis for efficient numerical treatment and computation. The results derived here form the basis for several of the design criteria considered in Chaps. 4 to 5 as well as for the applications and numerical results of Chap. 9.

The chapter is structured similarly to the previous one. Section 3.1 introduces the additional assumptions and the associated notation and formalism. Sections 3.2 and 3.3 contain the technical derivations of the Kullback-Leibler information criterion (KLIC), the likelihood, information matrices and related concepts under these assumptions.

Maximum-likelihood estimation is considered in Sec. 3.4, with a focus on the quantification of parameter uncertainty. As one of the central results of this thesis, a novel "robust" formula for the asymptotic covariance of parameter maximum-likelihood estimator (PMLE) is proposed and examined. It is valid even for models that are *both* incorrect *and* nonlinear, much in contrast to its typically used "classic" counterpart. This new results are the basis for the novel parameter-robust design criteria introduced in Chap. 5. Section 3.5 considers Bayesian inference under the additional assumptions.

3.1. Preliminaries: Central Assumptions

The following scenario summarizes the fundamental assumptions made in this chapter.

Scenario 3.1 (Statistical Inference)

(i) A process according to Def. 1.2 is given.

(ii) The function q characterizing the process is unknown.

(iii) Data is available from the process, consisting of observations from the observation domain $\mathcal{Y}$, obtained from a sequence of statistically independent experiments numbered $1, 2, \ldots$ performed under known conditions from the experimental domain $\mathcal{X}$.

For all $n \in \mathbb{N}$, the n-experiment exact design describing experiments 1 to n is denoted ξ_n, the data resulting from these experiments is denoted $d_n \in \mathcal{Y}^n$, and the corresponding sample is denoted $\mathcal{D}_n$.

(iv) A model family from Def. 1.3 is available for describing the process.

For each model $\mu \in \mathcal{M}$, the parameter domain $\mathcal{Q}^\mu$ is compact (and thus Lebesgue-measurable), and $p(y \mid x, \mu, \theta^\mu)$ is twice continuously differentiable and Lebesgue-measurable with respect to θ^μ for all $y \in \mathcal{Y}$ and all $x \in \operatorname{supp}(\xi_n)$ for all $n \in \mathbb{N}$.

(v) As a consequence of assumption (ii), it is not known whether the model family is correct for any of the designs $(\xi_n : n \in \mathbb{N})$ and the Kullback-Leibler information criterion (KLIC)-best models and KLIC-best parameters for these designs are unknown.

This scenario is identical to scenario 2.1 considered in the last chapter, it is repeated here for completeness.

3.1.1. Notation and Terminology

We continue to use the notation introduced in Sec. 1.3.1. In particular, $r_n(x) := n\xi_n(x)$ denotes the number of replications of the experiment under condition $x \in \operatorname{supp}(\xi_n)$, and $y_j(x) \in \mathcal{Y}$ denotes the observation resulting from replication no. $j \in \{1, \dots, r_n(x)\}$ of the experiment under $x \in \operatorname{supp}(\xi_n)$, for all $n \in \mathbb{N}$. Furthermore, $q(d_n \mid \xi_n)$ denotes the probability density function (PDF) of the sample $\mathcal{D}_n$, and $p(d_n \mid \xi_n, \mu, \theta^\mu)$ denotes the corresponding PDF specified by model μ with parameter θ^μ, for all $n \in \mathbb{N}$.

In addition, we use the following definitions. The OBSERVATION MEAN is

$$\bar{\eta}(x) := \int_{\mathcal{Y}} y q(y \mid x) \, \mathrm{d}y = \mathbb{E}\left[\mathcal{Y}_x\right], \tag{3.1}$$

and the RESPONSE of model $\mu \in \mathcal{M}$ is

$$\eta^\mu(x, \theta^\mu) := \int_{\mathcal{Y}} y p(y \mid x, \mu, \theta^\mu) \, \mathrm{d}y, \tag{3.2}$$

supposed that the expectations exist. Let $\eta^\mu_l(x, \theta^\mu)$ denote the l-th component of vector $\eta^\mu(x, \theta^\mu)$, and ∇ and ∇^2 denote the gradient and the Hessian differential operator, respectively, with respect to θ^μ. The RESPONSE JACOBIAN of model

$\mu \in \mathcal{M}$ is the $n_y \times n_{\theta^\mu}$ matrix

$$\boldsymbol{J}^\mu(x, \theta^\mu) := \nabla \eta^\mu(x, \theta^\mu), \tag{3.3}$$

and response Hessians of model μ are the $n_{\theta^\mu} \times n_{\theta^\mu}$ matrices

$$\boldsymbol{H}_l^\mu(x, \theta^\mu) := \nabla^2 \eta_l^\mu(x, \theta^\mu), \text{ with } l \in \{1, \dots, n_y\}. \tag{3.4}$$

supposed that the response is sufficiently differentiable with respect to θ^μ.

3.1.2. Known Observation Covariances

Definition 3.2 (Known Observation Covariances)

The observation covariances are known under design ξ, if for all $x \in \operatorname{supp}(\xi)$, the observation covariance $\mathbb{C}[\mathcal{Y}_x]$ exists, has full rank and is known. Without loss of generality (wlog), we then assume that

$$\mathbb{C}[\mathcal{Y}_x] = \mathbb{C}\left[\tilde{\mathcal{Y}}(x, \mu, \theta^\mu)\right] = \boldsymbol{I} \tag{3.5}$$

under all $x \in \operatorname{supp}(\xi)$, for all $\mu \in \mathcal{M}$, and all $\theta^\mu \in \mathcal{Q}^\mu$. The observation covariances are known, iff they are known under all designs $\xi \in \Xi$.

This is an assumption about an *actual property of the process*. In practice, we do not know how well it is actually met. If experimental data is available, violations of this assumption can in principle be detected (in a probabilistic sense) using statistical hypothesis tests. Any practical application of a result or method derived under this assumption should be accompanied by suitable statistical tests to detect (in a probabilistic sense) if it is violated. Performing such tests are standard tasks from applied statistics that we shall not mention anymore.

Note that if the observation covariances are known under ξ, then the observation mean $\bar{\eta}(x)$ exists under all $x \in \operatorname{supp}(\xi)$.

Equation (3.5) mandates some explanation: If the observation covariances $\boldsymbol{\Omega}(x) := \mathbb{C}[\mathcal{Y}_x]$ are known under some design, the generalized standard deviations $\boldsymbol{\Omega}^{½}(x)$ and their inverses $\boldsymbol{\Omega}^{-½}(x)$ exist and are known, too. The matrix square root is defined in Thm. A.2. Instead of $\mathcal{Y}_x$, one can then consider the normalized observables

$$\tilde{\mathcal{Y}}(x) := \boldsymbol{\Omega}^{-½}(x)\mathcal{Y}_x, \tag{3.6}$$

which have unit covariance by definition, since

$$\begin{aligned}\mathbb{C}\left[\tilde{\mathcal{Y}}(x)\right] = \mathbb{C}\left[\boldsymbol{\Omega}^{-1/2}(x)\mathcal{Y}_x\right] &= \boldsymbol{\Omega}^{-1/2}(x)\,\mathbb{C}\left[\mathcal{Y}_x\right]\boldsymbol{\Omega}^{-1/2\,\top}(x)\\ &= \boldsymbol{\Omega}^{-1/2}(x)\boldsymbol{\Omega}(x)\boldsymbol{\Omega}^{-1/2\,\top}(x) = \boldsymbol{I}. \quad (3.7)\end{aligned}$$

Any result based on $\tilde{\mathcal{Y}}(x)$ can be generalized to an observable $\mathcal{Y}_x$ with known, full-rank covariance $\boldsymbol{\Omega}(x)$ using the inverse of transformation (3.6). The corresponding transformation formulas for central quantities of this chapter are summarized at the end of the chapter in Tab. 3.1 on p. 113.

Expressions (3.6) and (3.7) justify to assume wlog that $\mathbb{C}\left[\mathcal{Y}_x\right] = \boldsymbol{I}$. If the observation covariance is known, this knowledge can be incorporated into the model formulation, such that one can wlog assume that $\mathbb{C}\left[\tilde{\mathcal{Y}}(x, \mu, \theta^{\mu})\right] = \boldsymbol{I}$.

3.1.3. Normal Processes and Models

Definition 3.3 (Normal Process under Known Observation Covariances)

Suppose the observation covariances are known under design ξ. A process q is normal under ξ, iff for all $y \in \mathscr{Y}$ and under all $x \in \operatorname{supp}(\xi)$,

$$q(y \mid x) = \phi(y \mid \bar{\eta}(x), \boldsymbol{I}) \overset{\text{(B.12b)}}{=} \exp\left(-\tfrac{1}{2}\left\|\bar{\eta}(x) - y\right\|_2^2 + n_y \ln(2\pi)\right). \tag{3.8}$$

The process is normal, iff it is normal under all designs $\xi \in \Xi$.

Like Def. 3.2, this is an assumption about the *property of the actual process* whose validity we do not know in practice. Any practical application of results based on Def. 3.3 should likewise be accompanied by suitable statistical tests to ensure that it properly reflects the actual process under consideration.

In many cases, assuming a normal distribution for an observable can be justified by the central limit theorem, which roughly says that the observable is approximately normal if its randomness has its source in many additive random contributions, regardless of their individual distribution laws.

Definition 3.4 (Normal Model under Known Observation Covariance)

Suppose the observation covariances are known under design ξ. Model $\mu \in \mathscr{M}$ is NORMAL UNDER ξ, iff for all $y \in \mathscr{Y}$, all $\theta^\mu \in \mathscr{Q}^\mu$ and under all $x \in \operatorname{supp}(\xi)$,

$$\begin{aligned} p(y\,|\,x, \mu, \theta^\mu) &= \phi(y\,|\,\eta^\mu(x, \theta^\mu), \boldsymbol{I}) \\ &\overset{(\text{B.12b})}{=} \exp\Bigl(-\tfrac{1}{2}\|\eta^\mu(x, \theta^\mu) - y\|_2^2 + n_y \ln(2\pi)\Bigr), \end{aligned} \tag{3.9}$$

and the response $\eta^\mu(x, \theta^\mu)$ is Lebesgue-measurable with respect to θ^μ for all $x \in \operatorname{supp}(\xi)$. The model is NORMAL, iff it is normal under all designs $\xi \in \Xi$.

This is an assumption about a *choice of the scientist,* that is twofold motivated by the previous assumptions.

If we actually assume that process *is* normal and the observation covariance is known, and regard a model as an attempt to describe the process, it would be plainly illogical to consider any other model class than that specified by Def. 3.4. Second, even if we do *not* assume the process to be normal, but still know the observation covariance, the principle of maximum entropy (see Appendix C on p. 302) suggests using a normal model, since that has minimal "prejudice" among all distributions with given covariance, see Prop. C.9.

3.1.4. Locally Affine-Linear Models

Definition 3.5 (Locally Affine-Linear Model)

Model $\mu \in \mathscr{M}$ is LOCALLY AFFINE-LINEAR AROUND PARAMETER $\tilde{\theta}^\mu \in \mathscr{Q}^\mu$ UNDER DESIGN ξ, iff under all $x \in \operatorname{supp}(\xi)$ and all θ^μ in a neighborhood of $\tilde{\theta}^\mu$, its response $\eta^\mu(x, \theta^\mu)$ exist, is Lebesgue-measurable and differentiable with respect to θ^μ, and the error of approximation

$$\eta^\mu(x, \theta^\mu) \approx \eta^\mu\bigl(x, \tilde{\theta}^\mu\bigr) + \boldsymbol{J}^\mu\bigl(x, \tilde{\theta}\bigr)\bigl(\theta^\mu - \tilde{\theta}^\mu\bigr) \tag{3.10}$$

is small (in a yet to be defined sense). It is LOCALLY AFFINE-LINEAR AROUND $\tilde{\theta}^\mu$, iff it is locally affine-linear around $\tilde{\theta}^\mu \in \mathscr{Q}^\mu$ under all designs $\xi \in \Xi$.

This definition purposely leaves open how to measure the error of the approximation, and when to consider it as "small". The technical details could easily be specified, but are not required in this thesis. If model μ is locally affine-linear around $\tilde{\theta}^{\mu}$, we write its approximate response as

$$\eta^{\mu}(x, \theta^{\mu}) \approx \boldsymbol{J}^{\mu}\big(x, \tilde{\theta}\big)\theta^{\mu} + h^{\mu}\big(x, \tilde{\theta}^{\mu}\big),$$
$$\text{with } h^{\mu}\big(x, \tilde{\theta}^{\mu}\big) := \eta^{\mu}\big(x, \tilde{\theta}^{\mu}\big) - \boldsymbol{J}^{\mu}\big(x, \tilde{\theta}^{\mu}\big)\tilde{\theta}^{\mu}. \tag{3.11}$$

3.2. Measuring the Process–Model Discrepancy

We are now equipped to consider the definitions related to process–model discrepancy from Sec. 1.4 under the additional assumptions of this chapter.

If the observation covariances are known and model $\mu \in \mathcal{M}$ are normal and correct, process and model can differ only in their means $\bar{\eta}$ and η^{μ}, respectively. A correct parameter (Def. 1.6) can thus be characterized in the following simplified way.

Corollary 3.6 (Correct Parameters under Known Observation Covariances and Normal Models)

Suppose that the observation covariances are known (under design ξ) and model $\mu \in \mathcal{M}$ is normal (under ξ). Then, parameter $\theta^{\mu} \in \mathcal{Q}^{\mu}$ is correct (under ξ), iff the process is normal (under ξ) and

$$\bar{\eta}(x) = \eta^{\mu}(x, \theta^{\mu}) \tag{3.12}$$

under all $x \in \mathcal{X}$ (under all $x \in \operatorname{supp}(\xi)$).

The definitions of correctness for models and model families (Items (ii) and (iii) of Def. 1.6) remain unchanged.

Definition 3.7 (Noncentrality)

Suppose the observation covariances are known under design ξ and the response $\eta^{\mu}(x, \theta^{\mu})$ of model $\mu \in \mathcal{M}$ exist under all $x \in \operatorname{supp}(\xi)$. The NONCENTRALITY OF MODEL μ is

$$\lambda^{\mu}(\theta^{\mu}, \xi) := \sum_{x \in \operatorname{supp}(\xi)} \xi(x) \left\| \eta^{\mu}(x, \theta^{\mu}) - \bar{\eta}(x) \right\|_2^2. \tag{3.13}$$

If both process and model are normal, the distributions specified by process and model can differ only in their respective means $\bar{\eta}$ and η^{μ}. Being a weighted sum of squares over their difference, the noncentrality $\lambda^{\mu}(\theta^{\mu}, \xi)$ measures the overall discrepancy between model μ with parameter θ^{μ} and the process under design ξ.

Corollary 3.8 (KLIC and KLIC-Best Parameters and Models under Known Observation Covariances and Normality)

Suppose that under design ξ, the observation covariances are known and the process and all models $\mu \in \mathcal{M}$ are normal. Then the following statements hold.

(i) The Kullback-Leibler information criterion (KLIC) from Def. 1.7 equals half the noncentrality,

$$\delta(\mu, \theta^{\mu}, \xi) = \tfrac{1}{2}\lambda^{\mu}(\theta^{\mu}, \xi). \tag{3.14}$$

(ii) Parameter $\theta^{\mu}(\xi)$ is KLIC-best under design ξ, iff

$$\theta^{\mu}(\xi) \in \operatorname*{argmin}_{\theta^{\mu} \in \mathcal{Q}^{\mu}} \lambda^{\mu}(\theta^{\mu}, \xi). \tag{3.15}$$

The parameter is IDENTIFIABLE, iff it is a *unique* minimizer.

(iii) Model $\mu(\xi)$ is KLIC-best under design ξ, iff

$$\mu(\xi) \in \operatorname*{argmin}_{\mu \in \mathcal{M}} \lambda^{\mu}(\theta^{\mu}(\xi), \xi). \tag{3.16}$$

The model is IDENTIFIABLE, iff it is a *unique* minimizer.

Proof Item (i) follows from the expression for the Kullback-Leibler distance (KLD) between normal distributions from (Thm. C.10) and the definition of the noncentrality (Def. 3.7). Based thereon, (ii) and (iii) follow immediately from the definition of best parameters and best models (Def. 1.9). □

Recall that the KLIC (like the KLD) is generally *not* a metric in the space of distributions. Under the considered additional assumptions, it reduces to the noncentrality, which notably *is* a metric between the model responses and the observation means under the given design.

3.3. Likelihood and Information Matrices

The central results from maximum-likelihood estimation and Bayesian inference considered in the previous chapter are expressed in terms of the likelihood and the information matrices introduced in Sec. 2.2. Here, we derive the particular simpler forms that these quantities take under the additional assumptions of this chapter. This section contains the technical derivations only, its results become meaningful in the context of maximum-likelihood and Bayesian inference considered in the subsequent Secs. 3.4 and 3.5.

3.3.1. Likelihood

Definition 3.9 (Sum of Squared Residuals (SSR))

Let $d_n \in \mathcal{Y}^n$ be the data obtained under the n-experiment exact design ξ_n and let $r_n(x) := n\xi_n(x)$ denote the number of replications of the experiment under $x \in \operatorname{supp}(\xi_n)$. Suppose that under ξ_n, the observation covariances are known and the response $\eta^\mu(x, \theta^\mu)$ exists. The SUM OF SQUARED RESIDUALS (SSR) OF MODEL μ is

$$s^\mu(\theta^\mu, d_n, \xi_n) := \frac{1}{n} \sum_{x \in \operatorname{supp}(\xi_n)} \sum_{j=1}^{r_n(x)} \left\| \eta^\mu(x, \theta^\mu) - y_j(x) \right\|_2^2. \tag{3.17}$$

A minimizer of the SSR with respect to $\theta^\mu \in \mathcal{Q}^\mu$ is a LEAST-SQUARES (LSQ) ESTIMATE. The SSR is the empirical counterpart of the noncentrality.

Corollary 3.10 (Likelihood and Parameter and Model MLEs under Known Observation Covariances and Normal Models)

Let d_n be the data obtained under the n-experiment exact design ξ_n. The following statements hold if the observation covariances under ξ_n are known and all models $\mu \in \mathcal{M}$ are normal under ξ_n.

(i) The log-likelihood is proportional to the negative SSR,

$$\ln p\left(d_n \,\middle|\, \xi_n, \mu, \theta^\mu\right) = -\tfrac{n}{2}\left(s^\mu\left(\theta^\mu, d_n, \xi_n\right) + n_y \ln(2\pi)\right). \tag{3.18}$$

(ii) Parameter $\hat{\theta}^\mu(d_n, \xi_n)$ is a parameter maximum-likelihood estimate (PMLE), iff

$$\hat{\theta}^\mu(d_n, \xi_n) \in \operatorname*{argmin}_{\theta^\mu \in \mathcal{Q}^\mu} s^\mu(\theta^\mu, d_n, \xi_n), \tag{3.19}$$

that is, iff it is a least-squares (LSQ) estimate.

(iii) Model $\hat{\mu}(d_n, \xi_n)$ is a model maximum-likelihood estimate (MMLE), iff

$$\hat{\mu}(d_n, \xi_n) \in \operatorname*{argmin}_{\mu \in \mathcal{M}} s^\mu\big(\hat{\theta}^\mu(d_n, \xi_n), d_n, \xi_n\big). \tag{3.20}$$

Proof Item (i) results from substituting the probability density function (PDF) of a normal distribution (B.12b) into the log-likelihood (2.5) and writing the result using the SSR (3.17). Items (ii) and (iii) follow immediately by applying (i) to Def. 2.13. □

3.3.2. Information Matrices

It is convenient to define the additional matrices. Let ξ be some design and $\mu \in \mathcal{M}$. Suppose that under all $x \in \operatorname{supp}(\xi)$, the observation mean $\eta(x)$ exists and the response $\eta^\mu(x, \theta^\mu)$ is twice differentiable with respect to θ^μ. Let $\bar{\eta}_l(\cdot)$ and $\eta_l^\mu(\cdot)$ denote the l-th component of $\bar{\eta}(x)$ and $\eta^\mu(\cdot)$, respectively. We define the following matrice:

$$\boldsymbol{M}^\mu(\theta^\mu, \xi) := \sum_{x \in \operatorname{supp}(\xi)} \xi(x) \boldsymbol{J}^{\mu\top}(x, \theta^\mu) \boldsymbol{J}^\mu(x, \theta^\mu), \tag{3.21}$$

$$\tilde{\boldsymbol{N}}^\mu(\theta^\mu, \xi) := \sum_{x \in \operatorname{supp}(\xi)} \xi(x) \sum_{l=1}^{n_y} \big(\eta_l^\mu(x, \theta^\mu) - \bar{\eta}_l(x)\big) \boldsymbol{H}_l^\mu(x, \theta^\mu), \tag{3.22}$$

$$\begin{aligned} \tilde{\boldsymbol{R}}^\mu(\theta^\mu, \xi) := {} & \big(\boldsymbol{M}^\mu(\theta^\mu, \xi) + \tilde{\boldsymbol{N}}^\mu(\theta^\mu, \xi)\big)^{-1} \boldsymbol{M}^\mu(\theta^\mu, \xi) \\ & \cdot \big(\boldsymbol{M}^\mu(\theta^\mu, \xi) + \tilde{\boldsymbol{N}}^\mu(\theta^\mu, \xi)\big)^{-1}. \end{aligned} \tag{3.23}$$

These matrices depend on the unknown observation mean, but do not involve any data.

Let ξ_n be a n-experiment exact design ξ_n, and let $r_n(x) := n\xi_n(x)$ denote the number of replications of the experiment under $x \in \operatorname{supp}(\xi_n)$. Suppose that

$\eta^\mu(x, \theta^\mu)$ is twice differentiable with respect to θ^μ under all $x \in \operatorname{supp}(\xi_n)$. Let d_n be data obtained under ξ_n, and $y_{jl}(x)$ be the l-th component of the observation made in the j-th repetition of the experiment under x and define

$$\boldsymbol{N}^\mu(\theta^\mu, d_n, \xi_n) := \tfrac{1}{n} \sum_{x \in \operatorname{supp}(\xi_n)} \sum_{j=1}^{r_n(x)} \sum_{l=1}^{n_y} \big(\eta_l^\mu(x, \theta^\mu) - y_{jl}(x)\big) \boldsymbol{H}_l^\mu(x, \theta^\mu). \tag{3.24}$$

This matrix is data-dependent, but does not involve any unknown function derived from the process.

All these matrices are $n_{\theta^\mu} \times n_{\theta^\mu}$ and symmetric. Furthermore, $\boldsymbol{M}^\mu$ and $\tilde{\boldsymbol{R}}^\mu$ are symmetric positive semi-definite (spsd), if they exist. If evaluated for an exact n-experiment design, all matrices are of $\mathcal{O}(1)$ with respect to n.

Theorem 3.11 (Information Matrices under Known Observation Covariances, Normal Models and Correctness)

Suppose that under design ξ and under exact design ξ_n, the observation covariances are known, model $\mu \in \mathcal{M}$ is normal, and its responses $\eta^\mu(x, \theta^\mu)$ are differentiable twice in θ^μ. Let d_n be the data obtained under ξ_n. Then,

$$\boldsymbol{F}_n^\mu(\theta^\mu, d_n, \xi_n) = \boldsymbol{M}^\mu(\theta^\mu, \xi_n) + \boldsymbol{N}^\mu(\theta^\mu, d_n, \xi_n), \tag{3.25}$$

$$\tilde{\boldsymbol{F}}^\mu(\theta^\mu, \xi) = \boldsymbol{M}^\mu(\theta^\mu, \xi) + \tilde{\boldsymbol{N}}^\mu(\theta^\mu, \xi), \tag{3.26}$$

$$\tilde{\boldsymbol{G}}^\mu(\theta^\mu, \xi) = \boldsymbol{M}^\mu(\theta^\mu, \xi), \tag{3.27}$$

and all matrices exist. If $\boldsymbol{M}^\mu(\theta^\mu, \xi) + \tilde{\boldsymbol{N}}^\mu(\theta^\mu, \xi)$ is invertible, also

$$\tilde{\boldsymbol{S}}^\mu(\theta^\mu, \xi) = \tilde{\boldsymbol{R}}^\mu(\theta^\mu, \xi). \tag{3.28}$$

If, in addition, the model is correct under ξ and $\bar{\theta}^\mu$ is a corresponding correct parameter, then

$$\tilde{\boldsymbol{F}}^\mu\big(\bar{\theta}^\mu, \xi\big) = \tilde{\boldsymbol{G}}^\mu\big(\bar{\theta}^\mu, \xi\big) = \boldsymbol{M}^\mu\big(\bar{\theta}^\mu, \xi\big), \tag{3.29}$$

and the matrices exit. If $\boldsymbol{M}^\mu\big(\bar{\theta}^\mu, \xi\big)$ has full rank (and is thus invertible), also

$$\tilde{\boldsymbol{S}}^\mu\big(\bar{\theta}^\mu, \xi\big) = \big(\boldsymbol{M}^\mu\big(\bar{\theta}^\mu, \xi\big)\big)^{-1}. \tag{3.30}$$

Proof For clarity, we omit the model index μ in the proof.

The l-th component of the vector $\mathcal{Y}_x$, y, $y_j(x)$, and $\eta(x)$ are denoted $\mathcal{Y}_l(x)$, y_l, $y_{jl}(x)$, and $\eta_l(x)$, respectively. As the observation covariances are known, $\bar{\eta}(x) = \mathbb{E}\left[\mathcal{Y}_x\right]$ exists and $\mathbb{C}\left[\mathcal{Y}_x\right] = \boldsymbol{I}$ for all the conditions from the support of ξ and ξ_n.

Since the model is normal, $\ln p(y\,|\,x,\theta) = -\frac{1}{2}\left\|\eta(x,\theta) - y\right\|_2^2 + \text{const}$. The corresponding gradient and Hessian with respect to the parameter are

$$\nabla \ln p(y\,|\,x,\theta) \overset{(A.7)}{=} -\boldsymbol{J}^{\top}(x,\theta)(\eta(x,\theta) - y) \text{ and} \tag{3.31}$$

$$\begin{aligned}\nabla^2 \ln p(y\,|\,x,\theta) \overset{(A.8)}{=} & -\boldsymbol{J}^{\top}(x,\theta)\boldsymbol{J}(x,\theta) \\ & - \sum_{l=1}^{n_y}(\eta_l(x,\theta) - y_l)\boldsymbol{H}_l(x,\theta),\end{aligned} \tag{3.32}$$

respectively. Equalities (3.25)–(3.27) are then derived as follows:

$$\begin{aligned}\boldsymbol{F}_n(\theta, d_n, \xi_n) & \overset{(2.12)}{=} -\tfrac{1}{n}\sum_{x\in\mathrm{supp}(\xi_n)}\sum_{j=1}^{r_n(x)}\nabla^2 \ln p\big(y_j(x)\,\big|\,x,\theta\big) \\ & \overset{(3.32)}{=} \tfrac{1}{n}\sum_{x\in\mathrm{supp}(\xi_n)}\sum_{j=1}^{r_n(x)}\left(\boldsymbol{J}^{\top}(x,\theta)\boldsymbol{J}(x,\theta) + \sum_{l=1}^{n_y}\big(\eta_l(x,\theta) - y_{jl}(x)\big)\boldsymbol{H}_l(x,\theta)\right) \\ & \overset{(3.21),(3.24)}{=} \boldsymbol{M}(\theta,\xi_n) + \boldsymbol{N}(\theta, d_n, \xi_n),\end{aligned} \tag{3.33}$$

$$\begin{aligned}\tilde{\boldsymbol{F}}(\theta,\xi) & \overset{(2.6)}{=} -\sum_{x\in\mathrm{supp}(\xi)}\xi(x)\,\mathbb{E}\left[\nabla^2 \ln p(\mathcal{Y}_x\,|\,x,\theta)\right] \\ & \overset{(3.32)}{=} \sum_{x\in\mathrm{supp}(\xi)}\xi(x)\left(\boldsymbol{J}^{\top}(x,\theta)\boldsymbol{J}(x,\theta) + \sum_{l=1}^{n_y}(\eta_l(x,\theta) - \mathbb{E}\left[\mathcal{Y}_l(x)\right])\boldsymbol{H}_l(x,\theta)\right) \\ & \overset{(3.21),(3.22)}{=} \boldsymbol{M}(\theta,\xi) + \tilde{\boldsymbol{N}}(\theta,\xi), \text{ and}\end{aligned} \tag{3.34}$$

$$\tilde{\boldsymbol{G}}(\theta,\xi) \overset{(2.7)}{=} \sum_{x\in\mathrm{supp}(\xi)}\xi(x)\,\mathbb{C}\left[\nabla \ln p(\mathcal{Y}_x\,|\,x,\theta)\right] \tag{3.35}$$

$$\overset{(3.31)}{=} \sum_{x\in\mathrm{supp}(\xi)}\xi(x)\,\mathbb{C}\left[-\boldsymbol{J}^{\top}(x,\theta)(\eta(x,\theta) - \mathcal{Y}_x)\right] \tag{3.36}$$

$$= \sum_{x\in\mathrm{supp}(\xi)}\xi(x)\big(\boldsymbol{J}^{\top}(x,\theta)\,\mathbb{C}\left[\mathcal{Y}_x\right]\boldsymbol{J}(x,\theta)\big) \tag{3.37}$$

$$= \sum_{x\in\mathrm{supp}(\xi)}\xi(x)\boldsymbol{J}^{\top}(x,\theta)\boldsymbol{J}(x,\theta) \tag{3.38}$$

$$\overset{(3.21)}{=} \boldsymbol{M}(\theta, \xi). \tag{3.39}$$

Based thereon, (3.28) follows from (2.8) and (3.23). Under the assumption of a correct model, $\eta_i\left(\bar{\theta}\right) - \bar{\eta}_i = 0$ for all $i \in \{1, \ldots, s\}$, see Cor. 3.6, and thus $\tilde{\boldsymbol{N}}\left(\bar{\theta}, \xi\right) \equiv \boldsymbol{0}$. Substitution into (3.25)–(3.28) leads to (3.29) and (3.30). □

It is important to realize that none of these results relies on the assumption that the observables are normally distributed.

Corollary 3.12 (Identifiability of KLIC-Best Parameters under Known Observation Covariances, Normal Models and Correctness)

Suppose that under design ξ, the observation covariances are known, model $\mu \in \mathcal{M}$ is normal, and the model response $\eta^\mu(\theta^\mu, x)$ is twice continuously differentiable with respect to θ^μ. Let $\bar{\theta}^\mu$ be an interior point of $\mathcal{Q}^\mu$. Then, under regularity conditions, the following statements hold.

(i) If $\bar{\theta}^\mu$ is an identifiable Kullback-Leibler information criterion (klic)-best parameter under ξ and $\boldsymbol{M}^\mu(\theta^\mu, \xi) + \tilde{\boldsymbol{N}}^\mu(\theta^\mu, \xi)$ has constant rank for all θ^μ in an open neighborhood of $\bar{\theta}^\mu$, then $\boldsymbol{M}^\mu\left(\bar{\theta}^\mu, \xi\right) + \tilde{\boldsymbol{N}}^\mu\left(\bar{\theta}^\mu, \xi\right)$ has full rank (and is thus invertible).

(ii) If $\bar{\theta}^\mu$ is a klic-best parameter under ξ and $\boldsymbol{M}^\mu\left(\bar{\theta}^\mu, \xi\right) + \tilde{\boldsymbol{N}}^\mu\left(\bar{\theta}^\mu, \xi\right)$ has full rank (and is thus invertible), then $\bar{\theta}^\mu$ is identifiable.

If model μ is correct under ξ and $\bar{\theta}^\mu$ is an interior point of $\mathcal{Q}^\mu$, also the next two statements hold.

(iii) If $\bar{\theta}^\mu$ is an identifiable correct parameter under ξ and $\boldsymbol{M}^\mu(\theta^\mu, \xi)$ has constant rank for all θ^μ from an open neighborhood of $\bar{\theta}^\mu$, then $\boldsymbol{M}^\mu\left(\bar{\theta}^\mu, \xi\right)$ has full rank (and is thus invertible).

(iv) If $\bar{\theta}^\mu$ is a correct parameter under ξ and $\boldsymbol{M}^\mu\left(\bar{\theta}^\mu, \xi\right)$ has full rank (and is thus invertible), then $\bar{\theta}^\mu$ is identifiable.

Proof The proof follows from applying Thm. 3.11 to Thm. 2.4. □

3.4. Maximum-Likelihood Estimation

This section deals with maximum-likelihood estimation in possibly incorrect models under the additional assumptions of known observation covariances and normal processes and models. In Sec. 3.4.1 we present formulas for the large-sample covariance of parameter maximum-likelihood estimators (PMLEs), which apply to models that may be *both* nonlinear *and* incorrect. They are essentially special cases of the general formulas given in Sec. 2.4.2, but have to the best of our knowledge not been stated explicitly so far. In Sec. 3.4.2 we examine the relation of these formulas to the "classic" ones which rely on assumptions of correctness or local linearity. As a main result of this thesis, we show in Sec. 3.4.3 that the asymptotic PMLEs covariance in possibly incorrect normal nonlinear models *can* in fact be consistently estimated – much in contrast to the general case discussed in Sec. 2.4. Section 3.4.4 describes how these results are applied in practice.

Throughout this section we make the following assumptions: The experiments are sampled from design ξ, such that the design sequence $\xi_1, \xi_2, \ldots$ converges to ξ. The Kullback-Leibler information criterion (KLIC)-best parameter $\bar{\theta}^\mu$ under ξ is identifiable in each model $\mu \in \mathcal{M}$, and the KLIC-best model $\bar{\mu}$ under ξ is identifiable. These are the assumptions made in the general treatment of maximum-likelihood estimation in Sec. 2.4. In addition, we assume in this section that under ξ and under all $(\xi_n : n \in \mathbb{N})$, the observation covariances are known and all models $\mu \in \mathcal{M}$ are normal.

Furthermore, $\hat{\mathcal{Q}}_n^\mu := \hat{\theta}^\mu(\mathcal{D}_n, \xi_n)$ denotes a PMLE of model $\mu \in \mathcal{M}$ and $\hat{\theta}_n^\mu := \hat{\theta}^\mu(d_n, \xi_n)$ a corresponding estimate. Likewise, $\hat{\mathcal{M}}_n := \hat{\mu}(\mathcal{D}_n, \xi_n)$ denotes a model maximum-likelihood estimator (MMLE) and $\hat{\mu}_n := \hat{\mu}(d_n, \xi_n)$ a corresponding estimate.

3.4.1. Large-Sample Properties of Parameter MLEs

Section 2.4.2 discussed the large-sample behavior of PMLEs in general. Its main result (2.21) is that under suitable regularity conditions, PMLEs are asymptotically normal with mean $\bar{\theta}^\mu$ and covariance $\frac{1}{n}\tilde{\mathbf{S}}^\mu\big(\bar{\theta}^\mu, \xi\big)$, where $\tilde{\mathbf{S}}^\mu$ is the expected sandwich information from (2.8). The main effect of the additional assumptions considered here are simplified expressions for the asymptotic PMLE covariance, direct consequences of Thm. 3.11.

Under these assumptions, $\tilde{\mathbf{S}}^\mu$ simplifies as stated in (3.28), leading to the large-sample approximation

$$\mathbb{C}\left[\hat{Q}_n^\mu\right] \overset{\infty}{\approx} \frac{1}{n}\tilde{\boldsymbol{R}}^\mu\left(\bar{\theta}^\mu, \xi\right) \tag{3.40}$$

for PMLE covariance, with $\tilde{\boldsymbol{R}}^\mu$ from (3.23). This approximation is justified even in models that are incorrect. In other words, it is robust with respect to systematical model errors. We refer to its right-hand side as the robust covariance formula for PMLEs. It depends on the unknown process both via the matrix $\tilde{\boldsymbol{R}}^\mu$ and via the best parameter $\bar{\theta}^\mu$.

Regularity Conditions

The regularity conditions for asymptotic normality of PMLE in general are discussed in Sec. 2.4.2, Under the considered assumptions, they simplify significantly. Apart from some technicalities, (3.40) requires that

(a) the experiments are sampled from design ξ, as already mentioned,

(b) the model response $\eta^\mu(\theta^\mu, x)$ is twice continuously differentiable with respect to θ^μ for all $x \in \operatorname{supp}(\xi)$,

(c) $\bar{\theta}^\mu \in \mathcal{Q}^\mu$ is an identifiable best parameter of model μ,

(d) $\bar{\theta}^\mu$ is an interior point of $\mathcal{Q}^\mu$,

(e) $\boldsymbol{M}^\mu(\theta^\mu, \xi) + \tilde{\boldsymbol{N}}^\mu(\theta^\mu, \xi)$ has constant rank for all θ^μ from an open neighborhood of $\bar{\theta}^\mu$, and

(f) $\boldsymbol{M}^\mu\left(\bar{\theta}^\mu, \xi\right)$ has full rank.

Condition (a) guarantees sufficient “repetition” in the sample, required for applying central limit theorems and laws of large numbers. Condition (b) is necessary to ensure that the involved information matrices exist and are continuous with respect to the parameter. Condition (c) makes the inference problem well posed. Condition (d) ensures that an open neighborhood of $\bar{\theta}$ exists. Together with the latter, condition (e) ensures that $\boldsymbol{M}^\mu\left(\bar{\theta}^\mu, \xi\right) + \tilde{\boldsymbol{N}}^\mu\left(\bar{\theta}^\mu, \xi\right)$ is invertible, according to Cor. 3.12(i). Finally, condition (f) is necessary to ensure that $\tilde{\boldsymbol{R}}^\mu\left(\bar{\theta}^\mu, \xi\right)$ has full rank and is thus a proper covariance matrix.

Note that a normal process is not required. If a normal process is assumed, the conditions can be simplified further based on the results summarized in Sec. 3.2.

Special Case: Correct Model

Suppose model μ is correct under ξ and let $\bar{\theta}^\mu$ be a corresponding correct parameter. Relation (2.24) states the asymptotic distribution of PMLEs for correct models in general. Under the additional conditions considered here, the involved information matrices simplify according to (3.29), leading to

$$\mathbb{C}\left[\hat{\mathcal{Q}}_n^\mu\right] \overset{\infty}{\approx} \left(n\boldsymbol{M}^\mu\left(\bar{\theta}^\mu,\xi\right)\right)^{-1}. \tag{3.41}$$

This approximation is a well-known result that can be found in many textbooks, for example in those of Pawitan [200, Chap. 9] or of Lehmann and Casella [170, Chap. 6]. We refer to its right-hand side as the CLASSIC COVARIANCE FORMULA for PMLEs. It can also be derived by replacing the correctness assumption by the assumption that the model is locally affine-linear around $\bar{\theta}^\mu$.

The correctness assumption implies that $\tilde{\boldsymbol{N}}^\mu\left(\bar{\theta}^\mu,\xi\right) = \boldsymbol{0}$. Accordingly, conditions (e) and (f) simplify to the requirement that $\boldsymbol{M}^\mu\left(\theta^\mu,\xi\right)$ has constant rank in an open neighborhood of $\bar{\theta}^\mu$.

3.4.2. Comparison of Classic and Robust Parameter MLE Covariance

If the model is correct or locally affine-linear around the best parameter, the asymptotic PMLE covariance is described by the classic covariance formula (3.41). If the model is *both* nonlinear *and* incorrect, this formula is not longer adequate and its robust counterpart (3.40) needs to be used. What error is made if the classic formula is applied in this case nevertheless?

For clarity, consider a particular model of the family and omit the model index μ and use the abbreviations $\bar{\boldsymbol{M}} := \boldsymbol{M}\left(\bar{\theta},\xi\right)$, $\bar{\tilde{\boldsymbol{N}}} := \tilde{\boldsymbol{N}}\left(\bar{\theta},\xi\right)$, and $\bar{\tilde{\boldsymbol{R}}} := \tilde{\boldsymbol{R}}\left(\bar{\theta},\xi\right)$. The question can then be restated as "What is the error when approximating $\bar{\tilde{\boldsymbol{R}}}$ through $\bar{\boldsymbol{M}}^{-1}$?" To answer this question we first introduce an alternative representation of $\tilde{\boldsymbol{R}}^\mu$.

Series Representation of Matrix $\tilde{\boldsymbol{R}}$

Define the NONLINEARITY[1]

$$\gamma(\theta,\xi) := \sum_{x\in\operatorname{supp}(\xi)} \xi(x) \sum_{j=1}^{n_y} \left\| \boldsymbol{H}_j(x,\theta)\right\|_{\mathrm{F}}^2. \tag{3.42}$$

For a model whose responses are affine-linear in θ^{μ}, $\gamma(\cdot,\xi) \equiv \boldsymbol{0}$. This definition, together with the definitions of the noncentrality λ from (3.13) and of matrix $\tilde{\boldsymbol{N}}$ from (3.22), imply the inequality

$$\left\|\tilde{\boldsymbol{N}}(\theta,\xi)\right\|_{\mathrm{F}}^2 \leqslant \lambda(\theta,\xi)\gamma(\theta,\xi), \text{ for all } \theta \in \mathcal{Q}. \tag{3.43}$$

Theorem 3.13 (Series Representation of Matrix $\tilde{\boldsymbol{R}}$)

Let ξ be some design. Suppose $\boldsymbol{M}(\theta,\xi) + \tilde{\boldsymbol{N}}(\theta,\xi)$ is invertible and $\boldsymbol{M}(\theta,\xi)$ has full rank (and is thus invertible) and

$$\gamma(\theta,\xi)\lambda(\theta,\xi) < \left\|\boldsymbol{M}^{-1}(\theta,\xi)\right\|_{\mathrm{F}}^{-2}. \tag{3.44}$$

Then, the matrix $\tilde{\boldsymbol{R}}$ from (3.23) can be expressed as the power series

$$\tilde{\boldsymbol{R}}(\theta,\xi) = \boldsymbol{M}^{-1}(\theta,\xi)\sum_{k=0}^{\infty}(k+1)\left(-\tilde{\boldsymbol{N}}(\theta,\xi)\boldsymbol{M}^{-1}(\theta,\xi)\right)^k. \tag{3.45}$$

Proof For brevity we omit the arguments θ and ξ in the proof. Since $\boldsymbol{M}$ is invertible, $\tilde{\boldsymbol{R}}$ can be rewritten as

$$\begin{aligned}\tilde{\boldsymbol{R}} &= \left(\boldsymbol{M}+\tilde{\boldsymbol{N}}\right)^{-1}\boldsymbol{M}\left(\boldsymbol{M}+\tilde{\boldsymbol{N}}\right)^{-1} = \left(\left(\boldsymbol{I}+\tilde{\boldsymbol{N}}\boldsymbol{M}^{-1}\right)\boldsymbol{M}\right)^{-1}\boldsymbol{M}\left(\left(\boldsymbol{I}+\tilde{\boldsymbol{N}}\boldsymbol{M}^{-1}\right)\boldsymbol{M}\right)\\ &= \boldsymbol{M}^{-1}\left(\boldsymbol{I}+\tilde{\boldsymbol{N}}\boldsymbol{M}^{-1}\right)^{-1}\boldsymbol{M}\boldsymbol{M}^{-1}\left(\boldsymbol{I}+\tilde{\boldsymbol{N}}\boldsymbol{M}^{-1}\right)^{-1} = \boldsymbol{M}^{-1}\left(\boldsymbol{I}+\tilde{\boldsymbol{N}}\boldsymbol{M}^{-1}\right)^{-2}.\end{aligned} \tag{3.46}$$

The Neumann series $\sum_{i=0}^{\infty}\left(-\tilde{\boldsymbol{N}}\boldsymbol{M}^{-1}\right)^i$ converges to $\left(\boldsymbol{I}+\tilde{\boldsymbol{N}}\boldsymbol{M}^{-1}\right)^{-1}$, if $\left\|\tilde{\boldsymbol{N}}\boldsymbol{M}^{-1}\right\|_{\mathrm{F}}^2 < 1$. Details can be found in most textbooks of functional analysis, for example that of Werner [264, Chap. 2]. If $\gamma\lambda = 0$, the latter condition is satisfied, since then $\tilde{\boldsymbol{N}} = \boldsymbol{0}$ and $\left\|\tilde{\boldsymbol{N}}\boldsymbol{M}^{-1}\right\|_{\mathrm{F}}^2 = 0$. For the case $\gamma\lambda \neq 0$ we get the upper bound $\left\|\tilde{\boldsymbol{N}}\boldsymbol{M}^{-1}\right\|_{\mathrm{F}}^2 \leqslant \left\|\tilde{\boldsymbol{N}}\right\|_{\mathrm{F}}^2\left\|\boldsymbol{M}^{-1}\right\|_{\mathrm{F}}^2 \leqslant \gamma\lambda\left\|\boldsymbol{M}^{-1}\right\|_{\mathrm{F}}^2$

[1] The provided definition is not a generally suitable measure of nonlinearity, as it only takes into account curvature and ignores all higher derivatives. Since the latter do not play a role in the considerations of this section, the provided definition suffices.

from the sub-multiplicativity of the norm and from (3.43). Therefore, the inequality $\gamma\lambda\left\|\boldsymbol{M}^{-1}\right\|_{\mathrm{F}}^{2} < 1$ from (3.44) is a sufficient condition for the convergence of the Neumann series. Substituting the series into (3.46) gives

$$\tilde{\boldsymbol{R}} = \boldsymbol{M}^{-1}\left(\sum_{i=0}^{\infty}\left(-\tilde{\boldsymbol{N}}\boldsymbol{M}^{-1}\right)^{i}\right)^{2} = \boldsymbol{M}^{-1}\sum_{i,j=0}^{\infty}\left(-\tilde{\boldsymbol{N}}\boldsymbol{M}^{-1}\right)^{i+j}. \tag{3.47}$$

To transform this double sum into a single one of the form $\sum_{k=0}^{\infty} c_k\left(-\tilde{\boldsymbol{N}}\boldsymbol{M}^{-1}\right)^{k}$, we have to appropriately weight each summand with its multiplicity in the double sum, denoted c_k. Calculating c_k amounts to answering the question “how many pairs $(i, j) \in \mathbb{N}_0 \times \mathbb{N}_0$ are there with $i + j = k$?” The answer $c_k = k + 1$ leading to (3.45) can also be formally derived from the multinomial theorem. □

Equipped with this theorem we can make some quick qualitative considerations considering the sought-after approximation error. We use the abbreviations $\bar{\gamma} := \gamma\left(\bar{\theta}, \xi\right)$ and $\bar{\lambda} := \lambda\left(\bar{\theta}, \xi\right)$. If theorem applies, it tells us that

$$\bar{\tilde{\boldsymbol{R}}} = \bar{\boldsymbol{M}}^{-1} - 2\bar{\boldsymbol{M}}^{-1}\bar{\tilde{\boldsymbol{N}}}\bar{\boldsymbol{M}}^{-1} + 3\bar{\boldsymbol{M}}^{-1}\bar{\tilde{\boldsymbol{N}}}\bar{\boldsymbol{M}}^{-1}\bar{\tilde{\boldsymbol{N}}}\bar{\boldsymbol{M}}^{-1} + \mathcal{O}\left(\left(\bar{\gamma}\bar{\lambda}\right)^{3}\right). \tag{3.48}$$

In other words, *the classic covariance formula (3.41) using $\bar{\boldsymbol{M}}^{-1}$ is an approximation of zeroth order in $\bar{\gamma}\bar{\lambda}$ of its robust counterpart (3.40) using $\bar{\tilde{\boldsymbol{R}}}$.* The classic formula can hence be expected to be adequate for models that are not too nonlinear in the parameter or do not exhibit too much systematical error in the sense that the product $\bar{\gamma}\bar{\lambda}$ is substantially smaller than $\left\|\bar{\boldsymbol{M}}^{-1}\right\|_{\mathrm{F}}^{-2}$. In other words, the asymptotic covariance of PMLEs is affected by systematical model errors only if its responses are nonlinear, and depends on second derivatives of the responses only if the model is incorrect.

Approximation Error

The relative error made when using $\bar{\boldsymbol{M}}^{-1}$ to approximate $\bar{\tilde{\boldsymbol{R}}}$ is

$$\frac{\left\|\bar{\boldsymbol{M}}^{-1} - \bar{\tilde{\boldsymbol{R}}}\right\|_{\mathrm{F}}}{\left\|\bar{\boldsymbol{M}}^{-1}\right\|_{\mathrm{F}}}. \tag{3.49}$$

A short calculation using the series representation (3.45), inequality (3.43), the triangle inequality and the sub-multiplicity of $\|\cdot\|_{\mathrm{F}}$ provides the series of inequalities

$$\begin{aligned}
\frac{\left\| \bar{\boldsymbol{M}}^{-1} - \bar{\tilde{\boldsymbol{R}}} \right\|_{\mathrm{F}}}{\left\| \bar{\boldsymbol{M}}^{-1} \right\|_{\mathrm{F}}} &\overset{(3.45)}{=} \frac{\left\| \bar{\boldsymbol{M}}^{-1} - \bar{\boldsymbol{M}}^{-1} \sum_{k=0}^{\infty} (k+1) \left(-\bar{\tilde{\boldsymbol{N}}} \bar{\boldsymbol{M}}^{-1} \right)^k \right\|_{\mathrm{F}}}{\left\| \bar{\boldsymbol{M}}^{-1} \right\|_{\mathrm{F}}} && (3.50) \\
&\leqslant \left\| \boldsymbol{I} - \sum_{k=0}^{\infty} (k+1) \left(-\bar{\tilde{\boldsymbol{N}}} \bar{\boldsymbol{M}}^{-1} \right)^k \right\|_{\mathrm{F}} && (3.51) \\
&= \left\| \boldsymbol{I} - \boldsymbol{I} - \sum_{k=1}^{\infty} (k+1) \left(-\bar{\tilde{\boldsymbol{N}}} \bar{\boldsymbol{M}}^{-1} \right)^k \right\|_{\mathrm{F}} && (3.52) \\
&\leqslant \sum_{k=1}^{\infty} (k+1) \left\| \left(-\bar{\tilde{\boldsymbol{N}}} \bar{\boldsymbol{M}}^{-1} \right)^k \right\|_{\mathrm{F}} && (3.53) \\
&\leqslant \sum_{k=1}^{\infty} (k+1) \left\| \bar{\tilde{\boldsymbol{N}}}^k \right\|_{\mathrm{F}} \left\| \bar{\boldsymbol{M}}^{-k} \right\|_{\mathrm{F}} && (3.54) \\
&\leqslant \sum_{k=1}^{\infty} (k+1) \left(\bar{\lambda} \bar{\gamma} \right)^{k/2} \left\| \bar{\boldsymbol{M}}^{-1} \right\|_{\mathrm{F}}^k . && (3.55)
\end{aligned}$$

If $\bar{\lambda}\bar{\gamma}$ is sufficiently smaller than $\left\| \bar{\boldsymbol{M}}^{-1} \right\|_{\mathrm{F}}^{-2}$, the first summand $2\left(\bar{\lambda}\bar{\gamma}\right)^{½} \left\| \bar{\boldsymbol{M}}^{-1} \right\|_{\mathrm{F}}$ dominates the sum in the last expression and is hence a good approximation for the error bound.

Remarkably, only the product of noncentrality $\bar{\lambda}$ and incorrectness $\bar{\gamma}$ enters the error bound. Hence, *a sufficiently small systematic error can compensate for large nonlinearity of the model and vice versa.* This might explain why the classic covariance formula often turns out to be adequate even if its premise of a correct model is violated.

The error bound (3.50) depends on quantities that are unknown practice. As discussed in the next section, $\bar{\boldsymbol{M}}^{-1}$ can under mild conditions be estimated consistently. Based on the strong consistency of the PMLE, it should be possible to show the same for $\bar{\lambda}$ and $\bar{\gamma}$.

3.4.3. Consistent Estimation of Parameter MLE Covariance

Neither the classic nor the robust formulas for the PMLEs covariance can be evaluated in practice. Under certain conditions examined in this section, they can be estimated consistently by their empirical counterparts. This topic was considered on a general level in Sec. 2.4.3.

This section uses the same setting and notation as the previous one, with the difference that $\hat{\mathcal{Q}}_n^\mu$ may be any strongly consistent estimator of the best parameter $\bar{\theta}^\mu$ under ξ, and is not necessarily a parameter maximum-likelihood estimate (PMLE).

Correct Models

Let us start with the classic assumption that model $\mu \in \mathcal{M}$ is correct under ξ and let $\bar{\theta}^\mu$ denote a corresponding correct parameter. The PMLE covariance is then asymptotically given by (3.41), and is unknown since $\bar{\theta}$ unknown. Under these assumptions, Thm. 3.11 can be applied to (2.25), leading to

$$\boldsymbol{M}^\mu\left(\hat{\mathcal{Q}}_n^\mu, \xi_n\right) + \boldsymbol{N}^\mu\left(\hat{\mathcal{Q}}_n^\mu, \mathcal{D}_n, \xi_n\right) \xrightarrow{\text{a.s.}} \boldsymbol{M}^\mu\left(\bar{\theta}^\mu, \xi\right) \quad \text{element-wise, for } n \to \infty. \tag{3.56}$$

In fact, all elements of $\boldsymbol{N}^\mu\left(\hat{\mathcal{Q}}_n^\mu, \mathcal{D}_n, \xi_n\right)$ converge almost surely to zero in a correct model, so that

$$\boldsymbol{M}^\mu\left(\hat{\mathcal{Q}}_n^\mu, \xi_n\right) \xrightarrow{\text{a.s.}} \boldsymbol{M}^\mu\left(\bar{\theta}^\mu, \xi\right) \text{ element-wise, for } n \to \infty. \tag{3.57}$$

This relation could also be derived more directly using the strong consistency of PMLEs and a generalized variant of Slutsky's theorem stated in Thm. B.4. The required regularity conditions are essentially a subset of those required for asymptotic normality, particularly conditions (a)–(d).

If $\boldsymbol{M}^\mu(\theta^\mu, \xi)$ has constant rank in vicinity of $\bar{\theta}^\mu$, it has full rank and is invertible Cor. 3.12(iii), so that the inverse of $\boldsymbol{M}^\mu\left(\hat{\mathcal{Q}}_n^\mu, \xi_n\right)$ exists asymptotically almost surely. Therefore, $\left(n\boldsymbol{M}^\mu\left(\hat{\mathcal{Q}}_n^\mu, \xi_n\right)\right)^{-1}$ is a strongly consistent estimator of the asymptotic PMLE covariance $\left(n\boldsymbol{M}^\mu\left(\bar{\theta}^\mu, \xi\right)\right)^{-1}$.

One often encounters the following alternative derivation for this classic result. Assume that the model is possibly incorrect, but locally affine-linear around $\hat{\mathcal{Q}}_n^\mu$ under ξ. It is a classic result of linear maximum-likelihood theory, found in most textbooks, that the *exact* PMLE covariance in the *linearized* model is $\left(n\boldsymbol{M}^\mu\left(\hat{\mathcal{Q}}_n^\mu, \xi_n\right)\right)^{-1}$, regardless of the sample size n. This expression can then be considered as *approximation* for the PMLE covariance in the actually *nonlinear* model.

Exchanging the correctness assumption with a linearity assumption thus leads to the same formula, a result that does not surprise considering the discussion of the last section.

Possibly Incorrect Models

The situation gets more complicated without the classic assumptions of correctness and/or local linearity. We know from Sec. 2.4.2 that the PMLE covariance is in general – without any assumptions of linearity, normality or correctness – asymptotically given by $\frac{1}{n}\tilde{\boldsymbol{S}}^{\mu}\big(\bar{\theta}^{\mu},\xi\big)$, see (2.21). Unfortunately, no *generally valid* consistent estimator for $\tilde{\boldsymbol{S}}^{\mu}\big(\bar{\theta}^{\mu},\xi\big)$ is available, as discussed at the end of Sec. 2.4.3.

In the following we show that in the *particular case* of known observation covariances and a normal model, where $\tilde{\boldsymbol{S}}^{\mu}\big(\bar{\theta}^{\mu},\xi\big) = \tilde{\boldsymbol{R}}^{\mu}\big(\bar{\theta}^{\mu},\xi\big)$, such an estimator does in fact exist, even if the model is nonlinear and incorrect. Applying Thm. 3.11 to (2.27) tells us that

$$\boldsymbol{M}^{\mu}\big(\hat{\mathcal{Q}}_n^{\mu},\xi_n\big)+\boldsymbol{N}^{\mu}\big(\hat{\mathcal{Q}}_n^{\mu},\mathcal{D}_n,\xi_n\big)\xrightarrow{\text{a.s.}}\boldsymbol{M}^{\mu}\big(\bar{\theta}^{\mu},\xi\big)+\tilde{\boldsymbol{N}}^{\mu}\big(\bar{\theta}^{\mu},\xi\big) \tag{3.58}$$

element-wise for $n\to\infty$, under regularity assumptions including conditions (a)–(d).

If we add condition (e), it follows from Cor. 3.12(i) that the sum of $\boldsymbol{M}^{\mu}(\theta^{\mu},\xi)$ and $\tilde{\boldsymbol{N}}^{\mu}(\theta^{\mu},\xi)$ is invertible in vicinity of $\bar{\theta}^{\mu}$. Therefore, the inverse of the sum $\boldsymbol{M}^{\mu}\big(\hat{\mathcal{Q}}_n^{\mu},\xi_n\big)+\boldsymbol{N}^{\mu}\big(\hat{\mathcal{Q}}_n^{\mu},\mathcal{D}_n,\xi_n\big)$ exists asymptotically almost surely and can thus be used to consistently estimate the first and the last factor of $\tilde{\boldsymbol{R}}^{\mu}\big(\bar{\theta},\xi\big)$ defined in (3.23).

It remains to find a strongly consistent estimator of the middle factor of $\tilde{\boldsymbol{R}}^{\mu}\big(\bar{\theta},\xi\big)$. One can show that first and second summand in the left-hand side of (3.58) converge *separately* to the corresponding summands in the right hand side. Or, alternatively, one can use the generalized variant of Slutsky's theorem (Thm. B.4) and the strong consistency of PMLEs to show that

$$\boldsymbol{M}^{\mu}\big(\hat{\mathcal{Q}}_n^{\mu},\xi_n\big)\xrightarrow{\text{a.s.}}\boldsymbol{M}^{\mu}\big(\bar{\theta}^{\mu},\xi\big)\ \text{element-wise, for}n\to\infty, \tag{3.59}$$

again under certain regularity conditions, notably conditions (a)–(d). Adding condition (f) ensures that $\boldsymbol{M}^{\mu}\big(\hat{\mathcal{Q}}_n^{\mu},\xi_n\big)$ and thus $\tilde{\boldsymbol{R}}^{\mu}\big(\bar{\theta},\xi\big)$ has full rank asymptotically. The following conjecture summarizes this argumentation.

Conjecture 3.14 (Consistent Estimation of PMLE Covariance in Possibly Incorrect Normal Models under Known Observation Covariances)

Suppose that under design ξ, the observation covariances are known and model $\mu \in \mathscr{M}$ is normal. Define the empirical counterpart of $\tilde{\boldsymbol{R}}^{\mu}(\theta^{\mu}, \xi)$ as

$$\begin{aligned}\boldsymbol{R}^{\mu}(\theta^{\mu}, d_n, \xi) := {} & (\boldsymbol{M}^{\mu}(\theta^{\mu}, \xi) + \boldsymbol{N}^{\mu}(\theta^{\mu}, d_n, \xi))^{-1}\boldsymbol{M}^{\mu}(\theta^{\mu}, \xi) \\ & \cdot (\boldsymbol{M}^{\mu}(\theta^{\mu}, \xi) + \boldsymbol{N}^{\mu}(\theta^{\mu}, d_n, \xi))^{-1},\end{aligned} \tag{3.60}$$

supposed the inverse exists. Under the same regularity conditions ensuring the asymptotic normality of PMLEs, notably conditions (a)–(f) on p. 97, $\boldsymbol{R}^{\mu}\big(\hat{\mathcal{Q}}^{\mu}_n, \mathcal{D}_n, \xi_n\big)$ exists asymptotically almost surely and element-wise

$$\boldsymbol{R}^{\mu}\big(\hat{\mathcal{Q}}^{\mu}_n, \mathcal{D}_n, \xi_n\big) \xrightarrow{\text{a.s.}} \tilde{\boldsymbol{R}}^{\mu}\big(\bar{\theta}^{\mu}, \xi\big), \text{ for } n \to \infty \tag{3.61}$$

This conjecture suggests that $\frac{1}{n}\boldsymbol{R}^{\mu}\big(\hat{\mathcal{Q}}^{\mu}_n, \mathcal{D}_n, \xi_n\big)$ is a strongly consistent estimator of the asymptotic PMLE covariance $\frac{1}{n}\tilde{\boldsymbol{R}}^{\mu}\big(\bar{\theta}^{\mu}, \xi_n\big)$ in normal models with known observation covariances. *Under the common assumptions of known observation covariances and a normal model, the asymptotic covariance of PMLEs can estimated consistently, even in models that are nonlinear and incorrect, and under independently but not identically distributed (INID) experiments.* To the best of our knowledge, this is a previously unstated result. It permits to empirically quantify the parameter uncertainty and to make parameter-robust predictions. We use it eventually to propose novel parameter-robust design criteria in Chap. 5.

The key to this result is the equality $\tilde{\boldsymbol{G}}^{\mu}(\theta^{\mu}, \xi) = \boldsymbol{M}^{\mu}(\theta^{\mu}, \xi)$ from (3.27) which relies on known observation covariances and a normal model. It can be seen in (3.35) and (3.37) that in fact both assumption are vital: Under normality, the covariance of the log-likelihood gradient reduces to a linear function of the observation covariances, and only if the latter are known the resulting formula can actually be evaluated.

3.4.4. Practical Application

To ease the following discussion presume that all regularity conditions are satisfied that ensure the consistency and asymptotic normality of parameter and MMLEs.

Empirical Approximations of Unknown KLIC-Best Parameters and Models

In practice, a PMLE $\hat{\theta}_n^\mu$ can be determined from the data d_n obtained under design ξ_n by minimizing the sum of squared residuals (SSR) $s^\mu(\theta^\mu, d_n, \xi_n)$ with respect to $\theta^\mu \in \mathcal{Q}^\mu$. It is an empirical approximation for the corresponding unknown best parameter,

$$\bar{\theta}^\mu \overset{\infty}{\approx} \hat{\theta}_n^\mu. \tag{3.62}$$

When we say that the approximation is EMPIRICAL, we mean that it can be evaluated based on the data, and does not depend on any unknown quantities. The PMLE is asymptotically normal with mean $\bar{\theta}^\mu$ and covariance given by $\frac{1}{n}$ times the matrix $\tilde{\boldsymbol{R}}^\mu(\bar{\theta}^\mu, \xi)$, see (3.40). This matrix is unknown, but can be approximated by its empirical counterpart,

$$\tilde{\boldsymbol{R}}^\mu(\bar{\theta}^\mu, \xi) \overset{\infty}{\approx} \hat{\boldsymbol{R}}_n^\mu := \boldsymbol{R}^\mu(\hat{\theta}_n^\mu, d_n, \xi_n), \tag{3.63}$$

see Conj. 3.14. Put together, these relations suggest the empirical approximation

$$p(\hat{\theta}^\mu \,|\, \mu, \xi_n) \overset{\infty}{\approx} \phi(\hat{\theta}^\mu \,|\, \hat{\theta}_n^\mu, \tfrac{1}{n}\hat{\boldsymbol{R}}_n^\mu) \tag{3.64}$$

for the unknown distribution of a PMLE. It is the counterpart of (2.33) for a normal model with known observation covariances. In contrast to the latter, *(3.64) is even valid if the model is incorrect.* Common characterizations of parameter uncertainty like confidence regions can be derived from it and remain meaningful regardless if the model if correct or not.

The matrix $\boldsymbol{R}^\mu$ explicitly depends on the data, and can thus be evaluated only *after* performing experiments. In contrast, the matrix $\boldsymbol{M}^\mu$ appearing in the classic counterpart of (3.64) is independent of the data, and can thus be evaluated even *before* performing experiments. It is thus not possible to use (3.64) directly for designing optimal experiments with the aim of identifying the best parameter. One can, however, use it to robustify sequential design criteria with respect to the current parameter uncertainty. We use (3.64) in Chap. 5 to derive such enhanced parameter-robust design criteria.

Given PMLEs $\hat{\theta}_n^\mu$ for all models $\mu \in \mathcal{M}$, a model maximum-likelihood estimate (MMLE) $\hat{\mu}_n$ can be determined in practice by minimizing $s^\mu(\hat{\theta}_n^\mu, d_n, \xi_n)$ with respect to $\mu \in \mathcal{M}$. It is an empirical approximation of the unknown best model,

$$\bar{\mu} \overset{\infty}{\approx} \hat{\mu}_n. \tag{3.65}$$

Due to the consistency of parameter and MMLE, these approximations improve with the sample size and are exact in the large-sample limit.

Empirical Approximations of Derived Quantities

The previous relations suggest further empirical approximations for functions of the unknown best parameter and/or model. In particular, applying (3.62) to the normal probability density function (PDF) of model μ from (3.9) leads to the approximation

$$p\big(y \,\big|\, x, \mu, \bar{\theta}^{\mu}\big) \overset{\infty}{\approx} \phi\big(y \,\big|\, \eta^{\mu}\big(\hat{\theta}_{n}^{\mu}, x\big), \boldsymbol{I}\big) \tag{3.66}$$

for the KLIC-best PDF of the model under experimental condition $x \in \mathcal{X}$, a special case of (2.31).

Based on (3.64) a parameter-robust counterpart of (3.66) can be derived using an expected value approach. The general form of this approximation, previously stated in (2.34), is

$$p\big(y \,\big|\, x, \mu, \bar{\theta}^{\mu}\big) \approx \int_{\mathcal{Q}^{\mu}} p\big(y \,\big|\, x, \mu, \hat{\theta}^{\mu}\big) p\big(\hat{\theta}^{\mu} \,\big|\, \mu, \xi_{n}\big)\, \mathrm{d}\hat{\theta}^{\mu}. \tag{3.67}$$

Under the considered assumptions, the parameter θ^{μ} enters the model only via the model responses $\eta^{\mu}(x, \theta^{\mu})$, see (3.8). Approximation (3.67) can thus be rewritten as

$$p\big(y \,\big|\, x, \mu, \bar{\theta}^{\mu}\big) \approx \int_{\mathcal{Q}^{\mu}} p(y \,|\, x, \mu, \hat{\eta}^{\mu}) p(\hat{\eta}^{\mu} \,|\, x, \mu, \xi_{n})\, \mathrm{d}\hat{\eta}^{\mu}, \tag{3.68}$$

where $p(y \,|\, x, \mu, \eta^{\mu})$ is the PDF under x specified by model μ for a *given* value $\eta^{\mu} \in \mathcal{Y}$ of the model response, and $p(\hat{\eta}^{\mu} \,|\, x, \mu, \xi_{n})$ is the PDF of $\eta^{\mu}\big(\hat{\mathcal{Q}}_{n}^{\mu}, x\big)$, that is, the PDF of the response under x evaluated at the PMLE $\hat{\mathcal{Q}}_{n}^{\mu} := \hat{\theta}^{\mu}(\mathcal{D}_{n}, \xi_{n})$.

Assume that the model is locally affine-linear around the PMLE $\hat{\theta}_{n}^{\mu}$, such that its response can be written as in (3.11). It then follows from the asymptotic normality of PMLEs (3.64) and the basic transformation rule (B.14) for affine-linear functions that

$$p(\hat{\eta}^{\mu} \,|\, x, \mu, \xi_{n}) \overset{\infty}{\approx} \phi\Big(\hat{\eta}^{\mu} \,\Big|\, \eta^{\mu}\big(x, \hat{\theta}_{n}^{\mu}\big), \tfrac{1}{n} \hat{\boldsymbol{J}}_{n}^{\mu}(x) \hat{\boldsymbol{R}}_{n}^{\mu} \hat{\boldsymbol{J}}_{n}^{\mu\top}(x)\Big). \tag{3.69}$$

Since the observation covariances are known and the model is normal,

$$p(y \,|\, x, \mu, \eta^{\mu}) = \phi(y \,|\, \eta^{\mu}, \boldsymbol{I}). \tag{3.70}$$

After substituting the last two relations, (3.68) turns into an integral over a product of two normal distributions. It can be solved analytically using (B.15), giving rise to the empirical parameter-robust approximation

$$p\big(y \,\big|\, x, \mu, \bar{\theta}^{\mu}\big) \overset{\infty}{\approx} \phi\Big(y \,\Big|\, \eta^{\mu}\big(x, \hat{\theta}^{\mu}_{n}\big), \boldsymbol{I} + \tfrac{1}{n}\hat{\boldsymbol{J}}^{\mu}_{n}(x)\hat{\boldsymbol{R}}^{\mu}_{n}\hat{\boldsymbol{J}}^{\mu\top}_{n}(x)\Big). \tag{3.71}$$

This approximation is analog to (2.34), yet remains valid even for incorrect models. It relies on a local linearization in vicinity of the PMLE for propagating the PMLE variability. It does, however, not make linearity assumptions for determining the PMLE variability in the first place.

Its validity relies on the accuracy of (3.69). Considering (3.67), we can expect (3.69) to be accurate if the responses are approximately linear in areas where the density of the PMLE $\phi\big(\hat{\theta}^{\mu} \,\big|\, \hat{\theta}^{\mu}_{n}, \frac{1}{n}\hat{\boldsymbol{R}}^{\mu}_{n}\big)$ is "large." As the sample size increases, this density accumulates in an arbitrary small area around the PMLE. One can thus expect that the crucial approximation (3.69) is accurate in sufficiently large samples.

Approximations for the unknown KLIC-best PDF $p\big(y \,\big|\, x, \bar{\mu}, \bar{\theta}\big)$ of the model family under x are obtained by evaluating the right-hand sides of (3.66) and (3.71) at the MMLE, that is, for $\mu = \hat{\mu}_{n}$.

3.5. Bayesian Inference

Bayesian inference was treated in general in Sec. 2.5. This section treats it under additional normality and/or linearity assumptions. The following formulas and approximations follow from substituting the equalities derived in Sec. 3.3 into the general counterparts from Sec. 2.5.

Throughout this section we consider scenario 3.1 and make the following additional assumptions: The experiments are sampled from design ξ, such that the design sequence $\xi_1, \xi_2, \ldots$ converges to ξ. The Kullback-Leibler information criterion (KLIC)-best parameter $\bar{\theta}^{\mu}$ under ξ is identifiable in each model $\mu \in \mathcal{M}$, and the KLIC-best model $\bar{\mu}$ under ξ is identifiable. Under all designs ξ and ξ_n, the observation covariances are known and all models $\mu \in \mathcal{M}$ are normal. In addition, all models have a normal parameter prior

$$p(\theta^\mu) := \phi\Big(\theta^\mu \,\Big|\, \theta_0^\mu, \boldsymbol{P}^{\mu^{-1}}\Big), \text{ for all } \theta^\mu \in \mathcal{Q}^\mu, \tag{3.72}$$

where $\theta_0^\mu \in \mathcal{Q}^\mu$ and $\boldsymbol{P}^\mu$ is a real-valued symmetric positive definite (spd) (and thus invertible) $n_{\theta^\mu} \times n_{\theta^\mu}$ matrix. The matrix $\boldsymbol{P}^\mu$ is a parameter-independent special case of the general prior information matrix $\boldsymbol{P}^\mu(\theta^\mu)$ defined in (2.47).

For all $n \in \mathbb{N}$, we write d_n for the data obtained under the n-experiment exact design ξ_n and $\hat{\theta}_n^\mu := \hat{\theta}^\mu(d_n, \xi_n) \in \mathcal{Q}^\mu$ for the corresponding parameter maximum-likelihood estimate (pmle) and use the abbreviations

$$\hat{\boldsymbol{M}}_n^\mu := \boldsymbol{M}^\mu\big(\hat{\theta}_n^\mu, \xi_n\big), \qquad \hat{\eta}_n^\mu(x) := \eta^\mu\big(x, \hat{\theta}_n^\mu\big), \tag{3.73}$$

$$\hat{\boldsymbol{N}}_n^\mu := \boldsymbol{N}^\mu\big(\hat{\theta}_n^\mu, d_n, \xi_n\big), \quad \hat{\boldsymbol{J}}_n^\mu(x) := \boldsymbol{J}^\mu\big(x, \hat{\theta}_n^\mu\big), \text{ and} \tag{3.74}$$

$$\hat{s}_n^\mu := s^\mu\big(\hat{\theta}_n^\mu, d_n, \xi_n\big). \tag{3.75}$$

Note that (3.72) contains the "little information" normal prior from (2.57) as special case for

$$\theta_0^\mu := \hat{\theta}_n^\mu \text{ and } \boldsymbol{P}^\mu := \hat{\boldsymbol{M}}_n^\mu + \hat{\boldsymbol{N}}_n^\mu, \tag{3.76}$$

supposed $\hat{\boldsymbol{M}}_n^\mu + \hat{\boldsymbol{N}}_n^\mu$ is invertible.

3.5.1. Single Regression Models

Consider a single regression model $\mu \in \mathcal{M}$.

Inference

Under the considered assumptions, the general empirical large-sample approximation for the parameter posterior (2.48) simplifies to

$$p(\theta^\mu \,|\, d_n, \xi_n) \approx \phi\Big(\theta^\mu \,\Big|\, \hat{\theta}_n^\mu, \tfrac{1}{n}\hat{\boldsymbol{B}}_n^{\mu^{-1}}\Big), \text{ where } \hat{\boldsymbol{B}}_n^\mu := \tfrac{1}{n}\boldsymbol{P}^\mu + \hat{\boldsymbol{M}}_n^\mu + \hat{\boldsymbol{N}}_n^\mu. \tag{3.77}$$

Like its general counterpart, approximation (3.77) remains adequate even if the models is *both* nonlinear *and* incorrect. The pmle $\hat{\theta}_n^\mu$ can be regarded as a point approximation for the unknown klic-best (or correct) parameter $\bar{\theta}^\mu$ of the model and the covariance $\frac{1}{n}\hat{\boldsymbol{B}}_n^{\mu^{-1}}$ as a quantification of the associated uncertainty.

Under the little informative normal prior defined by (3.76),

$$\hat{\boldsymbol{B}}_n^\mu = \tfrac{n+1}{n}\left(\hat{\boldsymbol{M}}_n^\mu + \hat{\boldsymbol{N}}_n^\mu\right). \tag{3.78}$$

The popular "classic" alternative

$$p(\theta^\mu \,|\, d_n, \xi_n) \approx \phi\left(\theta^\mu \,\middle|\, \hat{\theta}_n^\mu, \tfrac{1}{n}\hat{\boldsymbol{M}}_n^{\mu^{-1}}\right) \tag{3.79}$$

can be interpreted analogously to (3.77). It can be derived in two ways. First, it is a special case of (3.77) under the assumptions that model μ is correct and that the sample size n is large. Then, the matrix $\hat{\boldsymbol{N}}_n^\mu$ vanishes, see Thm. 3.11, and the $\mathcal{O}\left(\frac{1}{n}\right)$ term $\frac{1}{n}\boldsymbol{P}^\mu$ is negligible compared to the $\mathcal{O}(1)$ term $\hat{\boldsymbol{M}}_n^\mu$, so that (3.77) reduces (3.79).

Alternatively, it can be justified for samples of any size $n \in \mathbb{N}$ based on the assumption that the model is locally affine-linear around $\hat{\theta}_n^\mu$ (but possibly incorrect) and that the parameter prior $p(\theta^\mu)$ is locally uniform[2] (but not necessarily normal). This derivation, supposedly first given by Box and Hill [42, (7.9)], is common in literature.

Both ways imply that (3.79) is inadequate for models that *both* significantly nonlinear *and* substantially incorrect, unlike (3.77).

Predictions

Under considered assumptions and together with the parameter posterior approximation (3.77), the posterior prediction (2.38) of the model for an observation under the experimental condition $x \in \mathcal{X}$ is approximately

$$p(y \,|\, x, \mu, d_n, \xi_n) \approx \int\limits_{\mathcal{Q}^\mu} \phi(y \,|\, \eta^\mu(x, \theta^\mu), \boldsymbol{I})\,\phi\left(\theta^\mu \,\middle|\, \hat{\theta}_n^\mu, \tfrac{1}{n}\hat{\boldsymbol{B}}_n^{\mu^{-1}}\right) \mathrm{d}\theta^\mu. \tag{3.80}$$

[2] A parameter prior is locally uniform if it does not change much over the regions in which the likelihood has non-diminishing values, and do not take on large values outside these regions, so that Bayes' theorem for the parameter posterior simplifies to $p(\theta^\mu \,|\, d_n, \xi_n) \approx c^\mu p(d_n \,|\, \xi_n, \mu, \theta^\mu)$.

If the model is also assumed to be locally affine-linear around $\hat{\theta}_n^\mu$, the integral has a closed-form solution, leading to

$$p(y\,|\,x,\mu,d_n,\xi_n) \approx \phi\Big(y \,\Big|\, \hat{\eta}_n^\mu(x), \boldsymbol{I} + \tfrac{1}{n}\hat{\boldsymbol{J}}_n^\mu(x)\hat{\boldsymbol{B}}_n^{\mu\,-1}\hat{\boldsymbol{J}}_n^{\mu\,\top}(x)\Big), \tag{3.81}$$

The derivation follows the same steps leading from (3.67) to (3.71). The response $\hat{\eta}_n^\mu(x)$ predicts the average outcome of an experiment under condition x, based on the model and on the n available experiments. The covariance matrix quantifies the total uncertainty about the outcome of an experiment under x, given the model and the experiments. The matrix is composed of the identity matrix $\boldsymbol{I}$ representing the experimental uncertainty (see (3.5)) and the matrix $\frac{1}{n}\hat{\boldsymbol{J}}_n^\mu(x)\hat{\boldsymbol{B}}_n^{\mu\,-1}\hat{\boldsymbol{J}}_n^{\mu\,\top}(x)$ which quantifies, in a locally linear approximation, the propagation of the parameter uncertainty onto the prediction $\hat{\eta}^\mu(x)$.

Approximation (3.81) relies on a local linearization for propagating the uncertainty described by the parameter posterior onto the model response, but makes no linearity assumptions for determining the parameter posterior itself. The "classic" counterpart of (3.81) based on (3.79) is

$$p(y\,|\,x,\mu,d_n,\xi_n) \approx \phi\Big(y \,\Big|\, \hat{\eta}_n^\mu(x), \boldsymbol{I} + \tfrac{1}{n}\hat{\boldsymbol{J}}_n^\mu(x)\hat{\boldsymbol{M}}_n^{\mu\,-1}\hat{\boldsymbol{J}}_n^{\mu\,\top}(x)\Big). \tag{3.82}$$

Compared to (3.81) it relies either on an additional correctness assumption or on an additional local linearization, as discussed in the previous section.

Approximation (3.82) is well known and has, for example, been used by Hill and Hunter [118, (2.4)] and Box and Hill [42, (4.12)] in the context of optimal experimental design (OED). For models that are both nonlinear and incorrect, it is likely to be less adequate than (3.81), which we use for in our novel misspecification-robust design criteria proposed in Chap. 5.

Distribution of Parameter Posteriors and PMLEs

As discussed in Sec. 2.5.4, the distributions of parameter posteriors and parameter maximum-likelihood estimators (PMLEs) are asymptotically equal if the model is correct, but are different otherwise due to different covariances. Under the assumptions considered in this section, the covariances are related as follows.

In large samples, the $\mathcal{O}\big(\frac{1}{n}\big)$ prior information matrix $\boldsymbol{P}^\mu$ in (3.77) can be neglected, the PMLE $\hat{\theta}_n^\mu$ can be replaced by its limit value $\bar{\theta}^\mu$ and the empirical information matrices $\hat{\boldsymbol{M}}_n^\mu$ and $\hat{\tilde{\boldsymbol{N}}}_n^\mu$ can according to (3.56) be replaced by $\bar{\boldsymbol{M}}^\mu := \boldsymbol{M}^\mu\big(\bar{\theta}^\mu,\xi\big)$ and $\bar{\tilde{\boldsymbol{N}}}^\mu := \tilde{\boldsymbol{N}}^\mu\big(\bar{\theta}^\mu,\xi\big)$, respectively. Taken together, these substitutions lead to the large-sample approximation

$$\tfrac{1}{n}\left(\bar{\boldsymbol{M}}^{\mu} + \bar{\tilde{\boldsymbol{N}}}^{\mu}\right)^{-1} \tag{3.83}$$

for covariance of the parameter posterior. This formula is generally different from the corresponding formula

$$\tfrac{1}{n}\bar{\tilde{\boldsymbol{R}}}^{\mu} = \tfrac{1}{n}\left(\bar{\boldsymbol{M}}^{\mu} + \bar{\tilde{\boldsymbol{N}}}^{\mu}\right)^{-1}\bar{\boldsymbol{M}}^{\mu}\left(\bar{\boldsymbol{M}}^{\mu} + \bar{\tilde{\boldsymbol{N}}}^{\mu}\right)^{-1} \tag{3.84}$$

for the large-sample PMLE covariance from (3.40). Therefore, maximum-likelihood inference and Bayesian inference do generally lead to different quantifications of parameter uncertainty. If the model is correct under ξ or is locally affine-linear around $\bar{\theta}^{\mu}$, however, both approaches are consistent, since then $\bar{\tilde{\boldsymbol{N}}}^{\mu} = \boldsymbol{0}$, so that formulas reduce to $\tfrac{1}{n}\bar{\boldsymbol{M}}^{\mu^{-1}}$.

3.5.2. Families of Regression Models

Now consider a family of regression models with indices from the finite model index set $\mathcal{M}$.

Inference

Under the assumptions considered here, the empirical large-sample approximation (2.59) for the model posterior reduces for all $\mu \in \mathcal{M}$ to

$$p(\mu \,|\, d_n, \xi_n) \overset{\infty}{\approx} c_n p(\mu) \exp\left(-\tfrac{n}{2}\hat{s}_n^{\mu}\right) n^{-n_{\theta^{\mu}}/2}, \tag{3.85}$$

a product of four easily interpretable factors. The normalization factor $c_n \in \mathbb{R}^+$ ensures that the probabilities sum up to one, and $p(\mu)$ is the prior probability of model μ. The third factor $\exp\left(-\tfrac{n}{2}\hat{s}_n^{\mu}\right)$ is an exponentially decreasing function of the sum of squared residuals (SSR) which penalizes the lack-of-fit of model μ. The fourth factor is a decreasing function of the number of parameters in the model which penalizes over-parameterized (or rewards parsimonious) models. As discussed in Sec. 2.5.4, approximation (3.85) is particular good under the "little informative" normal prior, which is here specified by (3.72) and (3.76).

Independently from the publications leading to (3.85), formulas for model posteriors in normal models have been derived by Box and Henson [39, 40] and Box and Hill [42] and Stewart, Henson, and Box [240], culminating in the approximation of Stewart, Shon, and Box [239] that is identical to (3.85) except that its last factor is $2^{-n_{\theta^\mu}/2}$ instead of $n^{-n_{\theta^\mu}/2}$. Their formula relies on various assumption (particularly locally uniform parameter priors) which are not required for (3.85).

Predictions

The posterior prediction of the model family for the experimental outcome under $x \in \mathcal{X}$, defined as

$$p(y\,|\,x, d_n, \xi_n) \overset{(2.42)}{=} \sum_{\mu\in\mathcal{M}} p(y\,|\,x, \mu, d_n, \xi_n)\, p(\mu\,|\,d_n, \xi_n), \tag{3.86}$$

can be approximated empirically using (3.81) or (3.82) and (3.85). The resulting probability density function (PDF) is then a convex combination of normal PDFs, a so-called “Gaussian mixture.” Such distributions can typically not be approximated well by a (single) normal PDF. That is, *even under the normality and linearity assumptions considered here, the posterior prediction of the model family for the outcomes of unperformed experiments remains non-normal.*

This non-normality complicates the formulation of Bayesian design criteria for model discrimination (MD). In Chap. 5 we describe the established techniques used to deal with this non-normality and introduce novel design criteria using enhanced techniques.

3.5.3. Regularity Conditions

As discussed in Sec. 2.5.3, parameter posteriors are asymptotically normal under essentially the same regularity conditions ensuring the asymptotic normality of PMLEs, plus the requirement of a non-vanishing prior around the best parameter. The latter is automatically met under a normal prior, whose support is the whole parameter domain. In addition, consistency of the model posterior requires an identifiable KLIC-best model and a prior that does not vanish there.

Therefore, the empirical large-sample approximation of this chapter can be expected to be valid under the given assumptions if conditions (a)–(e) on p. 97 are met, the best model $\bar{\mu}$ is identifiable and $p(\bar{\mu}) > 0$.

Table 3.1.: Central quantities of statistical inference under normality with known unit and non-unit observation covariance.

	Definition for	
Quantity	$\mathbb{C}\left[\mathcal{Y}_x\right] = \boldsymbol{I}$	$\mathbb{C}\left[\mathcal{Y}_x\right] = \boldsymbol{\Omega}(x)$
$f(\cdot)$	—	$\tilde{f}(\cdot) := \boldsymbol{\Omega}^{-1/2}(x) f(\cdot)$ for any $f \in \left\{\mathcal{Y}_x, y_j(\cdot), \bar{\eta}(\cdot), \eta^\mu(\cdot), h^\mu(\cdot), \boldsymbol{J}^\mu(\cdot)\right\}$
$\boldsymbol{H}_j^\mu(x, \theta^\mu)$	(3.4)	$\tilde{\boldsymbol{H}}_j^\mu\left(x, \theta^\mu\right) := \sum_{l=1}^{n_y} \sigma_{jl}(x) \boldsymbol{H}_l^\mu\left(x, \theta^\mu\right)$ where $\sigma_{jl}(x)$ is the (j, l)-th element of $\boldsymbol{\Omega}^{-1/2}(x)$
$\lambda^\mu(\theta^\mu, \xi)$	(3.13)	$\sum_{x \in \mathrm{supp}(\xi)} \xi(x) \left\lVert \eta^\mu\left(x, \theta^\mu\right) - \bar{\eta}(x) \right\rVert_{\boldsymbol{\Omega}^{-1}(x)}^2$
$s^\mu(\theta^\mu, d_n, \xi_n)$	(3.17)	$\frac{1}{n} \sum_{x \in \mathrm{supp}(\xi_n)} \sum_{j=1}^{r_n(x)} \left\lVert \eta^\mu\left(x, \theta^\mu\right) - y_j(x) \right\rVert_{\boldsymbol{\Omega}^{-1}(x)}^2$
$\boldsymbol{M}^\mu(\theta^\mu, \xi)$	(3.21)	$\sum_{x \in \mathrm{supp}(\xi)} \xi(x) \boldsymbol{J}^{\mu\top}\left(x, \theta^\mu\right) \boldsymbol{\Omega}^{-1}(x) \boldsymbol{J}^\mu\left(x, \theta^\mu\right)$
$\tilde{\boldsymbol{N}}^\mu(\theta^\mu, \xi)$	(3.22)	$\sum_{x \in \mathrm{supp}(\xi)} \xi(x) \sum_{l=1}^{n_y} \left(\tilde{\eta}_l^\mu\left(x, \theta^\mu\right) - \tilde{\bar{\eta}}_l(x)\right) \tilde{\boldsymbol{H}}_l^\mu\left(x, \theta^\mu\right)$
$\boldsymbol{N}^\mu(\theta^\mu, d_n, \xi_n)$	(3.24)	$\frac{1}{n} \sum_{x \in \mathrm{supp}(\xi_n)} \sum_{j=1}^{r_n(x)} \sum_{l=1}^{n_y} \left(\tilde{\eta}_l^\mu\left(x, \theta^\mu\right) - \tilde{y}_{jl}(x)\right) \tilde{\boldsymbol{H}}_l^\mu\left(x, \theta^\mu\right)$

Part II.

Optimal Experimental Design (OED) for Model Discrimination (MD)

Of the two, design and analysis, the former is undoubtedly of greater importance. The damage of poor design is irreparable; no matter how ingenious the analysis, little information can be salvaged from poorly planned data. On the other hand, if the design is sound, then even quick and dirty methods of analysis can yield a great deal of information.

Box and Hunter [41, Sec. 3]

4. Fundamentals and Frequentist Strategies of Optimal Experimental Design

Contents

The previous two chapters concerned the questions of what and how knowledge about an unknown process can be obtained empirically, that is, from *given* experimental data. This and the next chapter deal with the question (already posed in (Q1.3) on p. 28) of experimental design:

(Q4.1) *Given a model family, under which conditions shall experiments be*

> *performed in order to improve the knowledge about the unknown process, or in in other terms, to reduce the structural uncertainty?*

Performing experiments is typically costly in terms of time, money, or other limited resources. One is thus interested in minimizing the number of experiments required to achieve a desired level of empirical knowledge, or in a fixed number of experiments that provide a maximal amount of knowledge. These aims lead to optimization problems know as optimal experimental design (OED) problems.

This chapter introduces fundamental concepts and considers strategies for solving OED problems that are based on frequentist inference, particularly on maximum-likelihood inference. OED strategies based on Bayesian inference are treated in the in next chapter.

Section 4.1 introduces the necessary basic concepts and examined properties of OED problems in general and their special cases of local and sequential OED problems. Section 4.2 introduces and discusses Kullback-Leibler (KL)-optimal designs and T-optimal designs, which are the theoretically best designs for model discrimination (MD). Albeit they depend on quantities that are unknown in practice, they define the aim that any practical approach for efficiently solving MD problems should strive for. Section 4.3 discusses two popular sequential strategies for MD: the Hunter-Reiner (HR)-strategy and the Buzzi-Ferraris (BF)-strategy.

A main result of this thesis is the new empirical formula for the covariance of parameter maximum-likelihood estimator (PMLE) for models that are both nonlinear and incorrect models that we proposed in Sec. 3.4. In Sec. 4.4 we show it can be used to derive new design criteria for MD with improved parameter-robustness, using the BF-criterion as example.

Section 4.1 and the KL-optimality in Sec. 4.2 make very few assumptions about the distributions of process and models family. T-optimality in Sec. 4.2, and Secs. 4.3 and 4.4 make the common assumptions of known covariance matrices and normal models that were considered in Chap. 3 in the context of statistical inference.

4.1. Optimal Experimental Design Problems

In the remaining chapter we consider the following scenario without further referencing it explicitly.

4.1.1. Problem Statement

Scenario 4.1 (Optimal Experimental Design)

(i) A process $q(y\,|\,x)$ according to Def. 1.2 is given. Its observation domain $\mathcal{Y}$ and its experimental domain $\mathcal{X}$ are compact.

(ii) The function q characterizing the process is unknown.

(iii) A model family is given for describing the process according to Def. 1.3. For each model $\mu \in \mathcal{M}$, the parameter domain $\mathcal{Q}^{\mu}$ is compact, and the function $p(y\,|\,x, \mu, \theta^{\mu})$ is continuous with respect to θ^{μ} for all $y \in \mathcal{Y}$ and all $x \in \mathcal{X}$, and is continuous with respect to x for all $y \in \mathcal{Y}$ and all $\theta^{\mu} \in \mathcal{Q}^{\mu}$.

(iv) The model family is correct. It contains an identifiable (that is, exactly one) correct model $\bar{\mu} \in \mathcal{M}$ which contains an identifiable correct parameter $\bar{\theta} \in \mathcal{Q}^{\bar{\mu}}$.

(v) The data $d_s \in \mathcal{Y}^s$ is available from the process, consisting of $s \in \mathbb{N}$ observations from the observation domain $\mathcal{Y}$, obtained in s statistically independent experiments performed under known conditions described by the s-experiment exact design ξ_s.

(vi) Additional experiments can be performed under arbitrary conditions from the experimental domain $\mathcal{X}$. Under given conditions, the corresponding observables are statistically independent of those from the previous experiments and among each other.

(vii) Data analysis and inference is possible following each individual additional experiment.

This scenario is a special case of scenario 1.5. It adds certain regularity assumptions about process and model family in (i) and (iii), the strong but essential assumption (iv) of correctness and identifiability, and the elementary prerequisite (vii) for a sequential data-adaptive approach. The role of assumptions (iv) and (vii) is discussed later.

Recall from Def. 1.6 that a parameter $\theta^{\mu} \in \mathcal{Q}^{\mu}$ of model $\mu \in \mathcal{M}$ is CORRECT, iff the corresponding model family member perfectly describes the process in the sense that

$$q(y\,|\,x) = p(y\,|\,x, \mu, \theta^{\mu}) \text{ for all } (y, x) \in \mathcal{Y} \times \mathcal{X}. \tag{4.1}$$

Furthermore, a model μ is correct, iff it contains a correct parameter exists in its parameter domain $\mathcal{Q}^{\mu}$, and a model family is correct, if it contains a correct model in the model index set $\mathcal{M}$. Items (ii) and (iv) hence imply that

(viii) neither $\bar{\mu}$ nor $\bar{\theta}$ are known.

Notation

We use the following notation in the context of scenario 4.1. The set of all designs (normed measures with finite support) over the experimental domain $\mathcal{X}$ is denoted Ξ.

Consider $n \in \mathbb{N}$ experiments performed under the exact design $\xi \in \Xi$. We write $x_i \in \mathcal{X}$ for condition of the i-th experiment (in an arbitrary enumeration) and $y_i \in \mathcal{Y}$ for the resulting observation, for all $i \in \{1, \ldots, n\}$.

Any observation y_i is considered as a realization of the corresponding observable $\mathcal{Y}_{x_i}$, a $\mathcal{Y}$-valued random variable. Likewise, the vector of data d is a realization of the sample $\mathcal{D} := \begin{bmatrix} \mathcal{Y}_{x_1} & \ldots & \mathcal{Y}_{x_n} \end{bmatrix}$, a random variable taking values in $\mathcal{Y}^n$.

The probability density function (PDF) of $\mathcal{D}$ is denoted $q(d\,|\,\xi)$, and the corresponding PDF specified by model μ with parameter θ^{μ} is denoted $p(d\,|\,\xi, \mu, \theta^{\mu})$. It follows from (vi) that the density assigned by the process to the data d obtained under ξ is

$$q(d\,|\,\xi) = \prod_{i=1}^{n} q(y_i\,|\,x_i), \tag{4.2}$$

and the corresponding PDF specified by model μ with parameter θ^{μ} is

$$p(d\,|\,\xi, \mu, \theta^{\mu}) = \prod_{i=1}^{n} q(y_i\,|\,x_i, \mu, \theta^{\mu}). \tag{4.3}$$

We write $\hat{\theta}^{\mu} \in \mathcal{Q}^{\mu}$ for a parameter maximum-likelihood estimate (PMLE) of model μ based on data d and ξ, and $\hat{\mathcal{Q}}^{\mu}$ for the corresponding estimator, see Def. 2.13.

OED for Model Discrimination and/or Parameter Estimation

Let Ξ be the set of designs over the experimental domain $\mathscr{X}$. If one is interested in the unknown correct model $\bar{\mu}$, scenario 4.1 gives rise to the question

(Q4.1) Under which design $\xi \in \Xi$ shall experiments be performed that are "best" for making empirical inferences about the unknown correct model $\bar{\mu}$?

The class of problems arising from this question is typically referred to as optimal experimental design (oed) for model discrimination (md). If one is interested in the unknown correct parameter $\bar{\theta}^{\mu} \in \mathscr{Q}^{\mu}$ of some model $\mu \in \mathscr{M}$ that one supposes to be correct, one is faced with the question

(Q4.2) Under which design $\xi \in \Xi$ shall experiments be performed that are "best" to make empirical inferences about the unknown correct parameter $\bar{\theta}^{\mu} \in \mathscr{Q}^{\mu}$ of a given model $\mu \in \mathscr{M}$, assuming that it is correct?

The class of problems arising from this question are usually referred to as optimal experimental design (oed) for parameter estimation (pe).

The names for these problem classes stem from the classic solution approaches: performing experiments under which the rival models make different predictions, so that one can *discriminate* between them by comparison with experimental data, and performing experiments that are beneficial for inference using parameter *estimation* techniques. The names are used, however, even when different approaches are used, like in Bayesian inference.

If one is interested in *both* the correct model *and* its correct parameter, one typically proceeds consecutively: first, experiments are performed until one finds a satisfactory well candidate for the correct model, then one focuses on that model and performs experiments to learn more about its correct parameter. This two-phase procedure is commonly considered in theory and applied in practice. For details we refer to the review of Franceschini and Macchietto [103] and the references given therein. In both phases additional data needs to be collected, so that suitable optimal experimental design (oed) methods for model discrimination (md) and parameter estimation (pe) answering (Q4.1) and (Q4.2) can reduce the required experimental effort.

This thesis focuses on oed strategies for md which deal with (Q4.1). Before we study related solution strategies in more detail, we discuss relevant general aspects of oed problems in the remaining section.

Correctness and Identifiability are Essential Assumptions

Items (Q4.1) and (Q4.2) are only well posed under assumption (iv), which ensures that the quantities of interest – the correct model and/or its correct parameter – exist and are unique. Without the assumption of identifiability, several correct models might exists with several correct parameters, leaving the ambiguity in which of them one is actually interested in.

Without assuming that the model family is correct, a correct model and/or a correct parameter might not exist at all. Then, one might be tempted to focus the interest on the model and/or the parameter that are best in the sense of the Kullback-Leibler information criterion (klic), see Sec. 1.4. Klic-best models and parameters, however, do generally depend on the chosen design, so that it is not possible to choose a design independent from the quantity of interest. There is hence no direct counterpart of (Q4.1) and (Q4.2) for klic-best models and parameters.

4.1.2. A General View on OED Problems

Besides identifying the correct model and/or the correct parameter, there is a plethora of goals that one might want to attain through experimentation. The general problem of finding experiments which are most "useful" for the particular goal can be formalized as follows.

Problem 4.2 (Optimal Experimental Design)

Let Ξ be the set of designs over the experimental domain $\mathcal{X}$. Given a subset $\Xi' \subseteq \Xi$ of admissible designs and a design criterion $\Psi : \Xi \mapsto \mathbb{R}$, find a Ψ-optimal design

$$\xi^\star \in \operatorname*{argmax}_{\xi \in \Xi'} \Psi(\xi). \tag{4.4}$$

The design criterion $\Psi(\xi)$ *anticipates* or *predicts* how "useful" the data obtained under design ξ will be for reaching the desired goal. Naturally, a design criterion is specific for this particular goal and for the methods of inference used for analyzing the data. A design criterion is model-based if it uses one or several models for predicting the process behavior. It is data-based or data-adaptive if it takes into account data available from performed experiments. If Ξ' is a proper subset of Ξ the OED problem is CONSTRAINED. In the remaining section we only consider unconstrained OED problems.

Kiefer and Wolfowitz [143] were probably the first to consider OED problems of this type. Fedorov [99] gives a brief an overview over the field of OED problems, and Atkinson and Bailey [17] survey its history up to the year 2001. Standard references for OED are the books of Atkinson and Donev [11], Fedorov [95], Fedorov and Hackl [96], and Pukelsheim [206] and Cox and Reid [74]. OED problems are optimization problems in the space of measures, since designs are normed measures with finite support on the experimental domain $\mathcal{X}$, see Def. 1.4. Molchanov and Zuyev [188] examine the problem on this general level and provide a general algorithm for solving it numerically.

Relaxation of OED Problems to Continuous Designs

One could think of limiting the maximization in (4.4) to *exact* n-experiment designs with $n \in \mathbb{N}$, since only they can be realized in practice. As a consequence, one would need to specify the number of experiments n *before* solving the OED problem, which might not be desirable or not even possible. Furthermore, the weights of an exact design take values in a discrete set, which introduces an integer aspect into the optimization problem (4.4), complicating both its theoretical analysis as well as its numerical solution.

For these reasons it is convenient to perform the optimization in OED problems over designs which might be non-exact. This relaxation goes back to Kiefer and Wolfowitz [143] and has since then become a de-facto standard in OED. We apply this relaxation in all OED problems considers in the remainder of this thesis.

In practice, exact n-experiment designs are used to *approximate* optimal designs obtained from the relaxed problem. By increasing the total number of experiments n, such approximations can be made arbitrarily precise. Suitable rounding strategies are described by Pukelsheim and Rieder [207] and Pukelsheim [206, Chap. 12] and references given therein.

OED as Optimization under Uncertainty

A typical way to derive a design criterion is to formulate a real-valued function $u(\xi, d)$ which measures how "useful" the data d obtained under the exact design ξ *actually* is for whatever goal one aims to achieve through experimentation. An ideal design would thus maximize $u(\xi, d)$ with respect to $\xi \in \Xi$. Since experiments are designed *before* they are performed, but the data d is known only *afterwards*, such an ideal design cannot be determined before experimentation.

The next step towards a practicable design criterion is thus to deal with this *experimental uncertainty* by formulating a *data-independent* function $U(\xi)$ that approximates or predicts $u(\xi, d)$,

$$U(\xi) \approx u(\xi, d) \text{ for all } \xi \in \Xi. \tag{4.5}$$

The function $U(\xi)$ might take into account the probabilities of obtaining particular data which are described by the function q characterizing the process. Then, U depends parametrically on q, written as $U(\xi; q)$. A typical example is the expected value approximation $U(\xi; q) := \int u(\xi, d) q(d \mid \xi) \, \mathrm{d}d$. Correct models and correct parameters, defined as solutions of equation (4.1), depend implicitly on q. When the goal is to identify one of them, the functions u and U will thus generally depend on q. In practice, however, also the function q is unknown, so that also designs maximizing $U(\xi; q)$ cannot be determined.

To obtain a practically evaluable design criterion one further needs to deal with this *structural uncertainty* by formulating a function $\Psi(\xi)$ that is *independent* of q and approximates $U(\xi; q)$,

$$\Psi(\xi) \approx U(\xi; q) \text{ for all } \xi \in \Xi. \tag{4.6}$$

The function Ψ does not depend on any unknown quantity and can thus be used in practice to determine optimal designs. To improve the quality of (4.6), the function Ψ might take into account all available knowledge about the process, expressed for example in terms of a model family and related parameter and model estimates or posteriors obtained from previous experiments.

In general, optimal experimental design problems appearing in practice are thus optimization problems under uncertainty, namely the experimental uncertainty and the structural uncertainty. The actual "usefulness" of an optimal design $\xi^\star \in \operatorname{argmax}_{\xi \in \Xi} \Psi(\xi)$ – given by $u(\xi^\star, d)$ – depends on the quality of approximations (4.5) and (4.6) dealing with these uncertainties.

Locally Optimal Designs

Assume that model $\mu \in \mathcal{M}$ and parameter $\theta^\mu \in \mathcal{Q}^\mu$ are correct. Then, the function $q(y \mid x)$ characterizing the process can be replaced by the model family member $p(y \mid x, \mu, \theta^\mu)$ under all $x \in \mathcal{X}$, see (4.1). Let $\Psi(\xi; \mu, \theta^\mu)$ be the function obtained by performing this substitution to the function $U(\xi; q)$ from the previous section. Then, $\Psi(\cdot; \mu, \theta^\mu)$ depends only on known quantities, so that its maximizer can actually be determined in practice, leading to the following class of OED problems.

Problem 4.3 (Local Optimal Experimental Design)

Let $\Psi(\xi;\mu,\theta^{\mu})$ be a LOCAL DESIGN CRITERION, a function that maps from Ξ to $\mathbb{R}$, depends parametrically on model $\mu \in \mathcal{M}$ and parameter $\theta^{\mu} \in \mathcal{Q}^{\mu}$, and does not depend directly or indirectly on the unknown process. Find a LOCALLY OPTIMAL DESIGN

$$\xi^{\star}(\mu,\theta^{\mu}) \in \operatorname*{argmax}_{\xi\in\Xi} \Psi(\xi;\mu,\theta^{\mu}). \tag{4.7}$$

In practice, it is of course not known whether the underlying assumption holds that model μ and parameter θ^{μ} are correct. Nevertheless, locally optimal designs are useful for examining the general structure of optimal designs and their dependency on the correct model and the correct parameter. Furthermore, local design criteria can be used to derive practically evaluable design criteria by using prior and/or empirically obtained partial knowledge about correct models and correct parameters.

4.1.3. Sequential Construction of Optimal Designs

Assume that model $\bar{\mu} \in \mathcal{M}$ and parameter $\bar{\theta} \in \mathcal{Q}^{\bar{\mu}}$ are correct and let $\Psi(\xi;\bar{\mu},\bar{\theta}) = U(\xi;q)$ be a corresponding local design criterion as defined in the preceding section. In practice, optimal designs $\xi^{\star}(\bar{\mu},\bar{\theta}) \in \operatorname{argmax}_{\xi\in\Xi} \Psi(\xi;\bar{\mu},\bar{\theta})$ are unknown. As discussed in the previous two chapters, uncertainty about $\bar{\mu}$ and $\bar{\theta}$ can be expressed empirically and tends to decrease (under regularity conditions) as more data is available. In turn, empirical approximations for $\Psi(\xi;\bar{\mu},\bar{\theta})$ tend to get better under reduced uncertainty. If it is possible to analyze the data and adapt the design once a new observation gets available, as assumed in (vi) and (vii), these relations suggest a sequential approach to determine $\xi^{\star}(\bar{\mu},\bar{\theta})$, in which experimentation, inference and design are repeated consecutively.

Sequential Procedure

Algorithm 4.1 on the next page outlines the essential elements of a sequential design procedure, using the symbol $\mathcal{F}$ to denote the model family and $\mathcal{U}_0$ to denote prior knowledge.

Such sequential procedures are commonly applied in practice and have, for example, been proposed by Asprey and Macchietto [10], Franceschini and Macchietto [103], and Kreutz and Timmer [152].

Algorithm 4.1: Sequential design procedure.

```
input  : experimental domain $\mathcal{X} \subseteq \mathbb{R}^{n_x}$, model family $\mathcal{F}$, knowledge $\mathcal{U}_0$
output : design $\xi_n$, data $d_n$, and knowledge $\mathcal{U}_n$ from $n \in \mathbb{N}$ experiments

$n \leftarrow 1$;
while not terminate($\mathcal{U}_{n-1}, \mathcal{F}$) do                    // termination check
    choose $x_n \in \mathcal{X}$ based on $\mathcal{F}$ and $\mathcal{U}_{n-1}$;            // design experiment
    get random variate $y_n$ from $q(y \mid x_n)$;               // perform experiment
    let $\xi_n$ be the design constituted by $x_1, \dots, x_n$;
    let $d_n^\top := \begin{bmatrix} y_1^\top & \dots & y_n^\top \end{bmatrix}$;
    infer empirical knowledge $\mathcal{U}_n$ from $\mathcal{U}_0$, $\xi_n$, $d_n$, and $\mathcal{F}$;    // analyze exps.
end
return $\xi_n$, $d_n$, $\mathcal{U}_n$
```

The procedure is terminated iteration $n - 1$ if the available empirical knowledge $\mathcal{U}_{n-1}$ suffices to consider the problem as solved. Otherwise, the procedure continues.

In the design phase, the condition x_n for the next experiment is chosen. The choice, possibly made by solving a sequential OED problem (discussed later), might take into account the available empirical knowledge $\mathcal{U}_{n-1}$, which itself might depend on the design ξ_{n-1} and the data d_{n-1}.

Subsequently, the experiment under x_n is performed and the resulting observation $y_n \in \mathcal{Y}$ is recorded. In the inference phase, the empirical knowledge $\mathcal{U}_n$ is updated from the design and data of *all* n available experiments and the prior knowledge. The procedure then continues with the termination check using the updated empirical knowledge.

The procedure is called DATA-ADAPTIVE, if and only if x_n depends in any iteration $n \in \mathbb{N}$ on at least one of the previous observations $y_1, \dots, y_{n-1}$.

In early stages of the procedure, little is known about the unknown correct model and the unknown correct parameter, such that designed experiments tend to be somewhat tentative. While the procedure continues and more observations are obtained, the empirical knowledge about these unknown gets more precise. This improved knowledge in turn allows to choose experiments which provide informative observations more reliably.

For all $n \in \mathbb{N}$, design ξ_n can be considered as a *predictor* for the unknown sought-after optimal design $\xi^\star\left(\bar{\mu}, \bar{\theta}\right)$ and x_{n+1} as a *corrector* that takes into account the new insight from the observation recently made under x_n and drives the resulting overall design ξ_{n+1} towards $\xi^\star\left(\bar{\mu}, \bar{\theta}\right)$.

Sequential OED Problem

The `design(·)` step in Alg. 4.1 on the facing page is crucial for the overall efficiency of the procedure. In the `inference(·)` step one can only try to extract as much knowledge from the data as possible, but it depends on the chosen experimental conditions how much information it contains in the first place.

Problem 4.4 (Sequential Optimal Experimental Design)

Let d_n be the data obtained from n previous experiments described by the exact design ξ_n. Let $\Psi_n(x; d_n, \xi_n)$ be a corresponding SEQUENTIAL DESIGN CRITERION, a continuous function that maps from the experimental domain $\mathcal{X}$ to $\mathbb{R}$ and depends parametrically on d_n and ξ_n. Find an experimental condition

$$x_{n+1} \in \operatorname*{argmax}_{x \in \mathcal{X}} \Psi_n(x; d_n, \xi_n). \tag{4.8}$$

Limitation to a Single Experiment

It is common and convenient to limit the maximization in sequential OED problems as in (4.8). A sequential procedure whose design criterion considers *only* the next experiment but ignores the possibility of further subsequent experiments is unlikely to be the most efficient procedure possible. Several early works, for example those of Bradt and Karlin [50] and DeGroot [78], indicate that sequential OED problems that take into account subsequent experiments are difficult to handle both theoretically and numerically. The sequential design criterion in (4.8), however, may in fact consider the possibility of further experiments. It simply yields only the one condition required for the next experiment as output, see (vi) and (vii).

Statistical Dependence between Data-Adaptively Designed Experiments

Adaptively choosing experimental conditions introduces statistical dependencies in the sample: The experimental condition x_{n+1} in (4.8) is determined based on the design ξ_n and the data d_n of all preceding experiments $1, \dots, n$. Since the data is subject to random fluctuations, so is x_{n+1}. The observable of experiment $n+1$ is hence not statistically independent of the preceding observables $\mathcal{Y}_{x_1}, \dots, \mathcal{Y}_{x_n}$.

The effect of these dependencies are, however, not too severe. In particular, the likelihood retains its additive form (2.5), since it involves only the conditional densities of the model family for *given* experimental conditions. Since x_{n+1} is typically chosen based on *all* preceding experiments, the dependency between the observable of experiment $n+1$ and any *particular* previous observable $\mathcal{Y}_{x_i}$ with $i \leqslant n$ tends to decrease as n goes to infinity.

4.1.4. Additional Normality Assumptions

In several sections of this and the next chapter we consider scenario 4.1 under the following additional normality assumptions. We list them here to avoid repetitions.

(ix) The observation covariance $\boldsymbol{\Omega}(x) := \mathbb{C}\left[\mathcal{Y}_x\right]$ exists, has full rank and is known under all experimental conditions $x \in \mathscr{X}$.

(x) The rival models are normal,

$$p(y \mid x, \mu, \theta^\mu) = \phi(y \mid \eta^\mu(x, \theta^\mu), \boldsymbol{\Omega}(x)) \qquad (4.9)$$
$$= \exp\Big(-\tfrac{1}{2}\|\eta^\mu(x, \theta^\mu) - y\|^2_{\boldsymbol{\Omega}^{-1}(x)} + n_y \ln(2\pi)\Big)$$

for all $y \in \mathscr{Y}$, all $x \in \mathscr{X}$, all $\mu \in \mathscr{M}$ and all $\theta^\mu \in \mathscr{Q}^\mu$.

Here and in the whole chapter, the symbol ϕ denotes the pdf of a normal distribution, see (B.12).

(xi) The response $\eta^\mu(x, \theta^\mu)$ of each model $\mu \in \mathscr{M}$ is continuous in x for all $\theta^\mu \in \mathscr{Q}^\mu$ and is twice continuously differentiable in θ^μ under all $x \in \mathscr{X}$.

Combined with the correctness assumption (iv) of scenario 4.1, these additional assumptions have the following implications.

(xii) The process is normal,

$$q(y \mid x) = \phi(y \mid \bar{\eta}(x), \boldsymbol{\Omega}(x)), \qquad (4.10)$$
$$= \exp\Big(-\tfrac{1}{2}\|\bar{\eta}(x) - y\|^2_{\boldsymbol{\Omega}^{-1}(x)} + n_y \ln(2\pi)\Big)$$

with observation mean $\bar{\eta}(x) = \eta^{\bar{\mu}}(x, \bar{\theta})$, for all $y \in \mathcal{Y}$ and all $x \in \mathcal{X}$.

(xiii) The observation mean $\bar{\eta}(x)$ is continuous in x.

Then, (ii) of scenario 4.1 implies that

(xiv) the observation mean $\bar{\eta}$ is unknown.

The observation mean is then in fact the only aspect of the process that remains unknown.

These assumptions are often good approximations of the situation found in practice. They permit to use the inference results from Chap. 3, which are beneficial for solving the arising OED problems numerically.

Notation

Under assumptions (ix) and (x), a PMLE of model $\mu \in \mathcal{M}$ based on data d and ξ reduces to a least-squares (LSQ) estimate, that is, a minimizer of the sum of squared residuals (SSR)

$$s^{\mu}(\theta^{\mu}, d, \xi) = \tfrac{1}{n} \sum_{i=1}^{n} \| \eta^{\mu}(x_i, \theta^{\mu}) - y_i \|^2_{\boldsymbol{\Omega}^{-1}(x_i)} \tag{4.11}$$

with respect to $\theta^{\mu} \in \mathcal{Q}^{\mu}$, see Def. 3.9 and Cor. 3.10 and Tab. 3.1 on p. 113. Let ∇ and ∇^2 denote the gradient and the Hessian differential operator, respectively, with respect to θ^{μ}. For all experimental conditions $x \in \mathcal{X}$, we write

$$\hat{\eta}^{\mu}(x) := \eta^{\mu}(x, \hat{\theta}^{\mu}), \tag{4.12}$$

$$\hat{\boldsymbol{J}}^{\mu}(x) := \nabla \eta^{\mu}(x, \theta^{\mu})|_{\theta^{\mu} = \hat{\theta}^{\mu}}, \text{ and} \tag{4.13}$$

$$\hat{\boldsymbol{H}}^{\mu}_{j}(x) := \nabla^2 \eta^{\mu}_{j}(x, \theta^{\mu})\Big|_{\theta^{\mu} = \hat{\theta}^{\mu}}, \tag{4.14}$$

for the response of model $\mu \in \mathcal{M}$, its Jacobian, and the Hessian of its j-th component $\eta^{\mu}_{j}(x, \theta^{\mu})$, respectively, evaluated at the PMLE.

Furthermore, we define the symmetric positive semi-definite (SPSD) $n_{\theta^{\mu}} \times n_{\theta^{\mu}}$ matrices

$$\hat{\boldsymbol{M}}^{\mu} := \tfrac{1}{n} \sum_{i=1}^{n} \hat{\boldsymbol{J}}^{\mu\top}(x_i) \boldsymbol{\Omega}^{-1}(x_i) \hat{\boldsymbol{J}}^{\mu}(x_i) \tag{4.15}$$

and the symmetric $n_{\theta^\mu} \times n_{\theta^\mu}$ matrix

$$\hat{\boldsymbol{N}}^\mu := \sum_{i=1}^{n}\sum_{j=1}^{n_y}\Big(\hat{\eta}_j^\mu(x_i) - y_{ij}\Big)\sum_{k=1}^{n_y}\sigma_{jk}(x)\hat{\boldsymbol{H}}_j^\mu(x), \tag{4.16}$$

where $\hat{\eta}_j^\mu(x)$ and y_{ij} are the j-th component of $\hat{\eta}^\mu(x)$ and y_i, respectively, and $\sigma_{jl}(x)$ is the (j,k)-th component of $\boldsymbol{\Omega}^{-1/2}(x)$. For details, see Sec. 3.3 and Tab. 3.1.

4.2. KL-Optimality and T-Optimality: Optimal Designs for Model Discrimination

This section introduces Kullback-Leibler (KL)-optimal designs, and their special case under normality assumptions known as T-optimal designs. From a widely accepted point of view, they are the best designs for model discrimination (MD) in the sense of (Q4.1) that are theoretically possible. KL-optimal designs and their local counterparts depend on unknown quantities and can thus not be directly determined in practice. The remaining sections of this chapter and the whole next chapter consider sequential procedures for approximating them based on experimental data.

4.2.1. History and Related Literature

Atkinson and Fedorov [21] introduce the concept of T-optimality for the problem of discriminating between two univariate normal nonlinear models, proposing as design criterion the special case of (4.30) for $n_y = 1$. Without using this name, Fedorov [93] introduces a similar concept. We shall refer to their ideas and results as classic T-optimality. Essentially, their strategy formalizes the intuitive ideas formulated by Hunter and Reiner [129]. Fedorov [94] and Dette and Titoff [79] analyze the T-criterion in more detail. Kuczewski [153] considers related computational aspects. A summary of classic T-optimality can be found in the book of Atkinson and Donev [11, Chap. 20].

To overcome the limitations of locally optimum designs, Leon and Atkinson [173], Leon [174] and Atkinson [14] studied the effects of prior information, resulting in a theory of Bayesian T-optimality. The relations of classic T-optimality to other strategies for optimal experimental design (OED) for MD, especially those proposed by Hunter and Reiner [129], Atkinson and Cox [19] and Box and Hill [42], are discussed and reviewed by Atkinson [15], Hill [116] and Atkinson [12]. The relations to OED for parameter estimation (PE) are examined by Fedorov and Khabarov [97].

In a series of publications, Uciński and Bogacka [248, 249, 250, 251, 252] generalize classic T-optimality to multivariate models, dynamic models, unknown observation covariance, sampling design, function-valued design variables and constrained designs. We refer to these extensions as GENERALIZED T-OPTIMALITY.

López-Fidalgo, Tommasi, and Trandafir [177] generalize the idea of T-optimality further to non-normal models, resulting in the concept of (two-model) KL-OPTIMALITY. They limit their considerations to two rival models and propose design criterion (4.25). Further results on the properties of KL-optimal designs and algorithms for their numerical construction are provided by Aletti, May, and Tommasi [4], May and Tommasi [183], and Tommasi, Santos-Martín, and Rodríguez-Díaz [246] and Aletti, May, and Tommasi [5]. Tommasi and López-Fidalgo [245] introduce a Bayesian variant of KL-optimality that takes into account prior information. These results are clarified and presented with some examples in Tommasi and López-Fidalgo [243].

The publications listed so far consider MD problems only between *two* models. Atkinson and Fedorov [20] used a worst-case approach to generalize the classic two-model T-criterion to several rival models, essentially proposing the univariate special case of design criterion (4.28). Tommasi [244] proceeded differently: they formulated a multi-model KL-criterion which relies on an extended meta-model which contains all rival models as special case.

A general algorithm for the numerical construction of locally optimal designs is described by Fedorov and Hackl [96]. López-Fidalgo, Tommasi, and Trandafir [177] describe an adaption of this algorithm for the computation of locally KL-optimal designs, Aletti, May, and Tommasi [4] refine it and examine necessary conditions for its convergence, as do Aletti, May, and Tommasi [5].

In contrast to the historical order, we first introduce KL-optimality and then the special case of T-optimality.

4.2.2. KL-Optimality

Consider scenario 4.1 and assume we are interested in identifying the unknown correct model $\bar{\mu} \in \mathcal{M}$. If we aim to gain knowledge about $\bar{\mu}$ based on experimental data we are faced with (Q4.1), which can be formalized as follows.

Recall from Sec. 1.4.2 that the Kullback-Leibler information criterion (KLIC)

$$\delta(\nu, \theta^\nu, \xi) := \sum_{x \in \operatorname{supp}(\xi)} \xi(x) \int_{\mathcal{Y}} q(y\,|\,x) \ln \frac{q(y\,|\,x)}{p(y\,|\,x, \nu, \theta^\nu)} \, \mathrm{d}y \tag{4.17}$$

measures the discrepancy of model ν with parameter θ^ν under design ξ to the process. Instead via the defining equation (4.1), correct and incorrect models can alternatively be characterized via the KLIC according to Prop. 1.8(ii) as follows.

If model $\nu \in \mathcal{M}$ is correct, then it contains a parameter (the correct one) at which it exhibits no discrepancy to the process under all possible designs,

$$\min_{\theta^\nu \in \mathcal{Q}^\nu} \delta(\nu, \theta^\nu, \xi) = 0 \text{ under all designs } \xi \in \Xi. \tag{4.18}$$

If model μ is *not* correct, then there exists at least one design $\xi \in \Xi$ under which it exhibits a non-vanishing discrepancy to the process at any parameter, even those minimizing the KLIC, so that

$$\min_{\theta^\nu \in \mathcal{Q}^\nu} \delta(\nu, \theta^\nu, \xi) > 0. \tag{4.19}$$

The key to the derivation of a design criterion for MD is that non-zero discrepancies can in principle be detected *empirically,* and that the larger the discrepancies are, the easier is it to detect them. Suitable methods are discussed later.

In order to empirically identify model ν as incorrect, one would thus perform experiments under a design maximizing (4.19). Such a design is specific for the considered model ν. If the aim is to find *one* design for empirically identifying *several* incorrect models, some kind of compromise is necessary. In the context of T-optimal designs, Atkinson and Fedorov [20, Sec. 1] advocated the following approach:

> Since the purpose of the experiment is to find the true model, and we are assuming that one model is true, then, at some stage, the problem will become that of discriminating between the true model and the model, or models, closest to it.

Applying this worst-case approach to the situation at hat, one would choose the design maximizing (4.19) for the incorrect model with the smallest KLIC. This approach gives rise to an OED problem with the following design criterion.

Definition 4.5 (KL-Criterion)

Assume there are at least two rival models, $n_{\mathcal{M}} \geqslant 2$. Suppose the KLIC $\delta(\mu, \theta^\mu, \xi)$ exists and is continuous in θ^μ for all $\mu \in \mathcal{M}$, all $\theta^\mu \in \mathcal{Q}^\mu$ and all $\xi \in \Xi$. The KL-CRITERION is the function $K\colon \Xi \mapsto \mathbb{R}$ defined for all $\xi \in \Xi$ as

$$K(\xi) := \min_{\substack{\nu \in \mathcal{M} \\ \nu \neq \bar{\mu}}} \min_{\theta^\nu \in \mathcal{Q}^\nu} \delta(\nu, \theta^\nu, \xi). \tag{4.20}$$

A maximizer of $K(\xi)$ over all $\xi \in \Xi$ is a KL-OPTIMAL DESIGN.

Note that without the limitation $\nu \neq \bar{\mu}$ in the outer minimization, the KL-criterion would be identically zero according to (4.18). Under a KL-optimal design, the "best" among the *incorrect* model family members exhibit the largest discrepancy to the process, and is thus easiest to detect as incorrect empirically. All other incorrect models have an even larger discrepancy, regardless of their parameter. In this sense is a KL-optimal design the sought-after "best" design for solving the MD problem in the sense of (Q4.1).

A KL-optimal design cannot be determined in practice, since the KL-criterion directly depends on the correct model $\bar{\mu}$ and on the process q, which are both unknown. *A KL-optimal design is thus a theoretical ideal case which one can aim to approximate in practice.*

Local KL-Optimality

The KLIC (4.17) depends explicitly on the unknown probability density functions (PDFs) $q(y\,|\,x)$ of the process. Replacing the latter by their counterparts specified by model μ with parameter θ^μ yields the function

$$\delta(\nu, \theta^\nu, \xi\,|\,\mu, \theta^\mu) := \sum_{x \in \operatorname{supp}(\xi)} \xi(x) \int_{\mathcal{Y}} p(y\,|\,x, \mu, \theta^\mu) \ln \frac{p(y\,|\,x, \mu, \theta^\mu)}{p(y\,|\,x, \nu, \theta^\nu)} \, \mathrm{d}y, \tag{4.21}$$

essentially a model family-based counterpart of the KLIC. It gives rise to the following local counterpart of Def. 4.5.

Definition 4.6 (Local KL-Criterion)

Assume that for a given model $\mu \in \mathscr{M}$ and a given parameter $\theta^{\mu} \in \mathscr{Q}^{\mu}$, the function $\delta(\nu, \theta^{\nu}, \xi \,|\, \mu, \theta^{\mu})$ exists and is continuous in θ^{ν} for all $\nu \in \mathscr{M}$, all $\theta^{\nu} \in \mathscr{Q}^{\nu}$ and all $\xi \in \Xi$. The LOCAL KL-CRITERION (FOR μ AND θ^{μ}) is

(4.22)

$$K(\xi \,|\, \mu, \theta^{\mu}) := \min_{\substack{\nu \in \mathscr{M} \\ \nu \neq \mu}} \min_{\theta^{\nu} \in \mathscr{Q}^{\nu}} \delta(\nu, \theta^{\nu}, \xi \,|\, \mu, \theta^{\mu}) \text{ for all } \xi \in \Xi. \tag{4.23}$$

A maximizer $\xi_{\mathrm{KL}}(\mu, \theta^{\mu})$ of $K(\xi \,|\, \mu, \theta^{\mu})$ over all $\xi \in \Xi$ is a LOCALLY KL-OPTIMAL DESIGN.

The function $\delta(\nu, \theta^{\nu}, \xi \,|\, \mu, \theta^{\mu})$ and hence the local KL-criterion depend solely on known or user-specified quantities, so that *locally* KL-optimal designs can in fact be determined in practice, in contrast to KL-optimal designs from Def. 4.5.

Suppose model $\mu \in \mathscr{M}$ is correct and has a correct parameter $\theta^{\mu} \in \mathscr{Q}^{\mu}$. Then, $q(y \,|\, x) = p(y \,|\, x, \mu, \theta^{\mu})$ for all $y \in \mathscr{Y}$ and all $x \in \mathscr{X}$, see Def. 1.6, and thus

$$\delta(\nu, \theta^{\nu}, \xi \,|\, \mu, \theta^{\mu}) = \delta(\nu, \theta^{\nu}, \xi) \text{ for all } \xi \in \Xi. \tag{4.24}$$

If, in addition, model μ is identifiable (excluding the possibility that the family contains other correct models), then $\xi_{\mathrm{KL}}(\mu, \theta^{\mu}) = \xi_{\mathrm{KL}}$. In other words, a locally KL-optimal design for model μ with parameter θ^{μ} is KL-optimal if μ and θ^{μ} are correct.

Suppose that there are only two rival models, $\mu \neq \nu$, and assume without loss of generality (WLOG) that model μ is the identifiable correct one. Then, the local KL-criterion reduces to

$$K(\xi, \mu, \theta^{\mu}) = \min_{\theta^{\nu} \in \mathscr{Q}^{\nu}} \delta(\nu, \theta^{\nu}, \xi \,|\, \mu, \theta^{\mu}), \tag{4.25}$$

which is design criterion originally introduced by López-Fidalgo, Tommasi, and Trandafir [177] under the name "KL-criterion." We use the same name for the multi-model variants in Defs. 4.5 and 4.6 since they are a straightforward generalizations.

Statistical Motivation

Model $\nu \in \mathcal{M}$ is correct if and only if the statistical hypothesis

$$\exists \theta^{\nu} \in \mathcal{Q}^{\nu}\ \forall x \in \mathcal{X} : \mathcal{Y}_x \sim p(y\,|\,x, \nu, \theta^{\nu}) \tag{H$^\nu$}$$

is true, see (4.1) and Def. 1.6. A conservative approach for identifying the correct model *empirically* is to consider all rival models as "tentatively" correct and try to *falsify* the corresponding hypothesis based on the available data.

In frequentist inference this is done using STATISTICAL HYPOTHESIS TESTS or simply TESTS, rules for deciding whether a statistical hypothesis is inconsistent with the available data and shall thus be "rejected" (=considered as false), or not. Due to the random nature of the data, the decisions of a test might be erroneous. The quality of a test can be assessed by the probability $\alpha \in (0,1)$ of a TYPE I ERROR that an actually true hypothesis is erroneously rejected, and by the probability $\beta \in (0,1)$ of a TYPE II ERROR that an actually false hypothesis is not rejected.

Among the many possible tests, one is interested in those with low probabilities for *both* error types. In most cases, however, one has to make a compromise, since reducing the one error probability leads to an increase of the other. Classically, one predefines an acceptable level for the probability of a type I error and then looks among the tests meeting this constraint for those with a small probability of a type II error.

Consider the special case of only two rival models $\mu \neq \nu$ with *fixed* parameters θ^{μ} and θ^{ν}, respectively, and assume WLOG that model μ is the correct one. Then, hypothesis H$^{\nu}$ is false, and a good statistical test should have a low probability of not rejecting it. Let $\beta^{\nu}_{\min}(\xi_n)$ be the *minimal* type II error probability among all possible tests for H$^{\nu}$ which are based on data obtained under the exact n-experiment design ξ_n. It follows from the results of Kullback [159, Sec. 4.3], that

$$\beta^{\nu}_{\min}(\xi_n) \overset{\infty}{\approx} \exp(-n\,\delta(\nu, \theta^{\nu}, \xi_n)). \tag{4.26}$$

The lowest possible type II error probability that can be achieved asymptotically by any test for H$^{\nu}$ thus drops exponentially with the associated KLIC of model ν. In the considered special case, a KL-optimal design is simply a maximizer of the KLIC of model ν, that is,

$$\xi_{\mathrm{KL}} \in \operatorname*{argmax}_{\xi \in \Xi} \delta(\nu, \theta^{\nu}, \xi). \tag{4.27}$$

In this case, a KL-optimal design minimizes the lowest possible probability of a type II error that any test for the false hypothesis H^{ν} can achieve asymptotically.

In the general case of several rival models with free parameters, a KL-optimal design maximizes the minimal value that the KLIC attains among all incorrect models and their parameters, see (4.20). Accordingly, *a KL-optimal design minimizes the lowest type II error probability that is asymptotically possible in the worst case.* Remarkably, this result does not depend on the significance level α that chosen for the applied test.

The *actual* type II error probability achieved in practice depends on the particular test that is applied. Based on the Neyman-Pearson lemma one can show that the type II error probability of a likelihood-ratio test asymptotically meets this lower bound under mild regularity conditions. In short, *KL-optimal designs maximize the probability of empirically detecting an actually incorrect model with a likelihood-ratio test,* supposed the sample is sufficiently large. KL-optimal designs are hence in fact the sought-after "ideal" designs for solving MD problems as stated in the introductory question.

A similar statistical justification is given by López-Fidalgo, Tommasi, and Trandafir [177, Sec. 2]. It is a consistent generalization of the argumentation used by Fedorov and Malyutov [98, Sec. 7], Fedorov [93, (4)], Dieses [80, Sec. 3.2.1] and Kuczewski [153, Sec. 3.2.1] for the classic T-criterion, and of that used by Uciński and Bogacka [252, Sec. 2] for the multivariate T-criterion.

4.2.3. T-Optimality: KL-Optimality under Normality

We can now introduce the well-known concept of T-optimality. Introduced by Atkinson and Fedorov [21] in 1975, it preceded the introduction of KL-optimality by López-Fidalgo, Tommasi, and Trandafir [177] by over 30 years. We shall see that T-optimality is a special case of KL-optimality under the common assumptions of known observation covariances and normality.

The following definitions comprise several of the generalizations that were proposed for the original T-optimality of Atkinson and Fedorov [21]. In contrast to some of the original literature, we explicitly distinguish between T-optimality and its local counterpart.

Definition 4.7 (T-Criterion)

Consider scenario 4.1 under the additional normality assumptions (ix) to (xi) and suppose there are at least two rival models, $n_{\mathcal{M}} \geqslant 2$. The T-criterion is the function $T\colon \Xi \mapsto \mathbb{R}$ defined for all $\xi \in \Xi$ as

$$T(\xi) := \min_{\substack{\nu \in \mathcal{M} \\ \nu \neq \bar{\mu}}} \min_{\theta^\nu \in \mathcal{Q}^\nu} \sum_{x \in \operatorname{supp}(\xi)} \xi(x) \| \eta^\nu(x, \theta^\nu) - \bar{\eta}(x) \|^2_{\Omega^{-1}(x)}. \tag{4.28}$$

A design ξ_T maximizing $T(\xi)$ over all $\xi \in \Xi$ is a T-optimal design.

Observe that the sum of squares in (4.28) is the noncentrality $\lambda^\nu(\theta^\nu, \xi)$ for known, non-unit observation covariances, see Tab. 3.1 on p. 113. A T-optimal design maximizes the noncentrality of the worst model with the worst parameter.

Definition 4.8 (Local T-Criterion)

Consider scenario 4.1 under the additional normality assumptions (ix) to (xi) and suppose there are at least two rival models, $n_{\mathcal{M}} \geqslant 2$. Let $\mu \in \mathcal{M}$ and $\theta^\mu \in \mathcal{Q}^\mu$. The local T-criterion (for μ and θ^μ) is for all $\xi \in \Xi$ defined as

$$T(\xi \mid \mu, \theta^\mu) := \min_{\substack{\nu \in \mathcal{M} \\ \nu \neq \mu}} \min_{\theta^\nu \in \mathcal{Q}^\nu} \sum_{x \in \operatorname{supp}(\xi)} \xi(x) \| \eta^\nu(x, \theta^\nu) - \eta^\mu(x, \theta^\mu) \|^2_{\Omega^{-1}(x)}. \tag{4.29}$$

A design $\xi_T(\mu, \theta^\mu)$ maximizing $T(\xi \mid \mu, \theta^\mu)$ over all $\xi \in \Xi$ is a locally T-optimal design (for μ and θ^μ).

This design criterion is an instance of a local design criterion from Prob. 4.3. T-optimality as defined here is a straightforward generalization of the univariate T-optimality considered by Atkinson and Fedorov [20, 21], and is consistent with the multivariate T-optimality of Uciński and Bogacka [248, 249, 250, 251, 252]. T-optimality is a special case of KL-optimality.

Corollary 4.9 (Consistency of KL-Optimality and T-Optimality)

Under the additional normality assumptions (ix) to (xi), $K(\xi) = \frac{1}{2} T(\xi)$ for all $\xi \in \Xi$, and $K(\xi \mid \mu, \theta^\mu) = \frac{1}{2} T(\xi \mid \mu, \theta^\mu)$ for all $\mu \in \mathcal{M}$, all $\theta^\mu \in \mathcal{Q}^\mu$ and all $\xi \in \Xi$. Accordingly, any (locally) KL-optimal design is (locally) T-optimal.

Proof The proof is essentially an application of Thm. C.10 to the KLIC from (4.17) and its model family-based counterpart (4.21). For the univariate special case an explicit proof is given by López-Fidalgo, Tommasi, and Trandafir [177, Thm. 2]. □

Consider the special case that there are only two rival models $\mu \neq \nu$. Then, the local T-criterion for model μ with parameter $\theta^\mu \in \mathcal{Q}^\mu$ reduces to

$$T(\xi \mid \mu, \theta^\mu) = \min_{\theta^\nu \in \mathcal{Q}^\nu} \sum_{x \in \operatorname{supp}(\xi)} \xi(x) \left\| \eta^\nu(x, \theta^\nu) - \eta^\mu(x, \theta^\mu) \right\|^2_{\Omega^{-1}(x)}. \tag{4.30}$$

The local T-criterion for model ν with parameter $\theta^\nu \in \mathcal{Q}^\nu$ is analog.

4.2.4. Discussion

To the best of our knowledge, the KL-criterion in the multi-model form stated in (4.23) has not been proposed so far. It is, however, fully consistent with the multi-model T-criterion of Atkinson and Fedorov [20], as shown in Cor. 4.9, and with the two-model local KL-criterion proposed by López-Fidalgo, Tommasi, and Trandafir [177], see (4.25). Technically, the local multi-model KL-criterion is a rather good-natured generalization of its two-model counterpart, since it extends it only by a minimization over the finite set $\mathcal{M} \setminus \{\mu\}$.

The two-model local KL-criterion is

(a) a concave function of ξ, as shown by Tommasi [244], and

(b) is upper semi-continuous in ξ if equipped with a proper metric for designs, as shown by May and Tommasi [183].

These properties ensure the existence of a KL-optimal design, supposed that the incorrect model has an identifiable best parameter under that design. Aletti, May, and Tommasi [5] show that under mild regularity conditions,

(c) the two-model local KL-criterion is also a continuous function of ξ if equipped with a proper metric for designs, and

(d) a corresponding optimal design it is invariant to a scale-position transformations of the experimental domain.

The discussion provided by Atkinson and Fedorov [20] suggest that these properties also hold for the multi-model KL-criterion under the following additional assumptions:

(xv) Each model $\mu \in \mathcal{M}$ has an identifiable (that is, unique) KLIC-best parameter $\bar{\theta}^\mu(\xi) \in \mathcal{Q}^\mu$ under ξ, which satisfies $\bar{\theta}^\mu(\xi) \in \operatorname{argmin}_{\theta^\nu \in \mathcal{Q}^\nu} \delta(\nu, \theta^\nu, \xi)$ by definition.

(xvi) The model family has an identifiable "second-best" model in the sense of the KLIC, meaning that $\operatorname{argmin}_{\nu \in \mathcal{M} \setminus \{\bar{\mu}\}} \delta\big(\nu, \bar{\theta}^{\nu}(\xi), \xi\big)$ is unique.

These assumptions are required to ensure that the minimums in (4.20) are unique.

Support

KL-optimal designs and T-optimal designs are not necessarily exact designs. Any non-exact design with s support points can, however, be approximated by an n-experiments exact design if $n \gg s$. The next theorem shows that T-optimal design have typically few support points.

Theorem 4.10 (Support of Locally T-Optimal Designs, Fedorov [94])

Consider Def. 4.8 for the special case of two rival models $\mu \neq \nu$ and univariate observables $n_y = 1$, and assume WLOG that model μ is the identifiable correct one. If (i) $\eta^{\nu}(\theta^{\nu}, x)$ is continuous on $\mathcal{Q} \times \mathcal{X}$, (ii) $\mathcal{Q}^{\nu}$ is convex, (iii) $\left(\eta^{\mu}(\theta^{\mu}, x) - \eta^{\nu}(\theta^{\nu}, x)\right)^2$ is a convex function of θ^{ν} for all $x \in \mathcal{X}$, and (iv) the minimum in (4.30) is unique, then a T-optimal design has not more than $n_{\theta^{\nu}} + 1$ support points.

If the incorrect model is affine-linear, $\eta^{\nu}(x, \theta^{\nu}) = \boldsymbol{J}^{\nu}(x)\theta^{\nu} + h^{\nu}(x)$, assumptions (iii) and (iv) always hold. Then, in fact, the number of support points is *equal* to $n_{\theta^{\nu}} + 1$ as shown by Dette and Titoff [79, Cor. 3.2]. According to Uciński and Bogacka [251, Rem. 3], the theorem can be extended to the multivariate T-criterion of the type stated in (4.29).

Dette and Titoff [79, Sec. 4.2] suggest that this result is also valid for KL-optimal designs, with $1 + n_{\theta^{\mu}} + n_{\theta^{\nu}}$ support points. In the examples regarded by López-Fidalgo, Tommasi, and Trandafir [177] and in follow-up papers by Tommasi [244] and Tommasi and López-Fidalgo [245] and Tommasi and López-Fidalgo [243], KL-optimal designs could in fact be well approximated by designs with few support points.

4.3. Two Popular Sequential Design Criteria

This section examines two popular sequential strategies for efficiently solving model discrimination (MD) problems that are based on frequentist inference.

Hunter and Reiner [129] are probably the first to propose a data-adaptive sequential procedure for designing optimal experiments for MD. Procedures of the same type are proposed by Fedorov [91] and Fedorov and Malyutov [98], Fedorov [93, Sec. IV] and Atkinson and Fedorov [21, Sec. 3]. In honor of their inventors, we refer to this approach as Hunter-Reiner (HR)-strategy. It is likely to be also the most-cited approach for optimal experimental design (OED) for MD and has found many applications, for example given in the publications of Asprey and Macchietto [10], Chen and Asprey [67], Dieses [80], and Espie and Macchietto [88] and Hoffmann [120].

Their design criterion has great intuitive appeal and is easy to grasp, but has several deficiencies from today's point of view. It is nevertheless still important, since it turns out to be a special case or a limit case of several more sophisticated design criteria and is comparably cheap to compute. Under mild conditions, the designs constructed in the HR-procedure converge to a T-optimal design.

For the problem of discriminating between several univariate nonlinear normal models, Buzzi-Ferraris and Forzatti [63] developed a sequential procedure and a corresponding design criterion. The idea was modified and extended to the multivariate case by Buzzi-Ferraris et al. [61] and Buzzi-Ferraris, Forzatti, and Canu [64]. Some details are clarified in the reply of Buzzi-Ferraris [62] to the work of Michalik, Stuckert, and Marquardt [186]. Since the authors did not name their approach, we shall refer to it as Buzzi-Ferraris (BF)-strategy. In this section we consider the BF-strategy for discriminating between two models. The multi-model extension suggested in the same paper is considered in Sec. 4.5.

At their core, the sequential procedures of the HR-strategy and the BF-strategy are similar. They mainly differ in their stop criteria and their design criteria. Both are based on the classic empirical formulas from maximum-likelihood inference summarized in the following.

4.3.1. Considered Scenario and Sequential Procedure

Both strategies are based on scenario 4.1 under the additional normality assumptions (ix) to (xi), and consider only *two* different rival models with indices $\mu \in \mathcal{M}$ and $\nu \in \mathcal{M}$, $\mu \neq \nu$.

Classic Empirical Formulas of Maximum-Likelihood Inference

Frequentist inference in the considered scenario is treated in detail in Sec. 3.4. We repeat it here for completeness and to introduce conveniently simplified notation.

Let $d \in \mathcal{Y}^n$ be the data obtained from n experiments performed under design $\xi \in \Xi$, and let $\hat{\theta}^\mu \in \mathcal{Q}^\mu$ be a corresponding parameter maximum-likelihood estimate (PMLE) of model μ. A well-known classic approximation for the distribution of a the corresponding estimator $\hat{\mathcal{Q}}^\mu$ is

$$\hat{\mathcal{Q}}^\mu \overset{a}{\sim} \phi\Big(\theta^\mu \,\Big|\, \hat{\theta}^\mu, n^{-1}\hat{\boldsymbol{M}}^{\mu^{-1}}\Big), \tag{4.31}$$

with the matrix $\hat{\boldsymbol{M}}^\mu$ from (4.15). The corresponding classic approximation for the distribution of experimental outcomes under condition $x \in \mathcal{X}$ is

$$q(y\,|\,x) \approx \phi\big(y\,\big|\,\hat{\eta}^\mu(x), \hat{\boldsymbol{T}}^\mu(x)\big), \text{ for all } y \in \mathcal{Y}. \tag{4.32}$$

The function $\hat{\eta}^\mu(x)$ (see (4.12)) predicts the *average* experimental outcome under $x \in \mathcal{X}$, based on model μ and the available experiments. The $n_y \times n_y$ matrix

$$\hat{\boldsymbol{T}}^\mu(x) := \boldsymbol{\Omega}(x) + \hat{\boldsymbol{V}}^\mu(x) \tag{4.33}$$

quantifies the total uncertainty about the *actual* outcome of an experiment under x, given model μ and the available experiments. It is composed of the matrix $\boldsymbol{\Omega}(x)$ representing the experimental uncertainty and the matrix

$$\hat{\boldsymbol{V}}^\mu(x) := n^{-1}\hat{\boldsymbol{J}}^\mu(x)\hat{\boldsymbol{M}}^{\mu^{-1}}\hat{\boldsymbol{J}}^{\mu^\top}(x). \tag{4.34}$$

which quantifies, in a locally linear approximation, the propagation of the parameter uncertainty onto the prediction $\hat{\eta}^\mu(x)$.

Approximation (4.31) can be motivated in two ways. Assuming that

(v) model μ is locally affine-linear (Def. 3.5) around the PMLE $\hat{\theta}^\mu$

justifies it for samples of any size $n \in \mathbb{N}$. Such a derivation is given, for example, by Buzzi-Ferraris, Forzatti, and Canu [64]. Alternatively, assuming that

(vi) model μ is correct,

justifies it for large sample sizes n. It is then a special case of (3.64). Approximation (4.32) relies in any case on assumption (v).

Sequential Procedure

The original BF-strategy is limited to observation covariances that are independent from the experimental condition. We took the additional (simple) steps to generalize it to observation covariances that might depend on the experimental condition. Furthermore, the BF-strategy is originally formulated for several models. Here, we only consider the comprised two-model case. Multi-model generalizations are discussed in Sec. 4.5.

Algorithm 4.2: Sequential procedure of the Hunter-Reiner strategy and the Buzzi-Ferraris strategy.

```
input  : two different rival models with indices $\mu \neq \nu$
         $s \in \mathbb{N}$ previous experiments, design $\xi_s$, data $d_s$
output : set of non-rejected models $\mathcal{M}_n$, either empty or singleton

for $n = s$ to $\infty$ do
    foreach $\lambda \in \{\mu, \nu\}$ do
        $\hat{\theta}_n^\lambda \leftarrow \operatorname{argmin}_{\theta^\lambda \in \mathcal{Q}^\lambda} s^\lambda(\theta^\lambda, d_n, \xi_n)$;    // lsq estimation
    end
    $\mathcal{M}_n := \{\lambda \in \{\mu, \nu\} \mid$ model $\lambda$ passed adequacy test using $d_n$ and $\xi_n\}$;
    if $|\mathcal{M}_n| < 2$ then return $\mathcal{M}_n$;                    // termination check
    $x_{n+1} \leftarrow \operatorname{argmax}_{x \in \mathcal{X}} \Psi(x; \mu, \nu, \hat{\theta}_n^\mu, \hat{\theta}_n^\nu, \xi_n, d_n)$;   // design experiment
    $y_{n+1} \leftarrow$ realization of $\mathcal{Y}_{x_{n+1}}$;                      // perform experiment
    $\xi_{n+1} \leftarrow \frac{n}{n+1}\xi_n + \frac{1}{n+1}\xi^{x_{n+1}}$;                    // update design
    $d_{n+1}^\top \leftarrow [d_n^\top \quad y_{n+1}^\top]$;                          // extend data vector
end
```

For efficiently solving the MD problem from (Q4.1) in this setting, both the HR-strategy and the BF-strategy propose a sequential approach. Its central steps are described by Alg. 4.2. Both strategies actually propose more sophisticated stop criteria than shown there. We do not discuss them here since our focus are the involved sequential OED problems.

Starting from two models and s previous experiments, Alg. 4.2 performs the following steps. Using all available experiments, it first calculates PMLE $\hat{\theta}_n^\mu$ for both models, which are least-squares (LSQ) estimates under the given assumptions.

Then, it assesses the adequacy of both models via adequacy tests based on available data and parameter estimates. If a model λ fails the test, the hypothesis that μ is correct is rejected. If only one test fails, the procedure stops and returns the index of the remaining model. If both tests fail, it stops and returns an empty set. Under certain regularity conditions (partially discussed later), the procedure terminates with probability tending to μ as $n \to \infty$.

If both tests succeed, the procedure continues to gather more data. It selects the experimental condition x_{n+1} for the next experiment by solving a sequential OED problem for MD with the design criterion Ψ (discussed in the next section), using the latest parameter estimates $\hat{\theta}_n^\mu$ and $\hat{\theta}_n^\nu$. The conditions of all available experiments are then described by the design ξ_{n+1}, which is determined from the corresponding design ξ_n from the previous iteration and the design $\xi^{x_{n+1}}$ which puts full weight at x_{n+1}.

It applies the new experimental condition x_{n+1} to the process and records the resulting data y_{n+1}. Then, it continues with parameter estimation in the hope that the previous and the newly gathered experimental results are sufficient to meet the stopping criterion.

4.3.2. The HR-Criterion and the BF-Criteria

The main difference between the HR-strategy and the BF-strategy is the particular form of the sequential design criterion Ψ used to determine the conditions of the next experimental design in Alg. 4.2.

Definition 4.11 (HR-Criterion)

The HR-CRITERION for discrimination between two models from $\mathcal{M}$ is the function $H: \mathcal{X} \times \mathcal{M} \times \mathcal{M} \mapsto \mathbb{R}$ defined under all $x \in \mathcal{X}$ and for all $\mu, \nu \in \mathcal{M}$ as

$$H(x; \mu, \nu) := \left\| \eta^\mu\left(x, \hat{\theta}_n^\mu\right) - \eta^\nu\left(x, \hat{\theta}_n^\nu\right) \right\|^2_{\Omega^{-1}(x)}. \tag{4.35}$$

Hunter and Reiner [129] derive this design criterion from log-likelihood statistic used in a likelihood ratio test for the MD problem. Their derivation rests upon the crucial assumption that exactly one of the two rival model is correct.

We write $H\left(x; \mu, \nu, \hat{\theta}^\mu, \hat{\theta}^\nu\right)$ to emphasize that the HR-criterion depends parametrically on the PMLEs of both rival models. Using this notation, the HR-procedure is the special case of Alg. 4.2 for the choice

$$\Psi\left(x; \mu, \nu, \hat{\theta}_n^\mu, \hat{\theta}_n^\nu, \cdot, \cdot\right) := H\left(x; \mu, \nu, \hat{\theta}_n^\mu, \hat{\theta}_n^\nu\right), \text{ for all } x \in \mathcal{X}. \tag{4.36}$$

Definition 4.12 (BF-Criteria [61, 64])

Let $\mathcal{X}'$ be the set of all $x \in \mathcal{X}$ under which $\hat{\boldsymbol{T}}^{\mu\nu}(x) := \hat{\boldsymbol{T}}^{\mu}(x) + \hat{\boldsymbol{T}}^{\nu}(x)$ is invertible. The BF-CRITERION and the MODIFIED BF-CRITERION for discrimination between two models from $\mathcal{M}$ are the functions $B\colon \mathcal{X}' \times \mathcal{M} \times \mathcal{M} \mapsto \mathbb{R}$ and $\tilde{B}\colon \mathcal{X}' \times \mathcal{M} \times \mathcal{M} \mapsto \mathbb{R}$, respectively, defined under all $x \in \mathcal{X}'$ and for all $\mu, \nu \in \mathcal{M}$ as

$$B(x;\mu,\nu) := \left\|\eta^{\mu}\left(x,\hat{\theta}^{\mu}\right) - \eta^{\nu}\left(x,\hat{\theta}^{\nu}\right)\right\|^2_{\left(\hat{\boldsymbol{T}}^{\mu\nu}(x)\right)^{-1}} \text{ and} \tag{4.37}$$

$$\tilde{B}(x;\mu,\nu) := B(x;\mu,\nu) + 2\,\mathrm{tr}\Big(\boldsymbol{\Omega}(x)\left(\hat{\boldsymbol{T}}^{\mu\nu}(x)\right)^{-1}\Big). \tag{4.38}$$

The BF-criterion was introduced by Buzzi-Ferraris et al. [61] and Buzzi-Ferraris and Forzatti [63] and the modified variant by Buzzi-Ferraris, Forzatti, and Canu [64]. Both variants are derived from the aim of finding the best experimental condition for falsifying the hypothesis that the responses of both rival models are equal, after averaging out the parameter uncertainty (in terms of the distribution of the parameter maximum-likelihood estimators (PMLEs)). The derivation does not require any of the models to be correct, but assumes that both models are locally affine-linear around their PMLEs. Furthermore, it makes the assumption that the PMLE of both models are statistically independent. The latter assumption is typically not met in practice, and particularly not in Alg. 4.2. The inventors are well aware of that and propose to consider the design criterion as a heuristic.

Backtracking the definition of $\hat{\boldsymbol{T}}^{\mu\nu}$ in Sec. 4.3.1 reveals that this matrix depends on the PMLEs of both models and on the underlying design. We write $B\left(x;\mu,\nu,\hat{\theta}^{\mu},\hat{\theta}^{\nu},\xi\right)$ to emphasize this dependency. With this notation, the BF-procedure is the special case of Alg. 4.2 for

$$\Psi\left(x;\mu,\nu,\hat{\theta}^{\mu}_n,\hat{\theta}^{\nu}_n,\xi_n,\cdot\right) := B\left(x;\mu,\nu,\hat{\theta}^{\mu}_n,\hat{\theta}^{\nu}_n,\xi_n\right) \tag{4.39}$$

for all $x \in \mathcal{X}$. Both Defs. 4.11 and 4.12 are instances of a sequential design criterion Prob. 4.4.

4.3.3. The Influence of Uncertainties on the Design Criteria

HR-Criterion

The HR-criterion can be considered as an one-experiment approximation of the two-model T-criterion (4.30), where the involved unknown parameters have been replaced by the available PMLES.

The response $\eta^{\mu}\big(x, \hat{\theta}_n^{\mu}\big)$ is the prediction of models μ for the average outcome of an experiment under x, based on the parameter estimate $\hat{\theta}_n^{\mu}$ obtained in n previous experiments. The observation covariance $\boldsymbol{\Omega}(x)$ quantifies the amount of random fluctuations of observations under x, that is, the uncertainty of an experiment under x.

The HR-criterion is thus the difference between the predictions of models μ and ν for the average outcome of an experiment under x, based on available PMLES, relative to the experimental uncertainty. In short, under an HR-optimal experimental setting, the difference between the model predictions is maximal relative to the experimental uncertainty.

The HR-criterion is *robust with respect to experimental uncertainty* in the sense that it takes into account the random fluctuations of the designed experiment. It is, however, *not parameter-robust,* since it does not take into account the variability of the PMLES.

BF-Criteria

The BF-criterion has the same shape as the T-criterion. It also measures the difference between the predictions of both models, but relative the sum of the total uncertainties of both rival models described by the matrix $\hat{\boldsymbol{T}}^{\mu\nu}$, see (4.33)

The first term of the modified BF-criterion is the BF-criterion, the second summand is the trace of a matrix product. This product can be rewritten as

$$\begin{aligned}\boldsymbol{\Omega}(x)\big(\hat{\boldsymbol{T}}^{\mu\nu}(x)\big)^{-1} &= \big(\hat{\boldsymbol{T}}^{\mu\nu}(x)\boldsymbol{\Omega}^{-1}(x)\big)^{-1}\\ &= \big(2 + \hat{\boldsymbol{V}}^{\mu}(x)\boldsymbol{\Omega}^{-1}(x) + \hat{\boldsymbol{V}}^{\nu}(x)\boldsymbol{\Omega}^{-1}(x)\big)^{-1}, \qquad (4.40)\end{aligned}$$

with $\hat{\boldsymbol{V}}^{\mu}(x)$ from (4.34). It quantifies the parameter-induced uncertainties in the model predictions represented by $\hat{\boldsymbol{V}}^{\mu}(x)$, relative to the experimental uncertainty represented by $\boldsymbol{\Omega}(x)$.

If $\boldsymbol{\Omega}(x)\big(\hat{\boldsymbol{T}}^{\mu\nu}(x)\big)^{-1}$ is “large” under x, the actual behavior of an experiment under x can be expected to differ from its model-based predictions in a magnitude exceeding that of the experimental uncertainty. It thus has a large probability of being simply inefficient for MD in practice. The additional term in the modified BF-criterion penalizes such undesirable experimental conditions.

Relation between HR-Criterion and BF-Criteria

Proposition 4.13 (BF-Criteria under Small Uncertainties)

If the parameter-induced uncertainties in the responses of both models vanish, $\hat{\boldsymbol{V}}^{\mu}(x) = \hat{\boldsymbol{V}}^{\nu}(x) = \boldsymbol{0}$ under all experimental conditions $x \in \mathcal{X}$, then $B(x;\mu,\nu) \propto H(x;\mu,\nu)$ and $\tilde{B}(x;\mu,\nu) \propto H(x;\mu,\nu)$ for all $x \in \mathcal{X}$. Consequentially, so that BF-optimal designs (original and modified) are HR-optimal.

Proof One can easily see from the given definitions that $\hat{\boldsymbol{V}}^{\mu}(x) = \hat{\boldsymbol{V}}^{\nu}(x) = \boldsymbol{0}$ implies that $\boldsymbol{T}^{\mu\nu}(x) = 2\boldsymbol{\Omega}(x)$. Applying this equation to the BF-criteria and substituting the definition of the HR-criterion yields the equality $B(x;\mu,\nu) = 2H(x;\mu,\nu)$ and $\tilde{B}(x;\mu,\nu) = 2H(x;\mu,\nu) + n_y$. □

The matrix $\hat{\boldsymbol{V}}^{\mu}(x)$ quantifies, in affine-linear approximation, the uncertainty in the response of model μ due to parameter uncertainty after n experiments in terms of $n^{-1}\hat{\boldsymbol{M}}^{\mu^{-1}}$, see (4.31) and (4.34). The HR-criterion is therefore a special case of both BF-criteria for vanishing parameter uncertainty. Reversely, the BF-criteria can be viewed as parameter-robust generalizations of the HR-criterion.

Based on Prop. 4.13 and some continuity arguments one can derive the following approximate results. If the parameter-induced uncertainty in the model responses is significantly smaller than the experimental uncertainty, that is, if

$$\left\|\hat{\boldsymbol{V}}^{\mu}(x)\right\| \ll \|\boldsymbol{\Omega}(x)\| \text{ and } \left\|\hat{\boldsymbol{V}}^{\nu}(x)\right\| \ll \|\boldsymbol{\Omega}(x)\| \text{ for all } x \in \mathcal{X}, \tag{4.41}$$

with some matrix norm $\|\cdot\|$, then B and $\tilde{B}$ are approximately proportional to the HR-criterion.

Under certain regularity conditions discussed in Secs. 3.4.1 and 3.4.3, the inverse of matrix $n\hat{\boldsymbol{M}}^{\mu}$ converges to zero as the sample size n tends to infinity, so that $\hat{\boldsymbol{V}}^{\mu}(x)$ vanishes asymptotically under all $x \in \mathcal{X}$. Then, (4.41) is satisfied in large samples. Therefore, both BF-criteria reduce asymptotically to the HR-criterion under certain regularity conditions.

4.3.4. Convergence to T-Optimal Designs

Theorem 4.14 (Convergence of the Hunter-Reiner Procedure)

Assume that the responses of both rival models μ and ν are linear in their parameter θ^μ and θ^ν, respectively. If the sequence of designs $\xi_s, \xi_{s+1}, \dots$ constructed by Alg. 4.2 on p. 142 with the HR-criterion (4.35) converges to a design under which the Kullback-Leibler information criterion (KLIC)-best parameters of both models are identifiable, then this design is almost surely T-optimal.

Proof Proof are given by Fedorov and Malyutov [98, Thm. 7.2] and Fedorov [93, Thm. 3].□

The HR-procedure is thus a data-adaptive procedure for the sequential construction of T-optimal designs. The available proofs of convergences are, however, limited to the univariate case and to linear models. A look at the proof of [93, Thm. 3] suggests, however, that a generalization to the multivariate case should easily be possible. The linearity assumption, however, cannot be easily circumvented. From a practical point of view, one thus uses the HR-procedure as a heuristic approach for efficiently solving MD problems.

To the best of our knowledge, the asymptotic behavior of the BF-procedure has not been examined rigorously. Suppose the sequence of designs constructed by the procedure converges to a limit design ξ, and assume that under this design, the regularity conditions for the consistency and asymptotic normality of PMLEs, discussed in Sec. 3.4.1, are met for each both rival models $\lambda \in \{\mu, \nu\}$. Then, the PMLE covariances given by the matrices $n^{-1}\hat{\boldsymbol{M}}^{\lambda^{-1}}$ vanish asymptotically, so that $\hat{\boldsymbol{T}}_n^{\mu\nu}$ converges to $2\boldsymbol{\Omega}^{-1}$, and both BF-criteria reduce to the HR-criterion. The BF-procedure is then asymptotically equivalent to the HR-procedure, which convergences to a T-optimal design under the previously given assumptions.

Since BF-optimal experiments are, however, not necessarily efficient for reducing the parameter covariance, a large number of experiments might be required in practice until this limiting behavior of the BF-procedure gets visible.

4.4. New Misspecification-Robust Sequential Design Criteria

In Sec. 3.4 we proposed a new empirical formula for the covariance of parameter maximum-likelihood estimator (PMLE) for models that are both nonlinear and incorrect models. Here, we show it can be used to derive new design criteria for model discrimination (MD) with improved parameter-robustness.

4.4.1. Classic and Robust Formulas for the Distribution of PMLEs

The classic empirical approximation

$$\hat{Q}^{\mu} \overset{a}{\sim} \mathcal{N}\left(\hat{\theta}^{\mu}, n^{-1}\hat{\boldsymbol{M}}^{\mu^{-1}}\right), \tag{4.42}$$

with $\hat{\boldsymbol{M}}^{\mu}$ from (4.15), for the distribution of a PMLE of model $\mu \in \mathcal{M}$ is based on the assumption that the model is correct and/or that is has locally affine-linear responses around around $\hat{\theta}^{\mu}$. As discussed in Sec. 3.4.2, this relation remains appropriate if the model is "almost" correct or "weakly" nonlinear. It is generally inappropriate, however, if the model is *both* significantly incorrect *and* properly nonlinear.

In Sec. 3.4 we proposed the new robust empirical approximation

$$\hat{Q}^{\mu} \overset{a}{\sim} \mathcal{N}\left(\hat{\theta}^{\mu}, n^{-1}\hat{\boldsymbol{R}}^{\mu}\right) \tag{4.43}$$

for the distribution of a PMLE, see (3.64), where

$$\hat{\boldsymbol{R}}^{\mu} := \left(\hat{\boldsymbol{M}}^{\mu} + \hat{\boldsymbol{N}}^{\mu}\right)^{-1}\hat{\boldsymbol{M}}^{\mu}\left(\hat{\boldsymbol{M}}^{\mu} + \hat{\boldsymbol{N}}^{\mu}\right)^{-1} \tag{4.44}$$

with $\hat{\boldsymbol{N}}^{\mu}(x)$ from (4.16). As discussed in Sec. 3.4.2, this approximation is justified even for models that are *both* incorrect *and* nonlinear. Furthermore, it is a consistent generalization of (4.42): if model μ is locally affine-linear around $\hat{\theta}^{\mu}$, then the Hessians of its response components almost vanish, so that $\hat{\boldsymbol{N}}^{\mu} \approx \boldsymbol{0}$ and thus $\hat{\boldsymbol{R}}^{\mu} \approx \hat{\boldsymbol{M}}^{\mu^{-1}}$. If the model is correct, these approximate equalities hold asymptotically.

In the class of MD problems arising from scenario 4.1, (a) all rival models may be nonlinear, (b) all of them are incorrect except for exactly one, and (c) this correct model is unknown. Therefore, the classic approximation (4.42) is generally inadequate for all models except one, which is unknown. Nevertheless, various frequentist design criteria use it for empirically quantifying the parameter uncertainty of the rival models. In the following, we show how such a design criterion can be reformulated to overcome this drawback. We use the Buzzi-Ferraris (BF)-criterion (Def. 4.12) as example. The argumentation can be applied likewise to any other frequentist design criterion for MD which quantifies the parameter uncertainty based on (4.42).

4.4.2. BF-Criteria with Improved Parameter-Robustness

We propose to replace the classic approximation (4.42) in both BF-criteria by its robust counterpart (4.43) for both rival models μ and ν. Analog to the matrix $\hat{\boldsymbol{T}}^{\mu\nu}$ used in the BF-criterion, the matrix

$$\hat{\boldsymbol{V}}^{\mu\nu}(x) := 2\boldsymbol{\Omega}(x) + n^{-1}\hat{\boldsymbol{J}}^{\mu}(x)\hat{\boldsymbol{R}}^{\mu}\hat{\boldsymbol{J}}^{\mu^{\top}}(x) + n^{-1}\hat{\boldsymbol{J}}^{\nu}(x)\hat{\boldsymbol{R}}^{\nu}\hat{\boldsymbol{J}}^{\nu^{\top}}(x)$$

describes the common total uncertainty of both models μ and ν about the outcome of an experiment under condition $x \in \mathcal{X}$. It takes into account both the experimental uncertainty in terms of $\boldsymbol{\Omega}(x)$ and the parameter uncertainty in terms of $\hat{\boldsymbol{R}}^{\mu}$ and $\hat{\boldsymbol{R}}^{\nu}$. Replacing $\hat{\boldsymbol{T}}^{\mu\nu}$ in both BF-criteria by $\hat{\boldsymbol{V}}^{\mu\nu}(x)$ leads to the following design criteria.

Definition 4.15 (Robust BF-Criteria)

Let $\mathcal{X}'$ be the set of all $x \in \mathcal{X}$ under which $\hat{\boldsymbol{V}}^{\mu\nu}(x)$ is invertible. The ROBUST counterparts of the BF-criteria for discrimination between two models from $\mathcal{M}$ are the functions $B' : \mathcal{X}' \times \mathcal{M} \times \mathcal{M} \mapsto \mathbb{R}$ and $\tilde{B}' : \mathcal{X}' \times \mathcal{M} \times \mathcal{M} \mapsto \mathbb{R}$, respectively, defined under all $x \in \mathcal{X}'$ and for all $\mu, \nu \in \mathcal{M}$ as

$$B'(x;\mu,\nu) := \left\|\eta^{\mu}\left(x,\hat{\theta}^{\mu}\right) - \eta^{\nu}\left(x,\hat{\theta}^{\nu}\right)\right\|^{2}_{\left(\hat{\boldsymbol{V}}^{\mu\nu}(x)\right)^{-1}} \text{ and} \tag{4.45}$$

$$\tilde{B}'(x;\mu,\nu) := B'(x;\mu,\nu) + 2\,\mathrm{tr}\left(\boldsymbol{\Omega}(x)\left(\hat{\boldsymbol{V}}^{\mu\nu}(x)\right)^{-1}\right). \tag{4.46}$$

It is obvious from the derivation that $B'(x;\mu,\nu) \approx B(x;\mu,\nu)$ and $\tilde{B}'(x;\mu,\nu) \approx \tilde{B}(x;\mu,\nu)$ for all $x \in \mathcal{X}$ if the responses of both models μ and ν are locally affine-linear around around their parameter maximum-likelihood estimates (PMLES). If model μ is correct,

$$\hat{\boldsymbol{V}}^{\mu\nu}(x) \overset{\infty}{\approx} 2\boldsymbol{\Omega}(x) + n^{-1}\hat{\boldsymbol{J}}^{\mu}(x)\hat{\boldsymbol{M}}^{\mu}\hat{\boldsymbol{J}}^{\mu^{\top}}(x) + n^{-1}\hat{\boldsymbol{J}}^{\nu}(x)\hat{\boldsymbol{R}}^{\nu}\hat{\boldsymbol{J}}^{\nu^{\top}}(x)$$

for large n. If model ν is correct, $\hat{\boldsymbol{V}}^{\mu\nu}(x)$ simplifies likewise. If both models are correct, then $\hat{\boldsymbol{V}}^{\mu\nu}(x) \overset{\infty}{\approx} \hat{\boldsymbol{T}}^{\mu\nu}$. This case, however, is excluded by the assumption that exactly one of the rival models is correct. Unless both rival models are affine-linear, the BF-criteria and its robust counterparts proposed here are different design criteria.

4.4.3. Discussion

The design criteria B' and $\tilde{B}'$ have largely similar properties as B and $\tilde{B}$, respectively, which are discussed in Secs. 4.3.3 and 4.3.4. In particular, they are robust with respect to the experimental uncertainty and the parameter uncertainty and reduce under regularity conditions asymptotically to the Hunter-Reiner (HR)-criterion.

The intent of using the robust PMLE covariance formula (4.44) instead of its classic counterpart is to capture the parameter uncertainty of the incorrect model more adequately. The aim is to make the optimal experiments obtained from the new design criteria B' and $\tilde{B}'$ less susceptible to fluctuations in the parameter estimates, compared to those obtained from the BF-criteria B and $\tilde{B}$.

The robust covariance formula is, however, based on an approximation that is valid only in the large-sample limit. It can thus not be determined a priori whether using it actually pays off in practice, where sample sizes are finite and possibly small. Using Monte-Carlo simulations, we compare the classic and the robust covariance formulas in Chap. 7, and the corresponding variants of the BF-criteria in Chap. 9.

4.5. Sequential Multi-Model Design Criteria

The sequential strategies for model discrimination (MD) considered so far are suited for discriminating between exactly *two* rival models. This section discusses various approaches for generalizing them to *several* models. Bayesian design criteria for MD, which typically come with an intrinsic support for several models, are discussed in detail in Chap. 5.

4.5.1. Problem Statement

Consider a family of $n_{\mathcal{M}} \geqslant 2$ models, without loss of generality (WLOG) distinguished by indices from $\mathcal{M} := \{1, \dots, n_{\mathcal{M}}\}$. Assume that exactly one of the models $\bar{\mu} \in \mathcal{M}$ is correct, but *which* one is not known to us. We aim to identify it empirically, that is, based on experimental data.

We proceed sequentially, designing, performing, and analyzing one experiment after another, as described by Sec. 4.5. The iteration index is denoted n. For all $n \in \mathbb{N}$, let $x_n \in \mathcal{X}$ be the condition under which the n-th experiment is performed, and let ξ_n be the exact design describing the conditions of experiments 1 to n.

We aim to keep the experimental effort for solving this problem low through optimal experimental design (OED). To that end, we require a SEQUENTIAL MULTI-MODEL DESIGN CRITERION $\tilde{\Psi}_n : \mathcal{X} \mapsto \mathbb{R}$ with the following property: if the experiments are performed under conditions maximizing it,

$$x_{n+1} \in \operatorname*{argmax}_{x \in \mathcal{X}} \Psi_n(x), \text{ for all } n \in \mathbb{N}_0, \tag{4.47}$$

then the resulting sequence of designs $\xi_1, \xi_2, \ldots$ converges, preferably fast, to a design $\xi^\star$ that is efficient for identifying the correct model. Examples for the latter are a Kullback-Leibler (KL)-optimal design, or a T-optimal design if the observation covariances are known and the models are normal.

4.5.2. Common Multi-Model Design Criteria

Efficient experiments for MD are model-dependent: Am experimental condition that is efficient for discriminating between model $\mu \in \mathcal{M}$ and model $\nu \in \mathcal{M}$ is not necessarily efficient for discriminating between model μ and a different model $\lambda \in \mathcal{M}$. A sequential multi-model design criterion must hence make a compromise between the influence of the several rival models from $\mathcal{M}$.

The relevant literature provides a plethora of sequential multi-model design criteria for MD. Those from the frequentist school, which we consider here, are typically heuristic generalizations of a sequential two-model design criterion that use different compromises for the influence of the several models. It seems that rigorous proofs for their asymptotic efficiency (like convergence to a KL-optimal or a T-optimal design) are rare. A notable exception is the multi-model Hunter-Reiner (HR)-criterion of Atkinson and Fedorov [20] considered in Sec. 4.5.3.

In the following we sketch some popular frequentist multi-model design criteria. Details can be found in the given references. For all $n \in \mathbb{N}$, let $\Psi_n : \mathcal{X} \times \mathcal{M} \times \mathcal{M} \mapsto \mathbb{R}$ be a sequential design criterion for discriminating between two models from $\mathcal{M}$ in iteration n. Suppose the design criterion is invariant under an exchange of the models, that is, $\Psi_n(x; \mu, \nu) = \Psi_n(x; \nu, \mu)$, for all $x \in \mathcal{X}$ and all $\mu, \nu \in \mathcal{M}$. Examples for Ψ_n are the HR-criterion (Def. 4.11), the Buzzi-Ferraris (BF)-criterion (Def. 4.12), and its misspecification-robust extension (Def. 4.15).

Pair of Seemingly Easiest/Hardest-to-Discriminate Models

The value $\Psi_n(x;\mu,\nu)$ is a *prediction* (based on n previous experiments) for the utility of an experiment under x for discriminating between models μ and ν. This interpretation can be used to derive sequential multi-model design criteria as follows.

One might perform that experiment under which the predicted utility is largest among selected pairs of different models, which corresponds to

$$\tilde{\Psi}_n(x) = \max_{\substack{\mu,\nu\in\mathcal{M}_n\\ \mu>\nu}} \Psi_n(x;\mu,\nu), \text{ for all } x\in\mathcal{X}, \tag{4.48}$$

where $\mathcal{M}_n \subseteq \mathcal{M}$. This choice focuses the experiment on the pair of models that seems easiest to discriminate. Alternatively, one might focus it on the model pair that seems hardest to discriminate by choosing

$$\tilde{\Psi}_n(x) = \min_{\substack{\mu,\nu\in\mathcal{M}_n\\ \mu>\nu}} \Psi_n(x;\mu,\nu), \text{ for all } x\in\mathcal{X}. \tag{4.49}$$

This first approach is, for example, suggested and used by Buzzi-Ferraris et al. [61], Buzzi-Ferraris and Forzatti [63], and Buzzi-Ferraris, Forzatti, and Canu [64] and by Chen and Asprey [67, Sec. 6] for the BF-criterion. The second approach is applied by Cooney and McDonald [72] to the HR-criterion.

Choosing $\mathcal{M}_n$ such that it includes only models that are “compliant” in some sense with the results of the n available experiments, one can reduce the risk of spending experimental effort on discriminating between models that are likely to be incorrect.

Generally, the computational effort for evaluating (4.48) and (4.49) is $n_{\mathcal{M}}^2$ times the effort required to evaluate Ψ_n.

Pair of Currently Best Models

Design criteria (4.48) and (4.49) choose a pair of models to discriminate between based on the *predicted* experimental utility. We discussed in Sec. 4.1.2 that such predictions are subject to various uncertainties, whose magnitude might be difficult to estimate. An less fragile variant is to choose a pair of models solely based on their compliance with the available data.

Following a worst-case approach, one might choose

$$\tilde{\Psi}_n(x) = \Psi(x;\mu_n,\nu_n), \text{ for all } x\in\mathcal{X}, \tag{4.50}$$

where μ_n and ν_n are the models which explain the currently available data from the n experiments "best" and "second-best" in some sense. This design criterion is essentially an a posteriori counterpart of (4.49).

Atkinson and Fedorov [20] propose this approach to generalize the HR-criterion to several models. It is considered in detail in Sec. 4.5.3.

Evaluating this multi-model design criterion requires the same computational effort as evaluating one two-model design criterion, plus the effort for identifying the best and second best model, which often comes for free since the necessary inferences are performed anyhow.

Equally Distributed Interest

If one wants to distribute the influence of the rival models on the resulting multi-model design criterion more evenly, one might choose

$$\tilde{\Psi}_n(x) = \sum_{\substack{\mu,\nu\in\mathcal{M}_n \\ \mu>\nu}} \Psi_n(x;\mu,\nu), \quad \text{for all } x \in \mathcal{X}. \tag{4.51}$$

As previously, the limitation to models from $\mathcal{M}_n \subseteq \mathcal{M}$ might reduce to the risk of wasting experimental effort on models which are likely to be incorrect.

Buzzi-Ferraris, Forzatti, and Canu [64] propose this approach as alternative to (4.48) for generalizing the BF-criterion to several models. This suggestion was used in the computations of Schwaab et al. [225, Exs. 1–4]. Asprey and Macchietto [10] and Espie and Macchietto [88] suggest it as multi-model generalization of the HR-criterion, using χ^2 lack-of-fit tests to determine the set of data-compliant models. This approach was used by Cooney and McDonald [72] for their computations.

The computational effort for evaluating it scales with $n_{\mathcal{M}}^2$ in general

Bilinear Weighting

Design criterion (4.51) can be further generalized by weighting the influence of each model pair onto the resulting multi-model design criterion. Suppose that for each model $\mu \in \mathcal{M}$ a weight $w_n^\mu \in \mathbb{R}_0^+$ is available which quantifies the "plausibility" that μ is the correct model, given the results of n experiments, and define $w^\top := \begin{bmatrix} w_n^1 & \dots & w_n^{n_\mathcal{M}} \end{bmatrix}^\top$. One might then perform the experiment maximizing the multi-model design criterion defined as

$$\tilde{\Psi}_n(x) = \sum_{\substack{\mu,\nu \in \mathcal{M} \\ \mu > \nu}} w_n^\mu w_n^\nu \Psi_n(x; \mu, \nu) = w_n^\top \boldsymbol{\Psi}_n(x) w_n, \text{ for all } x \in \mathcal{X}, \tag{4.52}$$

where $\boldsymbol{\Psi}_n(x)$ is an $n_{\mathcal{M}} \times n_{\mathcal{M}}$ matrix that has the component $\Psi_n(x; \mu, \nu)$ in row μ and column ν on the upper or lower triangle and is zero otherwise.

This design criterion is a bilinear form in the model weights. The influence of each pair of rival models onto this design criterion increases linearly in their joint plausibility $w_n^\mu w_n^\nu$. Design criterion (4.52) comprises (4.50) as special case, if $w_n^{\mu_n} = w_n^{\nu_n} = 1$ and all other weights are zero. Furthermore, it reduces to (4.51) if $w_n^\mu = 1$ for all $\mu \in \mathcal{M}_n$, and the remaining weights are zero.

Schwaab et al. [225, Exs. 1–4] proposes such an approach to generalize the BF-criterion to several models. They use weights derived from the χ^2-lack-of-fit test statistic. We shall see in Chap. 5 that certain Bayesian design criteria for MD considered in Chap. 5 also fall into this class.

The computational effort for evaluating (4.52) generally scales with $n_{\mathcal{M}}^2$.

4.5.3. Multi-Model Design Criteria Considered in this Thesis

Definition 4.16 (Multi-Model HR-Criterion)

Consider scenario 4.1 under the additional normality assumptions (ix) to (xi). For all $\mu \in \mathcal{M}$, let $\hat{\theta}^\mu \in \operatorname{argmin}_{\theta^\mu \in \mathcal{Q}^\mu} s^\mu(\theta^\mu, d, \xi)$. Suppose that $s^\mu\left(\hat{\theta}^\mu, d, \xi\right)$ has a unique minimum $\hat{\mu}$ on $\mathcal{M}$ and a unique minimum $\hat{\nu}$ on $\mathcal{M} \setminus \{\hat{\mu}\}$. Let H be the (two-model) HR-criterion from Def. 4.11. The HR-CRITERION for discriminating among *all* models from $\mathcal{M}$ is the function defined as $H(x; \hat{\mu}, \hat{\nu})$ for all $x \in \mathcal{X}$.

This design criterion is a straightforward generalization of the two-model HR-criterion from Def. 4.11 to several models. Both design criteria are identical in the case of two rival models.

Under the given assumptions, $\hat{\theta}^\mu$ is parameter maximum-likelihood estimates (MLEs) of model μ for each $\mu \in \mathcal{M}$. Furthermore $\hat{\mu}$ is a model maximum-likelihood estimate (MMLE) and $\hat{\nu}$ is a MMLE on the set of rival models excluding $\hat{\mu}$, The parameter-minimal sum of squared residuals (SSR) $s_\mu\left(\hat{\theta}^\mu, d, \xi\right)$ it the lack of fit of model μ with respect to the data d obtained under design ξ. Therefore,

$\hat{\mu}$ and $\hat{\nu}$ are the models which fit the data best and second-best, respectively. Definition 4.16 is thus an instance of (4.50), using the negative lack of fit to measure the "goodness" of the rival models.

If the observation covariances are known and the model is normal, the design resulting from a sequential application of the multi-model HR-criterion is under certain regularity assumptions asymptotically T-optimal, as shown by Atkinson and Fedorov [20, Sec. 3]. This is in fact the only convergence result for multi-model design criteria known to us.

Applying the same approach to the BF-criterion from Def. 4.12 or its misspecification-robust counterpart from Def. 4.15 yields the following design criteria.

Definition 4.17 (Multi-Model BF-Criterion)

Consider the same setting as in Def. 4.16 and let B be the (two-model) BF-criterion from Def. 4.12. The BF-CRITERION for discriminating among *all* models from $\mathcal{M}$ is the function defined as $B(x; \hat{\mu}, \hat{\nu})$ for all $x \in \mathcal{X}$.

Definition 4.18 (Robust Multi-Model BF-Criterion)

Consider the same setting as in Def. 4.16 and let B' be the robust (two-model) BF-criterion from Def. 4.15. The robust BF-CRITERION for discriminating among *all* models from $\mathcal{M}$ is the function defined as $B'(x; \hat{\mu}, \hat{\nu})$ for all $x \in \mathcal{X}$.

These multi-model design criteria are theoretically well justified, computationally tractable, and simple to implement. We shall hence use it for our computations in Chap. 9.

5. Bayesian Strategies of Optimal Experimental Design for Model Discrimination

In designing an experiment, decisions must be made before data collection, and data collection is restricted by limited resources. Because specific information is usually available prior to experimentation and, indeed, often motivates the experiment, Bayesian methods can play an important role.

Chaloner and Verdinelli [66, 1. Introduction]

Contents

Having discussed frequentist strategies for optimal experimental design (OED) for model discrimination (MD) in the last chapter, the focus is now set on corresponding Bayesian strategies.

After some preparative steps in Sec. 5.1, we consider in Sec. 5.2 the de-facto standard strategy of Box-Hill-Hunter (BHH), which is based on the information-theoretic concept of entropy. The BHH-criterion makes few assumptions and can in principle be applied to a wide range MD problems. Yet even under the comfortable assumptions of known observation covariances and normal models, it still leads to problems without that have no closed-form solution and may be numerically difficult to solve.

The classic remedy is to switch to the closed-form approximation of BHH-criterion described in Sec. 5.3. Albeit popular among practitioners, it has some serious drawbacks that might significantly reduce its practical efficiency. The remaining sections are dedicated to the derivation of new design criteria for MD under normality assumptions. Section 5.4 presents information-theoretic inequalities discovered in the only recent years. Based thereon, new design criteria are derived and discussed in Sec. 5.5.

Even if they are derived as closed-form approximations of the BHH-criterion, these new design criteria have intuitive interpretations for themselves. They take into account parameter and model uncertainty, have intrinsic support for more than two rival models, yet remain consistent with the Hunter-Reiner (HR)-criterion. Albeit similar in structure to the classic approximation of the BHH-criterion, they overcome several of its drawbacks.

5.1. Assumptions and Notation

This chapter uses the optimal experimental design (OED)-related notation introduced in Chap. 4. For clarity, it applies a simplified notation for Bayesian inference that slightly differs from that used in Sec. 2.5.

In particular, prior distributions and derived quantities are marked by the subscript 0, and the arguments d_n (data) and ξ_n (design) in posterior (predictive) distributions are omitted. Accordingly, $p_0(\theta^\mu)$ is the parameter prior of model $\mu \in \mathcal{M}$, $p(\theta^\mu)$ is the corresponding parameter posterior, and $p(y|x, \mu)$ is its posterior prediction for an observation under experimental condition $x \in \mathcal{X}$. Likewise, $p_0(\mu)$ is the model prior, $p(\mu)$ is the model posterior, and $p(y|x)$ is the posterior prediction of the model family for the experimental outcome under x.

Furthermore, $\mathcal{Q}^\mu$ is a continuous $\mathscr{Q}^\mu$-valued random variable distributed according to probability density function (PDF) $p(\theta^\mu)$, $\mathcal{M}$ is a discrete $\mathscr{M}$-valued random variable distributed with probability mass function (PMF) $p(\mu)$, and $\tilde{\mathcal{Y}}_x$ is a continuous $\mathscr{Y}$-valued random variable distributed with PDF $p(y\,|\,x)$. Then, $p(y\,|\,x, \mu)$ is the PDF of $\tilde{\mathcal{Y}}_x \,\big|\, \mathcal{M} = \mu$ and $p(y\,|\,x, \mu, \theta^\mu)$ is the PDF of $\tilde{\mathcal{Y}}_x \,\big|\, \mathcal{M} = \mu, \mathcal{Q}^\mu = \theta^\mu$.

The whole chapter makes the same assumptions as the previous one, summarized in scenario 4.1. In addition, it is assumed that a model prior $p_0(\mu)$ and parameter priors $p_0(\theta^\mu)$ for all models $\mu \in \mathscr{M}$ are given.

5.2. The Box-Hill-Hunter (BHH) Strategy

Box and Hill [42] propose a sequential strategy for efficiently solving model discrimination (MD) problems between several regression models with an arbitrary distribution, and apply it to univariate normal models, assuming a homoscedastic process with known observation variance. Hill and Hunter [117, 118] generalize the strategy to the multivariate situation and to unknown observation covariances, respectively. Box [46] extends the multivariate strategy to heteroscedastic models. We refer to this body of work as the Box-Hill-Hunter (BHH)-strategy.[1]

The BHH-strategy is a sequential data-adaptive optimal experimental design (OED) strategy for reducing the experimental effort required to solve MD problems in families of two or more multivariate and possibly non-normal regression models. Its features a design criterion that is takes into account experimental uncertainty, parameter uncertainty and model uncertainty.

The BHH-strategy is an early and deeply influential approach for OED for MD. It gave rise to a huge body of follow-up works which both advanced the underlying theory and applied the strategy in practice. At May 19, 2015, Web of Science (http://www.webofscience.com) lists more than 260 citations of the work of Box and Hill [42], with 38 publications being from the year 2010 or later. Overviews of the related literature are given by Hill [116], Burke [59, Sec. "Model Discrimination Methods"] and Franceschini and Macchietto [103, Sec. 3.2].

5.2.1. Setting and Procedure

For efficiently solving the MD problem from (Q4.1) in this setting, the BHH-strategy follows the data-adaptive sequential procedure described by Alg. 5.1 on

[1] It seems that the contributions of Hill and Hunter [117, 118] are not well known, so that most publications omit Hunter's name and simply speak of the "Box-Hill-strategy."

Algorithm 5.1: Sequential data-adaptive design procedure of the BHH-strategy.

```
input  : model family $(p(y\,|\,x,\mu,\theta^\mu) : \mu \in \mathcal{M}, \theta^\mu \in \mathscr{Q}^\mu)$
         parameter priors $p_0(\theta^1),\dots,p_0(\theta^{n_\mathcal{M}})$, model prior $p_0(\mu)$
         previous experiments: data $d_s$, design $\xi_s$, with $s \in \mathbb{N}$
output : model posterior $p(\mu)$
for $n = s$ to $\infty$ do
    foreach $\mu \in \mathcal{M}$ do
        $p(\theta^\mu) \leftarrow c^\mu_n p_0(\theta^\mu) p(d_n\,|\,\xi_n,\mu,\theta^\mu)$;  // update parameter posterior
    end
    $p(\mu) \leftarrow c_n p_0(\mu) p(d_n\,|\,\xi_n,\mu)$;                  // update model posterior
    if dostop$(p(\theta^1),\dots,p(\theta^{n_\mathcal{M}}),p(\mu))$ then // check stopping criterion
        return $p(\mu)$;
    end
    $x_{n+1} \leftarrow \operatorname{argmax}_{x\in\mathcal{X}} \Lambda(x;p(\theta^1),\dots,p(\theta^{n_\mathcal{M}}),p(\mu))$;     // design exp.
    $\xi_{n+1} \leftarrow \frac{n}{n+1}\xi_n + \frac{1}{n+1}\xi^{x_{n+1}}$;                       // update design
    $y_{n+1} \leftarrow$ realization of $\mathcal{Y}_{x_{n+1}}$;                  // perform experiment
    $d^\top_{n+1} \leftarrow [d^\top_n \quad y^\top_{n+1}]$;                  // record observation
end
```

the next page. It consists of the following steps.

Using the available model and parameter priors and all n available previous experiments, the procedure applies Bayes' theorem (2.37) and (2.40) (with suitable normalizing constants c^μ and c) to determine parameter posteriors $p(\theta^\mu)$ for all models $\mu \in \mathcal{M}$ and a model posterior $p(\mu)$, respectively.

Then, it checks the stopping criterion. If the boolean function "dostop$(\cdot)$" returns *true*, the problem is considered as solved sufficiently well and the procedure stops, returning the current model posterior $p(\mu)$. For our considerations, which focus on the efficiency of the design criterion, that particular formulation of the stopping criterion is irrelevant.

Otherwise, the procedure continues to gather more data. The conditions x_{n+1} for the next experiment are determined by solving a sequential OED problem with design criterion Λ, which takes into account the current parameter and model posteriors. The conditions of all available experiments are then described by the design ξ_{n+1}, which is determined from the corresponding design ξ_n from the previous iteration and the design $\xi^{x_{n+1}}$ which puts full weight at x_{n+1}. After performing an experiment under this experimental condition and recording the resulting data y_{n+1} it continues with updating the parameter and model posteriors, hoping that whole data available now suffices to meet the stopping criterion.

The behavior of this procedure is strongly dependent by the properties of the design criterion Λ, which we discuss in the next section. We come back to the overall behavior of Alg. 5.1 in Sec. 5.2.3.

5.2.2. The BHH-Criterion

A closer look into the publications of Box, Hill and Hunter reveals that they actually propose three design criteria: (a) the general one Λ used in Alg. 5.1, (b) an upper-bound approximation of the latter, and (c) a special case of the upper-bound approximation for normal and locally affine-linearizable models. This and the next section deal with the general design criterion Λ, the other two are considered in Sec. 5.3.

Definition 5.1 (BHH-Criterion [42, 118])

The BHH-CRITERION is a function $\Lambda : \mathcal{X} \mapsto \mathbb{R}$, defined for all $x \in \mathcal{X}$ as

$$\Lambda(x) := \sum_{\mu \in \mathcal{M}} p(\mu) \int_{\mathcal{Y}} p(y \mid x, \mu) \ln \frac{p(y \mid x, \mu)}{p(y \mid x)} \,\mathrm{d}y, \tag{5.1}$$

supposed the integrals exists.

The BHH-criterion depends on the parameter posteriors of all models and on the model posterior. We explicitly listed them as arguments in Alg. 5.1 to clarify dependencies. We use the more compact notation from (5.1) in the following.

Derivation

The POSTERIOR MODEL ENTROPY

$$\mathbb{H}[\mathcal{M}] \overset{(C.1)}{=} -\sum_{\mu \in \mathcal{M}} p(\mu) \ln p(\mu) \tag{5.2}$$

is a scalar non-negative measure for the amount of the uncertainty about the unknown correct model $\bar{\mu}$ that remains after n experiments. Details about the entropy are summarized in Appendix C.

The larger $\mathbb{H}[\mathcal{M}]$, the larger the model uncertainty. It attains its maximal value of $\ln n_{\mathscr{M}}$ if and only if $\mathcal{M}$ if uniformly distributed, so that it assigns a probability of $1/n_{\mathscr{M}}$ to each model and thus represents maximal model uncertainty. It achieves its minimal value of 0 if and only if $\mathcal{M}$ is subject to a degenerate distribution which assigns a probability of 1 to a single model $\nu \in \mathscr{M}$ and thus represents minimal model uncertainty.

Due to its consistency (2.52), the model posterior follows asymptotically a degenerate distribution assigning full weight to the sought-after correct model. Therefore, the posterior model entropy attains is minimal value of 0 if the MD problem is solved. The aim of solving an MD problem can thus be formalized as reducing the posterior model entropy. From this perspective, a sequential design criterion can be derived as follows.

Suppose one additional experiment shall be designed. Let $x \in \mathscr{X}$ be the condition under which it is performed, let $y \in \mathscr{Y}$ be its outcome, and let $p(\mu \mid y, x)$ denote the resulting model posterior. The posterior model entropy based on the previous experiments and the additional experiment is then

$$\mathbb{H}[\mathcal{M} \mid \mathcal{Y}_x = y] \overset{(C.1)}{=} -\sum_{\mu \in \mathscr{M}} p(\mu \mid y, x) \ln p(\mu \mid y, x). \tag{5.3}$$

Therefore, taking into account the additional experiment reduces the posterior model entropy by

$$\mathbb{H}[\mathcal{M}] - \mathbb{H}[\mathcal{M} \mid \mathcal{Y}_x = y]. \tag{5.4}$$

The larger this difference, the better is this particular experiment for solving the MD problem, which suggests to perform an experiment maximizing it.

Unfortunately, the observation y is unknown while designing the experiment, before it is performed. As remedy one can consider the expected value approximation

$$\begin{aligned} \mathbb{H}[\mathcal{M} \mid \mathcal{Y}_x = y] &\approx \int_{\mathscr{Y}} q(y \mid x)\, \mathbb{H}[\mathcal{M} \mid \mathcal{Y}_x = y]\, \mathrm{d}y \\ &\overset{(C.2)}{=} \mathbb{H}[\mathcal{M} \mid \mathcal{Y}_x], \end{aligned} \tag{5.5}$$

whose right-hand side does not depend on unknown observations. It does, however, involve their distribution $q(y \mid x)$, which is also unknown in practice.

Replacing it through the posterior prediction $p(y\,|\,x)$ of the model family,

$$\begin{aligned}\mathbb{H}[\mathcal{M}\,|\,\mathcal{Y}_x] &= \int_{\mathcal{Y}} q(y\,|\,x)\,\mathbb{H}[\mathcal{M}\,|\,\mathcal{Y}_x = y]\,\mathrm{d}y \\ &\overset{(2.43)}{\approx} \int_{\mathcal{Y}} p(y\,|\,x)\,\mathbb{H}[\mathcal{M}\,|\,\mathcal{Y}_x = y]\,\mathrm{d}y \overset{(\mathrm{C}.2)}{=} \mathbb{H}\big[\mathcal{M}\,|\,\tilde{\mathcal{Y}}_x\big]\end{aligned} \tag{5.6}$$

removes this dependency and leads to the approximation

$$\mathbb{H}[\mathcal{M}] - \mathbb{H}\big[\mathcal{M}\,|\,\tilde{\mathcal{Y}}_x\big] \overset{\text{Prop. C.7}}{=} \mathbb{D}\big[\tilde{\mathcal{Y}}_x\,|\,\mathcal{M}\big\|\tilde{\mathcal{Y}}_x\big] \tag{5.7}$$

for reduction of the posterior model entropy (5.4) due to an additional experiment under x. This expression involves only known quantities and can thus be in principle evaluated in practice. Explicitly writing out the Kullback-Leibler distance (KLD) leads to the BHH-criterion (5.1).

Interpretation

The integral in the BHH-criterion $\Lambda(x)$ is the KLD from $p(y\,|\,x,\mu)$ to $p(y\,|\,x)$. It measures the average discrepancy between the prediction of model μ for an outcome of an experiment under x to the corresponding prediction of the whole model family. Details about the KLD are summarized in Appendix C. The BHH-criterion is an average of these discrepancies over all models, weighted with the respective posterior model probabilities. That is, the larger $\Lambda(x)$, the larger the average discrepancy between the predictions of the individual models for an experiment under x to their average prediction.

Recall from (2.38) and (2.42) that $p(y\,|\,x,\mu)$ is a parameter-robust prediction of model $\mu \in \mathscr{M}$ for an observation under x, and that $p(y\,|\,x)$ is the corresponding model-robust and parameter-robust prediction of the whole model family, respectively. The BHH-criterion takes into account the parameter uncertainty via $p(y\,|\,x,\mu)$ and $p(y\,|\,x)$, and the model uncertainty via the weighted sum in (5.1) and via $p(y\,|\,x)$. By integrating over the (approximate) distribution of the as-yet-unperformed observation, it also takes into account the experimental uncertainty. In this sense it the BHH-criterion robust with respect to the parameter uncertainty, model uncertainty and experimental uncertainty.

An Information-Theoretic Point of View

Information theory provides a consistent framework for quantifying the information and uncertainties involved Bayesian probabilities, an idea going back to Lindley [176] and Stone [241]. Let $\mathbb{H}[\cdot]$, $\mathbb{D}[\cdot\|\cdot]$ and $\mathbb{I}[\cdot\|\cdot]$ denote the (conditional) entropy, the (conditional) Kullback-Leibler distance and the (conditional) mutual information from Defs. C.1, C.3 and C.5, respectively. The information-theoretic identities summarized in Prop. C.7 provide the representations

$$\Lambda(x) = \mathbb{D}\big[\tilde{\mathcal{Y}}_x \,|\, \mathcal{M} \big\| \tilde{\mathcal{Y}}_x\big] = \mathbb{H}\big[\tilde{\mathcal{Y}}_x\big] - \mathbb{H}\big[\tilde{\mathcal{Y}}_x \,|\, \mathcal{M}\big] = \mathbb{I}\big[\tilde{\mathcal{Y}}_x \big\| \mathcal{M}\big] \tag{5.8}$$

for the BHH-criterion (Def. 5.1). Since the entropy measures the uncertainty in a random variable, and the mutual information the information that one random variable carries about another (and vice versa), the last equality in (5.8) can be interpreted as follows:

> As a matter of fact, the amount of information which we get when we observe the result of an experiment (depending on chance) can be taken numerically equal to the amount of uncertainty concerning the outcome of the experiment before carrying it out. (Rényi [212, Sec. 3])

Information-theoretic concepts have been frequently applied to assess, compare and design experiments, important contributions coming from Blackwell [32, 33], Lindley [176], and Stone [241] and Rényi [212]. Many of these ideas have entered the field of Bayesian experimental design, see the excellent reviews of Chaloner and Verdinelli [66] and von Toussaint [259]. The BHH-criterion is a particular popular representative of this class of design strategies.

Behavior for Vanishing Uncertainties

If there is no model uncertainty in the sense that $p(\mu) = 1$ for some model $\mu \in \mathscr{M}$, then $p(y\,|\,x) = p(y\,|\,x,\mu)$ and thus

$$\Lambda(x) = \int_{\mathscr{Y}} p(y\,|\,x,\mu) \ln \frac{p(y\,|\,x,\mu)}{p(y\,|\,x,\mu)} \,\mathrm{d}y \equiv 0. \tag{5.9}$$

That is, if the MD problem is solved, the BHH-criterion correctly predicts that no experiment can further reduce the model uncertainty.

If there is no parameter uncertainty in the sense that all models $\mu \in \mathcal{M}$ have a degenerate parameter posterior putting full mass at some parameter θ^{μ}, then $p(y|x,\mu) = p(y|x,\mu,\theta^{\mu})$, so that

$$\Lambda(x) = \sum_{\mu \in \mathcal{M}} p(\mu) p(y|x,\mu,\theta^{\mu}) \ln \frac{p(y|x,\mu,\theta^{\mu})}{p(y|x)}, \tag{5.10}$$

where $p(y|x) = \sum_{\mu \in \mathcal{M}} p(y|x,\mu,\theta^{\mu})$.

5.2.3. Behavior of the BHH-Procedure

We can now make some general observations about the overall behavior of the BHH-procedure shown in Alg. 5.1.

Under suitable regularity conditions, the posterior probability $p(\bar{\mu})$ of the sought-after correct model $\bar{\mu}$ approaches 1 as n increases, see (2.52). Supposed these regularity conditions are met in its course, the BHH-procedure will thus identify $\bar{\mu}$ in the large-sample limit and thus solve the MD problem. By choosing the experiments which maximize the BHH-criterion, the procedure aims to reduce the uncertainty about $\bar{\mu}$ quickly, and thus improve the rate with which $p(\bar{\mu})$ converges to 1.

The reduction of posterior model entropy (5.4) measures the *actual* utility of an observation $y \in \mathcal{Y}$ obtained from an experiment under $x \in \mathcal{X}$ for reducing the model uncertainty. The BHH-criterion is a *predictor* for this entropy reduction, which is based on the model family and the previous knowledge,

$$\Lambda(x) \approx \mathbb{H}[\mathcal{M}] - \mathbb{H}[\mathcal{M}|\mathcal{Y}_x = y]. \tag{5.11}$$

The reliability of this prediction depends, among others, on the quality of the underlying approximation

$$p(y|x) \overset{(2.42)}{=} \sum_{\mu \in \mathcal{M}} p(\mu) \int_{\mathcal{Q}^{\mu}} p(\theta^{\mu}) p(y|x,\mu,\theta^{\mu})\, \mathrm{d}\theta^{\mu} \overset{(2.43)}{\approx} q(y|x), \tag{5.12}$$

which enters via (5.6). Due to the consistency of parameter and model posteriors, it tends to get better with the amount of available data, and is under regularity conditions exact in the large sample limit according to (2.60).

In early stages of experimentation, when model and parameter posteriors are vague, (5.12) and thus (5.11) cannot be expected to be particularly good. Accordingly, BHH-optimal experiments will thus not reliably lead to a large reduction the posterior model entropy. As long as the conditions for posterior consistency (2.44) and (2.52) are met, however, additional experiments tend to sharpen the parameter and model posteriors, albeit possibly slowly.

In later stages of experimentation, the quality of approximations (5.12) and (5.11) will thus increase so that BHH-optimal experiments will more reliably lead to a significant reduction of the posterior model entropy, which closes a positive feedback loop for reducing the model uncertainty. Algorithm 5.1 is thus in a sense "self-enhancing" with respect to the model uncertainty.

Such a self-enhancing behavior can be expected, however, only with respect to the model uncertainty, since there BHH-optimal experiments do not necessarily reduce the parameter uncertainty particularly well. In contrary, it is generally believed that optimal designs for MD are usually particularly *inefficient* for parameter estimation (PE), as noted, for example, by Atkinson, Bogacka, and Bogacki [18] or Atkinson [13].

5.2.4. No Closed-Form Representation under Normality

Identity (5.8) can also be written as

$$\Lambda(x) = \sum_{\mu\in\mathcal{M}} p(\mu)\, \mathbb{D}\big[\tilde{\mathcal{Y}}_x \,\big|\, \mathcal{M} = \mu \big\| \tilde{\mathcal{Y}}_x\big] \tag{5.13}$$

$$= \mathbb{H}\big[\tilde{\mathcal{Y}}_x\big] - \sum_{\mu\in\mathcal{M}} p(\mu)\, \mathbb{H}\big[\tilde{\mathcal{Y}}_x \,\big|\, \mathcal{M} = \mu\big]. \tag{5.14}$$

Under certain assumptions on process, model and data, $\tilde{\mathcal{Y}}_x \,\big|\, \mathcal{M} = \mu$ is approximately normally distributed. Such approximations are discussed in Secs. 2.5 and 3.5. If such a normal approximation applies to all models $\mu \in \mathcal{M}$, then $\tilde{\mathcal{Y}}_x$ is approximately subject to a normal mixture distribution, that is, a distribution described by a convex combination of normal probability density functions (PDFs). Then, $\mathbb{D}\big[\tilde{\mathcal{Y}}_x \,\big|\, \mathcal{M} = \mu \big\| \tilde{\mathcal{Y}}_x\big]$ is the KLD from a normal distribution to a normal mixture distribution, $\mathbb{H}\big[\tilde{\mathcal{Y}}_x\big]$ is the entropy of a normal mixture distribution, and $\mathbb{H}\big[\tilde{\mathcal{Y}}_x \,\big|\, \mathcal{M} = \mu\big]$ is the entropy of a normal distribution.

The latter has a closed-form representation as stated in Prop. C.8. Unfortunately, there seems to be no closed-form solution for the other expressions. Therefore, *the BHH-criterion has no closed-form representation even when the predictions of all models have the comfortable property of being normally distributed.* For evaluating it one needs to resort to approximations. Numerical approximations for the integrals involved in the KLD and the entropy are possible, but suffer from the curse of dimensionality. Numerical methods maximizing a design criterion typically evaluate it many times and thus multiply the computational effort required for such numerical integrations. The resulting problems might thus quickly get computationally intractable as the problem dimensions increase. Closed-form approximations are hence desirable.

The remaining chapter considers sequential design criteria for MD that can be interpreted as closed-form approximations of the BHH-criterion under normality assumptions.

5.3. The Classic Approximation of the BHH-Criterion

This section considers a classic approximation of the Box-Hill-Hunter (BHH)-criterion proposed by Box and Hill [42] and Hill and Hunter [118] themselves. A brief discussion of its underlying assumptions the resulting limitations motivates novel approximations that we propose in Sec. 5.5. The design criterion is based on the following "classic" Bayesian approximations.

5.3.1. Classic Empirical Bayesian Formulas

Consider scenario 4.1 under the additional normality assumptions (ix) to (xi) on p. 128. As discussed in Sec. 3.5, these assumptions justify for each model $\mu \in \mathscr{M}$ the approximation

$$p(\theta^\mu) \overset{(3.79)}{\approx} \phi\Big(\theta^\mu \,\Big|\, \hat{\theta}^\mu, \tfrac{1}{n}\hat{\boldsymbol{M}}^{\mu^{-1}}\Big) \tag{5.15}$$

for the parameter posterior and the approximation

$$p(y\,|\,x,\mu) \overset{(3.82)}{\approx} \phi\big(y\,\big|\,\hat{\eta}^\mu(x), \hat{\boldsymbol{T}}^\mu(x)\big) \tag{5.16}$$

for the posterior prediction of model μ for an observation under $x \in \mathscr{X}$, where

$$\hat{\boldsymbol{T}}^\mu(x) := \boldsymbol{\Omega}(x) + \tfrac{1}{n}\hat{\boldsymbol{J}}^\mu(x)\hat{\boldsymbol{M}}^{\mu^{-1}}\hat{\boldsymbol{J}}^{\mu^\top}(x), \tag{5.17}$$

see Tab. 3.1 on p. 113. Furthermore, they lead to the approximation

$$p(\mu) \overset{(3.85)}{\approx} \pi^\mu := c_n p_0(\mu) \exp\left(-\tfrac{n}{2}\hat{s}_n^\mu\right) n^{-n_{\theta^\mu}/2} \tag{5.18}$$

for the model posterior. We summarize the posterior model probabilities in the tuple $\pi^\top := \begin{bmatrix}\pi^1 & \dots & \pi^{n_\mathcal{M}}\end{bmatrix}$. Combining (5.16) and (5.18) yields the approximation

$$p(y\,|\,x) \approx c_n \sum_{\mu\in\mathcal{M}} \pi^\mu \phi\left(y\,\middle|\,\hat{\eta}^\mu(x), \hat{\boldsymbol{T}}^\mu(x)\right) \tag{5.19}$$

for the posterior prediction of the model family for an observation under x.

5.3.2. Classic Upper Bound of the BHH-Criterion

Theorem 5.2 (Classic Upper Bound of the BHH-Criterion)
Assume that the matrix $\hat{\boldsymbol{T}}^\mu(x)$ exists and is invertible for all models $\mu \in \mathcal{M}$ under all experimental conditions $x \in \mathcal{X}$, and that the approximations (5.16) and (5.18) are *exact*. For all $x \in \mathcal{X}$, define

$$U(x) := \tfrac{1}{2}\pi^\top \boldsymbol{U}(x)\pi, \tag{5.20a}$$

where $\boldsymbol{U}(x)$ is a $n_\mathcal{M} \times n_\mathcal{M}$ matrix which has for all $\mu, \nu \in \mathcal{M}$ the element

$$\|\hat{\eta}^\mu(x) - \hat{\eta}^\nu(x)\|^2_{\hat{\boldsymbol{T}}^{\nu^{-1}}(x)} + \operatorname{tr}\left(\hat{\boldsymbol{T}}^\mu(x)\hat{\boldsymbol{T}}^{\nu^{-1}}(x)\right) - n_y \tag{5.20b}$$

in row μ and column ν. Then, $U(x) \geqslant \Lambda(x)$ under all $x \in \mathcal{X}$.

Proof The following proof summarizes the arguments given by Box and Hill [42] for the univariate case $n_y = 1$ and by Hill and Hunter [118] for the multivariate case $n_y > 1$. For clarity, it uses the notation based on random variables described in Sec. 5.1. The convexity of the Kullback-Leibler distance (KLD) (Prop. C.4,property (vii)) implies that

$$\int_\mathcal{Y} p(y\,|\,x,\mu) \ln \frac{p(y\,|\,x,\mu)}{\sum_{\nu\in\mathcal{M}} p(\nu)p(y\,|\,x,\nu)}\,\mathrm{d}y \leqslant \sum_{\nu\in\mathcal{M}} p(\nu) \int_\mathcal{Y} p(y\,|\,x,\mu) \ln \frac{p(y\,|\,x,\mu)}{p(y\,|\,x,\nu)}\,\mathrm{d}y, \tag{5.21}$$

which is equivalent to

$$\mathbb{D}\big[\tilde{\mathcal{Y}}_x \,\big|\, \mathcal{M} = \mu \big\| \tilde{\mathcal{Y}}_x\big] \leqslant \sum_{\nu\in\mathscr{M}} p(\nu)\,\mathbb{D}\big[\tilde{\mathcal{Y}}_x \,\big|\, \mathcal{M} = \mu \big\| \tilde{\mathcal{Y}}_x \,\big|\, \mathcal{M} = \nu\big]. \tag{5.22}$$

After multiplication with $p(\mu)$ and summation over all $\mu \in \mathscr{M}$ the left-hand side equals the KLD-based representation of the BHH-criterion (5.13), so that

$$\Lambda(x) \leqslant \sum_{\mu,\nu\in\mathscr{M}} p(\mu)p(\nu)\,\mathbb{D}\big[\tilde{\mathcal{Y}}_x \,\big|\, \mathcal{M} = \mu \big\| \tilde{\mathcal{Y}}_x \,\big|\, \mathcal{M} = \nu\big]. \tag{5.23}$$

The normal approximation (5.16) can be expressed in terms of the random variables as

$$\tilde{\mathcal{Y}}_x \,\big|\, \mathcal{M} = \mu \sim \phi\big(y \,\big|\, \hat{\eta}^{\mu}(x), \hat{\boldsymbol{T}}^{\mu}(x)\big). \tag{5.24}$$

Since it is assumed to be exact, the KLDs in the right-hand side of (5.23) have according to Thm. C.10 the closed-form representations

$$\begin{aligned}\mathbb{D}\big[\tilde{\mathcal{Y}}_x \,\big|\, \mathcal{M} = \mu \big\| \tilde{\mathcal{Y}}_x \,\big|\, \mathcal{M} = \nu\big] = \tfrac{1}{2}\Big(&\big\|\hat{\eta}^{\mu}(x) - \hat{\eta}^{\nu}(x)\big\|^2_{\hat{\boldsymbol{T}}^{\nu-1}(x)} \\ &+ \operatorname{tr}\Big(\hat{\boldsymbol{T}}^{\mu}(x)\hat{\boldsymbol{T}}^{\nu^{-1}}(x)\Big) - \ln\det\Big(\hat{\boldsymbol{T}}^{\mu}(x)\hat{\boldsymbol{T}}^{\nu^{-1}}(x)\Big) - n_y\Big).\end{aligned} \tag{5.25}$$

The third summand vanishes when summed over all pairs of models,

$$\begin{aligned}&\sum_{\mu,\nu\in\mathscr{M}} p(\mu)p(\nu)\ln\det\Big(\hat{\boldsymbol{T}}^{\mu}(x)\hat{\boldsymbol{T}}^{\nu^{-1}}(x)\Big) \\ &\qquad = \sum_{\mu\in\mathscr{M}} p(\mu)\ln\det\hat{\boldsymbol{T}}^{\mu}(x) - \sum_{\nu\in\mathscr{M}} p(\nu)\ln\det\hat{\boldsymbol{T}}^{\nu}(x) \equiv 0.\end{aligned} \tag{5.26}$$

Since approximation (5.18) for the model posterior is assumed to exact, $p(\mu) = \pi^{\mu}$. The claimed inequality follows from substituting the latter equality and (5.25) into (5.23) and using (5.26). □

Box, Hill and Hunter propose to use this upper bound under the considered assumptions as *approximation* of the BHH-criterion,

$$U(x) \approx \Lambda(x) \text{ for all } x \in \mathscr{X}. \tag{5.27}$$

We refer to U as the CLASSIC UPPER-BOUND APPROXIMATION of the BHH-criterion. In practice, the Bayesian approximations (5.16) and (5.18) are typically not exact, as assumed in the theorem, so that U is to be considered as a heuristic design criterion. It can be interpreted as follows.

The first summand in (5.20) measures the difference between the predictions of models μ and ν for the observation mean under x, relative to the total uncertainty of this prediction under model ν. Under an experimental condition maximizing it, the systematic discrepancy between the model predictions is maximal with respect to the total uncertainty under model ν.

The second summand measures total uncertainty about an experimental outcome under x of model μ relative to that of model ν. Under an experimental condition maximizing it, the uncertainties of both models about the outcome of an observation are maximally different. Such an experimental condition, under which one prediction is more reliable than the other, is more helpful for recognizing a difference between these models than an experimental condition under which all model predictions are equally uncertain.

Since the design criterion comprises a sum over all model pairings, the asymmetries in the discussed terms with respect to the models cancel out. Maximizing this design criterion thus provides an experimental condition under which both the responses as well as their uncertainty are maximally different among the models.

5.3.3. Behavior under Small Uncertainties

Proposition 5.3 (Classic Upper-Bound Approximation under Small Uncertainties)

For all $x \in \mathcal{X}$ and all $\mu, \nu \in \mathcal{M}$, let $H(x;\mu,\nu) := \|\hat{\eta}^{\mu}(x) - \hat{\eta}^{\nu}(x)\|^2_{\boldsymbol{\Omega}^{-1}(x)}$ be the Hunter-Reiner (HR)-criterion from Def. 4.11 for discrimination between models μ and ν, and let $\boldsymbol{H}(x)$ the $n_{\mathcal{M}} \times n_{\mathcal{M}}$ matrix with element $H(x;\mu,\nu)$ in row μ and column ν.

If the parameter-induced uncertainties in the responses vanish in the sense that

$$\tfrac{1}{n}\hat{\boldsymbol{J}}^{\mu}(x)\hat{\boldsymbol{M}}^{\mu^{-1}}\hat{\boldsymbol{J}}^{\mu^{\top}}(x) = \boldsymbol{0} \tag{5.28}$$

for all models $\mu \in \mathcal{M}$ under all experimental conditions $x \in \mathcal{X}$, then

$$U(x) = \tfrac{1}{2}\pi^{\top}\boldsymbol{H}(x)\pi. \tag{5.29}$$

If, in addition, the model posterior focuses only on two different models $\mu \neq \nu$

such that $\pi^{\mu} + \pi^{\nu} = 1$, then

$$U(x) = H(x; \mu, \nu). \tag{5.30}$$

Proof Condition 5.28 implies $\hat{\boldsymbol{T}}^{\mu}(x) = \boldsymbol{\Omega}(x)$ for all $x \in \mathcal{X}$, so that

$$(5.20b) = \left\| \hat{\eta}^{\mu}(x) - \hat{\eta}^{\nu}(x) \right\|^2_{\boldsymbol{\Omega}^{-1}(x)} + \operatorname{tr}\left(\boldsymbol{\Omega}(x)\boldsymbol{\Omega}^{-1}(x)\right) - n_y \tag{5.31}$$

$$= \left\| \hat{\eta}^{\mu}(x) - \hat{\eta}^{\nu}(x) \right\|^2_{\boldsymbol{\Omega}^{-1}(x)} + n_y \qquad - n_y \tag{5.32}$$

$$= \left\| \hat{\eta}^{\mu}(x) - \hat{\eta}^{\nu}(x) \right\|^2_{\boldsymbol{\Omega}^{-1}(x)}, \tag{5.33}$$

which leads to (5.29). Since $\sum_{\mu \in \mathcal{M}} \pi^{\mu} = 1$ by definition, the condition $\pi^{\mu} + \pi^{\nu} = 1$ implies that $\pi^{\lambda} = 0$ for all $\lambda \in \mathcal{M} \setminus \{\mu, \nu\}$. Omitting the corresponding summands in (5.29) and substituting the expression for the HR-criterion from (4.35) into the result yields (5.30).□

The matrix $\frac{1}{n}\hat{\boldsymbol{J}}^{\mu}(x)\hat{\boldsymbol{M}}^{\mu^{-1}}\hat{\boldsymbol{J}}^{\mu^{\top}}(x)$ quantifies, in affine-linear approximation, the uncertainty in the response of model μ under x due to parameter uncertainty after n experiments in terms of $\frac{1}{n}\hat{\boldsymbol{M}}^{\mu^{-1}}$. The vector π of the model posteriors quantifies the corresponding model uncertainty. The HR-criterion H can therefore be considered as a special case of U for vanishing parameter uncertainty and almost vanishing model uncertainty. Reversely, U can be viewed as a multi-model, model-robust and parameter-robust generalization of the HR-criterion.

Based on Prop. 5.3 and some continuity arguments one can derive the following approximate results. If the parameter-induced uncertainty in the model responses is significantly smaller than the experimental uncertainty in the sense that

$$\left\| \tfrac{1}{n}\hat{\boldsymbol{J}}^{\mu}(x)\hat{\boldsymbol{M}}^{\mu^{-1}}\hat{\boldsymbol{J}}^{\mu^{\top}}(x) \right\| \ll \left\| \boldsymbol{\Omega}(x) \right\| \text{ for all } x \in \mathcal{X}, \tag{5.34}$$

with some matrix norm $\|\cdot\|$, then $U(x)$ has approximately the representation given in (5.29). If, in addition, there is a pair of different models $\mu \neq \nu$ with dominant posterior probabilities in the sense that

$$\pi^{\mu} \gg \pi^{\lambda} \text{ and } \pi^{\nu} \gg \pi^{\lambda}, \text{ for all } \lambda \in \mathcal{M} \setminus \{\mu, \nu\}, \tag{5.35}$$

which implies that $\pi^{\mu}\pi^{\nu} \gg \pi^{i}\pi^{j}$ for all $(i, j) \neq (\mu, \nu)$, then it has approximately the representation given in (5.30).

Under certain regularity conditions discussed in Secs. 3.4.1 and 3.4.3, the inverse of matrix $n\hat{\boldsymbol{M}}^{\mu}$ converges to zero as the sample size n tends to infinity, so that $\frac{1}{n}\hat{\boldsymbol{J}}^{\mu}(x)\hat{\boldsymbol{M}}^{\mu^{-1}}\hat{\boldsymbol{J}}^{\mu^{\top}}(x)$ vanishes asymptotically under all $x \in \mathcal{X}$. Then, (5.34) is satisfied in large samples. If the family contains a unique model that has a non-vanishing prior and is "second-best" in terms of the Kullback-Leibler information criterion (KLIC) under ξ, it it likely that the model posterior asymptotically concentrates at this second-best model and at the correct model. Inequalities (5.35) are then satisfied in large samples.

Therefore, the classic upper-bound approximation of the BHH-criterion reduces in the large-sample limit under regularity conditions to the HR-criterion. Under assumptions discussed in Sec. 4.3.4, the designs constructed by the latter converge to a T-optimal design, the best design that is theoretically possible for model discrimination (MD).

5.3.4. Discussion

The classic upper-bound approximation of the BHH-criterion has received considerable attention in theoretical work and has been applied to various problems in academia and industry. In fact, most literature dealing with the BHH-strategy actually uses this design criterion. Further references can be found in the overviews given by Hill [116], Burke [59, Sec. "Model Discrimination Methods"] and Franceschini and Macchietto [103, Sec. 3.2].

The original upper-bound approximation proposed by Box and Hill [42] and Hill and Hunter [118] used an slightly different expression than (5.18) for the model posterior. Their formula was criticized, for example by Atkinson and Cox [19, Sec. 6] and Atkinson [16], for being valid only if all rival models have the same number of parameters. Improved formulas not sharing this drawback were quickly developed, as discussed in Sec. 3.5.2. The formula (5.18) used here is one of these improved formulas. Besides this solved objection, at least two additional points of critique may be formulated.

First, it seems strange to *maximize* an *upper* bound, as already noted by Meeter, Pirie, and Blot [185, Sec. 1] and Fedorov and Malyutov [98, Sec. 6]. Such an approach leads to overly optimistic predictions for the *actual* reduction of the model uncertainty resulting from an additional experiment. As pointed out by Fedorov [92], a maximizer of the upper-bound approximation $U(x)$ might not even be close to a maximizer of the actual BHH-criterion $\Lambda(x)$ which it approximates.

Second, the posterior approximations (5.15) and (5.16) are likely to be inadequate for models that *both* nonlinear *and* incorrect, as discussed in Secs. 2.5 and 3.5. By assumption, however, all except one of the rival models are incorrect. The alternative improved approximations discussed in Sec. 3.5 have so far not been used in design criteria for MD.

In the remaining sections we shall derive new closed-form approximations of the BHH-criterion that aim to overcome all these drawbacks.

5.4. Lower Bounds for Entropy and KLD of Normal Mixtures

In our convention, design criteria for model discrimination (MD) are *maximized*. Whenever it is necessary to approximate them, it is thus reasonable to choose an approximation that *underestimates* their actual values.

Hershey and Olsen [115] discuss several possible approximations for the Kullback-Leibler distance (KLD) between Gaussian mixture models. From all given approaches, the "variational lower bound" approximation seems most suitable. It can be summarized as follows.

Theorem 5.4 (Lower Bound for the KLD between Normal Mixtures [115]) Let $\mathcal{I}$ and $\mathcal{K}$ be finite sets, let $\pi_i \in [0,1]$ for all $i \in \mathcal{I}$ with $\sum_{i\in\mathcal{I}} \pi_i = 1$, and $\rho_k \in [0,1]$ for all $k \in \mathcal{J}$ with $\sum_{k\in\mathcal{K}} \rho_k = 1$. Let $n \in \mathbb{N}$. For all $i \in \mathcal{I} \cup \mathcal{K}$, let $\mu_i \in \mathbb{R}^n$ and let $\boldsymbol{C}_i$ be a real-valued symmetric positive definite (SPD) $n \times n$ matrix. Let $\mathcal{U}$ and $\mathcal{V}$ be random variables distributed according to the normal mixture probability density functions (PDFs)

$$\sum_{i\in\mathcal{I}} \pi_i \phi_n(u \,|\, \mu_i, \boldsymbol{C}_i) \text{ and } \sum_{k\in\mathcal{K}} \rho_k \phi_n(u \,|\, \mu_k, \boldsymbol{C}_k), \tag{5.36}$$

respectively. Then, the KLD $\mathbb{D}[\mathcal{U}\|\mathcal{V}]$ satisfies the inequality

$$\mathbb{D}[\mathcal{U}\|\mathcal{V}] \geqslant \sum_{i\in\mathcal{I}} \pi_i \ln \frac{\sum_{j\in\mathcal{I}} \pi_j \exp(-f_{ij})}{\sum_{k\in\mathcal{K}} \rho_k \exp(-f_{ik})}, \text{ where}$$

$$f_{ij} := \tfrac{1}{2}\Big(\left\|\mu_i - \mu_j\right\|^2_{\boldsymbol{C}_j^{-1}} + \operatorname{tr}\big(\boldsymbol{C}_i \boldsymbol{C}_j^{-1}\big) - \ln\det\big(\boldsymbol{C}_i \boldsymbol{C}_j^{-1}\big) - n\Big)$$

is the KLD from the normal distribution with mean $\mu_i \in \mathbb{R}^n$ and covariance $\boldsymbol{C}_i$ to the normal distribution with mean $\mu_j \in \mathbb{R}^n$ and covariance $\boldsymbol{C}_j$, see Thm. C.10.

Proof Given by Hershey and Olsen [115, Sec. 7]. □

We require the following special case of Thm. 5.4.

Corollary 5.5 (Lower Bound for the KLD from a Normal Distribution to a Normal Mixture Distribution)

Let $\mathcal{V}$, $\mathcal{K}$, and all ρ_n, μ_k and $\boldsymbol{C}_k$ with $k \in \mathcal{K}$ be defined as in Thm. 5.4, and let $\mathcal{U}$ be a normally distributed random variable with mean μ and SPD covariance matrix $\boldsymbol{C}$. Then,

$$\mathbb{D}[\mathcal{U}\|\mathcal{V}] \geqslant -\ln \sum_{k\in\mathcal{K}} \rho_k \exp\left(-\tfrac{1}{2} f_k\right), \text{ where}$$
$$f_k := \|\mu - \mu_k\|^2_{\boldsymbol{C}_k^{-1}} + \operatorname{tr}\left(\boldsymbol{C}\boldsymbol{C}_k^{-1}\right) - \ln\det\left(\boldsymbol{C}\boldsymbol{C}_k^{-1}\right) - n. \tag{5.37}$$

Huber et al. [125] provides different approximations for the entropy of normal mixtures. The following lower bound is particularly useful for our purposes.

Theorem 5.6 (Lower Bound for the Entropy of a Normal Mixture [125])

Let $\mathcal{I}$ be a finite index set, let $\pi_i \in [0,1]$ for all $i \in \mathcal{I}$ with $\sum_{i\in\mathcal{I}} \pi_i = 1$. Let $n \in \mathbb{N}$. For all $i \in \mathcal{I}$, let $\mu_i \in \mathbb{R}^n$ and let $\boldsymbol{C}_i$ be a SPD real-valued $n \times n$ matrix. Let $\mathcal{U}$ be a random variable distributed according to the normal mixture PDF $\sum_{i\in\mathcal{I}} \pi_i \phi_n(u \,|\, \mu_i, \boldsymbol{C}_i)$. Then, the entropy $\mathbb{H}[\mathcal{U}]$ satisfies the inequality

$$\mathbb{H}[\mathcal{U}] \geqslant -\sum_{i\in\mathcal{I}} \pi_i \ln \sum_{j\in\mathcal{I}} \pi_j \phi_n\left(\mu_i \,\middle|\, \mu_j, \boldsymbol{C}_i + \boldsymbol{C}_j\right). \tag{5.38}$$

Proof Given by Huber et al. [125, Thm. 2]. □

By substituting the expression (B.12b) for a normal PDF, (5.38) can be rewritten as

$$\mathbb{H}[\mathcal{U}] \geqslant -\sum_{i\in\mathcal{I}} \pi_i \ln \sum_{j\in\mathcal{I}} \pi_j \exp\left(-\tfrac{1}{2} f_{ij}\right), \text{ with} \tag{5.39a}$$
$$f_{ij} := \left\|\mu_i - \mu_j\right\|^2_{(\boldsymbol{C}_i+\boldsymbol{C}_j)^{-1}} + \ln\det\left(\boldsymbol{C}_i + \boldsymbol{C}_j\right) + n_y \ln(2\pi). \tag{5.39b}$$

In the next sections we shall frequently meet expressions of the form appearing in the right-hand side of (5.39a). The next lemma summarizes some of its properties.

Lemma 5.7 (Properties of the Function ρ)
Let $\pi^\top := \begin{bmatrix} \pi^1 & \dots & \pi^n \end{bmatrix}$ with $\pi_i \in [0,1]$ for all $i \in \{1,\dots,n\}$ and with $\sum_{i\in\mathcal{I}} \pi_i = 1$. Let $\boldsymbol{C}$ be a symmetric $n \times n$ matrix with elements $c_{ij} \in \mathbb{R}_0^+$. Then, the function

$$\rho(\boldsymbol{C},\pi) := -\sum_{i=1}^{n} \pi_i \ln \sum_{j=1}^{n} \pi_j \exp\bigl(-c_{ij}\bigr) \tag{5.40}$$

has the following properties for all $i, j \in \mathcal{I}$:

(i) $\rho(\boldsymbol{C},\pi) \geqslant 0$

(ii) $\rho(\boldsymbol{C},\pi)$ increases in each component c_{ij} of $\boldsymbol{C}$

(iii) $\rho(\boldsymbol{C},\pi)$ is a concave function of each component c_{ij} of $\boldsymbol{C}$

(iv) $\rho(\boldsymbol{C}+c\mathbf{1},\pi) = \rho(\boldsymbol{C},\pi) + c$, for all $c \in \mathbb{R}$

(v) $\frac{\partial}{\partial c_{ij}}\rho(\boldsymbol{C},\pi)$ increases in π_i and π_j

(vi) $\pi_k = 1$ implies $\rho(\boldsymbol{C},\pi) = c_{kk}$.

Proof We only sketch the proofs, which are based on basic algebra and simple differential calculus, but are laborious in parts.

(i) Follows from the non-negativity of the components of π and $\boldsymbol{C}$.

(ii) Obviously, $-\rho(-\boldsymbol{C},\pi)$ is a concatenation of increasing functions of all c_{ij}, which implies that $\rho(\boldsymbol{C},\pi)$ is strictly increasing, too.

(iii) The function $\ln \sum_{j=1}^n \pi_j \exp\bigl(-c_{ij}\bigr) = \ln \sum_{j=1}^n \exp\bigl(\ln \pi_j - c_{ij}\bigr)$ is convex in all c_{ij}, since the corresponding Hessian is positive semidefinite. The function (5.40) is the negative of a convex combination of the latter function, and is thus concave.

(iv) Follows from writing out $\rho(\boldsymbol{C}+c\mathbf{1},\pi)$ according to the definition and from tedious but simple algebra.

(v) Can be seen by explicitly calculating the partial derivative.

(vi) If $\pi_k = 1$, then $\pi_i = 0$ for all $i \neq k$. Then, the sums in (5.39a) both reduce to a single term and the claim is obvious. □

The function $\rho(\boldsymbol{C},\pi)$ is best understood in comparison to the quadratic form $\pi^\top \boldsymbol{C}\pi$.

Lemma 5.8

Under the assumptions of Lem. 5.7, the function $\pi^\top \boldsymbol{C}\pi$ shares properties (i), (ii) and (iv) to (vi). In contrast to $\rho(\boldsymbol{C},\pi)$, however, the function $\pi^\top \boldsymbol{C}\pi$ is *convex* in c_{ij} for all $i, j \in \mathcal{I}$. In addition, $\pi^\top \boldsymbol{C}\pi \geqslant \rho(\boldsymbol{C},\pi)$.

Proof The proofs of properties (i), (ii) and (iv) to (vi) and the convexity is trivial. The claimed inequality follows from a variant Jensen's inequality states that $g(\sum_{i\in\mathcal{I}} \pi_i c_i) \leqslant \sum_i g(c_i)$ for all convex functions $g\colon \mathbb{R} \mapsto \mathbb{R}$. Since the exponential function is such a convex function,

$$\exp\left(\sum_j -\pi_i c_{ij}\right) \leqslant \sum_j \pi_i \exp\left(-c_{ij}\right) \Leftrightarrow \sum_j \pi_i c_{ij} \geqslant -\ln \sum_j \pi_i \exp\left(-c_{ij}\right)$$

for all $j \in \mathcal{I}$. Multiplying the latter inequality with π_j and summation over j leads to the claimed inequality. □

5.5. New Sequential Design Criteria for Model Discrimination

We are now prepared to derive new design criteria for model discrimination (MD). They are closed-form approximations of the Box-Hill-Hunter (BHH)-criterion which are based on the misspecification-robust Bayesian approximations discussed in Sec. 3.5 and the information-theoretic inequalities introduced in the previous section.

Throughout this section we consider scenario 4.1 under the additional assumptions (ix) to (xi) and the assumption that for any given $n \in \mathbb{N}$ each model $\mu \in \mathcal{M}$ has the "little informative" normal parameter prior

$$p_0(\theta^\mu) := \phi\left(\theta^\mu \,\middle|\, \hat{\theta}^\mu_n, \left(\hat{\boldsymbol{M}}^\mu_n + \hat{\boldsymbol{N}}^\mu_n\right)^{-1}\right), \text{ for all } \theta^\mu \in \mathcal{Q}^\mu, \tag{5.41}$$

from (3.72) and (3.76).

5.5.1. Advanced Empirical Bayesian Formulas

As discussed in Sec. 3.5, this considered setting justifies for each model $\mu \in \mathcal{M}$ the approximation

$$p(\theta^\mu) \overset{(3.77)}{\approx} \phi\Big(\theta^\mu \,\Big|\, \hat{\theta}^\mu, \tfrac{1}{n+1}\big(\hat{\boldsymbol{M}}^\mu + \hat{\boldsymbol{N}}^\mu\big)^{-1}\Big) \tag{5.42}$$

for the parameter posterior, the approximation

$$p(y \,|\, x, \mu) \overset{(3.80)}{\approx} \phi\big(y \,|\, \hat{\eta}^\mu(x), \hat{\boldsymbol{W}}^\mu(x)\big) \tag{5.43}$$

for the posterior prediction for observations under $x \in \mathcal{X}$, where

$$\hat{\boldsymbol{W}}^\mu(x) := \boldsymbol{\Omega}(x) + \tfrac{1}{n+1}\hat{\boldsymbol{J}}^\mu(x)\big(\hat{\boldsymbol{M}}^\mu + \hat{\boldsymbol{N}}^\mu\big)^{-1}\hat{\boldsymbol{J}}^{\mu\top}(x), \tag{5.44}$$

and the approximation

$$p(\mu) \overset{(3.85)}{\approx} \pi^\mu := c p_0(\mu) \exp\big(-\tfrac{n}{2}\hat{s}_n^\mu\big) n^{-n_{\theta^\mu}/2} \tag{5.45}$$

for the model posterior. As previously, we summarize the posterior model probabilities in the tuple $\pi^\top := \begin{bmatrix}\pi^1 & \dots & \pi^{n_{\mathcal{M}}}\end{bmatrix}$. Combining (5.43) and (5.45) yields the approximation

$$p(y \,|\, x) \approx c \sum_{\mu \in \mathcal{M}} \pi^\mu \phi\big(y \,|\, \hat{\eta}^\mu(x), \hat{\boldsymbol{W}}^\mu(x)\big) \tag{5.46}$$

for the posterior prediction of the model family for an observation under x.

These approximations are all justified in sufficiently "large" samples. Approximation (5.42) does neither assume that the model is correct nor that its responses are locally linear with respect to the parameter. Approximation (5.43) relies on a local linearization with respect to the parameter.

5.5.2. New Bounds for the BHH-Criterion

A new type of *lower* bound is obtained from the Kullback-Leibler distance (KLD)-based representation of the BHH-criterion (5.13) and from Cor. 5.5.

Theorem 5.9 (KLD-Based Lower Bound for the BHH-Criterion)

Assume that approximations (5.43) and (5.45) are exact and that the matrix $\hat{\boldsymbol{W}}^{\mu}(x)$ exists and is invertible for all models $\mu \in \mathscr{M}$ and under all experimental conditions $x \in \mathscr{X}$. For all $x \in \mathscr{X}$, define

$$\Gamma(x) := \rho\left(\tfrac{1}{2}\boldsymbol{\Gamma}(x), \pi\right) \tag{5.47a}$$

with the $n_{\mathscr{M}} \times n_{\mathscr{M}}$ matrix $\boldsymbol{\Gamma}(x)$ which has for all $\mu, \nu \in \mathscr{M}$ the element

$$\begin{aligned}&\|\hat{\eta}^{\mu}(x) - \hat{\eta}^{\nu}(x)\|^2_{\hat{\boldsymbol{W}}^{\nu}(x)^{-1}} + \operatorname{tr}\left(\hat{\boldsymbol{W}}^{\mu}(x)\hat{\boldsymbol{W}}^{\nu}(x)^{-1}\right)\\ &\qquad - \ln\det\left(\hat{\boldsymbol{W}}^{\mu}(x)\hat{\boldsymbol{W}}^{\nu}(x)^{-1}\right) - n_y.\end{aligned} \tag{5.47b}$$

in row μ and column ν. Then, $\Gamma(x) \leqslant \Lambda(x)$ under all $x \in \mathscr{X}$.

Proof Consider the BHH-criterion in the KLD-based form

$$\Lambda(x) = \sum_{\mu\in\mathscr{M}} p(\mu)\,\mathbb{D}\left[\tilde{\mathcal{Y}}_x \,\middle|\, \mathcal{M} = \mu \,\middle\|\, \tilde{\mathcal{Y}}_x\right] \tag{5.48}$$

from (5.13). Since approximation (5.43) is assumed to be exact, $\tilde{\mathcal{Y}}_x \mid \mathcal{M} = \mu$ is subject to a normal distribution, $\tilde{\mathcal{Y}}_x$ is subject to normal mixture distribution. Corollary 5.5 then provides for all $\mu \in \mathscr{M}$ the inequality

$$\mathbb{D}\left[\tilde{\mathcal{Y}} \,\middle|\, \mathcal{M} = \mu \,\middle\|\, \tilde{\mathcal{Y}}\right] \geqslant -\ln \sum_{\nu\in\mathscr{M}} p(\nu)\exp\left(-\tfrac{1}{2}f^{\mu\nu}\right), \tag{5.49}$$

where $f^{\mu\nu}$ stands for the expression (5.47b). Since approximation (5.18) for the model posterior is assumed to exact, $p(\mu) = \pi^{\mu}$ for all $\mu \in \mathscr{M}$. The claimed inequality follows from applying this equality together with (5.48) and (5.49) and writing the result using the function ρ_{π} defined in Lem. 5.7. □

An alternative new lower bound results from entropy-based representation (5.14) of the BHH-criterion and from Thm. 5.6.

Theorem 5.10 (Entropy-Based Lower Bound of the BHH-Criterion)

Assume that approximations (5.43) and (5.45) are exact and that the matrix $\hat{\boldsymbol{W}}^{\mu}(x)$ exists for all models $\mu \in \mathscr{M}$ and under all experimental conditions

$x \in \mathcal{X}$. For all $x \in \mathcal{X}$, define $\hat{\boldsymbol{W}}^{\mu\nu}(x) := \hat{\boldsymbol{W}}^{\mu}(x) + \hat{\boldsymbol{W}}^{\nu}(x)$ for all $\mu, \nu \in \mathcal{M}$ and

$$L(x) := \rho\big(\tfrac{1}{2}\boldsymbol{L}(x), \pi\big) - \tfrac{1}{2}\sum_{\mu \in \mathcal{M}} \pi^{\mu} \ln\det \hat{\boldsymbol{W}}^{\mu}(x) \tag{5.50a}$$

with the $n_{\mathcal{M}} \times n_{\mathcal{M}}$ matrix $\boldsymbol{L}(x)$ which has for all $\mu, \nu \in \mathcal{M}$ the element

$$\|\hat{\eta}^{\mu}(x) - \hat{\eta}^{\nu}(x)\|^2_{\hat{\boldsymbol{W}}^{\mu\nu}(x)^{-1}} + \ln\det \hat{\boldsymbol{W}}^{\mu\nu}(x) - n_y \tag{5.50b}$$

in row μ and column ν. Then, $L(x) \leqslant \Lambda(x)$ under all $x \in \mathcal{X}$.

Proof Consider the BHH-criterion in the entropy-based form

$$\Lambda(x) = \mathbb{H}\big[\tilde{\mathcal{Y}}_x\big] - \sum_{\mu \in \mathcal{M}} p(\mu)\, \mathbb{H}\big[\tilde{\mathcal{Y}}_x \,\big|\, \mathcal{M} = \mu\big] \tag{5.51}$$

from (5.14). Since approximation (5.43) is assumed to be exact, $\tilde{\mathcal{Y}}_x \,\big|\, \mathcal{M} = \mu$ is subject to a normal distribution, $\tilde{\mathcal{Y}}_x$ is subject to normal mixture distribution. Then, Prop. C.8 tells us that

$$\mathbb{H}\big[\tilde{\mathcal{Y}}_x \,\big|\, \mathcal{M} = \mu\big] = \tfrac{1}{2}\ln\det\big(\hat{\boldsymbol{W}}^{\mu}(x)\big) + \tfrac{1}{2} n_y (\ln(2\pi) + 1), \tag{5.52}$$

for all $x \in \mathcal{X}$ and all $\mu \in \mathcal{M}$, and Cor. 5.5 provides the inequality

$$\mathbb{H}\big[\tilde{\mathcal{Y}}_x\big] \geqslant -\sum_{\mu \in \mathcal{M}} p(\mu) \ln \sum_{\nu \in \mathcal{M}} p(\nu)\, \phi\big(\hat{\eta}^{\mu}(x) \,\big|\, \hat{\eta}^{\nu}(x), \hat{\boldsymbol{W}}^{\mu\nu}(x)\big) \tag{5.53}$$

$$\begin{aligned} &= -\sum_{\mu \in \mathcal{M}} p(\mu) \ln \sum_{\nu \in \mathcal{M}} p(\nu) \exp\Big(-\tfrac{1}{2}\big(\|\hat{\eta}^{\mu}(x) - \hat{\eta}^{\nu}(x)\|^2_{\hat{\boldsymbol{W}}^{\mu\nu}(x)^{-1}} \\ &\qquad + \ln\det \hat{\boldsymbol{W}}^{\mu\nu}(x) + n_y \ln(2\pi)\big)\Big), \end{aligned} \tag{5.54}$$

for all $x \in \mathcal{X}$ and all $\mu \in \mathcal{M}$. The equality in the second line results from substituting the expression (B.12b) for a normal probability density function (PDF). Since approximation (5.45) for the model posterior is assumed to exact, $p(\mu) = \pi^{\mu}$ for all $\mu \in \mathcal{M}$. Applying these relations to (5.51), writing the result using the function ρ_{π} defined in Lem. 5.7 and summarizing all constants yields the claimed inequality. □

5.5.3. Resulting New Design Criteria

Under the considered assumptions, we propose to use the bounds provided by Thms. 5.9 and 5.10 as *approximations* for the BHH-criterion:

$$L(x) \approx \Lambda(x) \text{ and } \Gamma(x) \approx \Lambda(x) \tag{5.55}$$

under all experimental conditions $x \in \mathcal{X}$. We refer to L and Γ as ENTROPY-BASED LOWER-BOUND CRITERION and KLD-BASED LOWER-BOUND CRITERION, respectively.

In practice, the Bayesian approximations (5.43) and (5.45) are typically *not* exact, as assumed in Thms. 5.9 and 5.10, so that the provided inequalities are also only of approximate nature. Therefore, Γ and L are to be considered as *heuristic* design criteria.

Interpretation of the KLD-Based Lower-Bound Criterion

Theorem 5.11 (Premetric for SPD Matrices [252])

Let $n \in \mathbb{N}$ and $\mathcal{P}_n$ be the set of real-valued symmetric positive definite (SPD) $n \times n$ matrices. The function $d\colon \mathcal{P}_n \times \mathcal{P}_n \mapsto \mathbb{R}$ defined as

$$d(\boldsymbol{A}, \boldsymbol{B}) := \operatorname{tr}\left(\boldsymbol{A}\boldsymbol{B}^{-1}\right) - \ln\det\left(\boldsymbol{A}\boldsymbol{B}^{-1}\right) - n \tag{5.56}$$

has the following properties.

(i) $d(\boldsymbol{A}, \boldsymbol{B}) \geqslant 0$ for all $\boldsymbol{A}, \boldsymbol{B} \in \mathcal{P}_n$,

(ii) $d(\boldsymbol{A}, \boldsymbol{B}) = 0 \Leftrightarrow \boldsymbol{A} = \boldsymbol{B}$, and

(iii) $d(\cdot, \cdot)$ is strictly convex on $\mathcal{P}_n \times \mathcal{P}_n$.

According to this lemma, provided by Uciński and Bogacka [252], the function d is a premetric for SPD matrices. It gives rise to a topology and thus to a notion of "closeness" on the set of SPD matrices.

Expression (5.47b) can hence be written as

$$f^{\mu\nu}(x) := \|\hat{\eta}^{\mu}(x) - \hat{\eta}^{\nu}(x)\|^2_{\hat{W}^{\nu}(x)^{-1}} + d\left(\hat{\boldsymbol{W}}^{\mu}(x), \hat{\boldsymbol{W}}^{\nu}(x)\right). \tag{5.57}$$

The first summand measures the systematic discrepancy between the predictions of both models for the outcome of an experiment under x, relative to the total uncertainty as predicted by model ν. The second summand measures the amount of uncertainty in the prediction of model μ for the outcome of an experiment under x, relative to the corresponding uncertainty of model ν. It rewards experimental conditions under which the prediction of one model is more reliable than that of the other model. Such experimental conditions are more likely to be helpful for recognizing a difference between these models than conditions under which both predictions are equally uncertain.

The KLD-based lower-bound criterion Γ combines the $n_{\mathcal{M}}^2$ functions $f^{\mu\nu}$ with $\mu, \nu \in \mathcal{M}$ into a single one using the function ρ from Lem. 5.7. Under experimental conditions maximizing Γ, both the predictions as well as the prediction uncertainties of all model pairs are maximally different.

The partial derivative of Γ with respect to $f^{\mu\nu}$ is an increasing function of the corresponding posterior probabilities π^μ and π^ν. Therefore, the larger the posterior probability of model μ, the larger the influence of the functions $(f^{\mu\nu} : \nu \in \mathcal{M})$ onto the design criterion.

Furthermore, Γ itself is increasing function of $f^{\mu\nu}$. Since ρ is concave, the relative influence of $f^{\mu\nu}$ decreases with larger values, so that extreme values receive relatively lesser attention than smaller ones.

The function $f^{\mu\nu}$ is not symmetric with respect to exchange of the model indices. The design criterion Γ is nevertheless symmetric, since it comprises for each term $f^{\mu\nu}$ also the term $f^{\nu\mu}$ with interchanged model indices.

Interpretation of the Entropy-Based Lower-Bound Criterion

The design criterion L is also composed of $n_{\mathcal{M}}^2$ functions measuring the dissimilarities between all model pairs using the function ρ, similar to Γ. In contrast to the latter and to U', it measures the dissimilarity between models μ and ν with expression (5.50b). Its first summand measures the systematic discrepancy between the predictions of both models for the outcome of an experiment under x, relative to the corresponding joined uncertainty of *both* predictions, measured in terms of $\hat{W}^\mu(x) + \hat{W}^\nu(x)$. The second summand quantifies the amount of this uncertainty, rewarding experimental conditions with low prediction uncertainty.

From a computational point of view, one might prefer L to Γ, since the former does neither involve inverses of $\hat{W}^\nu(x)$ nor the $\operatorname{tr}(\cdot)$-terms appearing in the latter.

5.5.4. Behavior under Small Uncertainties

Theorem 5.12 (New Design Criteria for MD under Small Uncertainties)

For all $x \in \mathcal{X}$ and all $\mu, \nu \in \mathcal{M}$, let $H(x;\mu,\nu) := \|\hat{\eta}^{\mu}(x) - \hat{\eta}^{\nu}(x)\|^2_{\boldsymbol{\Omega}^{-1}(x)}$ be the Hunter-Reiner (HR)-criterion from Def. 4.11 for discrimination between two models $\mu, \nu \in \mathcal{M}$, and let $\boldsymbol{H}(x)$ the $n_{\mathcal{M}} \times n_{\mathcal{M}}$ matrix with element $H(x;\mu,\nu)$ in row μ and column ν.

(i) If the parameter-induced uncertainties in the responses of all models vanish in the sense that

$$\tfrac{1}{n+1}\hat{\boldsymbol{J}}^{\mu}(x)\left(\hat{\boldsymbol{M}}^{\mu} + \hat{\boldsymbol{N}}^{\mu}\right)^{-1}\hat{\boldsymbol{J}}^{\mu\top}(x) = \boldsymbol{0} \tag{5.58}$$

for all $\mu \in \mathcal{M}$ and under all experimental conditions $x \in \mathcal{X}$, then it holds for all $x \in \mathcal{X}$ that

$$\Gamma(x) = \rho\left(\tfrac{1}{2}\boldsymbol{H}(x), \pi\right), \text{ and} \tag{5.59}$$

$$L(x) = \rho(\boldsymbol{H}(x), \pi) + \text{const.} \tag{5.60}$$

(ii) If, in addition to (i), the model posterior focuses only on two different models $\mu \neq \nu$ such that $\pi^{\mu} + \pi^{\nu} = 1$, then

$$\Gamma(x) = \pi^{\nu}\left(1 - \exp\left(-\tfrac{1}{2}H(x;\mu,\nu)\right) + \tfrac{1}{2}H(x;\mu,\nu)\right) + \mathcal{O}\left((\pi^{\nu})^2\right) \tag{5.61}$$

and also

$$L(x) = \pi^{\nu}(1 - \exp(-H(x;\mu,\nu)) + H(x;\mu,\nu)) + \mathcal{O}\left((\pi^{\nu})^2\right), \tag{5.62}$$

for all $x \in \mathcal{X}$. The last two equations remain true if μ is replaced by ν.

(iii) If there is no model uncertainty in the sense that $\pi^{\mu} = 1$ for some model $\mu \in \mathcal{M}$, then

$$\Gamma(x) \equiv 0 \text{ and } L(x) \equiv \text{const.} \tag{5.63}$$

The equations in the following proofs of (i) to (iii) hold for all $x \in \mathcal{X}$, so that we omit the argument x for clarity.

Proof (of Thm. 5.12(i)) The assumption (5.58) implies the simplifications

$$\hat{\boldsymbol{W}}^{\mu} = \boldsymbol{\Omega} \text{ and } \hat{\boldsymbol{W}}^{\mu\nu} = 2\boldsymbol{\Omega} \text{ for all } \mu, \nu \in \mathcal{M}. \tag{5.64}$$

Applying these simplifications to the KLD-based lower bound (5.47) yields

$$\begin{aligned}
(5.47b) &= \left\|\hat{\eta}^{\mu} - \hat{\eta}^{\nu}\right\|^2_{\boldsymbol{\Omega}^{-1}} + \operatorname{tr}\left(\boldsymbol{\Omega}\boldsymbol{\Omega}^{-1}\right) - \ln\det\left(\boldsymbol{\Omega}\boldsymbol{\Omega}^{-1}\right) - n_y &(5.65)\\
&= \left\|\hat{\eta}^{\mu} - \hat{\eta}^{\nu}\right\|^2_{\boldsymbol{\Omega}^{-1}} + \operatorname{tr}(\boldsymbol{I}) - \ln\det(\boldsymbol{I}) - n_y &(5.66)\\
&= \left\|\hat{\eta}^{\mu} - \hat{\eta}^{\nu}\right\|^2_{\boldsymbol{\Omega}^{-1}} + n_y - n_y = H(\cdot;\mu,\nu), &(5.67)
\end{aligned}$$

which proves (5.59). Applying them to the entropy-based lower bound (5.50) leads to

$$\begin{aligned}
(5.50b) &= \left\|\hat{\eta}^{\mu} - \hat{\eta}^{\nu}\right\|^2_{2\boldsymbol{\Omega}^{-1}} + \ln\det 2\boldsymbol{\Omega} - n_y &(5.68)\\
&= 2\left\|\hat{\eta}^{\mu} - \hat{\eta}^{\nu}\right\|^2_{\boldsymbol{\Omega}^{-1}} + \ln\det\boldsymbol{\Omega} + n_y\ln(2) - n_y &(5.69)\\
&= 2H(\cdot;\mu,\nu) + \ln\det\boldsymbol{\Omega} + n_y(\ln(2) - 1), &(5.70)
\end{aligned}$$

so that

$$\begin{aligned}
L(x) &= \rho\left(\boldsymbol{H} + \tfrac{1}{2}\ln\det(\boldsymbol{\Omega})\boldsymbol{I} + \tfrac{1}{2}n_y(\ln(2)-1)\boldsymbol{I}, \pi\right) - \tfrac{1}{2}\sum_{\mu\in\mathcal{M}}\pi^{\mu}\ln\det\boldsymbol{\Omega},\\
&\overset{\text{Lem. 5.7(iv)}}{=} \rho(\boldsymbol{H},\pi) + \tfrac{1}{2}\ln\det\boldsymbol{\Omega} + \tfrac{1}{2}n_y(\ln(2)-1) - \tfrac{1}{2}\ln\det\boldsymbol{\Omega},\\
&= \rho(\boldsymbol{H},\pi) + \tfrac{1}{2}n_y(\ln(2)-1),
\end{aligned}$$

proving (5.60). □

Proof (of Thm. 5.12(ii)) For proving (5.61) and (5.62), first note that assuming $\pi^{\mu} + \pi^{\nu} = 1$ implies $\pi^{\lambda} = 0$ for all $\lambda \in \mathcal{M} \setminus \{\mu,\nu\}$, since $\sum_{\mu\in\mathcal{M}}\pi^{\mu} = 1$ by definition. To prove (5.61), we apply this relation to (5.59) with adequately relabeled indices,

$$\begin{aligned}
\varGamma &= -\sum_{i\in\mathcal{M}}\pi^{i}\ln\sum_{j\in\mathcal{M}}\pi^{j}\exp\left(-\tfrac{1}{2}H(\cdot;\nu,\nu)\right) &(5.71)\\
&= -\sum_{i\in\mathcal{M}}\pi^{i}\ln\left(\pi^{\mu}\exp\left(-\tfrac{1}{2}H(\cdot;i,\mu)\right) + \pi^{\nu}\exp\left(-\tfrac{1}{2}H(\cdot;i,\mu)\right)\right) &(5.72)\\
&= -\pi^{\mu}\ln\left(\pi^{\mu}\exp\left(-\tfrac{1}{2}H(\cdot;\mu,\mu)\right) + \pi^{\nu}\exp\left(-\tfrac{1}{2}H(\cdot;\mu,\nu)\right)\right)\\
&\qquad - \pi^{\nu}\ln\left(\pi^{\mu}\exp\left(-\tfrac{1}{2}H(\cdot;\nu,\mu)\right) + \pi^{\nu}\exp\left(-\tfrac{1}{2}H(\cdot;\nu,\nu)\right)\right) &(5.73)\\
&= -\pi^{\mu}\ln\left(\pi^{\mu} + \pi^{\nu}\exp\left(-\tfrac{1}{2}H(\cdot;\mu,\nu)\right)\right)\\
&\qquad - \pi^{\nu}\ln\left(\pi^{\mu}\exp\left(-\tfrac{1}{2}H(\cdot;\nu,\mu)\right) + \pi^{\nu}\right). &(5.74)
\end{aligned}$$

To make the following calculations clearer, we use the abbreviations

$$f := \exp\bigl(-\tfrac{1}{2}H(\cdot;\mu,\nu)\bigr) = \exp\bigl(-\tfrac{1}{2}H(\cdot;\nu,\mu)\bigr) \text{ and } \epsilon := \pi^{\nu}, \tag{5.75}$$

and consider the design criterion Γ as function of ϵ,

$$\Gamma(\epsilon) = (\epsilon - 1)\ln(1 - \epsilon + \epsilon f) - \epsilon \ln((1-\epsilon)f + \epsilon). \tag{5.76}$$

We aim to expand the design criterion in a Taylor series around $\epsilon = 0$. To that end we require the derivative of $\Gamma(\epsilon)$ with respect to ϵ. The derivative of its first summand is

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\epsilon}(\epsilon - 1)\ln(1-\epsilon+\epsilon f) && (5.77)\\
={}& \ln(1-\epsilon+\epsilon f) + (\epsilon-1)\frac{\mathrm{d}}{\mathrm{d}\epsilon}\ln(1-\epsilon+\epsilon f) && (5.78)\\
={}& \ln(1-\epsilon+\epsilon f) + (\epsilon-1)(1-\epsilon+\epsilon f)^{-1}(f-1), && (5.79)
\end{aligned}$$

and the derivative of its second summand is

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\epsilon}\epsilon\ln((1-\epsilon)f+\epsilon) && (5.80)\\
={}& \ln((1-\epsilon)f+\epsilon) + \epsilon\frac{\mathrm{d}}{\mathrm{d}\epsilon}\ln((1-\epsilon)f+\epsilon) && (5.81)\\
={}& \ln((1-\epsilon)f+\epsilon) + \epsilon((1-\epsilon)f+\epsilon)^{-1}(1-f). && (5.82)
\end{aligned}$$

Therefore,

$$\Gamma(0) = -\ln(1) - 0\ln(f) = 0 \text{ and } \left.\frac{\mathrm{d}}{\mathrm{d}\epsilon}\Gamma(\epsilon)\right|_{\epsilon=0} = 1 - f - \ln(f). \tag{5.83}$$

The Taylor series expansion of $\Gamma(\epsilon)$ around $\epsilon = 0$ thus has the simple shape

$$\begin{aligned}
\Gamma(\epsilon) &= \Gamma(0) + \left.\frac{\mathrm{d}}{\mathrm{d}\epsilon}\Gamma(\epsilon)\right|_{\epsilon=0}\epsilon + \mathcal{O}\bigl(\epsilon^2\bigr) && (5.84)\\
&= \epsilon(1 - f - \ln(f)) + \mathcal{O}\bigl(\epsilon^2\bigr). && (5.85)
\end{aligned}$$

This equation is identical to (5.61) when written in the original notation. The proof of (5.62) is analog. □

Proof (of Thm. 5.12(iii)) If $\pi^{\mu} = 1$ for some model $\mu \in \mathcal{M}$, it follows from Lem. 5.7(vi) that

$$\begin{aligned}
\Gamma &\overset{(5.47b)}{=} \tfrac{1}{2}\Bigl(\bigl\|\hat{\eta}^{\mu} - \hat{\eta}^{\mu}\bigr\|^2_{\hat{W}^{\mu-1}} + \operatorname{tr}\Bigl(\hat{\boldsymbol{W}}^{\mu}\hat{\boldsymbol{W}}^{\mu^{-1}}\Bigr) - \ln\det\Bigl(\hat{\boldsymbol{W}}^{\mu}\hat{\boldsymbol{W}}^{\mu^{-1}}\Bigr) - n_y\Bigr)\\
&= \tfrac{1}{2}\bigl(0 + n_y - 0 - n_y\bigr) = 0, \text{ and}
\end{aligned}$$

$$\begin{aligned} L &\overset{(5.50\text{b})}{=} \tfrac{1}{2}\Big(\big\|\hat{\eta}^{\mu} - \hat{\eta}^{\mu}\big\|^{2}_{2\hat{W}^{\mu\mu-1}} + \ln\det 2\hat{\boldsymbol{W}}^{\mu} - n_y - \ln\det\hat{\boldsymbol{W}}^{\mu}\Big) \\ &= \tfrac{1}{2}\big(0 + \ln\det 2\hat{\boldsymbol{W}}^{\mu} - \ln\det\hat{\boldsymbol{W}}^{\mu} - n_y\big) \\ &= \tfrac{1}{2}\big(n_y \ln(2) - n_y\big) = \text{const}, \end{aligned}$$

which proves (5.63). □

The following approximate results follow from Thm. 5.12 and some continuity arguments. If the parameter-induced uncertainty in the model responses is substantially smaller than the experimental uncertainty, that is, if

$$\Big\|\tfrac{1}{n+1}\hat{\boldsymbol{J}}^{\mu}(x)\big(\hat{\boldsymbol{M}}^{\mu} + \hat{\boldsymbol{N}}^{\mu}\big)^{-1}\hat{\boldsymbol{J}}^{\mu\top}(x)\Big\| \ll \|\boldsymbol{\Omega}(x)\| \text{ for all } x \in \mathcal{X}, \tag{5.86}$$

with some matrix norm $\|\cdot\|$, then (5.59) and (5.60) hold approximately. If the posterior focuses strongly on the one model μ and to a lesser degree on the other model ν so that

$$\pi^{\mu} \gg \pi^{\nu} \gg \pi^{\lambda} \text{ for all } \lambda \notin \{\mu, \nu\}, \tag{5.87}$$

then the terms of $\mathcal{O}\Big(\big(\pi^{\nu}\big)^{2}\Big)$ in (5.61) and (5.62) can be neglected without introducing too much error, so that approximately

$$\Gamma(x) \propto 1 - \exp\big(-\tfrac{1}{2}H(x;\mu,\nu)\big) + \tfrac{1}{2}H(x;\mu,\nu), \text{ and} \tag{5.88}$$

$$L(x) \propto 1 - \exp(-H(x;\mu,\nu)) + H(x;\mu,\nu). \tag{5.89}$$

The function $1 - \exp(-cx) + cx$ is strictly convex in x for all $c \in \mathbb{R}_0^+$. Therefore, Γ and L are strictly convex transformations of the HR-criterion and thus have the same maximizers. If the parameter and model uncertainty is sufficiently small in the sense of (5.86) and (5.87), the proposed new design criteria provide approximately HR-optimal designs. They can hence be regarded as multi-model, model-robust and parameter-robust generalizations of the HR-criterion.

If there the MD problem is almost solved in the sense that

$$\pi^{\mu} \gg \pi^{\lambda} \text{ for all } \lambda \neq \mu, \tag{5.90}$$

then (5.63) holds approximately, and the proposed design criteria are almost independent of the experimental condition, as one would expect from a design criterion for MD.

5.5.5. Discussion

The proposed new design criteria for MD overcome several drawbacks of the classic upper-bound approximation of the BHH-criterion discussed in Sec. 5.3.4. Both use a formula for the parameter posteriors which does not rely on questionable assumptions of local linearity and a formula for the model posterior that is valid even for models with parameter vectors of different size. They are also both approximately *lower* bounds of the BHH-criterion and thus avoid loss of efficiency in the designed experiments due to overestimation.

They eventually reduce to forms that provide HR-optimal designs if parameter and model uncertainties are sufficiently small. Under certain regularity conditions discussed in Sec. 3.5, these uncertainties can be reduced below any bound by taking more data. If they are applied in a sequential design procedure, the newly proposed design criteria will converge to forms providing HR-optimal experiments, supposed the regularity conditions are met. Under comparably mild conditions, designs composed of HR-optimal experiments are asymptotically T-optimal, see Sec. 4.3, and thus the best designs theoretically possible for model discrimination. In contrast to the HR-criterion, the proposed design criteria take into account the current parameter uncertainty and the model uncertainty. This additional information used for selecting experimental condition should increase the rate with which the constructed sequential designs converge to a T-optimal ones.

The new design criteria are based on approximations that are exact only in the large-sample limit. How they perform in practice for finite and possibly small samples can be determined through numerical simulations, which we describe in Chap. 9.

Part III.

Numerical Methods and Results

[…] we cannot know that any statistical technique we develop is useful unless we use it. Major advances in science and in the science of statistics in particular, usually occur, therefore, as the result of the theory-practice iteration. The researcher hoping to break new ground in the theory of experimental design should involve himself in the design of actual experiments.

Box [43, p. 792]

6. Numerical Methods

> *[…] our central mission is to compute quantities that are typically uncomputable, from an analytical point of view, and to do it with lightning speed.*
>
> Trefethen, *The Definition of Numerical Analysis* [247]

Contents

This chapter deals with numerical methods required in the context of optimal experimental design for model discrimination. Section 6.1 discusses methods for least-squares problems that result from maximum-likelihood estimate problems in the context of model discrimination. Section 6.2 examines optimization problems arising from optimal experimental design for model discrimination, with a focus on their computational complexity and the consequences for their numerical treatment. Section 6.3 introduces low-discrepancy sequences, also known as pseudo-random numbers. In this thesis, they are to generate start values for local optimization techniques, and to generate space-filling experimental designs.

6.1. Parameter Estimation in Possibly Incorrect Models

Several of the strategies for solving model discrimination (MD) problems discussed in Chaps. 4 and 5 involve parameter maximum-likelihood estimates (PMLEs). This section discusses numerical methods for finding such estimates under the usual normality assumptions. It focuses on the special demands that arise in the context of MD problems, where the models might be incorrect. A major result is that the Gauss-Newton method is *not* appropriate in this scenario.

The section omits algorithmic details. They can be found in relevant textbooks, for example in that of Nocedal and Wright [194].

6.1.1. Problem Statement

The following scenario is encountered, for example, when one of the sequential strategies from Secs. 4.3, 4.4, 5.3 and 5.5 is applied to solve a MD problem. All these strategies involve PMLEs at some point, albeit parameter inference is not their central aim.

Considered Scenario

Suppose the observations $y_1, \ldots, y_n \in \mathbb{R}^{n_y}$ are available, realizations of the continuous $\mathbb{R}^{n_y}$-valued independent random variables $\mathcal{Y}_1, \ldots, \mathcal{Y}_n$, respectively. For all $i \in \{1, \ldots, n\}$, the random variable $\mathcal{Y}_i$ is normally distributed with mean $\bar{\eta}_i \in \mathbb{R}^{n_y}$ and a full-rank (and thus symmetric positive definite (SPD) and invertible) covariance matrix $\boldsymbol{\Omega}_i \in \mathbb{R}^{n_y \times n_y}$.

For all parameters $\theta \in \mathcal{Q} \subseteq \mathbb{R}^{n_\theta}$, a model is available which specifies for all $i \in \{1, \ldots, n\}$ a normal distribution mean $\eta_i(\theta) \in \mathbb{R}^{n_y}$ that is twice continuously differentiable and covariance $\boldsymbol{\Omega}_i$.

The functions $\bar{\eta}_1, \ldots, \bar{\eta}_n$ are unknown, which implies that it is not known if the model is correct, see Cor. 3.6.

Parameter Maximum-Likelihood Estimates

We define $y^\top := \begin{bmatrix} y_1^\top & \ldots & y_n^\top \end{bmatrix}$ and $\eta^\top(\theta) := \begin{bmatrix} \eta_1^\top(\theta) & \ldots & \eta_n^\top(\theta) \end{bmatrix}$, and write $\boldsymbol{\Omega}$ for the SPD block diagonal matrix composed of $\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_n$. In the considered scenario, a parameter is a PMLE, iff it minimizes over $\mathcal{Q}$ the sum of squared residuals (ssr) $s\colon \mathcal{Q} \mapsto \mathbb{R}_0^+$, defined as

$$s(\theta) := \tfrac{1}{2} \sum_{i=1}^{n} \| \eta_i(\theta) - y_i \|_{\boldsymbol{\Omega}_i^{-1}}^2 = \tfrac{1}{2} \left\| \boldsymbol{\Omega}^{-1/2} (\eta(\theta) - y) \right\|_2^2 \text{ for all } \theta \in \mathcal{Q}. \tag{6.1}$$

For details, see Cor. 3.10. For clarity, the given definition of the ssr use a normalization factor of $1/2$ instead of the factor $1/n$ used in previous chapters.

The problem of minimizing the ssr is an instance of a least-squares (lsq) problem.

6.1.2. Least-Squares (LSQ) Problems

Problem 6.1 (Least-Squares (LSQ))

Given the feasible set $\mathcal{V} \subseteq \mathbb{R}^{n_v}$ and the twice continuously differentiable residual function $r\colon \mathbb{R}^{n_v} \mapsto \mathbb{R}^{n_r}$, find a point $v^\star$ that minimizes over $\mathcal{V}$ the objective function $f\colon \mathbb{R}^{n_v} \mapsto \mathbb{R}_0^+$ defined by

$$f(v) := \tfrac{1}{2} \| r(v) \|_2^2 = \tfrac{1}{2} \sum_{i=1}^{n_r} r_i^2(v) \text{ for all } v \in \mathcal{V}. \tag{6.2}$$

This problem is sometimes called *nonlinear* lsq problem to emphasize that it allows r to be a nonlinear function of v. For all $v \in \mathbb{R}^{n_v}$, we write $\nabla f(v) \in \mathbb{R}^{n_v}$ and $\nabla^2 f(v) \in \mathbb{R}^{n_v \times n_v}$ for the gradient and the Hessian of f at v, respectively.

It is usually difficult to find a global solution $v^\star$ of this problem at which $f(v^\star) \leqslant f(v)$ for all $v \in \mathcal{V}$. It is significantly easier, and often sufficient in practice, to find a local solution $v^\star$, which satisfies this inequality for all $v \in (\mathcal{B} \cap \mathcal{V})$, where $\mathcal{B}$ is a neighborhood of $v^\star$.

Several of the many numerical methods available for finding local solutions of this well-examined problem class are discussed by Nocedal and Wright [194, Chap. 10]. In the following two sections we sketch and discuss some popular methods.

6.1.3. Newton-Type Methods for Unconstrained LSQ Problems

We first consider Prob. 6.1 without constraints, $\mathcal{V} = \mathbb{R}^{n_v}$. Then, a necessary condition that $v^\star \in \mathcal{V}$ is a local solution is that

$$\nabla f(v^\star) = 0. \tag{6.3}$$

The numerical methods of choice for solving this equation are Newton-type methods. Starting from a point $v^0 \in \mathbb{R}^{n_v}$, such a method determines a sequence of iterates $v^1, v^2, \dots$ in $\mathbb{R}^{n_v}$ according to

$$v^{k+1} := v^k + \alpha^k p^k, \text{ and } p^k := -\left(\boldsymbol{B}^k\right)^{-1} \nabla f\left(v^k\right), \text{ for all } k \in \mathbb{N}_0, \tag{6.4}$$

with step lengths $\alpha^k \in (0,1]$ and $n_v \times n_v$ matrices $\boldsymbol{B}^k$. It terminates the sequence if the current iterate is supposed to be sufficiently close to a local solution or if no further progress seems to be possible.

Different Newton-type methods arise from difference choices for α^k and $\boldsymbol{B}^k$. If $\alpha^k = 1$ for all $k \in \mathbb{N}$, one speaks of full-step method. In the following we consider three important Newton-type methods that arise from different choices for $\boldsymbol{B}^k$. For a discussion of step length selection algorithms we refer to Nocedal and Wright [194, Secs. 3.1 and 3.5].

Newton's Method

The eponymous Newton's method (sometimes called Newton-Raphson method) is not restricted to objective functions of the lsq type, but can be applied to all objective functions that are sufficiently smooth. Newton's method is defined by (6.4) with

$$\boldsymbol{B}^k = \nabla^2 f\left(v^k\right) \text{ for all } k \in \mathbb{N}_0. \tag{6.5}$$

If the Hessian $\nabla^2 f\left(v^k\right)$ is positive definite (and thus invertible), this choice ensures that the search direction p^k minimizes the second-order Taylor series approximation of $f\left(v^k + p\right)$ among all $p \in \mathcal{V}$.

If v^0 is sufficiently close to a local solution $v^\star$ and if $\nabla^2 f(v)$ is positive definite and Lipschitz continuous in a neighborhood of $v^\star$, then the iterates of Newton's method with unit step length converge towards $v^\star$ with a quadratic rate. A proof is given by Nocedal and Wright [194, Thm. 3.5].

Newton's method has attractive convergence properties, but also suffers from two main drawbacks: First, computing sufficiently precise second derivatives for the Hessian is often too costly or too error-prone. Second, the Hessian might not be positive definite. Then, it is possibly not invertible and the search direction p^k is not defined. Yet even if it *is* invertible, the resulting search direction might not lead to a decrease in the objective function.

The two popular Quasi-Newton methods discussed in the following avoid these problems by using suitable approximations of the exact Hessian.

Gauss-Newton Method

The highly popular Gauss-Newton method gains its efficiency by exploiting the special structure of the LSQ objective function (6.2). For all $v \in \mathcal{V}$ and all $i \in \{1, \dots, n_r\}$, let $\boldsymbol{J}(v) \in \mathbb{R}^{n_r \times n_v}$ be the Jacobian of the residual function r at v and let $\nabla^2 r_i(v) \in \mathbb{R}^{n_v \times n_v}$ be the Hessian of its i-th component at v. Using the explicit quadratic form of the objective function (6.2) and applying Prop. A.5, its Hessian can be written as

$$\nabla^2 f(v) \overset{\text{Prop. A.5}}{=} \boldsymbol{J}^\top(v)\boldsymbol{J}(v) + \boldsymbol{N}(v), \text{ where } \boldsymbol{N}(v) := \sum_{i=1}^{n_r} r_i(v)\nabla^2 r_i(v). \tag{6.6}$$

The Gauss-Newton method is defined by (6.4) with the Hessian approximation

$$\boldsymbol{B}^k = \boldsymbol{J}^\top\left(v^k\right)\boldsymbol{J}\left(v^k\right) \text{ for all } k \in \mathbb{N}_0, \tag{6.7}$$

which arises from (6.6) if $\boldsymbol{N}(v)$ is ignored.

If $\boldsymbol{J}^\top(v^\star)\boldsymbol{J}(v^\star)$ is positive definite at a local solution $v^\star$ and dominates the Hessian (6.6) in the sense that

$$\left\|\left(\boldsymbol{J}^\top(v^\star)\boldsymbol{J}(v^\star)\right)^{-1}\boldsymbol{N}(v^\star)\right\|_2 \ll 1 \tag{6.8}$$

and some regularity conditions are met, then the iterates of the Gauss-Newton method with unit step length converge locally towards $v^\star$ with a superlinear rate. If $\boldsymbol{N}(v^\star) = \boldsymbol{0}$, the rate is even quadratic. A proof can be found in the book of Nocedal and Wright [194, Thm. 10.1ff].

The Hessian approximation (6.7) involves only first derivatives of the objective function, and is positive definite whenever rank $\boldsymbol{J}\left(v^k\right) \geqslant n_v$. In LSQ problems where (6.8) is met, the Gauss-Newton method overcomes the two main drawbacks of Newton's method at the cost of a reduced, yet still high, rate of convergence. If (6.8) is not satisfied, an attractive alternative is the Quasi-Newton method discussed next.

BFGS Method

There are numerous Quasi-Newton methods which sequentially update a Hessian approximation from the information gained in each iteration. Among them, the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method is particularly popular because of its outstanding efficiency and robustness. It can applied to any type of sufficiently smooth objective function, not only to those of the LSQ type (6.2).

The BFGS method is defined by (6.4) and the Hessian update rule

$$\boldsymbol{B}^{k+1} := \boldsymbol{B}^k - \frac{\boldsymbol{B}^k s^k {s^k}^\top {\boldsymbol{B}^k}^\top}{{s^k}^\top \boldsymbol{B}^k s^k} + \frac{g^k {g^k}^\top}{{g^k}^\top s^k}, \tag{6.9a}$$

$$\text{with } s^k := v^{k+1} - v^k \text{ and } g^k := \nabla f\big(v^{k+1}\big) - \nabla f\big(v^k\big) \tag{6.9b}$$

for all $k \in \mathbb{N}_0$. The BFGS method thus requires an initial $n_v \times n_v$ Hessian approximation $\boldsymbol{B}^0$. Different derivations of this formula are given in the original publications of Broyden [53], Fletcher [100], Goldfarb [107], and Shanno [228] and in textbooks, for example in that of Nocedal and Wright [194, Sec. 6.1].

The BFGS Hessian approximations (6.9) involve only first derivatives of the objective function, and are positive definite if the initial matrix $\boldsymbol{B}^0$ is positive definite. Locally superlinear convergence of the BFGS method can be proved under different sets of mild assumptions, and practical implementations of the BFGS method in fact often converge at a superlinear rate. Details are given, for example, by Nocedal and Wright [194, Sec. 6.4].

The BFGS method thus avoids the two main drawbacks of Newton's method, at the cost of a reduced, yet still high, rate of convergence. In contrast to the Gauss-Newton method, it is also applicable to LSQ problems which fail to satisfy (6.8).

6.1.4. SQP Methods for Constrained LSQ Problems

Suppose the feasible set is characterized through the constraint functions $g\colon \mathbb{R}^{n_v} \mapsto \mathbb{R}^{n_g}$ and $h\colon \mathbb{R}^{n_v} \mapsto \mathbb{R}^{n_h}$ according to

$$\mathcal{V} := \big\{v \in \mathbb{R}^{n_v} : g(v) = 0 \wedge h(v) \geqslant 0\big\}, \tag{6.10}$$

with component-wise inequalities. Necessary conditions for a local solution in this case are the Karush-Kuhn-Tucker conditions. They are named after their discoverers Karush [138] and Kuhn and Tucker [156] and can be found in textbooks on constrained optimization, for example in that of Nocedal and Wright [194, Chap. 12]. The Karush-Kuhn-Tucker conditions are the counterpart of condition (6.3) for the constrained case. In the absence of inequality constraints, they can also be written in the form $F(v^\star) = 0$.

A popular and powerful approach for such problems are sequential quadratic programming (SQP) methods, first proposed by Wilson [268]. They quickly became one of the favorite methods for nonlinear constrained optimization. Overviews over the vast field of related publications are given by Boggs and Tolle [37] and Gould, Orban, and Toint [108]. The details mentioned in the following can be found in the in-depth discussion of Nocedal and Wright [194, Chap. 18].

In the absence of inequality constraints, the SQP method with full steps and exact Hessians is equivalent to Newton's method for the Karush-Kuhn-Tucker conditions. If equality constraints are present, it behaves at least locally like Newton's method under regularity conditions. As such, the full-step exact Hessian SQP method shares several properties with Newton's method for unconstrained problems, in particular the previously discussed difficulties that arise from using an exact Hessian. Sequential quadratic programming methods that use the Gauss-Newton or the BFGS Hessian approximation can avoid some of these problems, analogously to the unconstrained case. Details about the CONSTRAINED GAUSS-NEWTON METHOD as a special case of an SQP method for solving constrained LSQ problems are given by Bock [35].

6.1.5. Choosing a Method in the Context of MD Problems

To find a PMLE in the scenario described in Sec. 6.1.1, we need to solve a LSQ problem in the variable θ with the feasible set $\mathcal{Q} \subseteq \mathbb{R}^{n_\theta}$ and the residual function

$$r(\theta \,|\, y) := \boldsymbol{\Omega}^{-1/2}\big(\eta(\theta) - y\big), \quad \text{for all } \theta \in \mathcal{Q}, \tag{6.11}$$

which depends parametrically on the data y. If the problem arises in the context of MD, the following points should be taken into account when choosing a numerical method:

(a) *No good starting point might be available for the PMLE,* since MD problems typically arise in early stages of model building when little is known about the data-generating process.

(b) *The function η may be nonlinear.* Albeit the considered scenario permits affine-linear models, we do not restrict our considerations to this special case.

(c) *It is unknown if the model is correct.* If we knew that the model was correct, we would not be dealing with an MD problem in the first place.

In the remaining section we discuss the effects of these points onto the previously discussed Newton-type methods.

To that end, realize that $r(\theta \,|\, y)$ measures the mismatch between the model prediction under parameter θ and the observations (in the form of the difference $\eta(\theta) - y$), *relative* to random variability of the observations (in the form of the multivariate standard deviation $\boldsymbol{\Omega}^{1/2}$). Define $\mathcal{Y}^\top := \begin{bmatrix} \mathcal{Y}_1^\top & \dots & \mathcal{Y}_n^\top \end{bmatrix}$ and $\bar{\eta}^\top(\theta) := \begin{bmatrix} \bar{\eta}_1^\top(\theta) & \dots & \bar{\eta}_n^\top(\theta) \end{bmatrix}$. The average value and standard deviation of $r(\cdot)$ are

$$\mathbb{E}\left[r(\theta \,|\, \mathcal{Y})\right] = \boldsymbol{\Omega}^{-1/2}\left(\eta(\theta) - \bar{\eta}\right) \text{ and } \mathbb{C}\left[r(\theta \,|\, \mathcal{Y})\right]^{1/2} = \boldsymbol{I}, \tag{6.12}$$

respectively, for all $\theta \in \mathcal{Q}$. The difference $\eta(\theta) - \bar{\eta}$ can be regarded as the *systematic* mismatch between the model prediction under θ and the data-generating process.

Newton's Method and Exact-Hessian SQP Methods

Newton's method (for an unconstrained problem) and an exact-Hessian SQP method (for the constrained case) may fail if they encounter a non-positive definite Hessian. This drawback may be practically unproblematic if one can choose a starting point close to a local solution, where the Hessian is often positive definite. Due to (a), this is usually not possible in the context of MD problems, so that one can expect that these methods will run into problems there.

Gauss-Newton Hessian Approximation

Let $r_i(\cdot)$ be the i-th scalar component of $r(\cdot)$. The Gauss-Newton Hessian approximation is applicable if condition (6.8) is satisfied at a local solution $\theta^\star$. This is the case if the components of the matrix

$$\boldsymbol{N}(\theta^\star) \overset{(6.6)}{=} \sum_{i=1}^{n_r} r_i(\theta^\star \,|\, y) \nabla^2 r_i(\theta^\star \,|\, y) \tag{6.13}$$

are sufficiently small. This is true if and only if for each $i \in \{1, \dots, n_r\}$ it holds that (i) the residual $r_i(\theta^\star \,|\, y)$ is small, or (ii) the components of the Hessian $\nabla^2 r_i(\theta^\star \,|\, d)$ are small.

According to (6.12), the residuals remain small in average as long as the systematic mismatch $\eta(\theta) - \bar{\eta}$ is small relative to the random variability in terms of $\boldsymbol{\Omega}^{1/2}$. As discussed in Sec. 3.2, the difference $\eta(\theta) - \bar{\eta}$ is zero if and only if the model is correct and θ is a correct parameter. In a MD problem, we do not know whether the considered model is correct, see (c). We also do not know the function $\bar{\eta}$, so that it remains hidden to us *how* incorrect the model is in terms of $\eta(\theta) - \bar{\eta}$.

It is evident from (6.11) that the Hessians have small components if the model response η is almost affine-linear close to $\theta^\star$. In considered scenario, this is not necessarily true, see (b).

Consequentially, *in the context of an MD problem, we do not know whether the Gauss-Newton Hessian approximation is adequate for a given model.* Actually, we expect that it is typically inadequate for several of the rival models, unless we are in the fortunate but unlikely situation, that all of them are "almost" correct.

Both theory and practical experience suggest that the BFGS method (in unconstrained problems) or an SQP method with a BFGS Hessian approximation (if the problem is constrained) can deal fairly well with (a) to (c).

6.2. Sequential OED Problems for Model Discrimination

Main contributions of this thesis are the novel sequential design criteria for model discrimination (MD) proposed in Secs. 4.4 and 5.5, advanced versions of established criteria treated in Secs. 4.3 and 5.3. Experiments performed under conditions maximizing such a design criterion are supposed to be particularly efficient for MD. This section discusses methods for solving such maximization problems, which we call SEQUENTIAL OPTIMAL EXPERIMENTAL DESIGN (OED) PROBLEMS FOR MD. We shall see that they are particularly difficult to solve numerically due to their intrinsic non-linearity and non-convexity.

6.2.1. Problem Statement

Considered Scenario

Suppose experiments can be performed under conditions from the compact experimental domain $\mathcal{X} \subseteq \mathbb{R}^{n_x}$ and yield observations in the observation domain $\mathcal{Y} \subseteq \mathbb{R}^{n_y}$. For all $x \in \mathcal{X}$, an observation obtained from an experiment under condition x is a realization the continuous $\mathbb{R}^{n_y}$-valued random variable $\mathcal{Y}_x$, whose distribution may depend on x and whose covariance matrix $\boldsymbol{\Omega}(x) := \mathbb{C}[\mathcal{Y}_x]$ has full rank and is thus symmetric positive definite (SPD) and invertible.

For all $x \in \mathcal{X}$, the distribution of $\mathcal{Y}_x$ is unknown. To cope with this lack of knowledge, several regression models are available. For all $\mu \in \mathcal{M} := \{1, \ldots, n_{\mathcal{M}}\}$ and all $\theta^\mu \in \mathcal{Q}^\mu \subseteq \mathbb{R}^{n_{\theta^\mu}}$, model μ with parameter θ^μ specifies for all experimental conditions $x \in \mathcal{X}$ an n_y-dimensional normal distribution with mean $\eta^\mu(\theta^\mu, x)$ and covariance $\boldsymbol{\Omega}(x)$.

A Simplified Problem

General sequential OED problems are discussed in Sec. 4.1.3. We first consider an instructive simplified problem that arises in the considered scenario under the following additional assumptions.

(a) The experimental domain is solely characterized by box constraints, $\mathcal{X} = \left\{x \in \mathbb{R}^{n_x} : l \leqslant x \leqslant u\right\}$, with $l, u \in \mathbb{R}^{n_v}$ and $l \leqslant u$. The inequalities are meant component-wise.

(b) Under all experimental conditions $x \in \mathcal{X}$, the observation covariance $\boldsymbol{\Omega}(x)$ has the same value, denoted by $\boldsymbol{\Omega}$.

(c) The models are affine-linear in x. That is, for all models $\mu \in \mathcal{M}$ and all parameters $\theta^\mu \in \mathcal{Q}^\mu$ there exists a matrix $\boldsymbol{J}^\mu(\theta^\mu) \in \mathbb{R}^{n_y \times n_x}$ and a vector $h^\mu(\theta^\mu) \in \mathbb{R}^{n_y}$ such that $\eta^\mu(\theta^\mu, x) = \boldsymbol{J}^\mu(\theta^\mu)x + h^\mu(\theta^\mu)$ for all $x \in \mathcal{X}$.

(d) The design criterion is the multi-model Hunter-Reiner (HR)-criterion from Def. 4.11.

For all $\mu \in \mathcal{M}$, let $\hat{\theta}^\mu \in \mathcal{Q}^\mu$ be a parameter maximum-likelihood estimate (PMLE) of model μ. Suppose without loss of generality (WLOG) that the models $\mu \in \mathcal{M}$ and $\nu \in \mathcal{M}$ have the smallest and second-smallest lack-of-fit (in terms of the sum of squared residuals (SSR)), respectively, based on available previous experiments. Under assumptions (b)–(d) and with the abbreviations $\boldsymbol{J} := \boldsymbol{J}^\mu\left(\hat{\theta}^\mu\right) - \boldsymbol{J}^\nu\left(\hat{\theta}^\nu\right)$ and $h := h^\mu\left(\hat{\theta}^\mu\right) - h^\nu\left(\hat{\theta}^\nu\right)$, the design criterion is

$$\Psi(x) = \left\|\eta^\mu\left(\hat{\theta}^\mu, x\right) - \eta^\nu\left(\hat{\theta}^\nu, x\right)\right\|^2_{\boldsymbol{\Omega}^{-1}} \tag{6.14a}$$

$$= \|\boldsymbol{J}x + h\|^2_{\boldsymbol{\Omega}^{-1}} \tag{6.14b}$$

$$= x^\top \boldsymbol{J}^\top \boldsymbol{\Omega}^{-1} \boldsymbol{J} x + 2h^\top \boldsymbol{\Omega}^{-1} \boldsymbol{J} x + \text{const} \tag{6.14c}$$

under all $x \in \mathcal{X}$. The additive constant term is irrelevant for OED and is hence omitted in the following. Assumptions (a)–(d) thus give rise to the following sequential OED problem, a special case of Prob. 4.4 on p. 127.

Problem 6.2 (Simplified Sequential OED for MD))

Given the bounds l and u and the design criterion $\Psi\colon \mathcal{X} \mapsto \mathbb{R}_0^+$ from (6.14), find a maximizer $x^\star$ of Ψ in $\mathcal{X} = \left\{x \in \mathbb{R}^{n_x} : l \leqslant x \leqslant u\right\}$. The inequalities are meant component-wise.

This problem is an instance of the class of optimization problems considered next.

6.2.2. Quadratic Programs

Problem 6.3 (Quadratic Program (QP))

Given $(\boldsymbol{Q}, c, \boldsymbol{A}, b) \in \mathbb{R}^{n_v \times n_v} \times \mathbb{R}^{n_v} \times \mathbb{R}^{m \times n_v} \times \mathbb{R}^{m}$, with symmetric $\boldsymbol{Q}$, find a point $v^\star$ that minimizes the OBJECTIVE FUNCTION $f\colon \mathbb{R}^{n_v} \mapsto \mathbb{R}$, defined as

$$f(v) := v^\top \boldsymbol{Q} v + c^\top v, \quad \text{for all } v \in \mathbb{R}^{n_v}, \tag{6.15}$$

over the FEASIBLE SET $\mathcal{V} := \left\{v \in \mathbb{R}^{n_v} : \boldsymbol{A} v \geqslant b\right\}$, where the inequalities are meant component-wise.

A quadratic program (QP) is BOX-CONSTRAINED, iff its feasible set can be written in the form $\mathcal{V} = \left\{v \in \mathbb{R}^{n_v} : l \leqslant v \leqslant u\right\}$, with $l, u \in \mathbb{R}^{n_v}$. The inequalities are meant component wise. It is CONVEX, iff $\boldsymbol{Q}$ is positive semidefinite (all eigenvalues non-negative), and NON-CONVEX otherwise. A non-convex QP is CONCAVE, iff $\boldsymbol{Q}$ is negative semidefinite (all eigenvalues non-positive), and is INDEFINITE, if $\boldsymbol{Q}$ has at least one positive and one negative eigenvalue.

A GLOBAL SOLUTION of this problem is a point $v^\star$ at which $f(v^\star) \leqslant f(v)$ for all $v \in \mathcal{V}$. A LOCAL SOLUTION $v^\star$ satisfies this inequality in the intersection of $\mathcal{V}$ and an open neighborhood of $v^\star$. A global or local solution is STRICT, iff it strictly satisfies the defining inequality.

Computational Complexity

To discuss the difficulty of quadratic programs, we require some key ideas from computational complexity theory. More details can be found in the classic work of Garey and Johnson [105] and the more recent book of Arora and Bara [8].

The level of difficulty of an optimization problem can be measured based on its worst-case time complexity, that is, based on the number of elementary computational operations required to solve it in the worst case. This number directly translates into actual running time when the operations are performed on a particular computer, thus the term "time" complexity.

A problem can be solved in polynomial time (exponential time), iff the time required to solve it is a polynomial (an exponential function) in the size of the quantities required to specify an instance of the problem.

The class of problems that can be solved in polynomial time is denoted $\mathcal{P}$. The class $\mathcal{NP}$ comprises all problems that can be solved in polynomial time by a "non-deterministic algorithm." Simply speaking, $\mathcal{P}$ contains the problems that are "easy" to solve, and $\mathcal{NP}$ the problems for which it is "easy" to verify the correctness of a supposed solution (which itself was possibly hard to compute). The hardest problems in $\mathcal{NP}$ are called $\mathcal{NP}$-complete. A problem that is as hard as an $\mathcal{NP}$-complete problem, but is not necessarily in $\mathcal{NP}$, is $\mathcal{NP}$-hard.

Obviously, $\mathcal{P}$ is a subset of $\mathcal{NP}$. Up to now, it remains one of the great unresolved mathematical problems if both classes are identical or not, that is, if $\mathcal{P} = \mathcal{NP}$. Many interesting and challenging problems are $\mathcal{NP}$-hard. Yet if $\mathcal{P} \neq \mathcal{NP}$, as believed by many researchers, then these problems cannot be solved in polynomial time, which can mean in practice that large they are computationally intractable.

In the following, results concerning the computational complexity of quadratic programs are taken from Horst and Pardalos [123] or from Vavasis [255], if no other reference is given.

Convex Quadratic Programs

Convex quadratic programs have certain properties that significantly simplify their numerical solution. In particular, any Karush-Kuhn-Tucker point (see Sec. 6.1.4) is a local solution, which in turn is a global solution. Such problems are in $\mathcal{P}$, that is, they can be solved in polynomial time, as proved by Kozlov, Tarasov, and Khachiyan [151]. For available solution methods we refer to the book of Nocedal and Wright [194, Chap. 16] and the references given therein.

Unfortunately, quadratic programs appearing in the context of OED for MD are typically non-convex.

Non-Convex Quadratic Programs

Solving quadratic programs that are non-convex and possibly indefinite is generally tough. Even checking if a given feasible point is a *local* solution, or if a *local* solution is strict, are $\mathcal{NP}$-complete problems, as shown by Murty [191] and Pardalos and Schnitger [197].

Not surprisingly, the problem of finding a *global* solution is also $\mathcal{NP}$-complete: Sahni [219] shows that it is $\mathcal{NP}$-hard, and Vavasis [256] that it is in $\mathcal{NP}$. Pardalos and Vavasis [199] prove that it is $\mathcal{NP}$-hard even in the simplest case that $\boldsymbol{Q}$ has only one negative eigenvalue. Many special cases of non-convex quadratic programs are also $\mathcal{NP}$-hard, for example non-convex box-constrained quadratic programs.

Pardalos [198] reviews algorithms for finding global optima in non-convex quadratic programs. A branch-and-bound algorithm was recently proposed by Burer and Vandenbussche [56, 57]. More details can be found in the books of Horst and Pardalos [123, Chap. 4] and Horst, Pardalos, and Van Thoai [122, Chap. 2].

Concave Quadratic Programs

We now turn to concave quadratic programs and assume that the feasible set is closed. Solving such a QP means to search for a minimizer of a negative quadratic function on a polytope. Any local or global solution, if it exists, is a vertex (the equivalent of a corner in several dimensions) of that polytope. A proof of this well-known property can be found, for example, in Horst and Pardalos [123, Sec. 3.4]. Since a polytope has a finite number of vertices, this property introduces an integer aspect into concave quadratic programs. The number of vertices may grow exponentially with the problem dimension. In the box-constrained case with non-degenerate constraints, the feasible domain is a n_v-dimensional cuboid, which has 2^{n_v} vertices.

Exploiting that local solutions are located on vertices, Pardalos and Schnitger [197, Rem. 3] show that local optimality in an concave QP can be verified in polynomial time, in contrast to indefinite quadratic programs. Finding a global solution, however, is $\mathcal{NP}$-hard, like in the indefinite case. Polynomial-time algorithms are known for certain special cases, like minimizing the Euclidean norm on a cuboid, as shown by Horst, Pardalos, and Van Thoai [122, Secs. 2.4.2].

Approximate Solutions of Quadratic Programs

In practice, it often suffices to solve difficult problems only approximately. Vavasis [254] shows that the effort for *approximating* the global solution of Prob. 6.3 with a compact feasible set and a matrix $\boldsymbol{Q}$ with t negative eigenvalues is

$$\mathcal{O}\left(\left\lceil n_v(n_v+1)/\sqrt{\epsilon}\right\rceil^t \ell\right), \tag{6.16}$$

where the approximation quality ϵ ranges from $\epsilon = 0$ (exact global solution) to $\epsilon = 1$ (arbitrary feasible point), and ℓ denotes the time required to solve a *convex* QP with the same dimensions as Prob. 6.3.

In an indefinite QP we have $t < n_v$, so that the effort for obtaining an approximate global solution grows polynomially in n_v. It might thus be possible to efficiently approximate the global solution of such a problem as long as $\boldsymbol{Q}$ has not too many negative eigenvalues.

In an concave quadratic programs we have $t = n_v$, so that the effort for approximating the global solution grows exponentially with n_v. In practice, this effort might be computationally intractable even for moderately large n_v.

6.2.3. The Challenges of Real-World OED Problems

Solving the Simplified Sequential OED Problem

Consider the simplified sequential OED Prob. 6.2. If we replace the maximization with minimization, and switch the sign of the objective function as compensation, we see that this problem is a QP in the variable x with objective function

$$x^\top\left(-\boldsymbol{J}^\top \boldsymbol{\Omega}^{-1}\boldsymbol{J}\right)x - 2h^\top \boldsymbol{\Omega}^{-1}\boldsymbol{J}x \tag{6.17}$$

and the feasible set $\mathcal{X} = \left\{x \in \mathbb{R}^{n_x} : l \leqslant x \leqslant u\right\}$, where the inequalities are meant component-wise. The matrix $-\boldsymbol{J}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{J}$ is negative semidefinite by construction. In the typical case that $\operatorname{rank}(\boldsymbol{J}) \geqslant n_x$, it is even negative definite. *It is hence a box-constrained concave* QP.

As argued in the previous section, such problems might be difficult to solve. In particular, finding a global solution *exactly* is $\mathcal{NP}$-hard, and the effort for computing it only *approximately* increases exponentially in n_x.

Finding a local solution seems to be easier: pick one of the 2^{n_x} vertices and verify local optimality, which can be done in polynomial time. Yet if the verification fails, one has to start afresh at a different vertex. Therefore, the required effort increases exponentially in n_x in the worst-case.

If we accept to make such an effort, we might also directly go for the global solution. Besides, it is unclear how useful we should consider a local solution given that we known that there are $2^{n_x} - 1$ other potential local solutions.

Solving Real-World OED Problems

We saw that even the simplified (some might say simplistic) sequential OED problem resulting from assumptions (a)–(d) is hard to solve. In practice, these assumptions are rarely met, which further complicates the resulting sequential OED problems.

If at least one of the assumptions (b)–(d) is violated, meaning that the covariance $\boldsymbol{\Omega}$ depends on x, or the model responses $\eta^\mu(\theta^\mu, x)$ are nonlinear in x, or the design criterion $\Psi(\cdot)$ is not a quadratic function of $\eta^\mu(\cdot) - \eta^\nu(\cdot)$, then the resulting sequential OED problem is not longer a QP, but a possibly non-convex nonlinear program.

The same is true if the admissible experimental conditions are characterized by nonlinear equality and component-wise inequality constraints of the form $g(x) = 0$ and $h(x) \geqslant 0$ instead of the box constraints from assumption (a). In general, non-convex nonlinear programs are even harder to solve than concave quadratic programs.

Local Methods for Real-World OED Problems

In lack of efficient method to solve real-world sequential OED problems globally, one might be tempted to apply a local method.

A popular approach is the sequential quadratic programming (SQP) approach, briefly considered in Sec. 6.1.4, which solves a sequence of "local" quadratic programs obtained from Taylor approximations of the nonlinear program around the iterates. When applied to a sequential OED problem without the simplifying assumptions assumptions (a)–(d), the local quadratic programs will typically be indefinite, even close to a local solution. Therefore, SQP variants that apply positive-definite Hessian approximations, like the Gauss-Newton method or the Broyden-Fletcher-Goldfarb-Shanno (BFGS) methods, are unsuited for these problems. Besides being computationally expensive, little can be said about the global convergence behavior exact-Hessian SQP methods, a notable exception being the results of Bardow et al. [24].

If a local method converges, then usually to a point at which necessary (but not sufficient) conditions for local optimality are met. As in the case of indefinite quadratic programs, the problem of verifying if such a point is a local solution is again $\mathcal{NP}$-complete.

6.2.4. A Practical Approach for Low Dimensions: Grid Search

The solution of a sequential OED problem describes the conditions under which the next experiment is performed. To be useful in practice, it often suffices if such a condition is just somewhat better than the available alternatives, like points from a factorial design or conditions selected based on expert knowledge. Therefore, it often suffices to use rough approximations to the actual global solution (like one of the many local solutions), particularly if the data-generating process is sufficiently complicated.

In Chap. 9, however, we aim to assess and compare the efficiency of different sequential design criteria for solving MD problems. To avoid arbitrariness, these conclusions should be based on *global* solutions, or at least on good approximations of it. In general, this requirement entails tremendous computational effort. The particular OED problems considered in Chap. 9, however, have low-dimensional experimental domains and can thus be addressed with GRID SEARCH, that is, with extensive sampling of the design criterion on the experimental domain.

A grid search approximates a global minimizer of the design criterion $\Psi : \mathcal{X} \mapsto \mathbb{R}$ over the whole experimental domain $\mathcal{X}$ by a global minimizer over a grid $\mathcal{G} \subset \mathcal{X}$. The grid is a finite set of points that "cover" the experimental domain $\mathcal{X}$ in some sense. To that end, the design criterion must be evaluated at all points in $\mathcal{G}$.

A grid $\mathcal{G}$ is EQUIDISTANT RECTANGULAR with distance $d \in \mathbb{R}^+$, iff

$$\mathcal{G} = \{x \in \mathcal{X} : \|x_0 - x\|_1 = nd, \text{ with } n \in \mathbb{N} \text{ and } x_0 \in \mathcal{X}\}. \tag{6.18}$$

Such a grid is finite if $\mathcal{X}$ is closed. In average, it contains $(1/d)^{n_x}$ points in the n_x-dimensional unit cube.

Grid search suffers from the "curse of dimensionality": The number of required grid points for sufficiently good approximations typically increases exponentially with n_x. The computational effort of evaluating the design criterion at these points may thus be intractable even for moderate dimensions.

Nevertheless, grid search has the advantage of a "global view" on the minimization problem. Under some regularity conditions, its approximations converge to the actual global solution as the number of grid points goes to infinity.

In addition, it provides these approximations without the need of a starting point, which is sometimes difficult to choose. In small dimensions (and only there), grid search has shown to be a simple and numerically robust approach for approximating a global solution.

We use grid search for solving the sequential OED problems in the case study considered in Chap. 9.

6.3. Low-Discrepancy Sequences

Low-discrepancy sequences are sequences whose members are placed highly evenly in space. Members of such a sequence are also called QUASI-RANDOM NUMBERS. Despite their name, there is nothing "random" about them. They are completely deterministic and can be generated algorithmically in a reproducible manner on a computer, like pseudo-random numbers. In contrast to the latter, however, quasi-random numbers typically fail tests for randomness and statistical independence. In many applications, however, being "random" is not the decisive feature:

> [...] instead of trying to cope with the impalpable concept of randomness, one should select points according to a deterministic scheme that is well suited for the problem at hand. (Niederreiter [192])

The outstanding feature of low-discrepancy sequences is that they are spread out highly uniformly in space, in fact, more uniformly than uniformly distributed pseudo-random numbers. Examples of this behavior can be seen in Fig. 6.1 on p. 207. This property makes quasi-random numbers attractive for several tasks:

Based on low-discrepancy sequences one can construct *experimental designs* that fill the design space evenly and integrate easily in sequential procedures. They are used to generate initial designs and model-independent reference designs in Chaps. 7 and 9.

In *numerical optimization*, the quality of the solution provided by local optimization strategies typically depends on some starting point. Low-discrepancy sequences can be used to determine promising start values in the absence of previous knowledge for choosing them. Low-discrepancy sequences are used in this manner for finding best parameters as described in Sec. 6.1 and applied in Chaps. 7 and 9.

After introducing the concept of DISCREPANCY, the relevant measure for the even-distributiveness of a sequence, we describe some popular low-discrepancy sequences. We shall see that under certain circumstances, particularly in high dimensions, they are not as evenly distributed as desired, and discuss techniques that have been proposed to improve their performance.

We restrict our considerations to those aspects of low-discrepancy sequences necessary to understand our numerical techniques. For more detailed information, we refer to the review article of Niederreiter [193] and the references given therein, and to the book of Niederreiter [192].

6.3.1. Discrepancy

The following is based on the works Niederreiter [192, Sec. 2] and Morokoff and Catflisch [189, Sec. 2]. Let λ denote the Lebesgue measure and let $\chi_{\mathcal{J}}: \mathcal{J} \mapsto \{0,1\}$ denote the indicator function on $\mathcal{J}$, defined as $\chi(x) := 1$, iff $x \in \mathcal{J}$ and $\chi(x) := 0$ otherwise. The sequence $(x_i : i \in \mathbb{N})$ taking values in the d-dimensional unit hypercube $\mathcal{I}^d$ is EQUIDISTRIBUTED, iff

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \chi_{\mathcal{J}}(x_i) = \lambda(\mathcal{J}) \tag{6.19}$$

for all Lebesgue-measurable subsets $\mathcal{J}$ of $\mathcal{I}^d$. In other words, the sequence is equidistributed if and only if the number of points falling into a measurable set is asymptotically proportional to the volume of that set. Any sequence of random numbers satisfying a strong law of large numbers is hence equidistributed with probability 1.

According to this definition, being equidistant tells us little about the behavior of *finite* sequences, which we encounter in practice. The counterpart of (6.19) for a *finite* subsequence $x_1, \ldots, x_n$ is

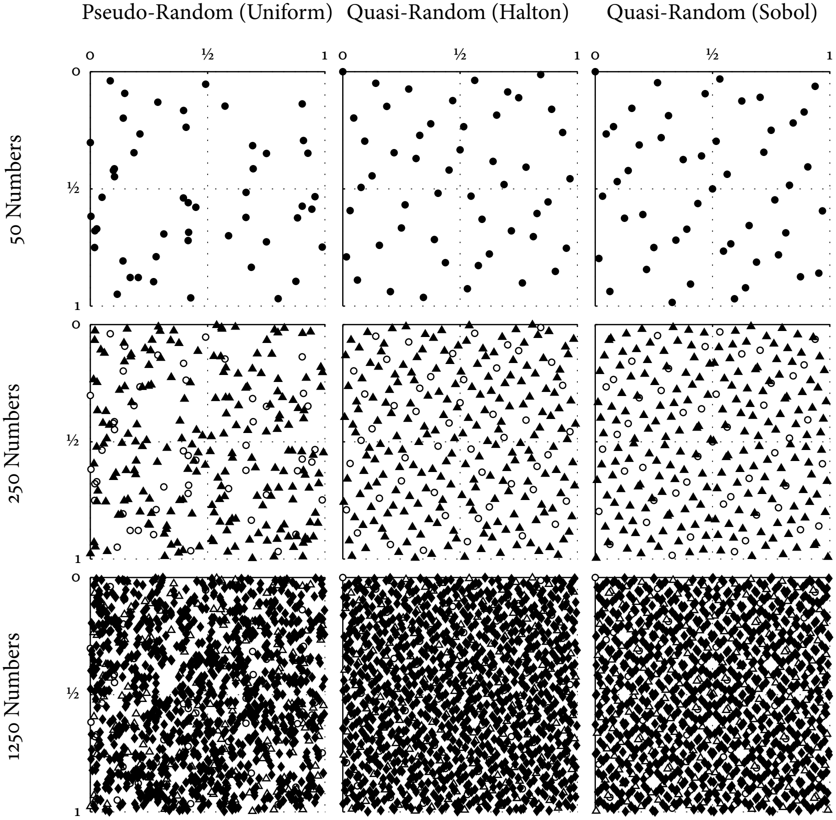
$$R_n(\mathcal{J}) := \frac{1}{n} \sum_{i=1}^{n} \chi_{\mathcal{J}}(x_i) - \lambda(\mathcal{J}), \tag{6.20}$$

defined for any Lebesgue-measurable set $\mathcal{J}$. How uneven the points of such a finite sequence are spread out in space can be measured by the DISCREPANCY and the STAR DISCREPANCY, defined as

$$D_n := \left| \sup_{\mathcal{J} \in \mathcal{A}} R_n(\mathcal{J}) \right| \text{ and } D_n^\star := \left| \sup_{\mathcal{J} \in \mathcal{A}^\star} R_n(\mathcal{J}) \right|, \tag{6.21}$$

respectively, where $\mathcal{A}$ is the set of all sub-hypercubes of $\mathcal{I}^d$, and $\mathcal{A}^\star$ the set

Figure 6.1.: Comparison of pseudo-random numbers and quasi-random numbers. Left column, top to bottom: first 50, 250 and 1250 members of a sequence of uniformly distributed two-dimensional pseudo-random numbers on the unit square, generated by the Mersenne twister algorithm of MATLAB with seed 1. Middle column: analogous members of a Halton sequence. Right column: analogous members of a Sobol sequence. Members 1–50 are represented by circles (○, ●), members 51–250 by triangles (△, ▲), and members 251–1250 by diamonds (◇, ◆). Members that are already present in preceding plot have "empty" markers (○, △, ◇).



of all sub-hypercubes of $\mathscr{I}^d$ having one corner at $0 \in \mathscr{I}^d$. It is easy to show that $D_n^\star \leqslant D_n \leqslant 2^n D_n^\star$. Other types of discrepancies can be defined likewise by restricting $\mathscr{J}$ to some class of subsets and taking a norm of R_n over this class. In general, the lower the discrepancy of a finite sequence, the more evenly are its points spread out over $\mathscr{I}^d$.

6.3.2. Low-Discrepancy Sequences

Here, we follow Niederreiter [192, Sec. 3] and Morokoff and Catflisch [189, Sec. 4]. The law of the iterated logarithms (Thm. B.7) implies that for a sequence of random numbers

$$D_n = \mathcal{O}\Big((\ln \ln n)^{½} / n^{½}\Big) \tag{6.22}$$

with probability 1. Halton [113] proved that for any dimension d there exist infinite sequences whose discrepancies satisfy

$$D_n = \mathcal{O}\Big((\ln n)^d / n\Big), \tag{6.23}$$

which is now regarded as the minimal asymptotic discrepancy possible for any infinite sequence. Halton's result is important since it shows that *there are in fact sequences whose points are more evenly spread out than uniformly distributed random numbers*. Such sequences, with an asymptotic discrepancy as in (6.23) are referred to as LOW-DISCREPANCY SEQUENCES. Note that such sequences have a low discrepancy *only asymptotically* – for any finite n, their discrepancy might well be above (6.23). We shall now describe some important representatives.

For all integers $p \geqslant 2$, any nonnegative integer n has a p-adic expansion $n = \sum_{i=0}^k c_i p^i$ with $0 \leqslant c_i \leqslant p$ for all $1 \leqslant i \leqslant k$. This expansion is unique except for summands with higher powers of p and coefficients of zero. The RADICAL INVERSE FUNCTION is

$$S_p(n) := \sum_{i=0}^k c_i p^{-i-1} = \frac{c_0}{p} + \frac{c_1}{p^2} + \ldots + \frac{c_k}{p^{k+1}}. \tag{6.24}$$

The definition implies that $S_p(n)$ takes values in $[0, 1)$ for all nonnegative n. It represents essentially a reflection at the decimal point: if the p-adic expansion of n is written as the string of digits $c_k c_{k-1} \ldots c_1$, then $\phi_p(n)$ is the p-adic fraction $0.c_0 c_1 \ldots c_k$. The sequence $S_p(1), S_p(2), \ldots$ is equidistributed.

Using the radical inverse function, the van der Corput sequence, introduced by van der Corput [253], can be expressed as $(S_2(i) : i \in \mathbb{N})$. It can be shown that both its discrepancy and its star discrepancy are of $\mathcal{O}((\ln n)/n)$.

Halton [113] generalized it to $d \geqslant 1$ dimensions. The Halton sequence is

$$(H_d(i) : i \in \mathbb{N}), \text{ where } H_d(i) := \big(S_{p_1}(i), \ldots, S_{p_d}(i)\big), \text{ for all } i \in \mathbb{N}, \tag{6.25}$$

where $p_1, \ldots, p_n$ are relatively prime integers, typically the first n primes. Its discrepancy satisfies

$$D_n^\star \leqslant \alpha_d \frac{(\ln n)^d}{n} + \mathcal{O}\left(\frac{(\ln n)^{d-1}}{n}\right), \tag{6.26}$$

where α_d is a dimension-dependent constant. A two-dimensional Halton sequence is shown in Fig. 6.1.

Also the family of Sobol sequences, due to Sobol [235], is based, at least indirectly, on p-adic expansions of the integers. Niederreiter [193] generalized this idea to the theory of so-called (t, s)-sequences. For details, we refer to the given original publications. Also Sobol sequences satisfy (6.26). An instance of a two-dimensional Sobol sequence is shown in Fig. 6.1 on p. 207.

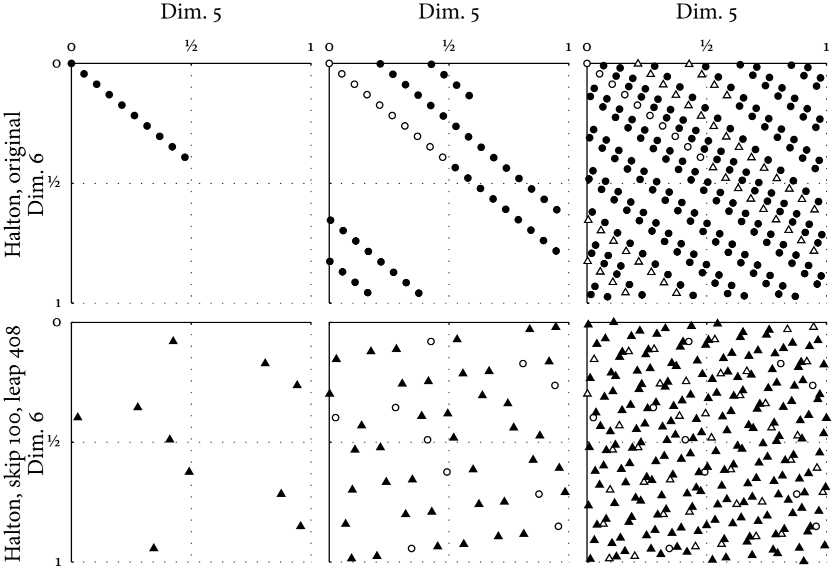
Implementing a generator for Halton sequences is conveniently simple. In fact, the algorithm proposed by Halton [112] requires less than a dozen pseudo-code statements. While the theory behind Sobol sequences is somewhat complex, the algorithms for their construction, for example those of Bratley and Fox [51] and Press et al. [205, Sec. 7.8], are surprisingly simple. Nowadays, implementations for both sequences are widely available in various programming languages, including Fortran 90, C, C++ and MATLAB.

6.3.3. Improving the Finite-Sample Discrepancy

The discrepancy bound (6.23) is asymptotic and does not necessarily describe the finite-sample behavior of a low-discrepancy sequence. For practical applications, however, the finite-sample behavior is decisive.

Halton sequences are notorious for their poor finite-sample performance, particularly in large dimensions, as discussed, for example, by Braaten and Weller [49] and Kocis and Whiten [146] and Morokoff and Catflisch [189]. The problem is shown in Fig. 6.2. The charts in the upper row show a 6-dimensional Halton

Figure 6.2.: Comparison of the original Halton sequence and a variant with skip and leap. Upper row, left to right: first 10, 50 and 250 members of the 6-dimensional original Halton sequence, orthogonal projection onto the 5th and 6th coordinate. Lower row: analogous members of the Halton sequence with skip 100 and leap 409. The markers have the same meaning as in Fig. 6.1 on p. 207.



sequence, orthogonally projected onto the 5th and 6th dimensions. Clearly, the points are far from being evenly distributed and seem to be highly correlated. Points are clustered in some areas, while others areas are. In general, the situation gets worse with increasing dimension.

This behavior of the Halton sequence is reflected in the discrepancy: several computations, for example those of Braaten and Weller [49], show that the discrepancy of the Halton sequence can even exceed that of a random sequence unless the number of points is sufficiently large. Based on their numerical results, Morokoff and Catflisch [189] estimate an exponential increase (6^d) in the number of points that are necessary before the discrepancy of a d-dimensional Halton sequence drops below the expected discrepancy of uniformly distributed random numbers.

The reason for the bad behavior of Halton sequences is well-understood.

Numerous variants of Halton sequences have been proposed as remedy, for example by Braaten and Weller [49], Chi and Jones [68], Kocis and Whiten [146], Matoušek [182], Morokoff and Catflisch [189], and Owen [196] and Faure and Lemieux [90], to mention a few. The performance of several of them for quasi Monte Carlo integration is compared in the numerical studies of Faure and Lemieux [90] and Schlier [222]. Without going into detail one can say that improved Halton sequences exists which do not, or at least only to a much lesser degree, exhibit the problems of the original Halton sequence.

Kocis and Whiten [146] propose a particularly simple and attractive variant: as i-th member of the sequence, choose the iL-th member of the original Halton sequence (6.25), where the leap L is a prime that is different from all used bases $p_1, \ldots, p_d$. The numerical results of Kocis and Whiten [146] suggest that this "Halton sequence leaped" performs significantly better than the original Halton sequence, at least for dimensions up to 400. It was was pointed out by Matoušek [182, Sec. 4] and Morokoff and Catflisch [189, Sec. 7] that the quality of a Halton sequence can be strongly improved by skipping some of its initial members. The Halton sequence with leap L and skip K is hence

$$(H'_d(i) : i \in \mathbb{N}), \text{ where } H'_d(i) := H_d(K + iL), \text{ for all } i \in \mathbb{N} \qquad (6.27)$$

The improvement of this sequence compared to the original Halton sequence can be clearly seen in Fig. 6.2.

7. Performance of Classic and Robust PMLE Covariance Approximations: Theory and Numerical Results

Contents

This chapter derives statistical measures and efficient algorithms for assessing and comparing empirical approximations for the covariance of a parameter maximum-likelihood estimators (PMLES).

Section 7.1 formally states the problem, Sec. 7.4 derives statistical measures for assessing and comparing empirical approximations for the covariance of a PMLES. Section 7.3 develops efficient algorithms for computing these measures and describes the implementation provided in the software package DoeSim.

Section 7.4 describes a model family for the water-gas shift reaction (WGSR) reaction, which is used for a case study. Section 7.5 describes the numerical results from a case study that compares the classic empirical PMLEs covariance approximation to its misspecification-robust alternative proposed in Sec. 3.4.

7.1. Problem Statement

The assumptions and concepts of this chapter are similar to those considered in Secs. 3.1 to 3.3. We use the same notation and terminology, with some simplifications to increase the readability.

7.1.1. Central Assumptions

Throughout the chapter we make the following assumptions.

(i) The OBSERVATIONS $y_1, y_2, \ldots$ from the compact OBSERVATION DOMAIN $\mathcal{Y} \subseteq \mathbb{R}^{n_y}$ are available, resulting from experiments performed under the known CONDITIONS $x_1, x_2, \ldots$, respectively, from the EXPERIMENTAL DOMAIN $\mathcal{X} \subseteq \mathbb{R}^{n_x}$.

For all $n \in \mathbb{N}$, we summarize the observations of the first n experiments in the DATA vector $d_n^\top := \begin{bmatrix} y_1^\top & \ldots & y_n^\top \end{bmatrix} \in \mathcal{Y}^n$.

(ii) The observations $y_1, y_2, \ldots$ are realizations of the respective OBSERVABLES $\mathcal{Y}_1, \mathcal{Y}_2, \ldots$, continuous $\mathcal{Y}$-valued random variables.

Accordingly, any vector of data d_n with $n \in \mathbb{N}$ is a realization of the continuous $\mathcal{Y}^n$-valued random variable $\mathcal{D}_n^\top := \begin{bmatrix} \mathcal{Y}_1^\top & \ldots & \mathcal{Y}_n^\top \end{bmatrix}$, the SAMPLE.

(iii) The observables $\mathcal{Y}_1, \mathcal{Y}_2, \ldots$ are statistically independent. For all $n \in \mathbb{N}$, observable $\mathcal{Y}_n$ is normally distributed with mean $\bar{\eta}(x_n) := \mathbb{E}[\mathcal{Y}_n]$ and full-rank (and thus invertible) covariance matrix $\boldsymbol{\Omega}(x_n) := \mathbb{C}[\mathcal{Y}_n]$.

(iv) A regression model (Def. 1.3) with a compact PARAMETER DOMAIN $\mathcal{Q} \subseteq \mathbb{R}^{n_\theta}$ is available. For all $n \in \mathbb{N}$, the model with parameter $\theta \in \mathcal{Q}$ specifies a normal distribution for observable $\mathcal{Y}_n$ with mean $\eta(x_n, \theta)$ and covariance $\boldsymbol{\Omega}(x_n)$. The RESPONSE $\eta(x_n, \theta)$ is twice continuously differentiable with respect to θ for all $n \in \mathbb{N}$.

Note that assumption (iv) implies the assumption that the observation covariances from assumption (iii) are known to whoever specifies the model.

7.1.2. Notation and Definitions

We use the following notation and definitions for all experiments $n \in \mathbb{N}$ and all parameters $\theta \in \mathcal{Q}$. The gradient and the Hessian differential operator with respect to θ are denoted ∇_θ and ∇_θ^2, respectively. We write

$$\boldsymbol{J}(x_n, \theta) := \nabla_\theta \eta(x_n, \theta) \tag{7.1}$$

for the $n_y \times n_\theta$ Jacobian matrix of the response and

$$\boldsymbol{H}_j(x_n, \theta) := \nabla_\theta^2 \eta_j(x_n, \theta) \tag{7.2}$$

for the $n_\theta \times n_\theta$ Hessian of the j-th response component $\eta_j(x_n, \theta)$. Based thereon, we define the symmetric positive semi-definite (SPSD) $n_\theta \times n_\theta$ matrix

$$\boldsymbol{M}_n(\theta) := \tfrac{1}{n} \sum_{i=1}^{n} \boldsymbol{J}^\top(x_i, \theta) \boldsymbol{\Omega}^{-1}(x_i) \boldsymbol{J}(x_i, \theta) \tag{7.3}$$

and its inverse $\boldsymbol{C}_n(\theta) := \boldsymbol{M}_n^{-1}(\theta)$, supposed that it exists. Further, we define the symmetric $n_\theta \times n_\theta$ matrix

$$\boldsymbol{N}_n(\theta, d_n) := \tfrac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n_y} \tilde{r}_j(x_i, \theta) \tilde{\boldsymbol{H}}_j(x_i, \theta), \tag{7.4}$$

where $\tilde{r}_j(\cdot)$ is the j-th component of the vector

$$\tilde{r}(x_i, \theta) := \boldsymbol{\Omega}^{-1/2}(x_i)(\eta(x_i, \theta) - y_i), \tag{7.5}$$

and

$$\tilde{\boldsymbol{H}}_j(x_i, \theta) := \sum_{k=1}^{n_y} \rho_{jk}(x_i) \boldsymbol{H}_k(x_i, \theta), \tag{7.6}$$

where $\rho_{jk}(x_i)$ is the component of matrix $\boldsymbol{\Omega}^{-1/2}(x_i)$ in row j and column k. The matrices defined in (7.3) and (7.4) are straightforward generalizations of their counterparts from Chap. 3 to the case of non-unit observation covariances $\boldsymbol{\Omega}(x_i)$, see Tab. 3.1 on p. 113.

7.1.3. Empirical Approximations for PMLE Covariance

Let us summarize some results from Secs. 3.1 to 3.4. Suppose assumptions (i)–(iv) hold. Then, a parameter maximum-likelihood estimate (PMLE) $\hat{\theta}_n := \hat{\theta}_n(d_n) \in \mathcal{Q}$ based on the first n experiments minimizes the sum of squared residuals (SSR)

$$s_n(\theta, d_n) := \tfrac{1}{n} \sum_{i=1}^{n} \| \eta(x_i, \theta) - y_i \|^2_{\Omega^{-1}(x_i)} \tag{7.7}$$

with respect to $\theta \in \mathcal{Q}$, see Def. 3.9, Cor. 3.10, and Tab. 3.1. The corresponding estimator $\hat{\theta}_n(\mathcal{D}_n)$ is a continuous $\mathcal{Q}$-valued random variable. We write

$$\boldsymbol{Q}_n := \mathbb{C}\left[\hat{\theta}_n(\mathcal{D}_n)\right] \tag{7.8}$$

for its ACTUAL COVARIANCE. In practice, the distribution of the sample $\mathcal{D}_n$ and thus also $\boldsymbol{Q}_n$ are typically *unknown*. If data is available, empirical (=data-based) approximations can be formulated for $\boldsymbol{Q}_n$. Under the given assumptions, the CLASSIC (EMPIRICAL) APPROXIMATION for the actual covariance is

$$\boldsymbol{Q}_n \approx \tfrac{1}{n} \boldsymbol{C}_n\left(\hat{\theta}_n\right). \tag{7.9}$$

It rests upon the assumption that the model is (a) correct or (b) locally affine-linear around the PMLE $\hat{\theta}_n$. If it is satisfied and certain regularity conditions are met, the error of this approximation error gets arbitrarily small (in a probabilistic sense) as the sample size n increases.

As alternative, we proposed the novel ROBUST (EMPIRICAL) APPROXIMATION

$$\begin{aligned} \boldsymbol{Q}_n &\approx \tfrac{1}{n} \boldsymbol{R}_n\left(\hat{\theta}_n, d_n\right) \\ &:= \tfrac{1}{n}\left(\boldsymbol{M}_n\left(\hat{\theta}_n\right) + \boldsymbol{N}_n\left(\hat{\theta}_n, d_n\right)\right)^{-1} \boldsymbol{M}_n\left(\hat{\theta}_n\right)\left(\boldsymbol{M}_n\left(\hat{\theta}_n\right) + \boldsymbol{N}_n\left(\hat{\theta}_n, d_n\right)\right)^{-1} \end{aligned} \tag{7.10}$$

in Conj. 3.14. It is a consistent generalization of its classic counterpart. For correct or affine-linear models, $\boldsymbol{R}_n\left(\hat{\theta}_n, d_n\right) = \boldsymbol{C}_n\left(\hat{\theta}_n\right)$, so that both approximations are identical. Yet even for models that are both nonlinear and incorrect, the error of the robust approximation gets arbitrarily small (in a probabilistic sense) with increasing sample size n, supposed certain regularity conditions are met.

The classic empirical approximation involves first derivatives of the model responses, its robust counterpart requires also second derivatives. Evaluating the former is therefore typically significantly cheaper than evaluating the latter.

7.1.4. Key Questions

Suppose the considered model is both nonlinear and incorrect. It is difficult to decide in practice whether the additional effort for evaluating the robust approximation is justified by its increased approximation quality, or if the cheaper, yet less precise classic approximation suffices. Furthermore, the error of both approximations tends to decrease with the sample size n, yet it is difficult to predict how large it has to be to reduce it to a practically acceptable level. In the remaining chapter we consider the following questions:

(Q7.1) How good are the classic approximation (7.9) and its robust counterpart (7.10) depending on the *amount* n of available data?

(Q7.2) How good are they depending on the *variability* of the data in terms of the covariance $\mathbb{C}\left[\mathcal{Y}_n\right]$?

Section 7.2 introduces measures for the quality of covariance approximations in general, and Sec. 7.3 deals with methods for their computation. Based thereon, (Q7.1) and (Q7.2) are studied for several different models in Sec. 7.4.

7.2. Quality of Empirical Approximations for PMLE Covariances

To examine the key questions, measures for assessing and comparing the quality of the relevant covariance approximations are required. Such measure are derived and discussed in the following.

7.2.1. Metrics for Covariance Matrices

The dissimilarity of two real-valued $m \times n$ matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ can be measured in terms of any matrix metric $d\colon \mathbb{R}^{m\times n} \times \mathbb{R}^{m\times n} \mapsto \mathbb{R}_0^+$, which by definition

(a) is positive definite, $d(\boldsymbol{A}, \boldsymbol{B}) \geqslant 0$, with equality iff $\boldsymbol{A} = \boldsymbol{B}$,

(b) is symmetric, $d(\boldsymbol{B}, \boldsymbol{A}) = d(\boldsymbol{A}, \boldsymbol{B})$, and

(c) satisfies the triangle inequality, $d(\boldsymbol{A}, \boldsymbol{B}) \leqslant d(\boldsymbol{A}, \boldsymbol{C}) + d(\boldsymbol{C}, \boldsymbol{B})$,

where $\boldsymbol{C}$ is also a real-valued $m \times n$ matrix. The larger the value of d, the more dissimilar the matrices. Any matrix norm $\|\cdot\|$ (for an overview, see, for example, Horn and Johnson [121, Chap. 5]) induces such a metric by $d(\boldsymbol{A}, \boldsymbol{B}) := \|\boldsymbol{A} - \boldsymbol{B}\|$.

Let a_{ij} and b_{ij} be the components in row i and column j of matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, respectively. A straightforward choice is the metric

$$d_{\mathrm{F}}(\boldsymbol{A},\boldsymbol{B}) := \frac{\|\boldsymbol{A}-\boldsymbol{B}\|_{\mathrm{F}}}{\sqrt{mn}} = \frac{1}{\sqrt{mn}}\left(\sum_{i=1}^{m}\sum_{j=1}^{n}\left(a_{ij}-b_{ij}\right)^{2}\right)^{½}, \tag{7.11}$$

a scaled variant of the metric induced by the Frobenius norm. The normalizing factor $1/\sqrt{mn}$ ensures that $d_{\mathrm{F}}(\cdot,\cdot)$ is $\mathcal{O}(1)$ with respect to both m and n, simplifying comparisons between matrices of different dimensions. This metric and similar ones induced by matrix norms are applicable even to non-symmetric positive definite (spd) and even to non-square matrices.

We are, however, interested in the special case that $\boldsymbol{A}$ and $\boldsymbol{B}$ are full-rank covariance matrices of parameter estimators or approximations thereof. The following metric respects some of the particular properties of such matrices.

Definition 7.1 (Riemannian Metric for SPD Matrices)

Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be real-valued spd $m \times m$ matrices and let $\lambda_i(\boldsymbol{A},\boldsymbol{B})$, $1 \leqslant i \leqslant m$, be the eigenvalues of $\boldsymbol{A}^{-1}\boldsymbol{B}$, or equivalently the inverse eigenvalues of $\boldsymbol{B}^{-1}\boldsymbol{A}$. The (normalized) Riemannian metric is

$$d_{\mathrm{R}}(\boldsymbol{A},\boldsymbol{B}) := \frac{1}{\sqrt{m}}\left\|\ln\left(\boldsymbol{A}^{-1}\boldsymbol{B}\right)\right\|_{\mathrm{F}} = \frac{1}{\sqrt{m}}\left(\sum_{i=1}^{m}(\ln\lambda_i(\boldsymbol{A},\boldsymbol{B}))^{2}\right)^{½}, \tag{7.12}$$

where $\ln(\cdot)$ denotes the matrix logarithm in the middle term and the usual logarithm in the last term.

The Riemannian metric measures the *relative* dissimilarity of two matrices on a logarithmic scale. In the univariate case $m = 1$ with $\boldsymbol{A} = a \in \mathbb{R}$ and $\boldsymbol{B} = b \in \mathbb{R}$, it simplifies to $d_{\mathrm{R}}(\boldsymbol{A},\boldsymbol{B}) = \|\ln(b/a)\|_2 = \|\ln(b) - \ln(a)\|_2$. The normalizing factor $m^{-½}$ in (7.12) ensures that $d_{\mathrm{R}}(\cdot,\cdot)$ is of $\mathcal{O}(1)$ with respect to the matrix dimension m, simplifying comparisons. Lang [161, Chap. XII, § 1], Förstner and Moonen [102] and Moakher and Batchelor [187, Sec. 17.2.1] treat this metric (without the factor $m^{-½}$) in more detail.

Theorem 7.2 (Properties of the Riemannian Metric for SPD Matrices)

The Riemannian metric d_R from Def. 7.1

(i) satisfies the characteristics (a) to (c) on p. 217 of a matrix metric,

(ii) is invariant with respect to affine transformations, meaning that $d_{\mathrm{R}}(\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^{\top}, \boldsymbol{X}\boldsymbol{B}\boldsymbol{X}^{\top}) = d_{\mathrm{R}}(\boldsymbol{A}, \boldsymbol{B})$ for all real-valued invertible $m \times m$ matrices $\boldsymbol{X}$, and

(iii) is invariant under inversion, $d(\boldsymbol{A}, \boldsymbol{B}) = d\left(\boldsymbol{A}^{-1}, \boldsymbol{B}^{-1}\right)$.

Proof A proof is given by Förstner and Moonen [102, Thm. 1]. □

Since $\mathbb{C}\left[\boldsymbol{X}\mathcal{Q} + a\right] = \boldsymbol{X}\,\mathbb{C}\left[\mathcal{Q}\right]\boldsymbol{X}^{\top}$ for any $\mathbb{R}^m$-valued random variable $\mathcal{Q}$ and any vector $a \in \mathbb{R}^m$, property (ii) ensures that d_R is invariant under reparameterizations of the type $\theta \mapsto \boldsymbol{X}\theta + a$. Essentially, this property provides a certain independence of d_R from details of the model implementation.

A full-rank covariance matrix is SPD by definition and thus has a unique inverse. This inverse, sometimes called “precision matrix,” carries exactly the same information concerning the variability of the underlying distribution as the covariance matrix itself. Property (iii) ensures that it does not matter whether covariance matrices or precision matrices are regarded in d_R.

In the general case, the metric d_{F} has neither property (ii) nor property (iii). Therefore, we use the Riemannian metric to measure the dissimilarity of covariance matrices.

7.2.2. Quality Measures

Using the Riemannian metric, the quality of the classic approximation (7.9) for a given model can be measured by $d_{\mathrm{R}}\left(\boldsymbol{Q}_n, \frac{1}{n}\boldsymbol{C}_n\left(\hat{\theta}_n(d_n)\right)\right)$, where smaller values correspond to better approximations. From this quantity, however, little can be inferred about the quality of the classic approximation *in general,* since the data d_n is subject to random fluctuations described by the distribution of $\mathcal{D}_n$.

A measure for the general quality of the classic approximation should take into account the distribution of the random variable

$$\Delta_{\mathrm{cl}}(n) := d_{\mathrm{R}}\left(\boldsymbol{Q}_n, \tfrac{1}{n}\boldsymbol{C}_n\left(\hat{\theta}_n(\mathcal{D}_n)\right)\right), \tag{7.13}$$

and a corresponding measure for the robust approximation should take into account the distribution of

$$\Delta_{\mathrm{rob}}(n) := d_{\mathrm{R}}\left(\boldsymbol{Q}_n, \tfrac{1}{n}\boldsymbol{R}_n\left(\hat{\theta}_n(\mathcal{D}_n), \mathcal{D}_n\right)\right). \tag{7.14}$$

We use the expected values $\mathbb{E}\left[\Delta_{\mathrm{cl}}(n)\right]$ and $\mathbb{E}\left[\Delta_{\mathrm{rob}}(n)\right]$ as measures for the average approximation error, and analogously the corresponding standard deviations $\mathbb{C}\left[\Delta_{\mathrm{cl}}(n)\right]^{½}$ and $\mathbb{C}\left[\Delta_{\mathrm{rob}}(n)\right]^{½}$ as measures for their variability. For comparing the quality of both approximations we consider the random variable

$$\Delta(n) := \Delta_{\mathrm{rob}}(n) - \Delta_{\mathrm{cl}}(n) \tag{7.15}$$

which takes value on the whole real line. The smaller (more negative) its expected value $\mathbb{E}\left[\Delta(n)\right]$, the better is the robust approximation in average compared to its classic counterpart. The standard deviation $\mathbb{C}\left[\Delta(n)\right]^{½}$ quantifies the associated variability.

We use these expected values and standard deviations to assess and compare the quality of the considered covariance approximations. Monitoring these quantities under increasing sample size $n \in \mathbb{N}$ allows to examine the influence of the amount of available data on to approximation quality (Q7.1). Observing them under data with different covariances makes it possible to study the effect of the variability of the data on the approximation quality (Q7.2).

7.3. Computational Methods

To examine (Q7.1) and (Q7.2) under controlled conditions, we *define* the distribution of the sample $\mathcal{D}_n$. Via the functional dependencies (7.13) and (7.14), this choice also determines the distributions of $\Delta_{\mathrm{cl}}(n)$ and $\Delta_{\mathrm{rob}}(n)$, respectively. Their expected values and standard deviations can typically not be represented in a closed form, but can be approximated computationally of replications of the data are available.

Suppose experiments 1 to n have been replicated $r \in \mathbb{N}$ times, and let $d_{n1}, \ldots, d_{nr}$ be the corresponding replicated data, independently and identically distributed (IID) realizations of the sample $\mathcal{D}_n$. The corresponding parameter maximum-likelihood estimates (PMLEs) $\hat{\theta}_{n1}, \ldots, \hat{\theta}_{nr}$ are then IID realizations of the estimator $\hat{\mathcal{Q}}_n := \hat{\theta}_n(\mathcal{D}_n)$.

7.3.1. Replication-Based Approximations of PMLE Covariance

Let us first consider two classes of approximations for the actual covariance $\boldsymbol{Q}_n$ that are based on the replicated estimates $\hat{\theta}_{n1}, \ldots, \hat{\theta}_{nr}$.

Sample Covariance Matrix

It is well known that the sample covariance

$$\tilde{\mathbf{Q}}_{nr} := \tfrac{1}{r-1}\sum_{l=1}^{r}\big(\hat{\theta}_{nl}-\bar{\theta}_{nr}\big)\big(\hat{\theta}_{nl}-\bar{\theta}_{nr}\big)^{\top}, \text{ with } \bar{\theta}_{nr} := \tfrac{1}{r}\sum_{l=1}^{r}\hat{\theta}_{nl},$$

consistently estimates $\mathbf{Q}_n$: the larger r, the better (in a probabilistic sense) is the approximation $\tilde{\mathbf{Q}}_{nr} \approx \mathbf{Q}_n$. The quality of this approximation is, however, very sensitive to the presence of outliers. That is, it tends to suffer significantly if some of the realizations $\hat{\theta}_{n1}, \ldots, \hat{\theta}_{nr}$ are "far-off" from the bulk of the others.

The probability of obtaining outliers in r realizations drops with increasing r. For any given covariance $\mathbf{Q}_n$, the quality (in a probabilistic sense) of approximation $\tilde{\mathbf{Q}}_{nr} \approx \mathbf{Q}_n$ can thus be improved by increasing r. Unfortunately, the probability of obtaining outlier also increases sharply with the magnitude of $\mathbf{Q}_n$. Even for covariances of moderate magnitude, the number of realizations r required to ensure a given approximation quality might hence be very large, as discussed by Gupta and Gupta [110] and others.

Preliminary computations revealed that using the sample covariance for approximating $\mathbf{Q}_n$ requires replication numbers r that are practically intractable. We therefore use a "robust" alternative that is less susceptible to outliers.

MCD Covariance Estimator

A covariance estimate that is more robust than the sample covariance with respect to outliers can be obtained from the minimum covariance determinant (mcd) method, which was introduced by Rousseeuw [215] and Rousseeuw [217] and recently reviewed by Hubert and Debruyne [127] and Hubert, Rousseeuw, and Van Aelst [128].

Given an integer number $h \leqslant r$, the mcd method determines h of the r realizations $\hat{\theta}_{n1}, \ldots, \hat{\theta}_{nr}$ whose sample covariance matrix has minimal determinant. The mcd estimate $\hat{\mathbf{Q}}_{nr}$ for $\mathbf{Q}_n$ is then a multiple of the sample covariance of these h realizations.

As shown by Butler, Davies, and Jhun [60], the mcd covariance estimator is consistent and converges to $\mathbf{Q}_n$ with a rate of $\mathcal{O}\big(r^{-1/2}\big)$. It is robust in the sense that up to $r-h$ replicates may be arbitrarily "far-off" from the bulk of the remaining ones without affecting the value of the estimate.

Rousseeuw and Van Driessen [216] describe an efficient algorithm (called "FAST-MCD") for the computation of the mcds covariance estimate. An implementation in MATLAB is available in the package LIBRA, described by Verboven and Hubert [257, 258].

7.3.2. Replication-Based Quality Measures

Given a replication-based estimate $\hat{\boldsymbol{Q}}_{nr} := \hat{\boldsymbol{Q}}_{nr}\big(\hat{\theta}_{n1}, \dots, \hat{\theta}_{nr}\big)$ for the actual covariance $\boldsymbol{Q}_n$ one can compute

$$\delta^l_{\mathrm{cl}}(n) := d_{\mathrm{R}}\Big(\hat{\boldsymbol{Q}}_{nr}, \tfrac{1}{n}\boldsymbol{C}_n\big(\hat{\theta}_{nl}\big)\Big) \text{ for all } l \in \{1, \dots, r\}. \tag{7.16}$$

If the covariance estimator is consistent, then $\hat{\boldsymbol{Q}}_{nr} \approx \boldsymbol{Q}_n$ for large r, so that $\delta^1_{\mathrm{cl}}(n), \dots, \delta^r_{\mathrm{cl}}(n)$ are *approximately* IID realizations of the random variable $\Delta_{\mathrm{cl}}(n)$. Under regularity conditions, the weak law of large numbers (see Def. B.5 and Thm. B.6) then provides the approximations

$$\mathbb{E}\left[\Delta_{\mathrm{cl}}(n)\right] \approx \bar{\delta}_{\mathrm{cl}}(n) := \tfrac{1}{r}\sum_{l=1}^{r} \delta^l_{\mathrm{cl}}(n) \text{ and} \tag{7.17}$$

$$\mathbb{C}\left[\Delta_{\mathrm{cl}}(n)\right]^{½} \approx \sigma_{\mathrm{cl}}(n) := \left(\tfrac{1}{r-1}\sum_{l=1}^{r}\big(\delta^l_{\mathrm{cl}}(n) - \bar{\delta}_{\mathrm{cl}}(n)\big)^2\right)^{½} \tag{7.18}$$

for large r. Analogously, the sample mean $\bar{\delta}_{\mathrm{rob}}(n)$ and the sample standard deviation $\sigma_{\mathrm{rob}}(n)$ of

$$\delta^l_{\mathrm{rob}}(n) := d_{\mathrm{R}}\Big(\hat{\boldsymbol{Q}}_{nr}, \tfrac{1}{n}\boldsymbol{R}_n\big(\hat{\theta}_{nl}, d_n\big)\Big), \text{ with } l \in \{1, \dots, r\}, \tag{7.19}$$

can approximate $\mathbb{E}\left[\Delta_{\mathrm{rob}}(n)\right]$ and $\mathbb{C}\left[\Delta_{\mathrm{rob}}(n)\right]^{½}$, respectively, and the sample mean $\bar{\delta}(n)$ and the sample standard deviation $\sigma(n)$ of

$$\delta^l(n) := \delta^l_{\mathrm{rob}}(n) - \delta^l_{\mathrm{cl}}(n) \text{ with } l \in \{1, \dots, r\}, \tag{7.20}$$

can approximate $\mathbb{E}\left[\Delta(n)\right]$ and $\mathbb{C}\left[\Delta(n)\right]^{½}$, respectively.

Approximations (7.17) and (7.18) rely on the weak law of large numbers, which is classically proven assuming statistically independent random variables. The quantities $\delta^1_{\mathrm{cl}}(n), \dots, \delta^r_{\mathrm{cl}}(n)$, however, are correlated since any of them is affected by *all* PMLES $\hat{\theta}_{n1}, \dots, \hat{\theta}_{nr}$ via the covariance estimate $\hat{\boldsymbol{Q}}_{nr}$. The amount of this correlation, however, is small for large r, because the influence of any particular PMLE $\hat{\theta}_{nl}$ on the covariance estimate $\hat{\boldsymbol{Q}}_{nr}$ quickly decreases with r. In fact, the (strong or weak) law of large numbers holds even for such dependent, but weakly correlated random variables. This case is comprised in the very general sufficient conditions for the strong law of large numbers in dependent random variables provided by Hu, Rosalsky, and Volodin [124] and Kuczmaszewska [154]. Therefore, (7.17) and (7.18) can be expected to remain valid for large r despite of the mentioned statistical dependencies. The same holds likewise for the corresponding quantities related to $\Delta_{\text{rob}}(n)$ and $\Delta(n)$.

7.3.3. A Monte Carlo Method

Algorithm 7.1 on the next page computes $\bar{\delta}_{\text{cl}}(n)$, $\sigma_{\text{cl}}(n)$, $\bar{\delta}_{\text{rob}}(n)$, $\sigma_{\text{rob}}(n)$, $\bar{\delta}(n)$, and $\sigma(n)$ for all experiments $n \in \{n_\theta, \ldots, n_{\max}\}$, where $n_\theta \in \mathbb{N}$ is the number of parameters in the model and $n_{\max} \geqslant n_\theta$ is a predefined maximum number of experiments. The algorithm is essentially a Monte Carlo (MC) method for the expectations and standard deviations of $\Delta_{\text{cl}}(n)$, $\Delta_{\text{rob}}(n)$, and $\Delta(n)$. We comment on some of its characteristics.

For computing a PMLE $\hat{\theta}_{nl}$, one needs to solve a least-squares (LSQ) problem. Since the considered model may both nonlinear and incorrect, the LSQ problem may also be nonlinear and may exhibit large residuals even in the solution. Suitable numerical methods are discussed in Sec. 6.1. The evaluation of $\hat{\boldsymbol{C}}_n$ and $\hat{\boldsymbol{R}}_{nl}$ requires first and second derivatives, respectively, of the model response.

The algorithm may be computationally demanding. To give some typical numbers, applying the algorithm $n_{\max} = 100$ experiments and $r = 10000$ MC runs involves the solution of approximately 1 million LSQ problems. Solving such a number of problems may take a considerable amount of time. Fortunately, the algorithm can be parallelized in large parts, which allows to reduce computation times on todays multi-processor multi-core hardware. In particular, the individual runs of all **foreach**-loops can be run concurrently, which includes the potentially expensive solutions of the LSQ problems.

Implementation in DoeSim

We implemented a variant of Alg. 7.1 in our software package DoeSim.

For generating observations, the implementation uses MATLAB's `mrg32k3a` pseudo-random number generator, which combines the 32-bit combined multiple recursive generator of L'Ecuyer [167] and the Ziggurat algorithm of Marsaglia and Tsang [180].

For solving LSQ problems, it applies the Broyden-Fletcher-Goldfarb-Shanno

Algorithm 7.1: Monte Carlo method for comparing classic and robust PMLE covariance approximations.

```
input  : a model with $n_\theta$ parameters satisfying assumption (iv) on p. 214
         experimental conditions $x_1, \dots, x_{n_\text{max}} \in \mathscr{X}$, with $n_\text{max} \geqslant n_\theta$
         observables $\mathcal{Y}_1, \dots, \mathcal{Y}_{n_\text{max}}$ satisfying assumption (iii) on p. 214
         number $r \geqslant n_\theta$ of Monte Carlo simulations
output : mean errors and variabilities $\bar{\delta}_\text{cl}(n)$, $\bar{\delta}_\text{rob}(n)$, $\sigma_\text{cl}(n)$, $\sigma_\text{rob}(n)$ $\bar{\delta}(n)$, and
         $\sigma(n)$, with $n \in \{n_\theta, \dots, n_\text{max}\}$

foreach $n \in \{1, \dots, n_\text{max}\}$ do
    generate observations $y_{n1}, \dots, y_{nr}$, independent realizations of $\mathcal{Y}_n$;
end
foreach $n \in \{n_\theta, \dots, n_\text{max}\}$ do
    foreach $l \in \{1, \dots, r\}$ do
        $d_{nl}^\top \leftarrow \begin{bmatrix} y_{1l}^\top & \dots & y_{nl}^\top \end{bmatrix}$;
        $\hat{\theta}_{nl} \leftarrow \operatorname{argmin}_{\theta \in \mathscr{Q}} s_n(\theta, d_{nl})$ ;        // see (7.7)
        $\hat{\boldsymbol{C}}_{nl} \leftarrow \boldsymbol{C}_n\big(\hat{\theta}_{nl}\big)$ ;        // see (7.9)
        $\hat{\boldsymbol{R}}_{nl} \leftarrow \boldsymbol{R}_n\big(\hat{\theta}_{nl}, d_{nl}\big)$ ;        // see (7.10)
    end
    compute covariance estimate $\hat{\mathbf{Q}}_{nr}$ from $\hat{\theta}_{n1}, \dots, \hat{\theta}_{nr}$;
    foreach $l \in \{1, \dots, r\}$ do
        $\delta_\text{cl}^l(n) \leftarrow d_\text{R}\big(\hat{\mathbf{Q}}_{nr}, \frac{1}{n}\hat{\boldsymbol{C}}_{nl}\big)$ ;        // see (7.16)
        $\delta_\text{rob}^l(n) \leftarrow d_\text{R}\big(\hat{\mathbf{Q}}_{nr}, \frac{1}{n}\hat{\boldsymbol{R}}_{nl}\big)$ ;        // see (7.19)
        $\delta^l(n) \leftarrow \delta_\text{rob}^l(n) - \delta_\text{cl}^l(n)$ ;        // see (7.20)
    end
    determine mean $\bar{\delta}_\text{cl}(n)$ and std. dev. $\sigma_\text{cl}(n)$ of $\delta_\text{cl}^1(n), \dots, \delta_\text{cl}^r(n)$ ;        // see (7.17),(7.18)
    determine mean $\bar{\delta}_\text{rob}(n)$ and std. dev. $\sigma_\text{rob}(n)$ of $\delta_\text{rob}^1(n), \dots, \delta_\text{rob}^r(n)$;
    determine mean $\bar{\delta}(n)$ and std. dev. $\sigma(n)$ of $\delta^1(n), \dots, \delta^r(n)$;
end
return $\bar{\delta}_\text{cl}(n)$, $\sigma_\text{cl}(n)$, $\bar{\delta}_\text{rob}(n)$, $\sigma_\text{rob}(n)$, $\bar{\delta}(n)$, and $\sigma(n)$, with $n \in \{n_\theta, \dots, n_\text{max}\}$
```

(BFGS) [53, 100, 107, 228] quasi-Newton method provided by the MATLAB function `fminunc`.

First derivatives of the model response – required by the BFGS method and for evaluating $\hat{\boldsymbol{C}}_{nl}$ – are computed in machine precision using the complex step differentiation introduced by Lyness and Moler [178], reviewed by Martins, Sturdza, and Alonso [181]. For computing the second derivatives required for evaluating $\hat{\boldsymbol{R}}_{nl}$, this technique is combined with finite central differences.

As replication-based estimate $\hat{\boldsymbol{Q}}_{nl}$ for the actual covariance (line 11), the implementation adopts the MCD method of Rousseeuw [215] and Rousseeuw [217], available in MATLAB through the function `mcdcov` provided by the package LIBRA [257, 258].

The implementation is parallelized with respect to solving the LSQ problems and evaluating the covariance approximations (lines 5–10 in Alg. 7.1). Also the matrix metrics (lines 12–16) are be evaluated concurrently.

7.4. Water-Gas Shift Reaction (WGSR)

For our case studies in this and the next chapter we consider the water-gas shift reaction (WGSR).

7.4.1. Data-Generating Process

The WGSR is a chemical equilibrium reaction between water and carbon monoxide on the one side and hydrogen and carbon dioxide on the other side,

$$CO + H_2O \overset{r(x)}{\rightleftharpoons} CO_2 + H_2. \tag{7.21}$$

The rate $r \in \mathbb{R}$ of this reaction depends on various external factors x. We consider it under the following assumptions.

All reactants of the WGSR are in the gas phase. The experimental conditions comprise the partial pressures of CO, H_2O, CO_2, and H_2, and the temperature, summarized (in that order) in the vector $x \in \mathbb{R}^5$. The partial pressures are limited to the interval $[0.05, 1]$ and the temperature is fixed at 473.16 Kelvin (200 degrees Celsius), so that the experimental domain is

$$\mathcal{X} := [0.05, 1] \times [0.05, 1] \times [0.05, 1] \times [0.05, 1] \times \{473.16\}. \tag{7.22}$$

Let $x_1, x_2, \dots$ be a sequence of experimental conditions from that domain, and let $y_1, y_2, \dots \in \mathbb{R}$ be the corresponding observed values of the reaction rate. In each experiment $n \in \mathbb{N}$, the observed value y_n is composed of the actual reaction rate $r(x_n)$ and an additive measurement error that is normally distributed with mean zero and a constant non-zero variance of σ^2. In other words, each y_n is a realization of the observable

$$\mathcal{Y}_n \sim \mathcal{N}\big(r(x_n), \sigma^2\big). \tag{7.23}$$

This setting satisfies assumptions (i)–(iii) on p. 214.

It is known that the observables are normally distributed with variance σ^2, but their mean $r(\cdot)$ is unknown. It remains for a model to describe the reaction rate $r(x)$.

7.4.2. Model Family

The following 13 models were collected by Schwaab et al. [225, Sec. 3.3] for testing model discrimination (md) strategies. Each model $\mu \in \mathcal{M} := \{1, \dots, 13\}$ involves a parameter θ^μ that can take values in $\mathcal{Q}^\mu := \mathbb{R}^{n_{\theta^\mu}}$. For each $\mu \in \mathcal{M}$ and all $\theta^\mu \in \mathcal{Q}^\mu$, model μ with parameter θ^μ specified for all experimental conditions $x \in \mathcal{X}$ a real-valued response $\eta^\mu(x, \theta^\mu)$ for predicting the reaction rate $r(x)$. Writing θ^μ_i and z_i for the i-th component of θ^μ and x, respectively, the responses of these models are for all $x \in \mathcal{X}$ defined as

$$\eta^1\big(x, \theta^1\big) := \alpha(x) \frac{z_1 z_2}{\left(\theta^1_1 + \theta^2_2 z_1 + \theta^1_3 z_2 + \theta^1_4 z_3 + \theta^1_5 z_4\right)^2}, \tag{7.24a}$$

$$\eta^2\big(x, \theta^2\big) := \alpha(x) \frac{z_1 z_2}{\theta^2_1 + \theta^2_2 z_1 + \theta^2_3 z_2 + \theta^2_4 z_3 + \theta^2_5 z_4}, \tag{7.24b}$$

$$\eta^3\big(x, \theta^3\big) := \alpha(x) \frac{z_1 \sqrt{z_2}}{\theta^3_1 + \theta^3_2 z_1 + \theta^3_3 z_2 + \theta^3_4 z_3 + \theta^3_5 z_4}, \tag{7.24c}$$

$$\eta^4\big(x, \theta^4\big) := \alpha(x) \frac{z_1 z_2}{\theta^4_1 + \theta^4_2 z_1 + \theta^4_3 z_2}, \tag{7.24d}$$

$$\eta^5\big(x, \theta^5\big) := \alpha(x) \frac{z_1}{\theta^5_1 + \theta^5_2 z_3 z_4 / z_2 + \theta^5_3 z_4 + \theta^5_4 z_2 + \theta^5_5 z_3}, \tag{7.24e}$$

$$\eta^6\big(x, \theta^6\big) := \alpha(x) \frac{z_2}{\theta^6_1 + \theta^6_2 z_2 / z_4}, \tag{7.24f}$$

$$\eta^7\big(x, \theta^7\big) := \alpha(x) \frac{z_1 z_2}{\theta^7_1 z_1 + \theta^7_2 z_2 + \theta^7_3 z_3}, \tag{7.24g}$$

$$\eta^8(x, \theta^8) := \alpha(x) \frac{z_1 z_2}{\theta_1^8 z_1 + \theta_2^8 z_2 + \theta_3^8 z_3 + \theta_4^8 z_4}, \tag{7.24h}$$

$$\eta^9(x, \theta^9) := \alpha(x) \frac{z_1 z_2}{\theta_1^9 z_1 + \theta_2^9 z_2}, \tag{7.24i}$$

$$\eta^{10}(x, \theta^{10}) := \alpha(x) \frac{z_2}{\theta_1^{10} z_1 + \theta_2^{10} z_2}, \tag{7.24j}$$

$$\eta^{11}(x, \theta^{11}) := \alpha(x) e^{-\theta_1^{11}} z_1^{\theta_2^{11}} z_2^{\theta_3^{11}} z_3^{\theta_4^{11}} z_4^{\theta_5^{11}}, \tag{7.24k}$$

$$\eta^{12}(x, \theta^{12}) := \alpha(x) e^{-\theta_1^{12}} z_1^{\theta_2^{12}} z_2^{\theta_3^{12}}, \text{ and} \tag{7.24l}$$

$$\eta^{13}(x, \theta^{13}) := \alpha(x) z_1 z_2 \theta_1^{13}. \tag{7.24m}$$

Assuming that the temperature z_5 is measured in Kelvin,

$$\alpha(x) := 1 - \frac{z_3 z_4}{z_1 z_2} \exp\left(4.33 - \frac{4577.8}{z_5}\right). \tag{7.25}$$

The responses are listed in the same order as by Schwaab et al. [225]. For numerical reasons they have been reparameterized to reduce correlations between parameter estimates. We refer to these models as the WGSR model family. Each of its models satisfies assumption (iv) on p. 214.

Some of the models in the family are special cases of others, as shown in Fig. 7.1 on the following page. Only models 3, 5, 6, and 10 are neither special cases nor generalizations of any other model.

7.5. Numerical Results for the WGSR Model Family

The computational results discussed in the following are obtained from the following setting.

Experiments are performed under a repeated sequence of $s := 10$ pairwise distinct conditions such that $x_{n+s} = x_n$ for all $n \in \mathbb{N}$. The conditions are listed in columns 2 to 4 of Tab. 7.1 on the next page.

The reaction rate is for all $n \in \mathbb{N}$ *defined* as

$$r(x_n) := \frac{\eta^1(x_n, \bar{\theta})}{c} \text{ for all , where}$$

$$\bar{\theta} := \begin{bmatrix} 1.6855 & 4.5947 & 0.94219 & 0.89669 & 2.4591 \end{bmatrix}^\top$$

Figure 7.1.: Hierarchy of the WGSR model family. The notation $\boxed{\mu}\xrightarrow{C}\boxed{\nu}$ means that model μ reduces to model ν under condition C.

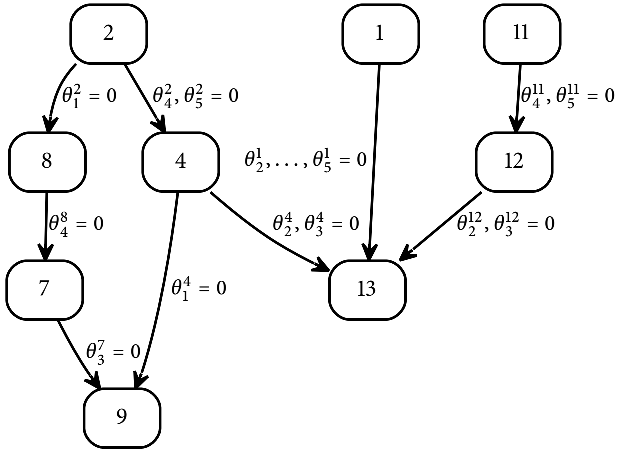

2
1
11
$\theta_1^2 = 0$
$\theta_4^2, \theta_5^2 = 0$
$\theta_4^{11}, \theta_5^{11} = 0$
8
4
$\theta_2^1, \dots, \theta_5^1 = 0$
12
$\theta_4^8 = 0$
$\theta_2^4, \theta_3^4 = 0$
$\theta_2^{12}, \theta_3^{12} = 0$
7
$\theta_1^4 = 0$
13
$\theta_3^7 = 0$
9

Table 7.1.: Experimental conditions used for comparing classic and robust PMLE covariance approximations in the WGSR model family and resulting reaction rates. Rounded to five digits.

n	x_n				$r(x_n)$
1	0.19102	0.72634	0.0728	0.32697	1.3086
2	0.66602	0.40967	0.8328	0.86983	0.67049
3	0.42852	0.30412	0.4528	0.19125	0.89556
4	0.90352	0.93745	0.2628	0.73411	1.7635
5	0.13164	0.62078	0.6428	0.46268	0.60993
6	0.60664	0.19856	0.2248	0.13309	0.71491
7	0.36914	0.83189	0.9848	0.67595	1.0797
8	0.84414	0.51523	0.6048	0.40452	1.2057
9	0.25039	0.08128	0.4148	0.94738	0.093776
$s = 10$	0.72539	0.71461	0.7948	0.26880	1.658

and $c := \frac{1}{s} \sum_{n=1}^{s} \eta^1(x_n, \bar{\theta})$.

That is, a rescaled variant of model 1 with parameter $\bar{\theta}$ is defined to be correct. The parameter value $\bar{\theta}$ is the same used by Schwaab et al. [225]. The factor c normalizes the reaction rate r to an average value of 1 under the given experimental conditions. The responses (7.24) of the water-gas shift reaction (WGSR) model family are normalized accordingly. This normalization allows to interpret the standard deviation σ directly as magnitude of the *relative* measurement error.

7.5.1. Dependency on the Amount of Data

In this section we discuss computational results dealing with (Q7.1) on p. 217. The upper chart in Fig. 7.2 on the next page shows the mean error $\bar{\delta}_{\mathrm{cl}}(n)$ of the classic approximation as function of the sample size n for models 1 to 12 of the WGSR model family under a moderate relative measurement error $\sigma = 12.8\%$. For clarity, the linear model 13 is omitted, and results are shown in steps of ten experiments. The results were computed with the DoeSim implementation of Alg. 7.1 on p. 224 using $r = 160\,000$ replications. The lower chart of Fig. 7.2 shows the analogous results for the average error $\bar{\delta}_{\mathrm{rob}}(n)$ of the robust approximation.

Except for a few models, the average errors of both the classic and the robust approximation behave qualitatively similarly. They decrease monotonically with the sample size, yet a quickly declining rate. The charts to not allow to judge whether the average errors converge to positive constants or converge very slowly to zero.

We use log-log plots in the following to reveal more details. In a log-log plot, a power law of the type $f(n) = \alpha n^{\beta}$ with $\alpha, \beta \in \mathbb{R}$ appears as a straight line with axis intercept $\log(\alpha)$ and slope β, since $\log(f(n)) = \log(\alpha) + \beta \log(n)$. Data points appearing on a line in a log-log plot may hence indicate an underlying power law. Clauset, Rohilla Shalizi, and Newman [71] describe statistical techniques for inferring whether error-corrupted data stems from a power law (and from which one) or from a function with a similar appearance.

Mean Error of the Classic Approximation

Figure 7.3 on p. 231 shows the same results as the upper chart in Fig. 7.3 in log-log scale. It permits to further differentiate the behavior of the classic approximation. In models 1 and 2, its average error lies approximately on parallel lines with the same negative slope, indicating a decrease according to power laws (with the same

Figure 7.2.: *Top:* mean error of the classic approximation as function of the sample size. Models 1 to 12 of the WGSR model family, relative measurement error $\sigma = 12.8\%$, $r = 160\,000$ replications. *Bottom:* analog results for the robust approximation. Abscissa scale applies to both charts, legend shown in Fig. 7.4 on p. 233.

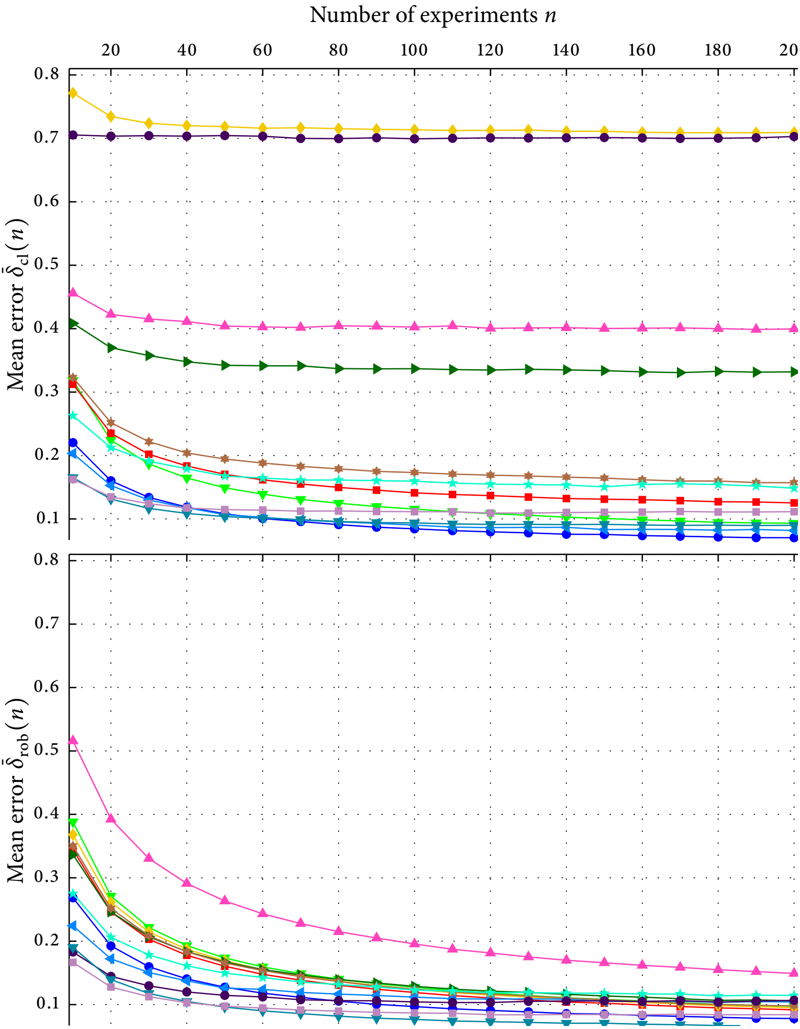
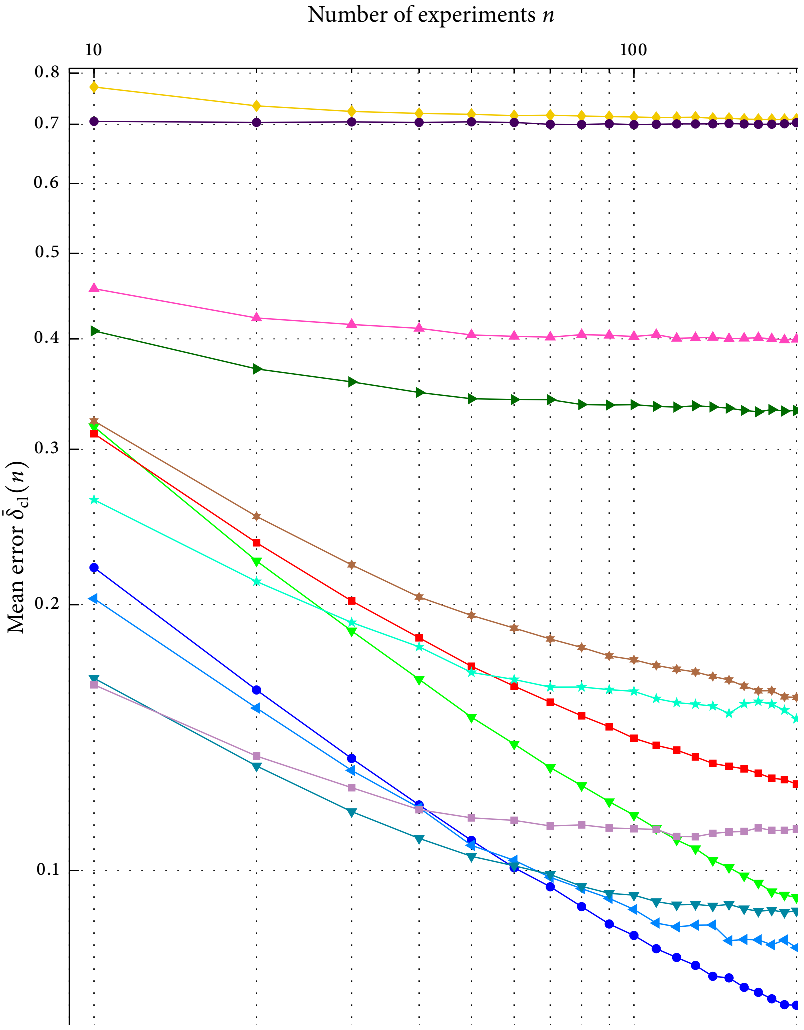


Figure 7.3.: Mean error of the classic approximation as function of the sample size, log-log counterpart of the upper chart in Fig. 7.2 on the preceding page. Models 1 to 12 of the WGSR family, relative measurement error $\sigma = 12.8\%$, $r = 160\,000$ replications. Legend shown in Fig. 7.4 on p. 233.



negative exponent). The average errors in models 3, 6, 8, 9 and 11 lie on graphs that are "bent up" compared to models 1 and 2, indicating they they decrease slower than a power law, possibly to constant positive limit values. In models 4, 5, 7, and 12, the average errors decrease slowly from the very start and seem to converge to constant positive values. In model 10, finally, the average error remains for all sample sizes at a constant positive value. Besides these differences in the rate of decrease, it is immediately visible that the absolute value of the mean approximation error in models 4, 5, 7, 10 is significantly larger than than of the other models.

Mean Error of the Robust Approximation

The mean error of the robust approximation, in contrast, behaves much more homogeneously among the models, as seen in the log-log plot in Fig. 7.4. First of all, there is less variation among the mean errors of the robust approximation between different models compared to its classic counterpart, there are no obvious "outliers." In models 1 to 5, 7, 8 and 11, the mean errors drop in good approximation according to power laws with the same exponent. The mean errors of models 6, 9, and 10 initially follow a power law with the same exponent (smaller than that of the previous group of models), but eventually flatten out and seem to approach a positive constant. The final average errors after $n = 200$ experiments are either very close to the corresponding values of the classic approximation or significantly smaller.

Variability of Classic and Robust Approximation

So far we compared considered only the average errors. Figure 7.5 on p. 234 shows the corresponding variabilities. The upper chart shows the standard deviation $\sigma_{\mathrm{cl}}(n)$ of the classic approximation as function of the sample size n for models 1 to 12 of the WGSR model family under a relative measurement error $\sigma = 12.8\%$. The lower chart shows the analogous standard deviation $\sigma_{\mathrm{rob}}(n)$ of the robust approximation. We observe little difference between the variabilities of the classic and the robust approximation. For both approximations, the variabilities decrease according to power laws with exponents around -0.1 in all models, and have very similar absolute values for all models except for models 4 and 5.

Figure 7.4.: Mean error of the robust approximation as function of the sample size. Models 1 to 12 of the WGSR family, relative measurement error $\sigma = 12.8\%$, $r = 160\,000$ replications. Log-log counterpart of the lower chart in Fig. 7.2 on p. 230.

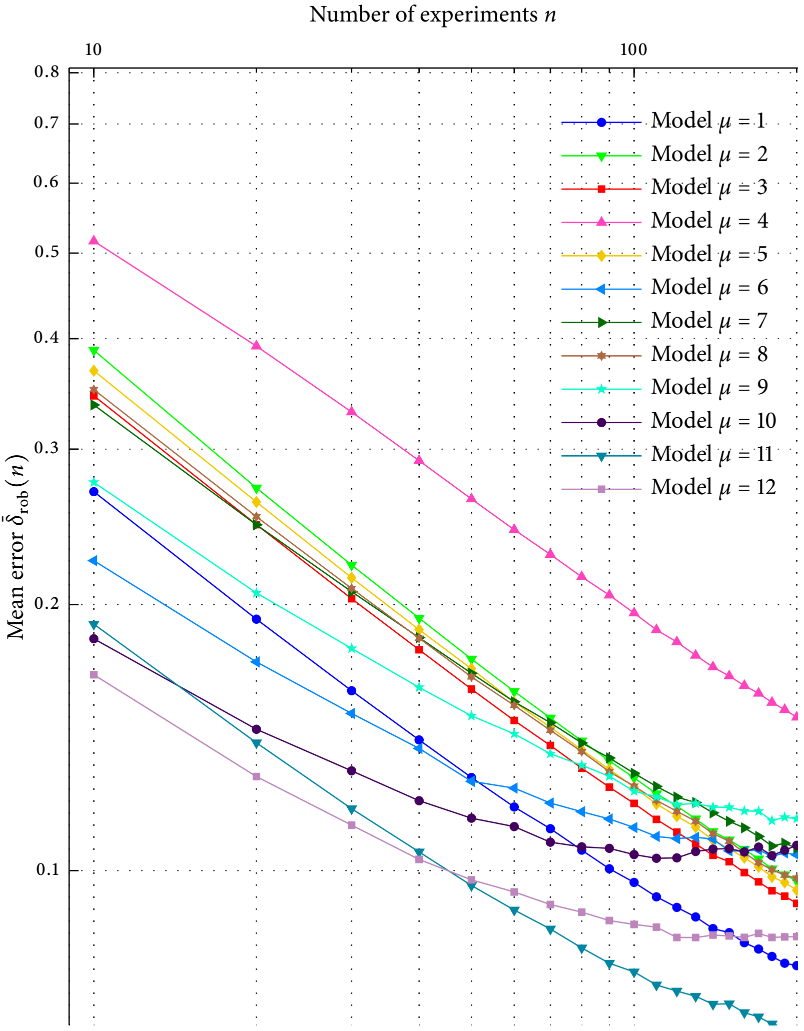


Figure 7.5.: *Top:* standard deviation of the error of the classic approximation as function of the sample size. Models 1 to 12 of the WGSR model family, relative measurement error $\sigma = 12.8\%$, $r = 160\,000$ replications. *Bottom:* analog results for the robust approximation. Abscissa scale applies to both charts, legend shown in Fig. 7.4 on the previous page.

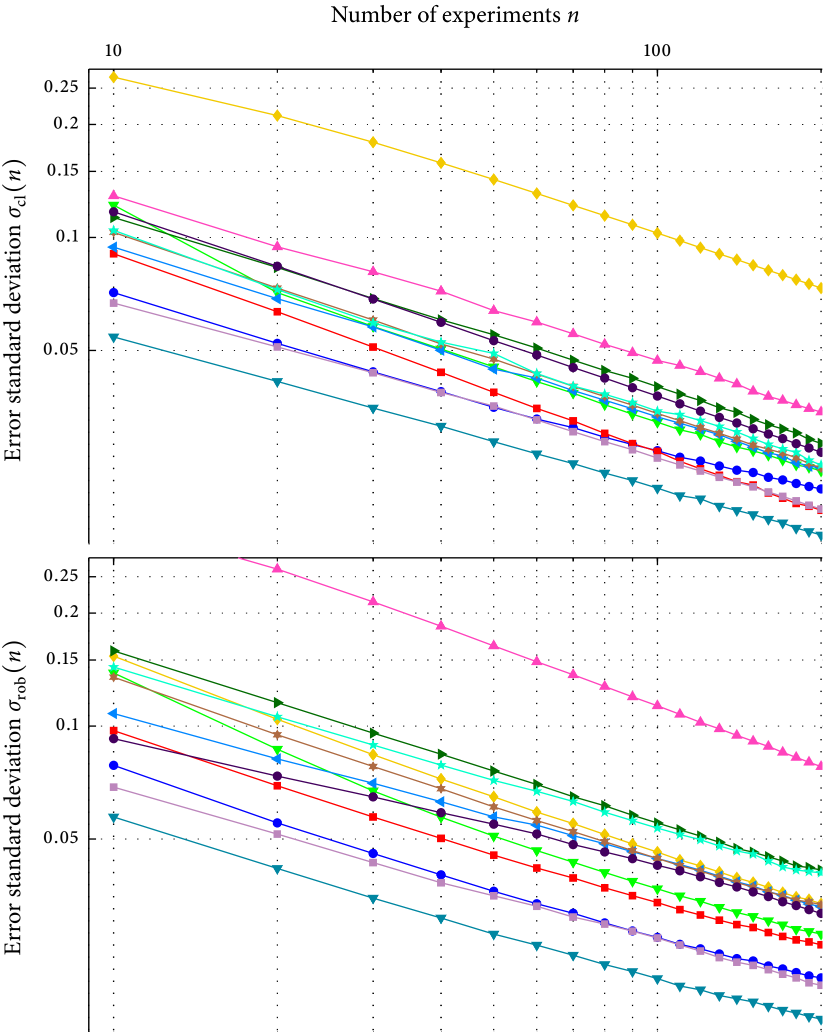
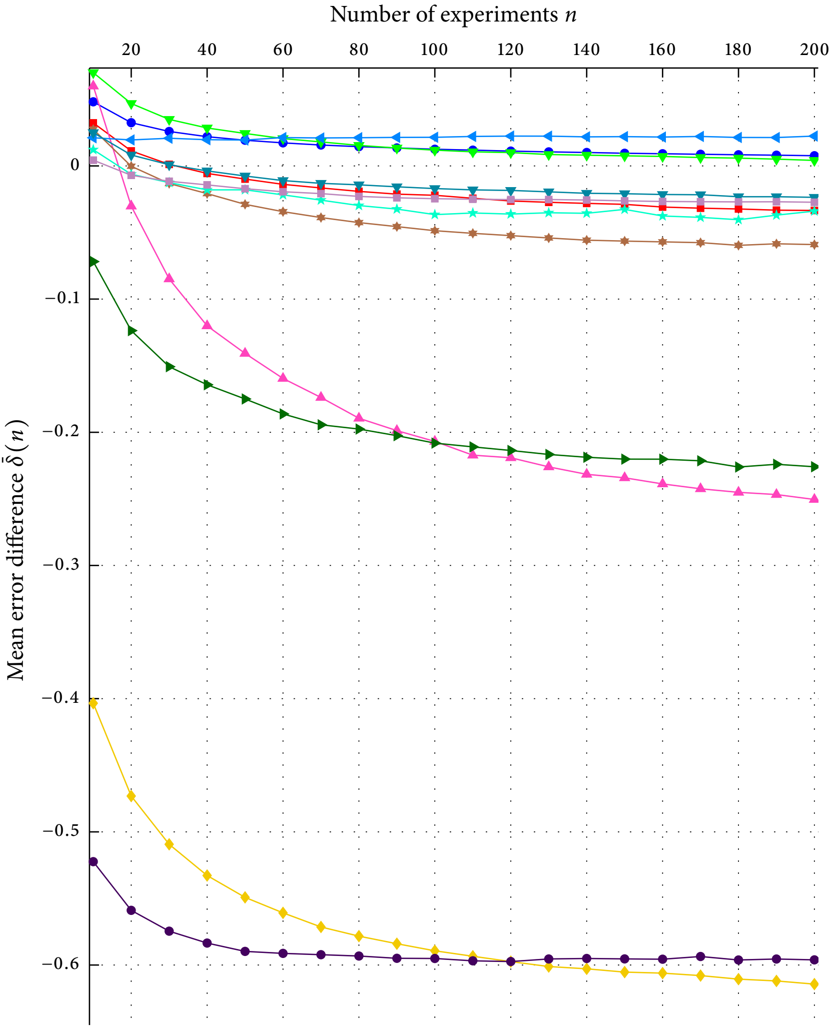


Figure 7.6.: Difference between then mean errors of the robust and the classic approximation as function of the sample size. Models 1 to 12 of the WGSR family, relative measurement error $\sigma = 12.8\%$, $r = 160\,000$ replications. Legend shown in Fig. 7.4 on p. 233.



Comparison of Both Approximations

Since both approximations exhibit very similar variabilities, a coarse comparison of can be based solely on their average errors. Figure 7.6 on the previous page shows the difference $\bar{\delta}(n) := \bar{\delta}_{\text{rob}}(n) - \bar{\delta}_{\text{cl}}(n)$ between the mean errors using a linear scale on both axes. The smaller (more negative) $\bar{\delta}(n)$, the better is the robust approximation compared to its classic counterpart in the sense of a smaller mean error. Initially, the classic approximation is slightly better than its robust counterpart in most models except for the "outliers" models 5, 7, and 10. This advantage quickly gets smaller as the sample size increases. Once the sample size reaches 20 experiments, the robust approximation is better for nine of the twelve models. For the other three models, the classic approximation remains better up to a sample size of $n = 200$, yet the advantage is small in absolute terms.

7.5.2. Dependency on Data Variability

Let us now discuss how the classic and the robust approximation are affected by the variability of the underlying data, as stated in (Q7.2) on p. 217. To that end, we repeated the computations discussed in the previous section for different standard deviation σ of the observables. The computations were performed with the DoeSim implementation of Alg. 7.1 on p. 224 using $r = 160\,000$ replications.

The results are summarized in Figs. 7.7 to 7.9 on pp. 237–239, using a separate chart for each model of the WGSR model family. To improve readbility, model with a similar magnitude of the mean error difference are grouped in one figure.

Each chart shows the difference $\bar{\delta}(n) := \bar{\delta}_{\text{rob}}(n) - \bar{\delta}_{\text{cl}}(n)$ between the mean error of the robust and the classic approximation as function of the sample size n, obtained from data with a relative measurement error σ of 1.6%, 3.2%, 6.4%, 12.8%, 25.6%, and 51.2%. Note that the charts in Fig. 7.9 use a larger scale than those in Figs. 7.7 and 7.8. As previously, the linear model 13 is omitted for clarity, and results are shown in steps of ten experiments.

Recall that the more $\bar{\delta}(n)$ is lower than zero, the better is the robust approximation compared to its classic counterpart in the sense of a smaller mean error, and vice versa for values greater zero. For most models and magnitudes of σ, the mean error difference $\bar{\delta}(n)$ shows qualitatively similar behavior to that seen in Fig. 7.6. As the sample size n gets larger, the mean error difference $\bar{\delta}(n)$ decreases almost monotonically.

When the measurement error σ is "small" (1.6%, 3.2%, and 6.4%) the robust approximation is – already for small sample sizes – better than its classic counterpart for most models (number 3 to 5 and 7 to 12), or is only slightly worse

Figure 7.7.: Difference between then mean errors of the classic and the robust approximation as function of the sample size under measurement errors of 1.6% (blue), 3.2%, 6.4%, 12.8%, 25.6%, and 51.2% (green). Models 1, 2, 3, and 6 of the WGSR family, $r = 160\,000$ replications.

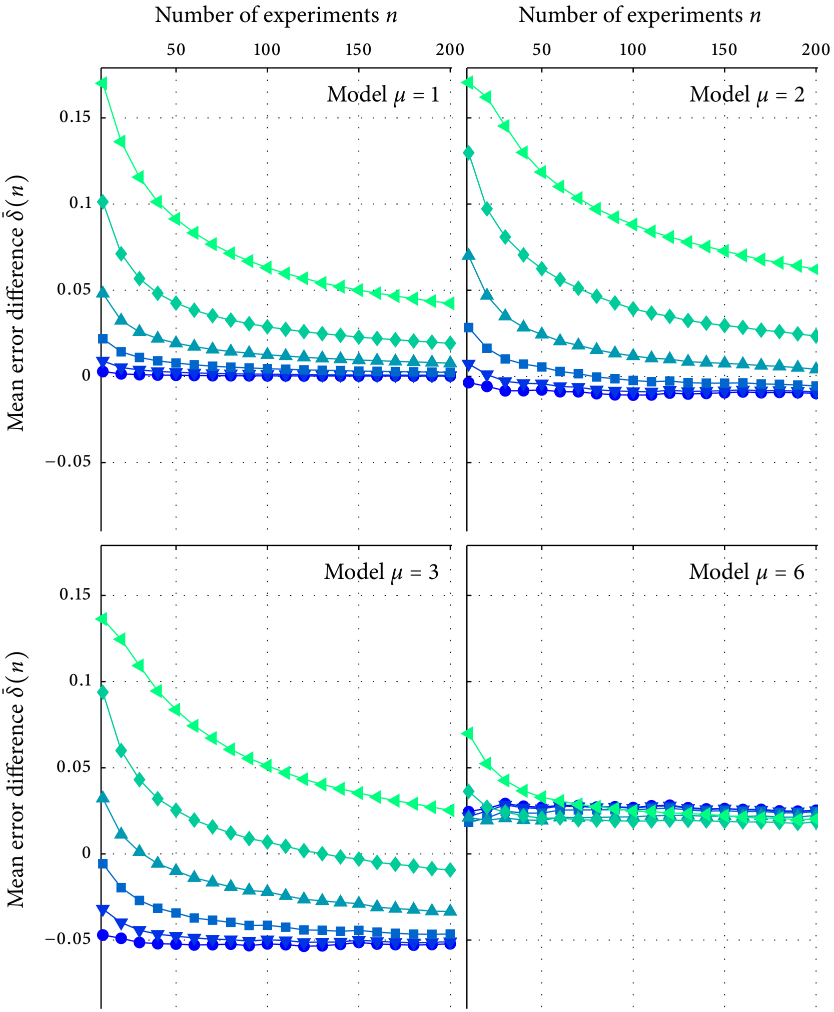


Figure 7.8.: Difference between then mean errors of the classic and the robust approximation as function of the sample size under measurement errors of 1.6% (blue), 3.2%, 6.4%, 12.8%, 25.6%, and 51.2% (green). Models 8, 9, 11, and 12 of the wgsr family, $r = 160\,000$ replications.

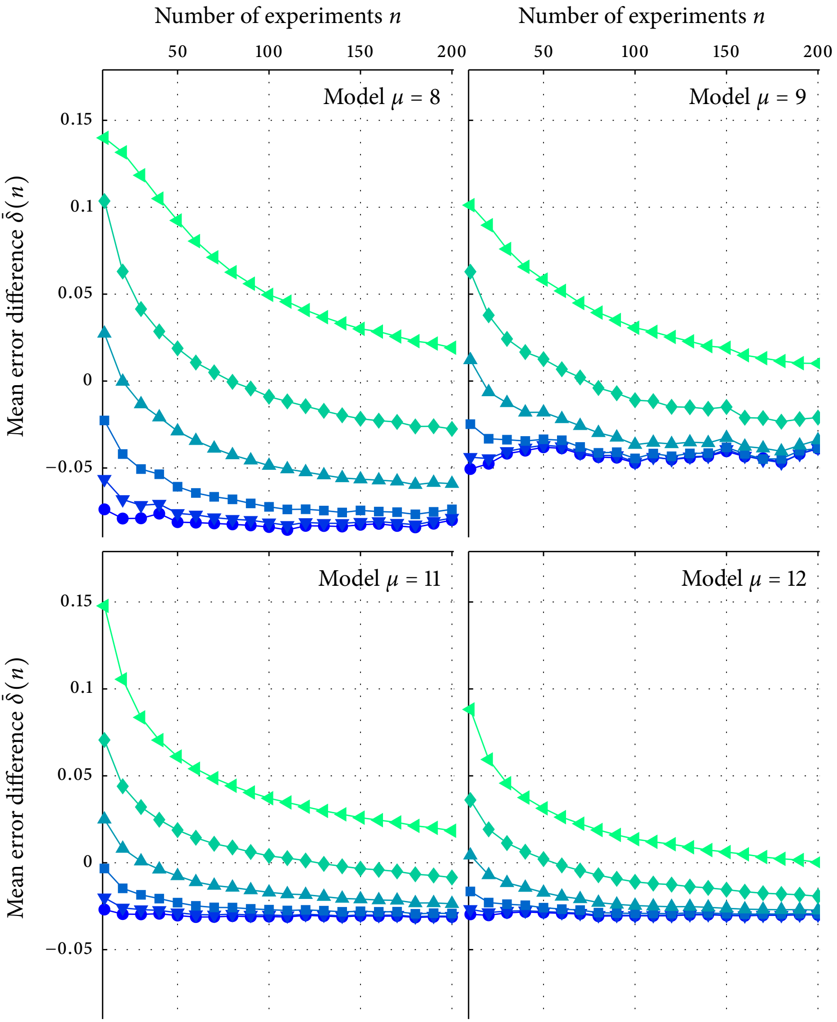
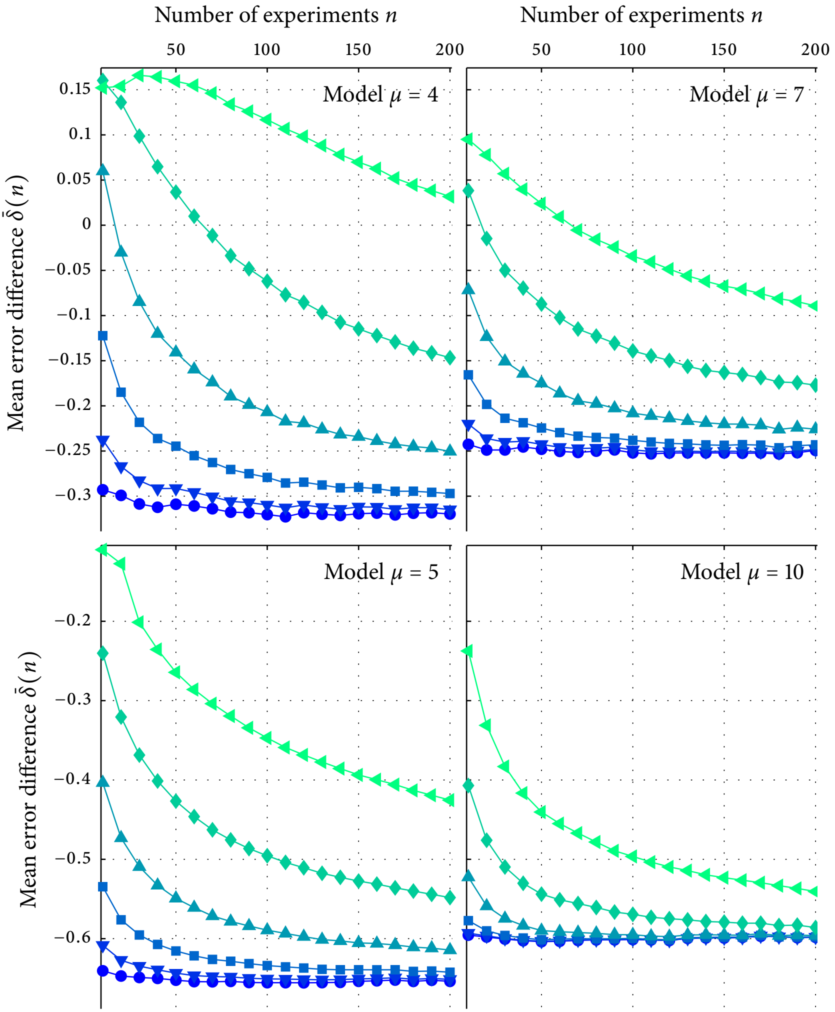


Figure 7.9.: Difference between then mean errors of the classic and the robust approximation as function of the sample size under measurement errors of 1.6% (blue), 3.2%, 6.4%, 12.8%, 25.6%, and 51.2% (green). Models 4, 5, 7, and 10 of the WGSR family, $r = 160\,000$ replications.



(number 1, 2, and 6). For larger measurement errors σ of 12.8%, 25.6%, and 51.2%, the classic approximation is initially better for most models (all except 5 and 10), yet this advantage drops with increasing n. Once the sample size is sufficiently large, $\bar{\delta}(n)$ switches to a negative sign in favor of the robust approximation. For some models this root can be seen in the charts. For the very large measurement error of 51.2%, however, $\bar{\delta}(n)$ drops so slowly that the classic approximation remains the better one up to $n = 200$ for all models except number 5, 7, and 10. For the latter models, the robust approximation has a very clear advantage.

The overall results can be summarizes as follows. The robust approximation is in most models initially a little worse than its classic counterpart, but is for a few models substantially better initially, and in large samples better for almost all models. In the examined WGSR scenarios, choosing the robust approximation instead of the classic one avoids gross missestimations of the parameter maximum-likelihood estimate (PMLE) covariance at the cost of a small loss of approximation quality for a large fraction models. Whether this observation can be generalized to other scenarios shall be the task of future work.

8. Performance of Design Criteria for MD: Theory and Algorithms

Contents

This chapter develops a framework that allows to assess and compare the practical performance of sequential design criteria for model discrimination (MD). The framework comprises statistical measures of performance, algorithms for their numerical computation, and an actual implementation.

In section Sec. 8.2 we derive two statistical measures for the performance of a design criterion. One of them is based on the concept of T-optimality introduced in Chap. 4, the other on Bayesian posterior probabilities discussed in Chap. 3. In Sec. 8.3, we briefly review various sequential design criteria for model discrimination and provide a unified representation for them. Based thereon, we describe a Monte Carlo algorithm in Sec. 8.4 that allows to efficiently compute the introduced performance measures. We also describe its implementation provided in the software package DoeSim.

The developed framework is used extensively in the case studies presented in

Chap. 9.

8.1. Problem Statement

8.1.1. Considered Scenario

Experiments can be performed under conditions or settings from the compact experimental domain $\mathcal{X} \subseteq \mathbb{R}^{n_x}$ and yield observations, results, or outcomes in the observation domain $\mathbb{R}^{n_y}$. For all $x \in \mathcal{X}$, an observation obtained from an experiment under condition x is a realization (or “variate”) of the continuous $\mathbb{R}^{n_y}$-valued random variable $\mathcal{Y}_x$, called observable. The observables $\mathcal{Y}_x$ and $\mathcal{Y}_{x'}$ of two experiments performed under – not necessarily different – conditions $x, x' \in \mathcal{X}$ are statistically independent.

The following model family is available. For all $\mu \in \mathcal{M} := \{1, \dots, n_{\mathcal{M}}\}$ and all $\theta^\mu \in \mathcal{Q}^\mu \subseteq \mathbb{R}^{n_{\theta^\mu}}$, model μ with parameter θ^μ specifies for all experimental conditions $x \in \mathcal{X}$ an n_y-dimensional normal distribution (Def. B.8) with mean $\eta^\mu(x, \theta^\mu)$ and symmetric positive definite (spd) covariance matrix $\boldsymbol{\Omega}$. The response $\eta^\mu(x, \theta^\mu)$ of each model $\mu \in \mathcal{M}$ is twice continuously differentiable in θ^μ for all $x \in \mathcal{X}$.

The model family contains a unique correct model $\bar{\mu} \in \mathcal{M}$, which has a unique correct parameter $\bar{\theta} \in \mathcal{Q}^{\bar{\mu}}$, so that

$$\mathcal{Y}_x \sim \mathcal{N}(\bar{\eta}(x), \boldsymbol{\Omega}), \text{ with } \bar{\eta}(x) := \eta^{\bar{\mu}}\big(x, \bar{\theta}\big), \text{ for all } x \in \mathcal{X}. \tag{8.1}$$

The observation covariance $\boldsymbol{\Omega}$ is known, but the correct model $\bar{\mu}$, its correct parameter $\bar{\theta}$, and hence the function $\bar{\eta}$ are unknown. Furthermore, the prior knowledge is minimal: the model prior is uniform, $p(\mu) := 1/n_{\mathcal{M}}$ for all $\mu \in \mathcal{M}$, and the parameter prior of each model is vague in the sense that it can be neglected compared to any empirical information.

8.1.2. Sequential Design Procedures for MD

In the considered scenario, we aim to identify the correct model $\bar{\mu}$ *empirically* in a sequential design procedure, that is, by designing, performing, and analyzing one experiment after the other.

We use the variable n to enumerate the iterations of such a procedure. For all $n \in \mathbb{N}$, the variable $x_n \in \mathcal{X}$ denotes the condition of the n-th experiment, $y_n \in \mathbb{R}^{n_y}$ denotes the corresponding observation, ξ_n stands for the exact design resulting constituted by the conditions $x_1, \dots, x_n$, and $d_n^\top := \begin{bmatrix} y_1^\top & \dots & y_n^\top \end{bmatrix}$ summarizes the corresponding DATA. The tuple d_n is thus a realization of the SAMPLE $\mathcal{D}_n^\top := \begin{bmatrix} \mathcal{Y}_{x_1}^\top & \dots & \mathcal{Y}_{x_n}^\top \end{bmatrix}$, a $\mathbb{R}^{n \cdot n_y}$-valued random variable.

A sequential design procedure can be regarded as an algorithm that takes as input a model family and determines a sequence of designs $\xi_1, \xi_2, \dots$ and corresponding data $d_1, d_2, \dots$ until it terminates after iteration n, providing the design ξ_n and the data d_n as output, see Alg. 4.1 on p. 126.

If we apply such a procedure to identify the correct model, we are in each iteration $n \in \mathbb{N}$ faced with the following model discrimination (MD) problem: "Given the design ξ_n and data d_n, find the correct model $\bar{\mu}$." Due to the random nature of the data, this problem can typically be solved only approximately. We discussed suitable methods of statistical inference in Chaps. 2 and 3. Optimal experimental design (OED) for MD, considered in Chaps. 4 and 5, aims to reduce the number of experiments n required to achieve a satisfactory approximation quality.

A sequential design criterion for MD is a function of the type $\Psi_n : \mathcal{X} \mapsto \mathbb{R}$, defined for all $n \in \mathbb{N}$, whose maximizers $x_{n+1} \in \operatorname{argmax}_{x \in \mathcal{X}} \Psi_n(x)$ are supposed to be particularly "efficient" experimental conditions for solving MD problems. To that end, such a design criterion typically takes into account the design and data of the available experiments and/or the inferred quantities representing empirical knowledge.

8.1.3. Key Questions

In Chaps. 4 and 5 we studied various sequential design criteria for MD that are all motivated *asymptotically.* That is, if applied in each iteration n of a sequential design procedure, they aim to provide designs and data that are particularly efficient for MD in the *limit* $n \to \infty$.

In this chapter, we assess and compare some of these design criteria for the practically relevant case of a finite and possibly small number n of iterations. To that end, we apply them in the sequential design procedure described by Alg. 8.1 on the next page. The procedure is a special case of Alg. 4.1 on p. 126 for the considered scenario which starts from s initial experiments under predefined conditions and terminates once it reaches a predefined maximal number $n_{\max} \in \mathbb{N}$ of experiments. The symbol $\mathcal{U}_n$ is a placeholder for any collection of quantities used to express the state of knowledge in iteration n.

We examine the following key questions. Suppose Alg. 8.1 is applied using a sequential design criterion Ψ_n.

Algorithm 8.1: Sequential design procedure for assessing design criteria for MD.

```
input  : dimension $n_y \in \mathbb{N}$ of observables, experimental domain $\mathcal{X} \subseteq \mathbb{R}^{n_x}$
         full-rank $n_y \times n_y$ observation covariance matrix $\boldsymbol{\Omega}$
         model index set $\mathcal{M}$, parameter domains $\mathcal{Q}^{\mu} \in \mathbb{R}^{n_{\theta^{\mu}}}$ for all $\mu \in \mathcal{M}$
         response function $\eta^{\mu} : \mathcal{X} \times \mathcal{Q}^{\mu} \mapsto \mathbb{R}^{n_y}$ for all $\mu \in \mathcal{M}$
         initial experimental conditions $x_1, \dots, x_s \in \mathcal{X}$
         maximal number of experiments $n_{\max} > s$
         sequential design criterion $\Psi_n : \mathcal{X} \mapsto \mathbb{R}$ for all $n \geqslant s$
         correct model $\bar{\mu} \in \mathcal{M}$, correct parameter $\bar{\theta} \in \mathcal{Q}^{\bar{\mu}}$
output : designs $\xi_1, \dots, \xi_{n_{\max}}$, data $d_1, \dots, d_{n_{\max}}$
let $\bar{\eta}(x) := \eta^{\bar{\mu}}\big(x, \bar{\theta}\big)$ for all $x \in \mathcal{X}$;
for $n = 1$ to $s$ do
    get random variate $y_n$ of $\mathcal{N}(\bar{\eta}(x_n), \boldsymbol{\Omega})$;            // experiment, see (8.1)
end
for $n = s$ to $n_{\max}$ do
    let $\xi_n$ be the design constituted by $x_1, \dots, x_n$;
    let $d_n^{\mathsf{T}} := \begin{bmatrix} y_1^{\mathsf{T}} & \dots & y_n^{\mathsf{T}} \end{bmatrix}$;
    determine knowledge $\mathcal{U}_n$ from $\xi_n$, $d_n$, $(\eta^{\mu})_{\mu\in\mathcal{M}}$, and $\boldsymbol{\Omega}$;          // inference
    find $x_{n+1} \in \operatorname{argmax}_{x\in\mathcal{X}} \Psi_n\Big(x; \mathcal{U}_n, (\eta^{\mu})_{\mu\in\mathcal{M}}, \boldsymbol{\Omega}\Big)$;          // sequential oed
    get random variate $y_{n+1}$ of $\mathcal{N}(\bar{\eta}(x_{n+1}), \boldsymbol{\Omega})$;        // experiment, see (8.1)
end
return $\xi_1, \dots, \xi_{n_{\max}}, d_1, \dots, d_{n_{\max}}$
```

(Q8.1) How efficient are the provided designs and data for MD, that is, for empirically identifying the correct model, depending on the *amount* of available experiments n?

(Q8.2) How is this efficiency affected by the number $n_{\mathcal{M}}$ of rival models?

(Q8.3) How does this efficiency depend on the *variability* of the data in terms of the observation covariance $\boldsymbol{\Omega}$?

8.1.4. Notation and Definitions

We use the following notation and definitions for all models $\mu \in \mathcal{M}$, all parameters $\theta^{\mu} \in \mathcal{Q}^{\mu}$, all experimental conditions $x \in \mathcal{X}$, and all observations $y \in \mathbb{R}^{n_y}$.

The noncentrality and the sum of squared residuals (ssr) are

$$\lambda_n^\mu(\theta^\mu, \xi_n) := \tfrac{1}{n}\sum_{i=1}^{n} \|\eta^\mu(x_i, \theta^\mu) - \bar{\eta}(x_i)\|^2_{\boldsymbol{\Omega}^{-1}}, \text{ and} \tag{8.2}$$

$$s_n^\mu(\theta^\mu, d_n, \xi_n) := \tfrac{1}{n}\sum_{i=1}^{n} \|\eta^\mu(x_i, \theta^\mu) - y_i\|^2_{\boldsymbol{\Omega}^{-1}}, \tag{8.3}$$

respectively, see Defs. 3.7 and 3.9 and Tab. 3.1. The inverse $\boldsymbol{\Omega}^{-1}$ exists since $\boldsymbol{\Omega}$ is spd by assumption.

We write $\boldsymbol{J}^\mu(x, \theta^\mu)$ for the $n_y \times n_{\theta^\mu}$ Jacobian of the response $\eta^\mu(x, \theta^\mu)$ with respect to θ^μ, and $\boldsymbol{H}_j^\mu(x, \theta^\mu)$ for the $n_{\theta^\mu} \times n_{\theta^\mu}$ Hessian of its j-th component $\eta_j^\mu(x, \theta^\mu)$ with respect to θ^μ. For all $j \in \{1, \ldots, n_{\theta^\mu}\}$, we write $\tilde{r}_j^\mu(y, x, \theta^\mu)$ for the j-th component of

$$\tilde{r}^\mu(y, x, \theta^\mu) := \boldsymbol{\Omega}^{-1/2}(\eta^\mu(x, \theta^\mu) - y), \tag{8.4}$$

and $\tilde{\boldsymbol{H}}_j^\mu(x, \theta^\mu)$ for the $n_{\theta^\mu} \times n_{\theta^\mu}$ Hessian of this component with respect to θ^μ. Since $\boldsymbol{\Omega}$ is assumed to be spd, the matrix $\boldsymbol{\Omega}^{-1/2}$ in (8.4) exists. We further define

$$\boldsymbol{M}_n^\mu(\theta^\mu, \xi_n) := \tfrac{1}{n}\sum_{i=1}^{n} \boldsymbol{J}^{\mu\top}(x_i, \theta^\mu)\boldsymbol{\Omega}^{-1}\boldsymbol{J}^\mu(x_i, \theta^\mu) \text{ and} \tag{8.5}$$

$$\boldsymbol{N}_n^\mu(\theta^\mu, d_n, \xi_n) := \tfrac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n_y} \tilde{r}_j^\mu(y_i, x_i, \theta^\mu)\tilde{\boldsymbol{H}}_j^\mu(x_i, \theta^\mu), \tag{8.6}$$

both symmetric $n_{\theta^\mu} \times n_{\theta^\mu}$ matrices.

In this scenario, a parameter maximum-likelihood estimate (pmle) $\hat{\theta}_n^\mu := \hat{\theta}_n^\mu(d_n, \xi_n) \in \mathcal{Q}^\mu$ based on the data d_n obtained under design ξ_n minimizes the ssr $s_n^\mu(\theta^\mu, d_n, \xi_n)$ with respect to $\theta^\mu \in \mathcal{Q}^\mu$, see Def. 3.9, Cor. 3.10, and Tab. 3.1. For clarity, we use the abbreviations $\hat{s}_n^\mu := s_n^\mu(\hat{\theta}_n^\mu, d_n, \xi_n)$, $\hat{\eta}_n^\mu(x) := \eta^\mu(x, \hat{\theta}_n^\mu)$, $\hat{\boldsymbol{J}}_n^\mu(x) := \boldsymbol{J}^\mu(x, \hat{\theta}_n^\mu)$, $\hat{\boldsymbol{M}}_n^\mu := \boldsymbol{M}_n^\mu(\hat{\theta}_n^\mu, \xi_n)$, and $\hat{\boldsymbol{N}}_n^\mu := \boldsymbol{N}_n^\mu(\hat{\theta}_n^\mu, d_n, \xi_n)$.

8.2. Efficiency of Sequential Design Procedures for MD

To examine the key questions, we require a measure that tells us how efficient a sequential design procedure is for the aim of identifying the correct model empirically.

8.2.1. T-Efficiency

Let us summarize some results from Secs. 3.2 and 4.2. In the considered scenario, the noncentrality $\lambda_n^\mu(\theta^\mu, \xi_n)$ measures the average systematic mismatch between the data-generating process under experimental conditions $x_1, \ldots, x_n$ (which constitute the design ξ_n) and model μ with parameter θ^μ. The corresponding TOTAL MISMATCH in experiments 1 to n is hence $n\lambda_n^\mu(\theta^\mu, \xi_n)$. It is non-negative, and is zero if and only if model μ is correct under design ξ_n and θ^μ is the corresponding correct parameter. Incorrect models have a non-zero noncentrality, which can be detected with a certain probability by means of a statistical analysis of the data. The larger the total mismatch, the easier is the detection.

The considered model discrimination (MD) problem is solved if it is known which of the models in $\mathscr{M}$ is correct, or equivalently, if all incorrect models – including that one closest to the correct model – from $\mathscr{M}$ are known. The efficiency of experiments 1 to n for solving the MD problem can hence be measured by the corresponding lowest possible total mismatch among the *incorrect* models, which can be written as

$$T_n(\xi_n) := n \min_{\substack{\mu \in \mathscr{M} \\ \mu \neq \bar{\mu}}} \min_{\theta^\mu \in \mathscr{Q}^\mu} \lambda_n^\mu(\theta^\mu, \xi_n). \tag{8.7}$$

Besides the factor n, this expression is the T-criterion Def. 4.7 of design ξ_n. We thus refer to $T_n(\xi_n)$ as T-EFFICIENCY.

Application in Data-Adaptive Sequential Procedures

In a data-adaptive sequential procedure like Alg. 8.1 on p. 244, the condition of the n-th experiment is chosen based on the observations of all preceding experiments 1 to $n-1$. Consequently, the design describing experiments 1 to n may depend on the data d_{n-1}, and is thus written $\xi_n(d_{n-1})$. The T-efficiency of these experiments for MD is thus $T_n(\xi_n(d_{n-1}))$. This particular value has a limited meaning since it depends on the data, which is subject to random fluctuations. More expressive measures can and should be derived from the distribution of the corresponding random variable

$$\mathcal{T}_n := T_n(\xi_n(\mathcal{D}_{n-1})). \tag{8.8}$$

Its expectation $\mathbb{E}\left[\mathcal{T}_n\right]$ is a suggestive "the-large-the-better" measure of the average efficiency of experiments 1 to n for solving the MD problem. Its standard deviation $\mathbb{C}\left[\mathcal{T}_n\right]^{½}$ measures the corresponding variability on a "the-smaller-the-better" scale.

8.2.2. Posterior Model Probability

The model posterior $p(\mu \mid \xi_n, d_n)$ is the Bayesian belief that model $\mu \in \mathcal{M}$ is the correct one, after taking into account the data d_n obtained under the design ξ_n, see Sec. 2.5. Under certain regularity conditions, it can in the considered scenario be approximated by

$$\pi(\mu \mid \xi_n, d_n) := c_n \exp\left(-\tfrac{n}{2}\hat{s}_n^{\mu}\right) n^{-n_{\theta^\mu}/2} \tag{8.9}$$

for all $\mu \in \mathcal{M}$, see (3.85). In this formula, the uniform model prior $1/n_{\mathcal{M}}$ is absorbed in the normalization constant $c_n \in \mathbb{R}^+$ which ensures that the posterior probabilities sum up to 1. We define $\pi_n^\top := \begin{bmatrix}\pi(1 \mid \xi_n, d_n) & \dots & \pi(n_{\mathcal{M}} \mid \xi_n, d_n)\end{bmatrix}$.

Application in Data-Adaptive Sequential Procedures

From a Bayesian point of view a sequential procedure that aims to solve an MD problem should be continued "until the posterior probabilities indicate that one model is clearly superior to the others." (Hill and Hunter [118], Box and Hill [42]) That model is then considered as the best guess for the solution, the sought-after correct model.

A simple practical formalization of this rule is to stop the procedure after iteration n, if there exists a model $\hat{\mu}_n \in \mathcal{M}$ whose approximate posterior $\pi(\hat{\mu}_n \mid \xi_n, d_n)$ reaches or exceeds a predefined threshold $\alpha \in (½, 1)$. Using this rule, the problem is then actually *solved* in iteration n, if $\hat{\mu}_n = \bar{\mu}$, or equivalently, if

$$\pi(\bar{\mu} \mid \xi_n, d_n) \geqslant \alpha. \tag{8.10}$$

The probability that the MD problem is solved in this sense in iteration n of a data-adaptive sequential procedure is

$$\mathbb{P}\left[\mathcal{P}_n \geqslant \alpha\right], \text{ where } \mathcal{P}_n := \pi(\bar{\mu} \mid \xi_n(\mathcal{D}_{n-1}), \mathcal{D}_n) \tag{8.11}$$

is a random variable that takes into account the random fluctuations of data and designs. The probability $\mathbb{P}\left[\mathcal{P}_n \geqslant \alpha\right]$ measures the efficiency of experiments 1 to n for solving the MD problem in a Bayesian sense.

8.3. Considered Sequential Design Criteria

We discussed various design criteria for model discrimination (MD) throughout Chaps. 4 and 5. In the following, we list those examined in this chapter under the assumptions of the considered scenario.

8.3.1. Covariances of Parameters and Responses

Let us summarize some key results of Chaps. 2 and 3. Under regularity conditions, empirical knowledge (or uncertainty) about the Kullback-Leibler information criterion (KLIC)-best parameter of a model $\mu \in \mathcal{M}$ can approximately be represented by a normal distribution around the parameter maximum-likelihood estimate (PMLE) $\hat{\theta}_n^\mu$ and a model-dependent covariance matrix that can be evaluated based on the current design ξ_n and possibly on the resulting data d_n. Under the CLASSIC assumption that the model is correct or affine-linear, the covariance is

$$\tfrac{1}{n}\hat{\boldsymbol{M}}_n^{\mu\,-1} \tag{8.12}$$

in both maximum-likelihood inference and Bayesian inference, see (3.41) and (3.79). Without these assumptions, the covariance is

$$\tfrac{1}{n}\hat{\boldsymbol{R}}_n^\mu := \tfrac{1}{n}\left(\hat{\boldsymbol{M}}_n^\mu + \hat{\boldsymbol{N}}_n^\mu\right)^{-1}\hat{\boldsymbol{M}}_n^\mu\left(\hat{\boldsymbol{M}}_n^\mu + \hat{\boldsymbol{N}}_n^\mu\right)^{-1} \tag{8.13}$$

in maximum-likelihood inference, and is

$$\tfrac{1}{n}\hat{\boldsymbol{B}}_n^{\mu\,-1} := \tfrac{1}{n}\left(\hat{\boldsymbol{M}}_n^\mu + \hat{\boldsymbol{N}}_n^\mu\right)^{-1}, \tag{8.14}$$

in Bayesian inference. In the last expression, the parameter prior is omitted according to our assumption of minimal prior knowledge. We introduced these MISSPECIFICATION-ROBUST parameter covariance formulas in (3.64) and (3.77).

In a locally linear approximation, the uncertainty associated with the prediction of model μ for the outcome of an experiment under $x \in \mathcal{X}$ can then be described by a normal distribution around $\hat{\eta}^{\mu}(x)$. The covariance is

$$\hat{\boldsymbol{T}}_n^{\mu}(x) := \boldsymbol{\Omega} + \tfrac{1}{n}\hat{\boldsymbol{J}}_n^{\mu}(x)\hat{\boldsymbol{M}}_n^{\mu\,-1}\hat{\boldsymbol{J}}_n^{\mu\,\top}(x) \tag{8.15}$$

in both maximum-likelihood inference and Bayesian inference under classic assumptions. The corresponding misspecification-robust covariances are

$$\hat{\boldsymbol{V}}_n^{\mu}(x) := \boldsymbol{\Omega} + \tfrac{1}{n}\hat{\boldsymbol{J}}_n^{\mu}(x)\hat{\boldsymbol{R}}_n^{\mu}\hat{\boldsymbol{J}}_n^{\mu\,\top}(x) \text{ and} \tag{8.16}$$

$$\hat{\boldsymbol{W}}_n^{\mu}(x) := \boldsymbol{\Omega} + \tfrac{1}{n}\hat{\boldsymbol{J}}_n^{\mu}(x)\hat{\boldsymbol{B}}_n^{\mu\,-1}\hat{\boldsymbol{J}}_n^{\mu\,\top}(x) \tag{8.17}$$

in maximum-likelihood inference and Bayesian inference, respectively.

8.3.2. Design Criteria

We are now prepared to introduce the design criteria considered in this chapter. The following definitions apply for all $n \in \mathbb{N}$ and all $x \in \mathcal{X}$.

Suppose that $s_n^{\mu}\big(\hat{\theta}_n^{\mu}, d_n, \xi_n\big)$ has a unique minimum $\hat{\mu}_n$ on $\mathcal{M}$ and a unique minimum $\hat{\nu}_n$ on $\mathcal{M} \setminus \{\hat{\mu}\}$. The multi-model Hunter-Reiner (hr)-criterion from Def. 4.16 is

$$H_n(x) := \left\| \hat{\eta}_n^{\hat{\mu}_n}(x) - \hat{\eta}_n^{\hat{\nu}_n}(x) \right\|_{\boldsymbol{\Omega}^{-1}}^2, \tag{8.18}$$

and the multi-model Buzzi-Ferraris (bf)-criterion from Def. 4.17 is

$$B_n(x) := \left\| \hat{\eta}_n^{\hat{\mu}_n}(x) - \hat{\eta}_n^{\hat{\nu}_n}(x) \right\|_{\hat{\boldsymbol{T}}_n^{-1}(x)}^2, \tag{8.19}$$

where $\hat{\boldsymbol{T}}_n(x) := \hat{\boldsymbol{T}}_n^{\hat{\mu}_n}(x) + \hat{\boldsymbol{T}}_n^{\hat{\nu}_n}(x)$. In Def. 4.18 we proposed the novel misspecification-robust variant of the bf-criterion

$$B_n'(x) := \left\| \hat{\eta}_n^{\hat{\mu}_n}(x) - \hat{\eta}_n^{\hat{\nu}_n}(x) \right\|_{\hat{\boldsymbol{V}}_n^{-1}(x)}^2, \tag{8.20}$$

with $\hat{\boldsymbol{V}}_n(x) := \hat{\boldsymbol{V}}_n^{\hat{\mu}_n}(x) + \hat{\boldsymbol{V}}_n^{\hat{\nu}_n}(x)$. The classic upper bound of the Box-Hill-Hunter (bhh)-criterion from Thm. 5.2 is

$$U_n(x) := \tfrac{1}{2}\pi_n^{\top}\boldsymbol{U}_n(x)\pi_n, \tag{8.21a}$$

where $\boldsymbol{U}_n(x)$ is a $n_{\mathcal{M}} \times n_{\mathcal{M}}$ matrix that has for each $\mu, \nu \in \mathcal{M}$ the component

$$\left\|\hat{\eta}_n^\mu(x) - \hat{\eta}_n^\nu(x)\right\|^2_{\hat{\boldsymbol{T}}_n^{\nu^{-1}}(x)} + \operatorname{tr}\left(\hat{\boldsymbol{T}}_n^\mu(x)\hat{\boldsymbol{T}}_n^{\nu^{-1}}(x)\right) - n_y \tag{8.21b}$$

in row μ and column ν. The novel Kullback-Leibler distance (KLD)-based lower-bound criterion introduced in Thm. 5.9 is

$$\Gamma_n(x) := \rho\left(\tfrac{1}{2}\boldsymbol{\Gamma}_n(x), \pi_n\right), \tag{8.22a}$$

where $\boldsymbol{\Gamma}_n(x)$ is a $n_{\mathcal{M}} \times n_{\mathcal{M}}$ matrix which has for all $\mu, \nu \in \mathcal{M}$ the component

$$\begin{aligned}\left\|\hat{\eta}_n^\mu(x) - \hat{\eta}_n^\nu(x)\right\|^2_{\hat{\boldsymbol{W}}_n^\nu(x)^{-1}} &+ \operatorname{tr}\left(\hat{\boldsymbol{W}}_n^\mu(x)\hat{\boldsymbol{W}}_n^\nu(x)^{-1}\right)\\ &- \ln\det\left(\hat{\boldsymbol{W}}_n^\mu(x)\hat{\boldsymbol{W}}_n^\nu(x)^{-1}\right) - n_y\end{aligned} \tag{8.22b}$$

in row μ and column ν. Finally, the new entropy-based lower-bound criterion proposed in Thm. 5.10 is

$$L_n(x) := \rho\left(\tfrac{1}{2}\boldsymbol{L}_n(x), \pi_n\right) - \tfrac{1}{2}\sum_{\mu\in\mathcal{M}} \pi_n^\mu \ln\det \hat{\boldsymbol{W}}_n^\mu(x), \tag{8.23a}$$

where $\boldsymbol{L}_n(x)$ is a $n_{\mathcal{M}} \times n_{\mathcal{M}}$ matrix which has for all $\mu, \nu \in \mathcal{M}$ the component

$$\left\|\hat{\eta}_n^\mu(x) - \hat{\eta}_n^\nu(x)\right\|^2_{\hat{\boldsymbol{W}}_n^{\mu\nu}(x)^{-1}} + \ln\det \hat{\boldsymbol{W}}_n^{\mu\nu}(x) - n_y \tag{8.23b}$$

in row μ and column ν, with $\hat{\boldsymbol{W}}_n^{\mu\nu}(x) := \hat{\boldsymbol{W}}_n^\mu(x) + \hat{\boldsymbol{W}}_n^\nu(x)$.

Unified Formulation

Tracing back the definitions reveals that these design criteria may depend on the data d_n and the design ξ_n through the following quantities: (i) the model posterior approximations $\pi(\cdot)$ and the model indices $\hat{\mu}$ and $\hat{\nu}$, which are all based on $\hat{s}_n^\mu$, (ii) the PMLEs $\hat{\theta}_n^\mu$, and (iii) the parameter covariance approximations $\frac{1}{n}\hat{\boldsymbol{M}}_n^{\mu^{-1}}$, $\frac{1}{n}\hat{\boldsymbol{R}}_n^\mu$, or $\frac{1}{n}\hat{\boldsymbol{B}}_n^{\mu^{-1}}$. They can hence all be written in the unified form

$$\Psi_n\left(x; \left(\hat{s}_n^\mu, \hat{\theta}_n^\mu, \boldsymbol{X}_n^\mu,\right)_{\mu\in\mathcal{M}}\right), \tag{8.24}$$

where $\boldsymbol{X}_n^\mu$ is a symmetric $n_{\theta^\mu} \times n_{\theta^\mu}$ matrix.

8.3.3. Reference Design Strategy

As standard of comparison for the sequential design criteria, we consider an design strategy that chooses experimental conditions *independently* from models and from data.

The strategy is realized using low-discrepancy sequences, discussed in Sec. 6.3. For all $n \in \mathbb{N}$, it chooses experimental condition x_n to be the $100 + n$-th member of the RR2-modified Halton sequence of Kocis and Whiten [146, Sec. 2.3]. We refer to it as LOW-DISCREPANCY (LD) STRATEGY.

It provides experimental conditions that are spread in a uniform manner across the experimental domain. Comparing it to one of the optimal experimental design (OED) strategies from Sec. 8.3 reveals how much MD efficiency the considered OED strategy gains by taking into account the models and the data that are available.

8.4. Computational Methods

To examine (Q8.1)–(Q8.3) on p. 244 under controlled conditions, we *define* and thus *know* the correct model $\bar{\mu}$ and its correct parameter $\bar{\theta}$. According to (8.1), we then also know for all $n \in \mathbb{N}$ the distribution of the observable $\mathcal{Y}_{x_n}$, of the sample $\mathcal{D}_n$, and hence the distribution of the random variables $\mathcal{T}_n$ and $\mathcal{P}_n$ of interest. In general, however, the distributions of the latter cannot be expressed in a closed form.

8.4.1. A Monte-Carlo Method

We use the following Monte-Carlo method to approximate the quantities of interest. Let $t_{n1}, \dots, t_{nr}$ be independently and identically distributed (IID) replications of $\mathcal{T}_n$. Their sample mean and their sample standard deviation

$$t_n := \tfrac{1}{r} \sum_{l=1}^{r} t_{nl} \quad \text{and} \quad \sigma_n^t := \left(\tfrac{1}{r-1} \sum_{l=1}^{r} \left(t_{nl} - t_n \right)^2 \right)^{½} \tag{8.25}$$

are unbiased and consistent estimates of the expectation $\mathbb{E}\left[\mathcal{T}_n\right]$ and the standard deviation $\mathbb{C}\left[\mathcal{T}_n\right]^{½}$, respectively. These well known relations follow essentially from the weak law of large numbers (Def. B.5 and Thm. B.6) and the continuous mapping theorem (Thm. B.3). Supposed r is sufficiently large, the sample mean t_n therefore measures the average model discrimination (MD) efficiency of experiments 1 to n performed in a sequential design procedure, and the sample standard deviation σ_n^t quantifies the corresponding variability.

Let $\bar{\pi}_{n1}, \dots, \bar{\pi}_{nr}$ be IID replications of $\mathcal{P}_n$. The relative frequency

$$f_n := \tfrac{1}{r}\left|\left\{\bar{\pi}_{nl} : l \in \{1, \dots, r\}, \bar{\pi}_{nl} \geqslant \alpha\right\}\right| \tag{8.26}$$

of the replications exceeding the threshold $\alpha \in (½, 1)$ is a consistent estimate of the probability $\mathbb{P}\left[\mathcal{P}_n \geqslant \alpha\right]$ that the MD problem is considered as solved in iteration n in the sense of (8.10). This relation is an immediate corollary of the property that defines "IID replications" in the first place. Accordingly, f_n measures the efficiency of experiments 1 to n of a sequential design procedure for solving the MD problems, supposed that r is sufficiently large.

Algorithm 8.2 on the next page generates the required replications for n between $s \in \mathbb{N}$ and $n_{\max} \in \mathbb{N}$. It uses the additional subscript $l \in \{1, \dots, r\}$ to indicate that a quantity refers to the l-th replication. The placeholder $\boldsymbol{X}_{nl}^{\mu}$ stands for one of the matrices $\frac{1}{n}\hat{\boldsymbol{M}}_{nl}^{\mu\,-1}$, $\frac{1}{n}\hat{\boldsymbol{R}}_{nl}^{\mu}$, or $\frac{1}{n}\hat{\boldsymbol{B}}_{nl}^{\mu\,-1}$. The algorithm can be applied to any design criterion that can be written in the form (8.24). Its output can be used directly to examine (Q8.1) on p. 244 using the discussed measures. Running it with different observation covariances $\boldsymbol{\Omega}$ and with model families of different size $n_{\mathcal{M}}$ allows to study (Q8.3) and (Q8.2), respectively.

The condition $s \geqslant \max_{\mu \in \mathcal{M}} \{n_{\theta^\mu}\}/n_y$ for the number of initial experiments s ensures that the sum of squared residuals (SSR) (8.3) and the noncentrality (8.2) have at least as many summands as the number of parameters n_{θ^μ} in all models $\mu \in \mathcal{M}$, a necessary condition for the uniqueness of their minimizers.

8.4.2. Implementation in DoeSim

An implementation of Alg. 8.2 is available in our software package DoeSim.

For generating random variates, it uses MATLAB's `mrg32k3a` pseudo-random number generator, which combines the 32-bit combined multiple recursive generator of L'Ecuyer [167] with the Ziggurat algorithm of Marsaglia and Tsang [180].

To compute a parameter maximum-likelihood estimate (PMLE) $\hat{\theta}_n^{\mu}$ one needs to solve a least-squares (LSQ) problem whose objective function is the SSR (8.3). Since the considered models may be both nonlinear and incorrect, the LSQ problem may also be nonlinear and may exhibit large residuals even in the solution. The problem of finding a parameter-minimizer of the noncentrality in the T-efficiency (8.7) has similar properties. Suitable numerical methods

Algorithm 8.2: Monte Carlo method for assessing MD efficiency of sequential design criteria.

```
input  : dimension $n_y \in \mathbb{N}$ of observables, experimental domain $\mathscr{X} \subseteq \mathbb{R}^{n_x}$
         full-rank $n_y \times n_y$ observation covariance matrix $\boldsymbol{\Omega}$
         model index set $\mathscr{M}$, parameter domains $\mathscr{Q}^{\mu} \in \mathbb{R}^{n_{\theta^{\mu}}}$ for all $\mu \in \mathscr{M}$
         response function $\eta^{\mu} : \mathscr{X} \times \mathscr{Q}^{\mu} \mapsto \mathbb{R}^{n_y}$ for all $\mu \in \mathscr{M}$
         initial experimental conditions $x_1, \dots, x_s \in \mathscr{X}$, with $s \geqslant \max_{\mu \in \mathscr{M}} \{n_{\theta^{\mu}}\} / n_y$
         maximal number of experiments $n_{\max} > s$
         sequential design criterion $\Psi_n : \mathscr{X} \mapsto \mathbb{R}$ for all $n \geqslant s$
         correct model $\bar{\mu} \in \mathscr{M}$, correct parameter $\bar{\theta} \in \mathscr{Q}^{\bar{\mu}}$
         number $r \in \mathbb{N}$ of replications
output : MD efficiencies $(t_{nl}, \bar{\pi}_{nl} : n \in \{s, \dots, n_{\max}\}, l \in \{1, \dots, r\})$

let $\bar{\eta}(x) := \eta^{\bar{\mu}}\big(x, \bar{\theta}\big)$ for all $x \in \mathscr{X}$;
foreach $l \in \{1, \dots, r\}$ do
    for $n = 1$ to $s$ do
        experiment: get random variate $y_{nl}$ of $\mathcal{N}(\bar{\eta}(x_n), \boldsymbol{\Omega})$;        // see (8.1)
    end
    for $n = s$ to $n_{\max}$ do
        let $\xi_{nl}$ be the design constituted by $x_1, \dots, x_n$;
        let $d_{nl}^{\top} := \begin{bmatrix} y_{1l}^{\top} & \dots & y_{nl}^{\top} \end{bmatrix}$;
        foreach $\mu \in \mathscr{M}$ do
            compute PMLE $\hat{\theta}_{nl}^{\mu}$ and $\hat{s}_{nl}^{\mu}$ from $\xi_{nl}$, $d_{nl}$, $\eta^{\mu}$, and $\boldsymbol{\Omega}$;    // see (8.3)
            compute $\boldsymbol{X}_{nl}^{\mu}$ from $\xi_{nl}$, $d_{nl}$, $\eta^{\mu}$, and $\boldsymbol{\Omega}$;        // see (8.12)--(8.14)
        end
        compute $t_{nl} := T_n(\xi_{nl})$ from $(\eta^{\mu})_{\mu \in \mathscr{M}}$, $\boldsymbol{\Omega}$, $\bar{\mu}$, and $\bar{\theta}$;         // see (8.7)
        compute $(\pi(\mu \mid \xi_{nl}, d_{nl}))_{\mu \in \mathscr{M}}$ from $\big(\hat{s}_{nl}^{\mu}\big)_{\mu \in \mathscr{M}}$;         // see (8.9)
        let $\bar{\pi}_{nl} := \pi(\bar{\mu} \mid \xi_{nl}, d_{nl})$;
        compute $x_{n+1,l} \in \operatorname*{argmax}_{x \in \mathscr{X}} \Psi_n\Big(x; \big(\hat{s}_{nl}^{\mu}, \hat{\theta}_{nl}^{\mu}, \boldsymbol{X}_{nl}^{\mu}\big)_{\mu \in \mathscr{M}}\Big)$;        // see (8.24)
        experiment: get random variate $y_{n+1,l}$ of $\mathcal{N}(\bar{\eta}(x_{n+1,l}), \boldsymbol{\Omega})$; // see (8.1)
    end
end
return $(t_{nl}, \bar{\pi}_{nl} : n \in \{s, \dots, n_{\max}\}, l \in \{1, \dots, r\})$
```

for such problems are discussed in Sec. 6.1. Our implementation applies the Broyden-Fletcher-Goldfarb-Shanno (BFGS) [53, 100, 107, 228] quasi-Newton method provided by the MATLAB function `fminunc`.

The BFGS method requires first derivatives of the model response with respect to the parameter. For evaluating the matrix $\boldsymbol{X}_{nl}^{\mu}$, first and possibly also second derivatives of the response with respect to the parameter are required. The implementation uses the complex step differentiation technique introduced by Lyness and Moler [178], reviewed by Martins, Sturdza, and Alonso [181] for computing first derivatives in machine precision. For computing second derivatives, the implementation combines this technique with finite central differences.

The considered design criteria listed in Sec. 8.3 are generally not convex in $x \in \mathcal{X}$. The optimal experimental design (OED) problem is thus a non-convex optimization problem, which is hard to solve numerically, as discussed in Sec. 6.2. In our implementation, the OED problem is solved by a grid search on an equidistant rectangular grid, as discussed in Sec. 6.2.4. This approach is computationally tractable for low-dimension experimental domains, like that of the water-gas shift reaction (WGSR) model family with $n_x = 4$.

Algorithm 8.2 may be computationally demanding. To give some typical numbers, applying the algorithm for $n_{\max} - s = 100$ additional experiments and $r = 1\,000$ Monte Carlo (MC) runs involves the solution of approximately 200 000 LSQ problems of increasing size and 100 000 OED problems, which may take a considerable amount of computing time.

Fortunately, the algorithm can parallelized to a large extent. In particular, the individuals runs of **foreach**-loops can be run concurrently, which includes the potentially expensive solutions of the LSQ problems and the OED problems. The implementation is parallelized with respect to the outermost **foreach**-loop, such that essentially all expensive computations arising from different replications are performed concurrently.

9. Performance of Design Criteria for MD: Numerical Case Study

Contents

This chapter examines the practical performance of established and newly proposed sequential design criteria for model discrimination (MD) from Chaps. 4 and 5 in a numerical case study. The study uses the numerical framework developed in the previous Chap. 8. The considered MD problems are based on the water-gas shift reaction (WGSR) model family introduced in Sec. 7.4.

Section 9.1 describes the general setting of the study. Section 9.2 we discuss in detail the observed behavior of the design criteria in discrimination problems among two models. Sections 9.3 and 9.4 contain the results from MD problems among three or more models, respectively. Their behavior under varying the magnitude of the data variability is described in Sec. 9.5.

9.1. Considered Scenario

Various model discrimination (MD) problems can be derived from the water-gas shift reaction (WGSR) family described in Sec. 7.4. For each subset $\mathcal{N} \subseteq \mathcal{M} = \{1, \dots, 13\}$ of at least $|\mathcal{N}| \geqslant 2$ models one can formulate a MD problem. We select some of them based on the following arguments.

For notational convenience, we identify and distinguish these problems by explicitly stating the model index set $\mathcal{N}$. For example, we write $\{1, 5, 9\}$ to refer to the problem of discriminating between the WGSR models 1, 5, and 9.

9.1.1. Correct Model

We *define* that a rescaled variant of model 1 is correct,

$$\bar{\eta}(x) := \eta^1\big(x, \bar{\theta}\big)/c \text{ for all } x \in \mathcal{X}, \tag{9.1a}$$

with the correct parameter

$$\bar{\theta} := \begin{bmatrix} 1.6855 & 4.5947 & 0.94219 & 0.89669 & 2.4591 \end{bmatrix}^\top. \tag{9.1b}$$

The factor

$$c := 165 \approx \frac{\int_{\mathcal{X}} \eta^1\big(x, \bar{\theta}\big)\, \mathrm{d}x}{\int_{\mathcal{X}} \mathrm{d}x} \tag{9.1c}$$

normalizes the observation mean $\bar{\eta}$ approximately to an average value of 1 on the experimental domain. The responses (7.24) of the WGSR model family are normalized accordingly. This normalization allows to interpret the standard deviation σ of the observation directly as magnitude of the *relative* measurement error. Schwaab et al. [225] uses the same correct model and correct parameter without the rescaling.

9.1.2. Well-Posed Discrimination Problems in the WGSR Family

A MD problem is well-posed only if (a) the model family contains a correct model, and (b) if the correct model is unique, see Sec. 4.1.1. Since we define model 1 to be correct, it must be one among of the rival models, leaving $n_{\mathcal{M}} - 1 = 12$ other models to choose from. Of the $2^{12} - 1 = 4095$ MD problems that are possible in total, not all are reasonable due to the following constraints.

Model $\mu \in \mathscr{M}$ is a SPECIAL CASE of model $\nu \in \mathscr{M}$, and model ν is a GENERALIZATION of model μ, iff for each parameter $\theta^\mu \in \mathscr{Q}^\mu$ there exists a parameter $\theta^\nu \in \mathscr{Q}^\nu$ such that $\eta^\mu(x, \theta^\mu) = \eta^\nu(x, \theta^\nu)$ for all $x \in \mathscr{X}$. It follows immediately from Cor. 3.6 that a generalization of a correct model is also correct, and that a special case of an incorrect model is incorrect.

When faced with MD problem in practice, it is unknown for each model whether it is incorrect or correct. If it is incorrect, then all special cases of it (if they exist) are also incorrect – and thus not of interest. If it is correct, however, then its special cases might also be correct. If such a correct special case exists, however, condition (b) is violated. To ensure that a MD problem is well-posed and to reduce its complexity, one would therefore consider only the most general ones among the rival models.

We respect these constraints in our choice of WGSR-based MD problems: The most general models in the WGSR family are models 1, 2, 3, 5, 6, 10, and 11, as shown in Fig. 7.1 on p. 228. Each of the remaining models 4, 7, 8, 9, 12 and 13 is a special case of at least one other model.

9.1.3. Initial Experiments

Table 9.1.: Initial experimental conditions for the WGSR model family, rounded to five digits.

n	x_n			
1	0.19102	0.72634	0.0728	0.32697
2	0.66602	0.40967	0.8328	0.86983
3	0.42852	0.30412	0.4528	0.19125
4	0.90352	0.93745	0.2628	0.73411
5	0.13164	0.62078	0.6428	0.46268
6	0.60664	0.19856	0.2248	0.13309
7	0.36914	0.83189	0.9848	0.67595
8	0.84414	0.51523	0.6048	0.40452
9	0.25039	0.08128	0.4148	0.94738
$s = 10$	0.72539	0.71461	0.7948	0.26880

The first $s = 10$ initial experiments are performed under the pairwise distinct conditions listed in columns 2 to 4 of Tab. 9.1. They are identical with the 10 first experimental conditions provided by low-discrepancy (LD) strategy.

9.1.4. Numerical Settings

All computations are performed with our DoeSim implementation of Alg. 8.2. The optimal experimental design (OED) grid search uses an rectangular equidistant grid (6.18) with a distance of $d = 0.05$, corresponding to $20^4 = 160\,000$ grid points in total. For computing the relative frequency (8.26) we used a threshold of $\alpha = 0.99$ for the posterior probability. All results are based on $r = 2\,000$ replications.

9.2. Two-Model Problems

We begin with two-model discrimination problems and a moderate measurement error of $\sigma = 12.8\%$.

9.2.1. Overview

For the discussion, we define $n^{\Psi}_{0.95}$ as the minimal n for which $f_{n+s} \geqslant 0.95$ for design criterion Ψ. That is, $n^{\Psi}_{0.95}$ is the smallest number of *additional* experiments (to the $s = 10$ initial ones) for which the problem is solved in the sense of (8.10) in 95% of the computed replications. We then say that the design criterion solves the problem RELIABLY with $n^{\Psi}_{0.95}$ experiments. None of our central conclusions depends crucially on the arbitrarily chosen threshold value of 0.95. They would remain unchanged, if a different sufficiently "high" value was used instead.

Table 9.2 on the next page lists the considered water-gas shift reaction (WGSR)-based two-model discrimination problems, approximately in order of decreasing difficulty. The last eight of the listed problems are particularly simple: already one additional experiment, or even the set of initial experiments alone, suffices to solve the problem reliably with any of the considered design criteria H_n, B_n, B'_n, U_n, L_n, or Γ_n.

Figures 9.1 to 9.4 on p. 260 and on pp. 263–265 show the efficiency of the considered design criteria for solving the reasonably difficult problems $\{1, 2\}$, $\{1, 8\}$, $\{1, 3\}$, and $\{1, 11\}$, respectively. The upper and the lower charts plot the percentage f_n of solved problems from (8.26) and the average T-efficiency t_n from (8.26), respectively, as functions of the number of available experiments n.

Table 9.2.: Two-model discrimination problems with $\sigma = 12.8$, approximately in order of decreasing difficulty.

	$n^{\Psi}_{0.95}$, for $\Psi = \dots$							
MD problem	LD	H_n	B_n	U_n	B'_n	L_n	Γ_n	shown in
$\{1,2\}$	–	6	4	6	4	4	6	Fig. 9.1
$\{1,8\}$	–	4	3	4	4	3	4	Fig. 9.2
$\{1,3\}$	–	4	4	3	4	3	3	Fig. 9.3
$\{1,11\}$	–	3	3	2	2	2	3	Fig. 9.4
$\{1,7\}$	6	1	1	1	1	1	1	
$\{1,9\}$	3	1	1	1	1	1	1	
$\{1,12\}$	7	1	1	1	1	1	1	
$\{1,6\}$	0	0	0	0	0	0	0	
$\{1,4\}$	0	0	0	0	0	0	0	
$\{1,5\}$	0	0	0	0	0	0	0	
$\{1,10\}$	0	0	0	0	0	0	0	

In both charts, the abscissa starts at after the $n = 10$ initial experiments have been performed. Since simple guessing solves a two-model model discrimination (MD) problem in 50% of the cases, the charts for f_n are limited to the interval $[½, 1]$. The ordinate axes for t_n have a different scale in each figure to increase readability. A dashed horizontal line in the lower chart indicates a solution percentage of 95%. Two dashed vertical lines in both charts indicate the smallest and the largest $n^{\Psi}_{0.95}$ among all considered design criteria.

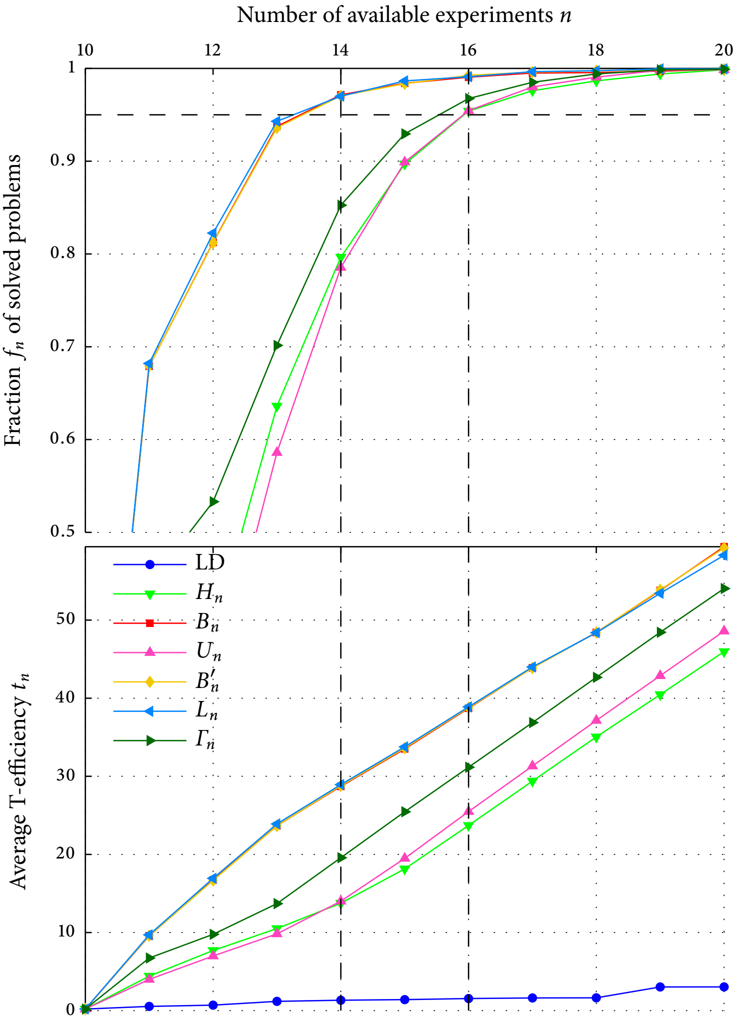
9.2.2. Model 1 vs. Model 2

We first consider the most difficult two-model problem $\{1, 2\}$. The computational results are shown in Fig. 9.1 on the next page.

Observed Performance of Design Criteria

Qualitatively, all considered design criteria H_n, B_n, B'_n, U_n, L_n, and Γ_n behave similarly: the fraction f_n of solved problems increases monotonically towards 1 with the number of available experiments n. This is an essential behavior that one expects from any reasonable MD strategy. Also the average T-efficiency t_n increases monotonically with n.

Figure 9.1.: Efficiency for discriminating between models 1 and 2, measurement error σ = 12.8%. *Top:* Fraction of problems solved based on experiments 1 to n, as function of n. *Bottom:* Analogous results for the average T-efficiency, abscissa scale as in top chart. Legend applies to both charts.



After an initial phase ending between $n = 13$ and $n = 15$, the T-efficiency t_n growths almost linearly for all design criteria, yet with slightly different slopes. Recall from (8.7) that t_n is the product of n and a design-dependent factor. The observed linear increases in n indicate that the underlying designs do not change significantly after the initial phase. Subsequent experiments can thus be seen as samples from those "fixed" designs. The different slopes tell us, however, that these fixed designs differ between the design criteria.

The results from H_n and U_n are similar: their f_n value differ less than 0.03, and their T-efficiencies differ by not more than 10% of their common mean. Design criterion Γ_n performs slightly better. All three of them solve the problem reliably after $n^{\Psi}_{0.95} = 6$ additional experiments.

The results from L_n, B_n, and B'_n are almost identical, and all are significantly better than H_n and U_n: they require only $n^{\Psi}_{0.95} = 4$ additional experiments to solve the problem reliably, and show an accordingly larger T-efficiency.

For the reference strategy low-discrepancy (LD), f_n is less than 50% for n up to 20, so that it is not visible the range of the upper chart. Its T-efficiency increases roughly linearly with the number of experiments, yet with a small slope. For any of the design criteria, the T-efficiency t_n at $n^{\Psi}_{0.95}$ is approximately 25. A rough linear extrapolation with an estimated slope of 0.25 tells us that LD strategy would require $n^{\mathrm{LD}}_{0.95} = 90$ experiments to reach $t_n = 25$ and solve the problem reliably.

Interpretation

In the absence of parameter uncertainty, both the Buzzi-Ferraris (BF)-criterion B_n and its misspecification-robust counterpart B'_n reduce to the Hunter-Reiner (HR)-criterion H_n, see Prop. 4.13 and Sec. 4.4.3. The observation that B_n and B'_n perform significantly better than H_n thus tells us that parameter uncertainty is actually relevant in the considered problem, and that taking it into account for the design of experiments can actually pay off.

The misspecification-robust parameter covariance formula (8.13) used by B'_n remains valid even if a model is nonlinear *and* incorrect, in contrast to the classic formula (8.12) used by B_n. Since B_n and B'_n perform almost equally, these factors do not seem to play an important role in the considered problem.

The classic upper-bound approximation U_n of the Box-Hill-Hunter (BHH)-criterion uses the same matrices (8.12) and (8.15) as B_n for quantifying parameter uncertainty and its effect on the uncertainty of predictions. Furthermore, U_n uses the same model posteriors as L_n for quantifying parameter uncertainty.

Nevertheless, U_n performs considerably worse than these two alternatives. That is, U_n does seem to be able to take advantage of the available uncertainty quantifications. We guess that the reason for its bad performance is the fact that it is actually a *lower* bound for a design criterion which is *maximized*.

9.2.3. Model 1 vs. Model 8, Model 3, and Model 11

Let us now examine the remaining reasonably difficult two-model problems $\{1, 8\}$, $\{1, 3\}$, and $\{1, 11\}$, shown in Figs. 9.2 to 9.4 on pp. 263–265, respectively.

Compared to problem $\{1, 2\}$, their difficulty decreases in the given order: all design criteria require fewer experiments to solve the problem reliably, and the scale of the T-efficiencies is increased accordingly. With a few exceptions, the relative performance of the design criteria remains similar to the one observed previously. In particular, design criteria L_n, B_n, and B'_n perform almost equally and better than all other design criteria. Design criterion H_n performs worst, apart from the LD strategy, and the results from Γ_n lie somewhere between these extremes.

As opposed to problem $\{1, 2\}$, however, design criterion U_n is slightly more efficient than H_n.

Furthermore, the average T-efficiency of Γ_n and L_n grows noticeably sublinear beyond $n^{\Psi}_{0.95}$, almost coming to a stop in $\{1, 11\}$. This behavior can easily be explained: For $n \geqslant s + n^{\Psi}_{0.95}$, the posterior probability of the correct model is close to 1 in most cases, see (8.26). In that case, both Γ_n and L_n are almost constant, see Thm. 5.12, and do not longer provide efficient experimental conditions for MD. In practice, however, this behavior will never be encountered, since a large posterior probability of the correct model means that the MD problem is solved, so that designing further experiments is not necessary.

Based on a linear extrapolation of the T-efficiency, we estimate that the LD strategy would solve problems $\{1, 8\}$, $\{1, 3\}$, and $\{1, 11\}$ reliably with 36, 26 and 30 experiments, respectively. By using any of the design criteria, the problem can be solved reliably with 15% of the experimental effort or less.

9.3. Three-Model Problems

More diverse results for the different design criteria might be seen in more difficult model discrimination (MD) problems.

Figure 9.2.: Efficiency for discriminating between models 1 and 8, measurement error $\sigma = 12.8\%$. *Top:* Fraction of problems solved based on experiments 1 to n, as function of n. *Bottom:* Analogous results for the average T-efficiency, abscissa scale as in top chart. Legend applies to both charts.

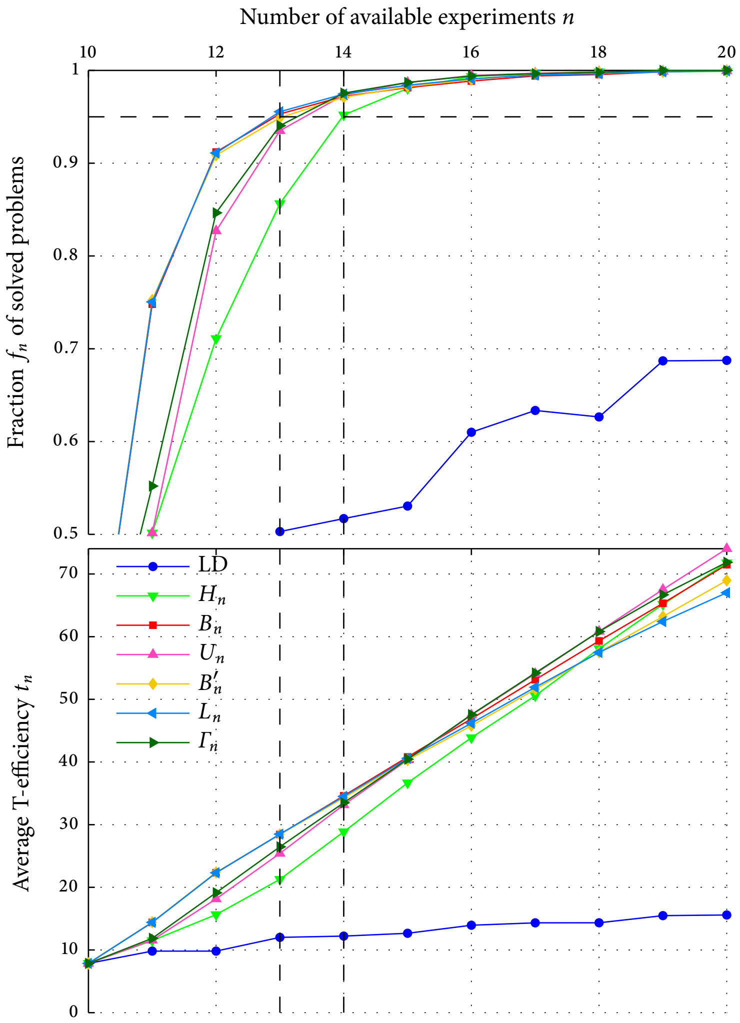


Figure 9.3.: Efficiency for discriminating between models 1 and 3, measurement error σ = 12.8%. *Top:* Fraction of problems solved based on experiments 1 to n, as function of n. *Bottom:* Analogous results for the average T-efficiency, abscissa scale as in top chart. Legend applies to both charts.

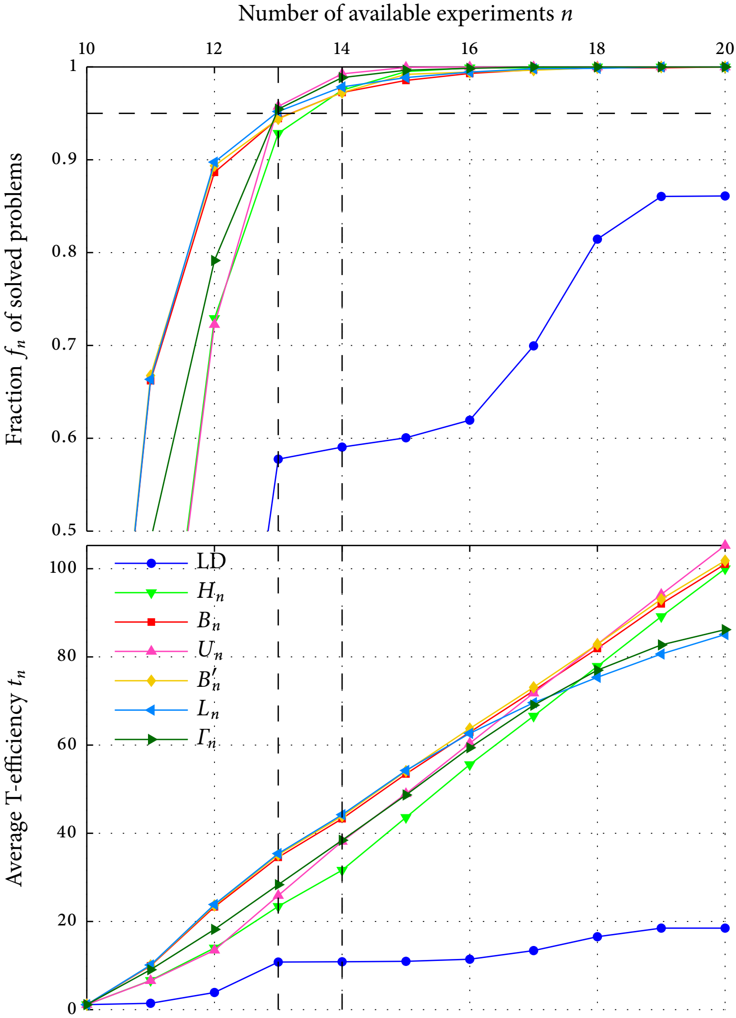
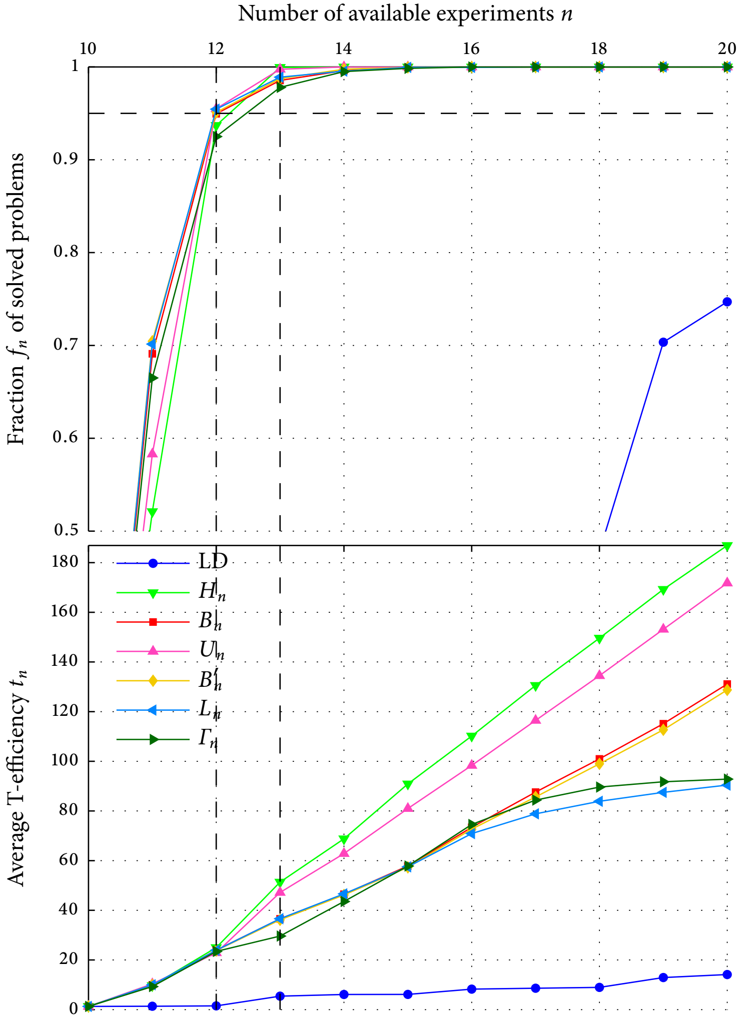


Figure 9.4.: Efficiency for discriminating between models 1 and 11, measurement error $\sigma = 12.8\%$. *Top:* Fraction of problems solved based on experiments 1 to n, as function of n. *Bottom:* Analogous results for the average T-efficiency, abscissa scale as in top chart. Legend applies to both charts.



9.3.1. Overview

Table 9.3.: Three-model discrimination problems with σ = 12.8, approximately in order of decreasing difficulty.

	$n_{0.95}^{\Psi}$							
MD problem	LD	H_n	B_n	U_n	B'_n	L_n	Γ_n	shown in
$\{1,2,3\}$	–	7	6	–	6	6	6	Fig. 9.5
$\{1,2,11\}$	–	6	5	10	6	5	6	Fig. 9.6
$\{1,2,5\}$	–	6	4	–	4	4	6	Fig. 9.7
$\{1,2,6\}$	–	6	4	–	4	4	6	
$\{1,2,10\}$	–	6	4	–	4	4	6	
$\{1,3,11\}$	–	4	4	4	4	3	4	Fig. 9.8
$\{1,3,5\}$	–	4	4	14	4	3	3	Fig. 9.9
$\{1,3,6\}$	–	4	4	–	4	3	3	
$\{1,3,10\}$	–	4	4	–	4	3	3	
$\{1,5,11\}$	–	3	3	13	2	2	3	Fig. 9.10
$\{1,6,11\}$	–	3	3	–	2	2	3	
$\{1,10,11\}$	–	3	3	–	2	2	3	
$\{1,5,6\}$	0	0	0	0	0	0	0	
$\{1,5,10\}$	0	0	0	0	0	0	0	
$\{1,6,10\}$	0	0	0	0	0	0	0	

Table 9.3 lists the water-gas shift reaction (WGSR)-based three-model discrimination problems that meet the requirements from Sec. 9.1, roughly in order of decreasing difficulty. The last three of them are so simple that they are already solved reliably based on the 10 initial experiments.

Figures 9.5 to 9.10 on pp. 267–272 give a representative overview of the MD performances encountered in three-model problems. The results of problems involving models 6 or 10 are not shown, since they are very similar to those from the corresponding problems with model 5. The figures have the same layout as the previous ones, described in Sec. 9.2.

The considered design criteria behave qualitatively similar as in two-model problems, with two notable exceptions discussed later. Their MD efficiency has the same ranking in all problems: design criteria L_n and B_n behave almost indistinguishable and perform best, followed by Γ_n and by H_n; all of them are substantially better than the low-discrepancy (LD) strategy.

Figure 9.5.: Efficiency for discriminating between models 1, 2, and 3, measurement error $\sigma = 12.8\%$. *Top:* Fraction of problems solved based on experiments 1 to n, as function of n. *Bottom:* Analogous results for the average T-efficiency, abscissa scale as in top chart. Legend applies to both charts.

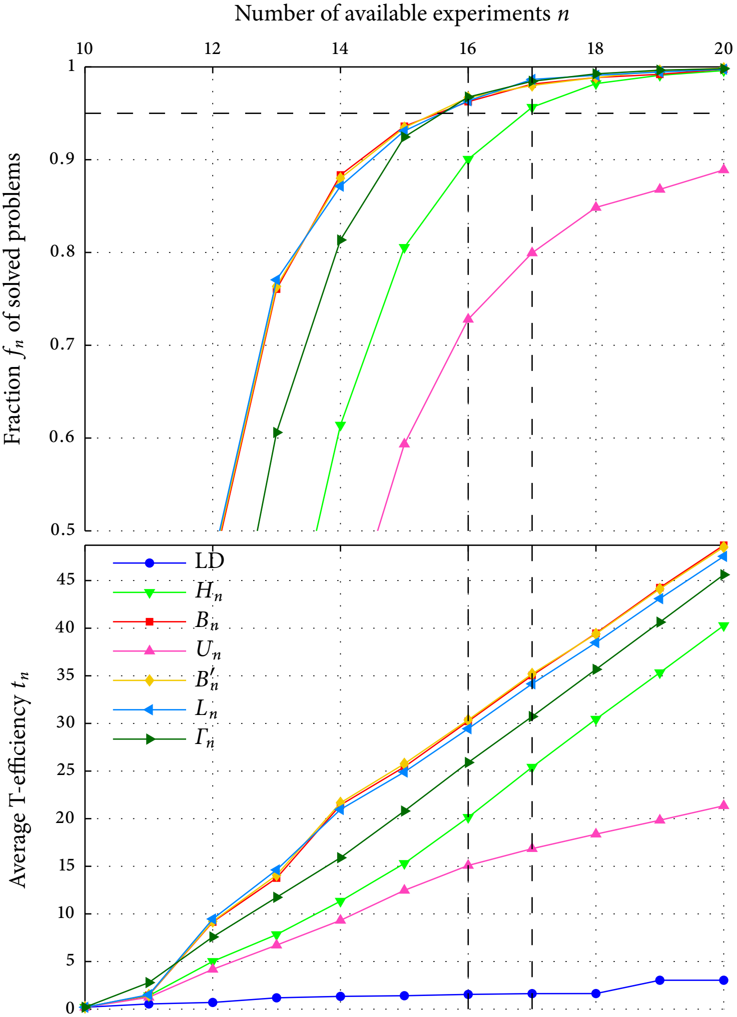


Figure 9.6.: Efficiency for discriminating between models 1, 2, and 11, measurement error σ = 12.8%. *Top:* Fraction of problems solved based on experiments 1 to n, as function of n. *Bottom:* Analogous results for the average T-efficiency, abscissa scale as in top chart. Legend applies to both charts.

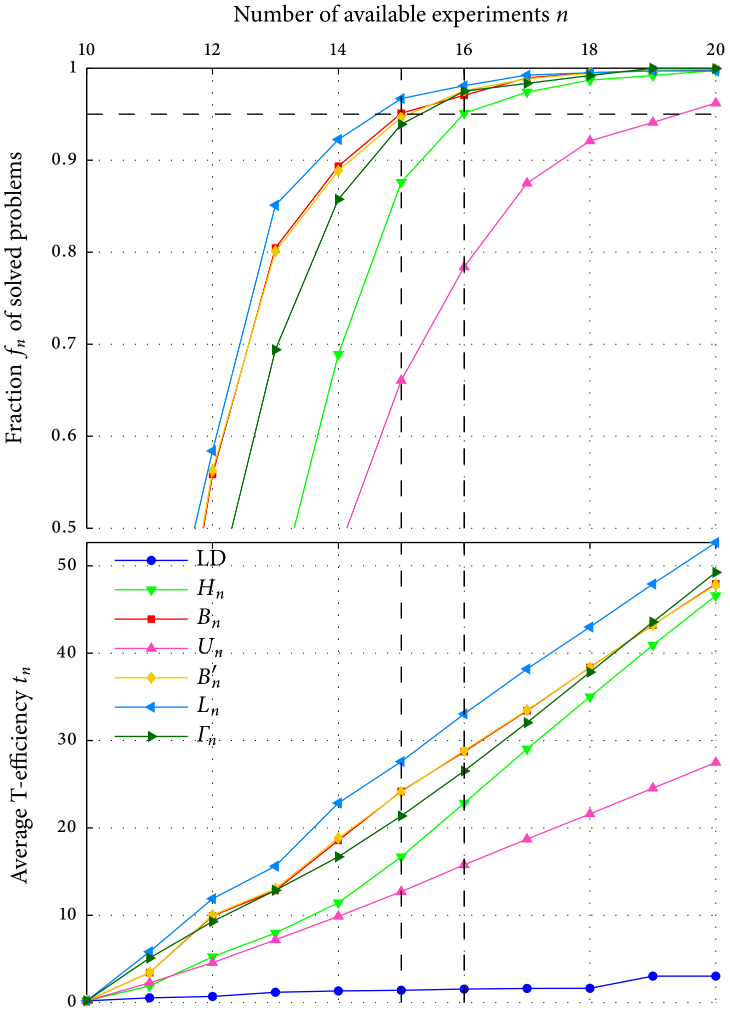


Figure 9.7.: Efficiency for discriminating between models 1, 2, and 5, measurement error $\sigma = 12.8\%$. *Top:* Fraction of problems solved based on experiments 1 to n, as function of n. *Bottom:* Analogous results for the average T-efficiency, abscissa scale as in top chart. Legend applies to both charts.

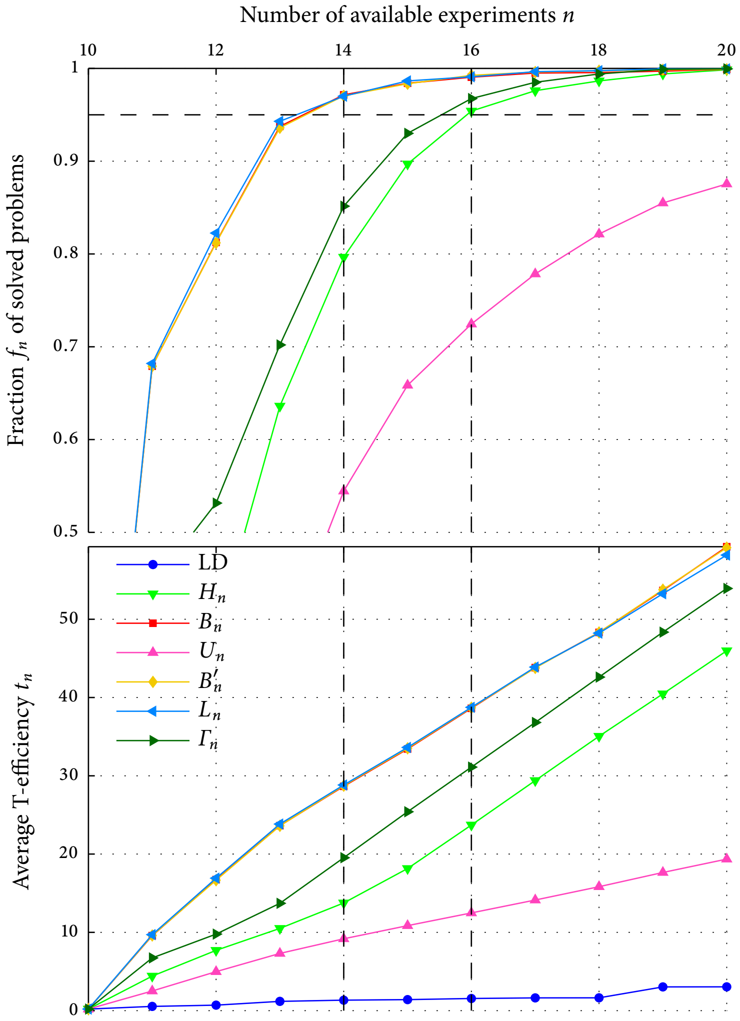


Figure 9.8.: Efficiency for discriminating between models 1, 3, and 11, measurement error $\sigma = 12.8\%$. *Top:* Fraction of problems solved based on experiments 1 to n, as function of n. *Bottom:* Analogous results for the average T-efficiency, abscissa scale as in top chart. Legend applies to both charts.

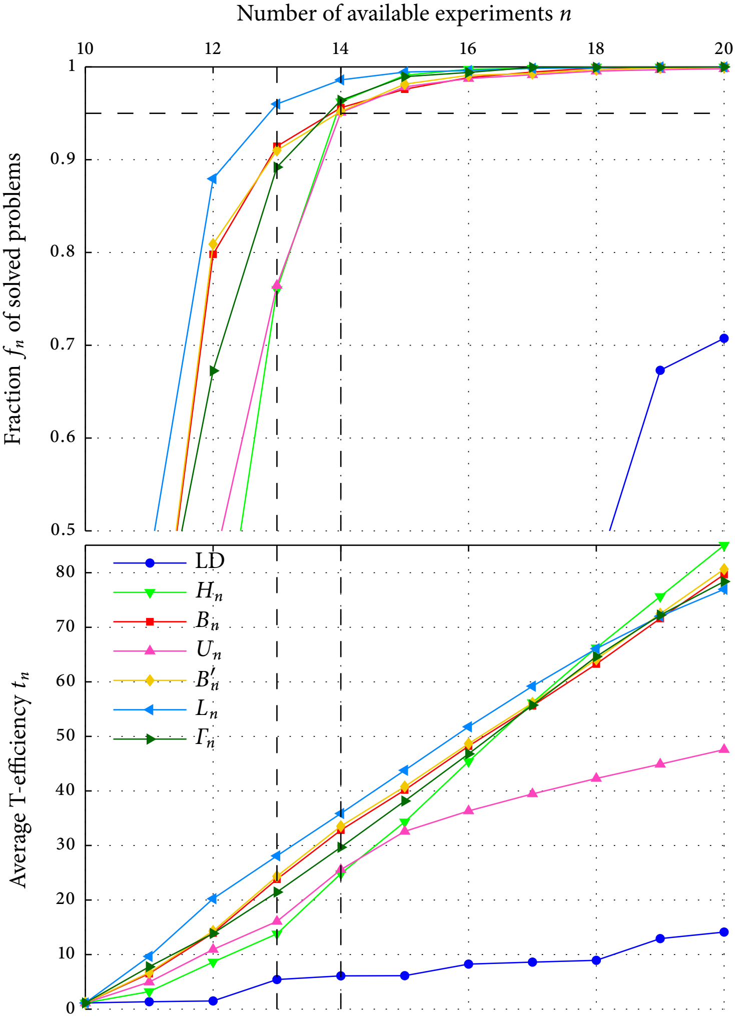


Figure 9.9.: Efficiency for discriminating between models 1, 3, and 5, measurement error σ = 12.8%. *Top:* Fraction of problems solved based on experiments 1 to n, as function of n. *Bottom:* Analogous results for the average T-efficiency, abscissa scale as in top chart. Legend applies to both charts.

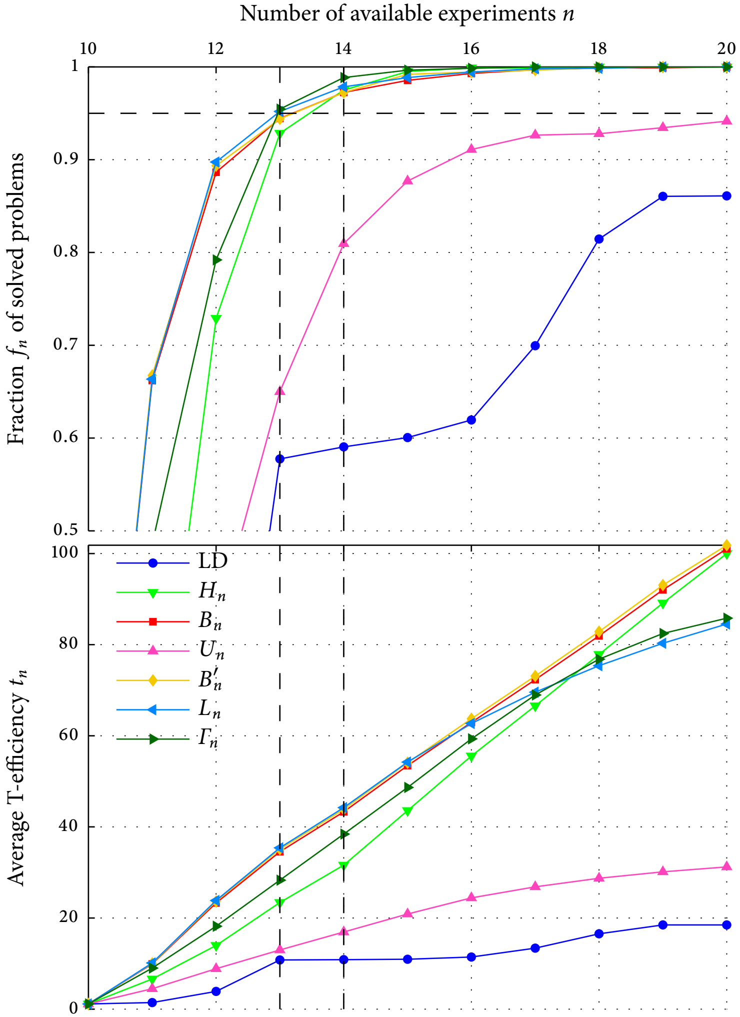
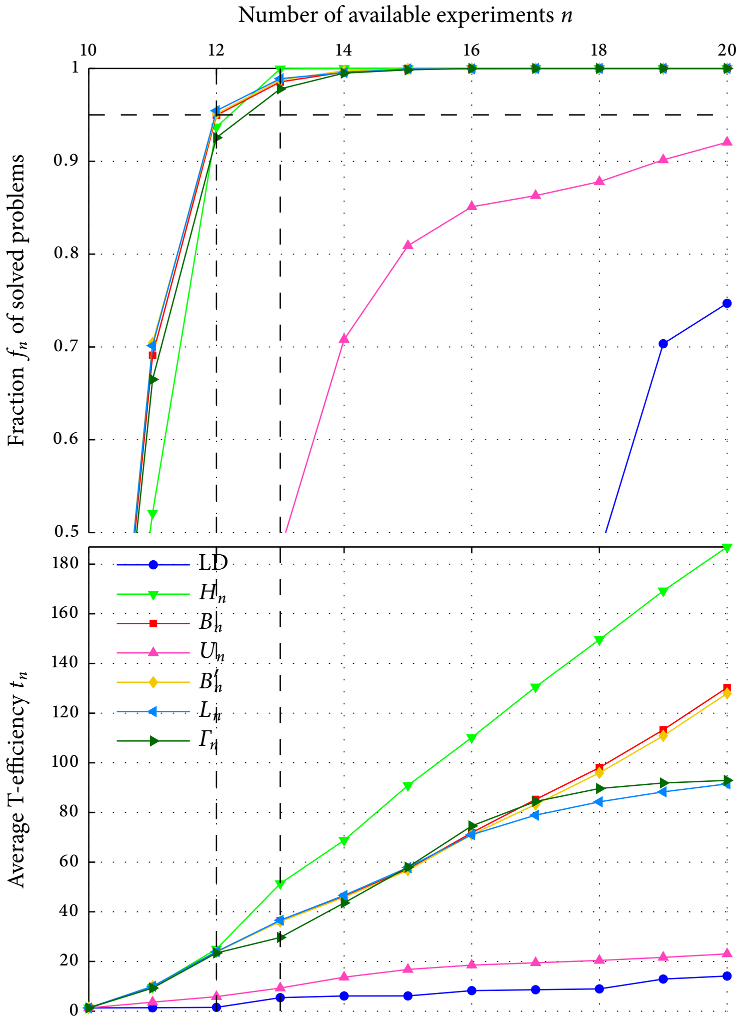


Figure 9.10.: Efficiency for discriminating between models 1, 5, and 11, measurement error $\sigma = 12.8\%$. *Top:* Fraction of problems solved based on experiments 1 to n, as function of n. *Bottom:* Analogous results for the average T-efficiency, abscissa scale as in top chart. Legend applies to both charts.



Differences can be observed in the behavior of the classic upper-bound approximation L_n of the Box-Hill-Hunter (BHH)-criterion and the entropy-based lower-bound approximation L_n.

9.3.2. Inefficiency of U_n

While it performed comparable to H_n in the two-model case, design criterion U_n is in all three-model problems *significantly* less efficient in terms of both f_n and t_n. It fails to solve eight problems reliably in the considered range of 25 experiments. In problem $\{1, 3, 5\}$, it is not even notably better than the model-independent LD strategy.

It is instructive to compare the behavior of U_n in problems $\{1, 2\}$ and $\{1, 3\}$ with that in the combined problem $\{1, 2, 3\}$. In the former, its MD efficiency is very similar to that of H_n, while it is significantly worse than in the latter. In the other three-model problems, the relations are similar. Obviously, it is the number of rival models, not the nature of the rival models themselves, which causes its bad performance. As mentioned previously, we conjecture that its nature as *lower bound* is the cause.

9.3.3. Improved Efficiency of L_n

In all considered two-model problems, the results from L_n are almost indistinguishable from those of B_n and B'_n. This is also the case for most three-model problems, with the notable exceptions in problems $\{1, 2, 11\}$ and $\{1, 3, 11\}$. There, L_n is more efficient than B_n and B'_n, and thus the most efficient of the considered design criteria. While the advantage is not tremendous – a gain of 5 in terms of T-efficiency – it arises already after one or two additional experiment are performed, and persists until the problem is solved reliably at $n = s + n^L_{0.95}$.

Problem $\{1, 2, 11\}$ can be regarded as the result of extending problem $\{1, 2\}$ by rival model 11. In the latter, all three design criteria perform comparably well. In $\{1, 2, 11\}$, the T-efficiency of B_n and B'_n is approximately 10 units lower than in $\{1, 2\}$, while it drops only by 5 units for L_n. Analogous observation is made when comparing $\{1, 3\}$. and $\{1, 3, 11\}$.

It seems that L_n is equally or less affected by the increased problem complexity due to an additional model than the other two design criteria.

9.4. Multi-Model Problems

We further examine how the model discrimination (MD) efficiency of the design criteria changes with the number of rival models. As examples, we consider the MD problems $\{1, 2, 3, 11\}$ and $\{1, 2, 3, 5, 6, 10, 11\}$, the presumably most difficult well-posed four-model problem and the largest well-posed problem in the water-gas shift reaction (WGSR) family, respectively In addition, we examine the problem $\{1, \ldots, 13\}$ of discriminating between *all* models of the WGSR family. Since it contains several nested models, see Fig. 7.1 on p. 228, this MD problem is not well-posed, as discussed in Sec. 9.1. The results for this case might nevertheless be valuable for a practitioner, who might find it hard to ensure the condition of non-nested models.

Table 9.4 summarizes the key results, Figs. 9.11, 9.12 and 9.13 on the next page, on p. 276 and on p. 277 show the MD efficiencies in detail.

Table 9.4.: Discrimination problems among more than three models with $\sigma = 12.8$.

	$n^{\Psi}_{0.95}$, for $\Psi = \ldots$							
MD problem	LD	H_n	B_n	U_n	B'_n	L_n	Γ_n	shown in
$\{1, 2, 3, 11\}$	–	7	7	–	7	6	6	Fig. 9.11
$\{1, 2, 3, 5, 6, 10, 11\}$	–	7	7	–	7	6	6	Fig. 9.12
$\{1, \ldots, 13\}$	–	8	7	–	7	6	6	Fig. 9.13

We observe that the results from $\{1, 2, 3, 11\}$ and $\{1, 2, 3, 5, 6, 10, 11\}$ are almost identical. In both problems, design criterion L_n shows the significantly better MD efficiency than all other design criteria first shown in problems $\{1, 2, 3\}$ and $\{1, 2, 11\}$. Problem $\{1, 2, 3, 11\}$ differs from the previously discussed problems $\{1, 2, 3\}$, $\{1, 2, 11\}$ and $\{1, 3, 11\}$ only by one additional rival model. Only the MD efficiency of U_n suffers significantly from extending the problem by the models 5, 6, and 10.

Besides these differences, the design criteria show no previously unseen behavior.

In all previous problems, the Kullback-Leibler distance (KLD)-based lower-bound criterion Γ_n is for all n less efficient than B_n and B'_n both in terms of f_n and t_n. In problem $\{1, \ldots, 13\}$, however, Γ_n is more efficient in terms of f_n for $n \geqslant 15$, yet remains worse as measured by t_n.

Figure 9.11.: Efficiency for discriminating between models 1, 2, 3, and 11, measurement error σ = 12.8%. *Top:* Fraction of problems solved based on experiments 1 to n, as function of n. *Bottom:* Analogous results for the average T-efficiency, abscissa scale as in top chart. Legend applies to both charts.

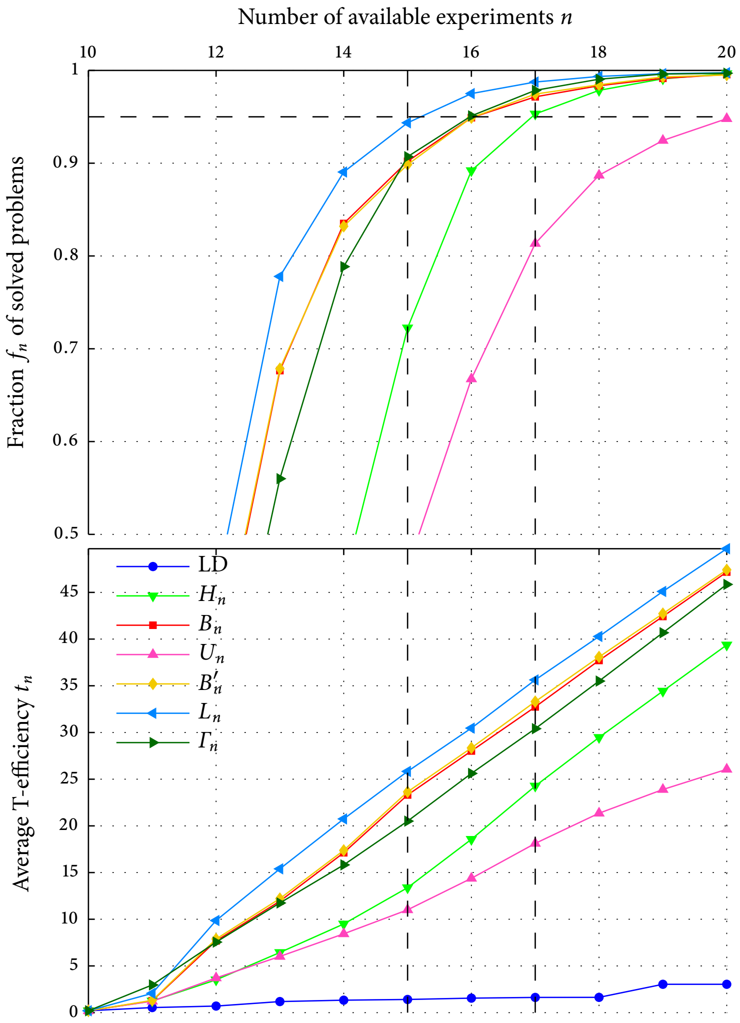


Figure 9.12.: Efficiency for discriminating between models 1, 2, 3, 5, 6, 10, and 11, measurement error $\sigma = 12.8\%$. *Top:* Fraction of problems solved based on experiments 1 to n, as function of n. *Bottom:* Analogous results for the average T-efficiency, abscissa scale as in top chart. Legend applies to both charts.

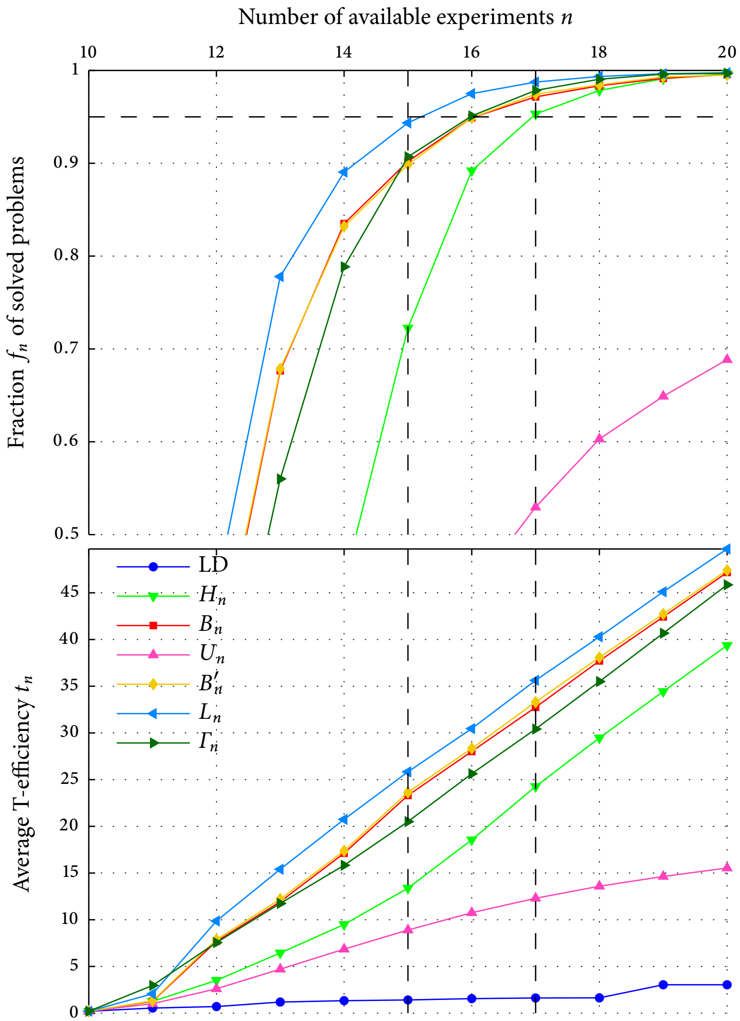
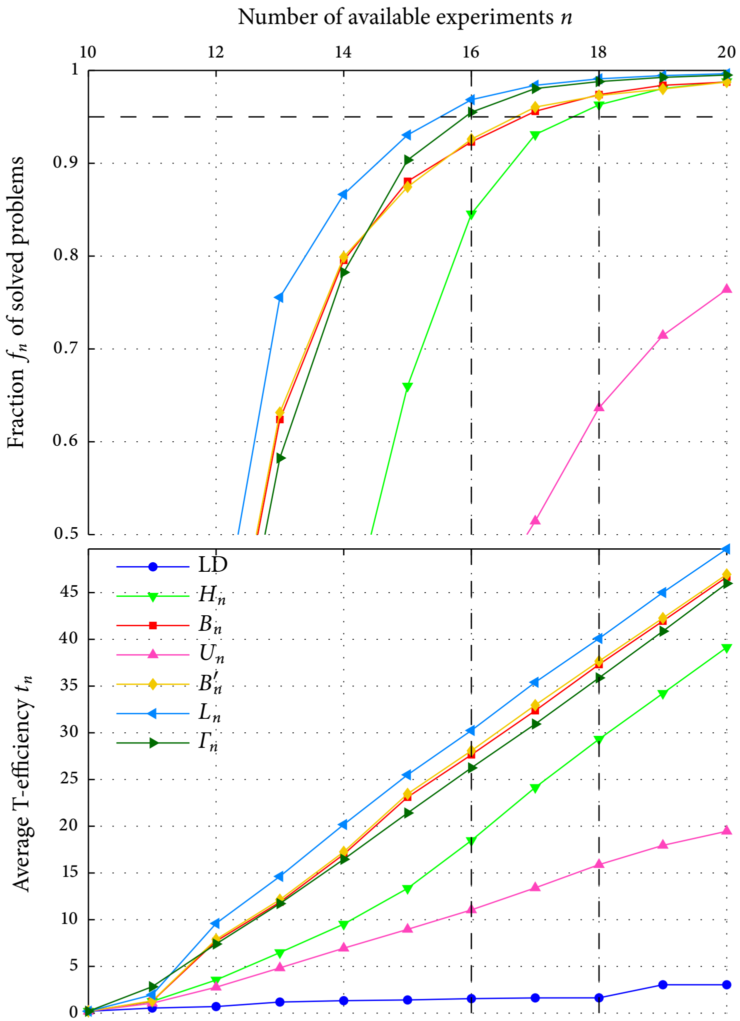


Figure 9.13.: Efficiency for discriminating between models 1 to 13, measurement error $\sigma = 12.8\%$. *Top:* Fraction of problems solved based on experiments 1 to n, as function of n. *Bottom:* Analogous results for the average T-efficiency, abscissa scale as in top chart. Legend applies to both charts.



9.5. Dependency on Data Variability

This section examines how the model discrimination (MD) efficiency of the design criteria depends on the variability of the data, as stated in (Q8.3). To that end, a subset of the previously discussed MD problems was considered under measurement errors σ of different magnitudes, as shown in Tab. 9.5.

Table 9.5.: Performance of design criteria under different magnitudes of measurement error.

	$n_{0.95}^{\Psi}$, for $\Psi = \dots$							
MD problem	LD	H_n	B_n	U_n	B'_n	L_n	Γ_n	shown in
Measurement error $\sigma = 6.4\%$								
$\{1,2\}$	–	2	1	2	1	1	1	
$\{1,2,3\}$	–	3	2	3	2	2	2	
$\{1,2,3,11\}$	–	3	3	4	2	2	2	
$\{1,2,3,5,6,10,11\}$	–	3	3	4	2	2	2	
$\{1,\dots,13\}$	–	3	3	4	2	2	2	
Measurement error $\sigma = 12.8\%$								
$\{1,2\}$	–	6	4	6	4	4	6	Fig. 9.1
$\{1,2,3\}$	–	7	6	–	6	6	6	Fig. 9.5
$\{1,2,3,11\}$	–	7	7	–	7	6	6	Fig. 9.11
$\{1,2,3,5,6,10,11\}$	–	7	7	–	7	6	6	Fig. 9.12
$\{1,\dots,13\}$	–	8	7	–	7	6	6	Fig. 9.13
Measurement error $\sigma = 25.6\%$								
$\{1,2\}$	–	22	22	24	22	22	24	
$\{1,2,3\}$	–	24	28	–	–	26	26	Fig. 9.14
$\{1,2,3,11\}$	–	24	28	–	28	26	25	
$\{1,2,3,5,6,10,11\}$	–	24	29	–	28	27	25	Fig. 9.15
$\{1,\dots,13\}$	–	–	–	–	–	27	26	Fig. 9.16
Measurement error $\sigma = 51.2\%$								
$\{1,2\}$	–	93	111	109	114	109	105	
$\{1,2,3\}$	–	108	–	–	–	120	116	
$\{1,2,3,11\}$	–	107	117	–	–	112	110	
$\{1,2,3,5,6,10,11\}$	–	107	119	–	–	113	111	Fig. 9.17
$\{1,\dots,13\}$	–	–	–	–	–	120	–	Fig. 9.18

Under $\sigma = 6.4\%$, the considered problems were solved reliably with four or less additional experiments by any design criterion. In these few steps, their behavior shows no significant differences to the case $\sigma = 12.8\%$, which was examined in detail in Secs. 9.2 to 9.4.

In all considered problems with more than two rival models, the design criterion U_n lags far behind all of its competitors, regardless of the measurement error. We thus do not consider it in the following discussion.

Under the large measurement errors of $\sigma = 25.6\%$ and $\sigma = 51.2\%$, we observe the following qualitatively new behavior.

In problems $\{1, 2, 3\}$, $\{1, 2, 3, 11\}$, and $\{1, 2, 3, 5, 6, 10, 11\}$, the relative performance of the design criteria substantially differs from that observed for 6.4% and 12, 8%. Design criterion H_n is most efficient both in terms of f_n and t_n, followed in short distance by L_n, Γ_n, B_n, and B_n'.

In problem $\{1, \ldots, 13\}$ however, the order again changes completely: there, design criterion L_n is by far most efficient, followed closely by Γ_n, as in the case of small-error case. Design criterion H_n is significantly less efficient, and B_n and B_n even more. Under $\sigma = 25.6\%$, the T-efficiency of design criterion L_n is 11.8% larger than that of H_n at $n = 36$, under $\sigma = 51.2\%$, it is 14.2% larger at $n = 120$.

9.5.1. Summary

Let us summarize our results for MD problems between three or more models.

The considered examples show that the newly developed design criteria L_n and Γ_n can provide substantially more efficient experiments for MD than common alternatives. In almost all of the considered cases, either L_n or Γ_n perform best in terms of the T-efficiency, in some cases together with B_n and B_n'. The advantage of L_n or Γ_n is particular big in the largest MD problem that involves *all* model of the water-gas shift reaction (WGSR) family, including the nested ones. Under small and moderately large measurement errors, L_n was outperformed by B_n.

Design criteria B_n and B_n' differ only in their parameter maximum-likelihood estimator (PMLE) covariance formulas, whose difference increases with the noncentrality. Both design criteria use a multi-model generalization which reduces the multi-model optimal experimental design (OED) problem to that of discriminating between the two best-fitting models. Therefore, both design criteria involve the parameter covariances of two models only, and these two models tend to have a *small* noncentrality. Consequentially, both design criteria behave similarly, as observed here. One might see larger discrepancies if one chooses a different multi-model generalization which involves the PMLE covariances of more models and/or with larger noncentrality.

Figure 9.14.: Efficiency for discriminating between models 1, 2, and 3, measurement error $\sigma = 25.6\%$. *Top:* Fraction of problems solved based on experiments 1 to n, as function of n. *Bottom:* Analogous results for the average T-efficiency, abscissa scale as in top chart. Legend applies to both charts.

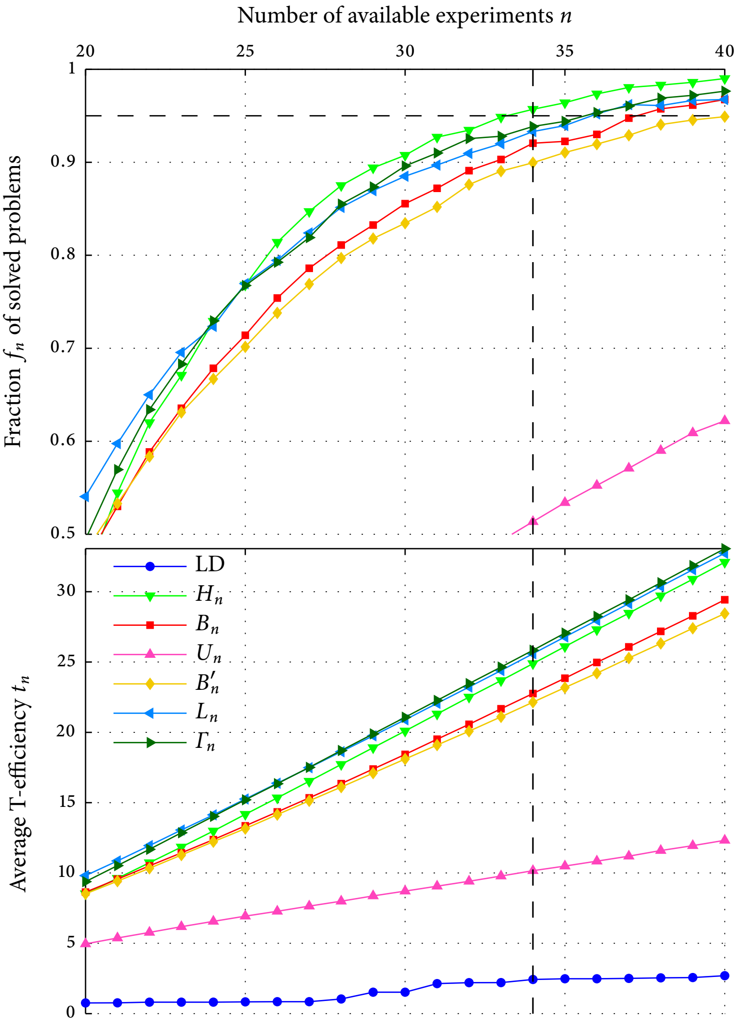


Figure 9.15.: Efficiency for discriminating between models 1, 2, 3, 5, 6, 10, and 11, measurement error $\sigma = 25.6\%$. *Top:* Fraction of problems solved based on experiments 1 to n, as function of n. *Bottom:* Analogous results for the average T-efficiency, abscissa scale as in top chart. Legend applies to both charts.

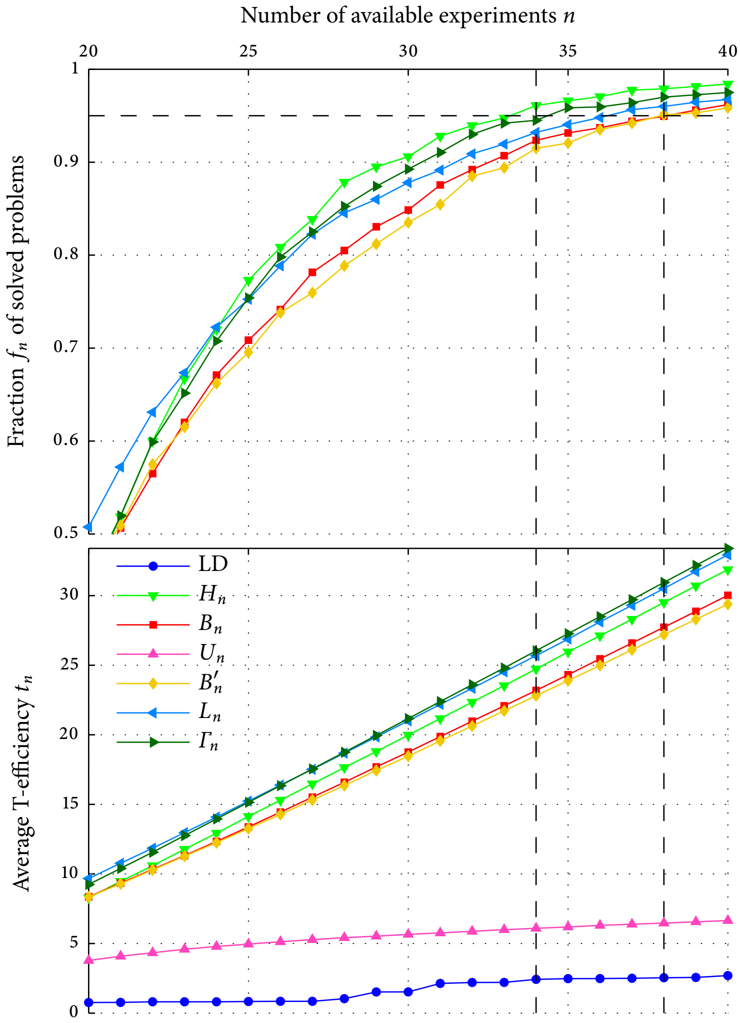


Figure 9.16.: Efficiency for discriminating between models 1 to 13, measurement error $\sigma = 25.6\%$. *Top:* Fraction of problems solved based on experiments 1 to n, as function of n. *Bottom:* Analogous results for the average T-efficiency, abscissa scale as in top chart. Legend applies to both charts.

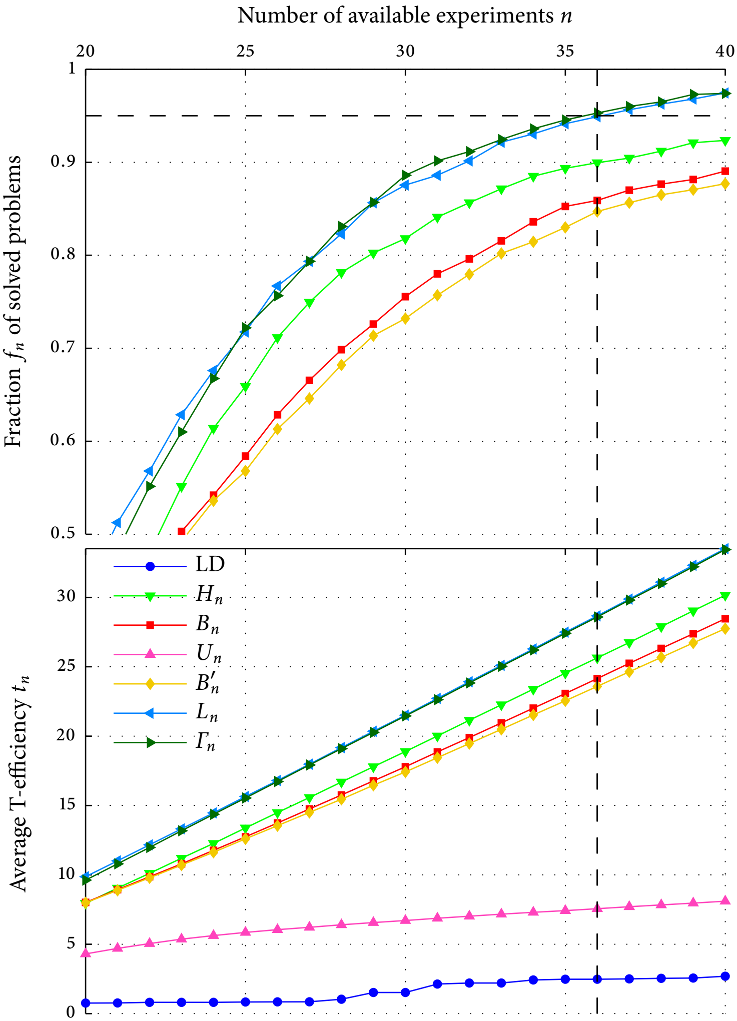


Figure 9.17.: Efficiency for discriminating between models 1, 2, 3, 5, 6, 10, and 11, measurement error σ = 51.2%. *Top:* Fraction of problems solved based on experiments 1 to n, as function of n. *Bottom:* Analogous results for the average T-efficiency, abscissa scale as in top chart. Legend applies to both charts.

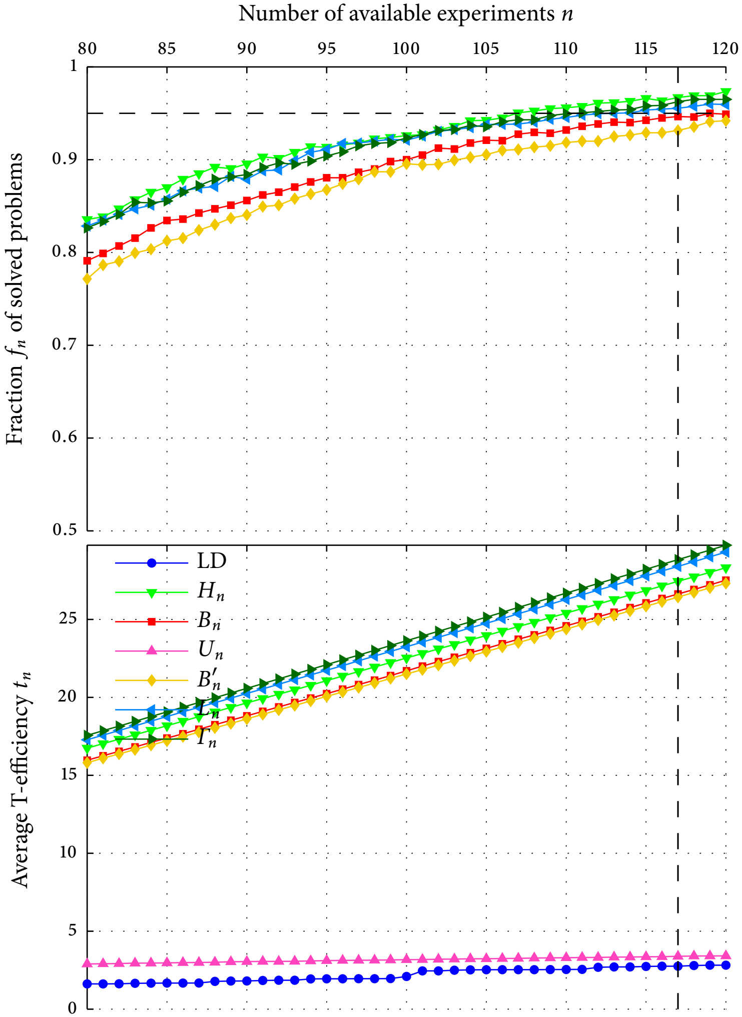
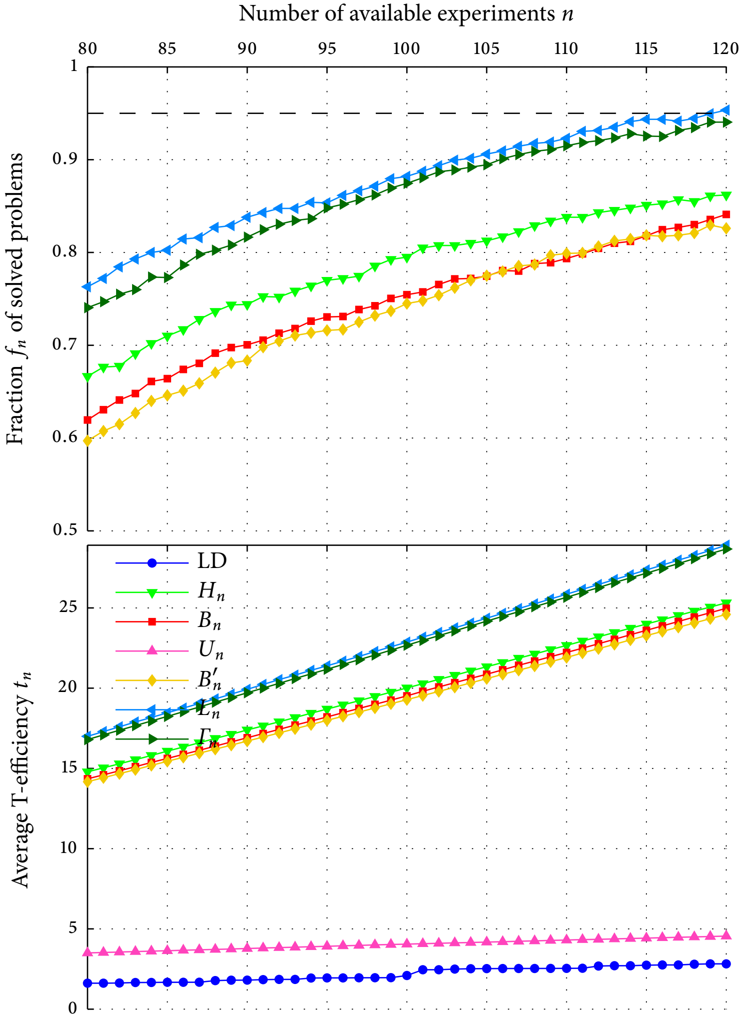


Figure 9.18.: Efficiency for discriminating between models 1 to 13, measurement error $\sigma = 51.2\%$. *Top:* Fraction of problems solved based on experiments 1 to n, as function of n. *Bottom:* Analogous results for the average T-efficiency, abscissa scale as in top chart. Legend applies to both charts.



The classic upper-bound approximation U_n of the Box-Hill-Hunter (BHH)-criterion has shown to be substantially less efficient than any other design criterion, and even worse than the low-discrepancy (LD) reference strategy. Although U_n is one of best-examined design criteria for MD, a similar observation has not been published before to the best of our knowledge.

Appendices

A. Supplements

In this appendix, we reprint some frequently used definitions and theorems from *other* fields than statistics and probability theory – the latter can be found in the next appendix.

Definition A.1 (Gradient, Hessian, and Jacobian)

Let $m, n \in \mathbb{N}$ and let $\mathcal{X} \subseteq \mathbb{R}^m$. The gradient of a differentiable scalar function $f : \mathcal{X} \mapsto \mathbb{R}$ is the function $\nabla f : \mathcal{X} \mapsto \mathbb{R}^m$ defined as

$$\nabla f(x) := \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} & \dots & \frac{\partial f(x)}{\partial x_m} \end{bmatrix} \text{ for all } x \in \mathcal{X}. \tag{A.1}$$

The Hessian (matrix) of a twice differentiable scalar function $f : \mathcal{X} \mapsto \mathbb{R}$ is the function $\nabla^2 f : \mathcal{X} \mapsto \mathbb{R}^{m \times m}$ defined as

$$\nabla^2 f(x) := \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1 \, \partial x_1} & \dots & \frac{\partial^2 f(x)}{\partial x_1 \, \partial x_m} \\ \vdots & & \vdots \\ \frac{\partial^2 f(x)}{\partial x_m \, \partial x_1} & \dots & \frac{\partial^2 f(x)}{\partial x_m \, \partial x_m} \end{bmatrix} \text{ for all } x \in \mathcal{X}. \tag{A.2}$$

Let $f_1, \dots, f_n$ be the scalar components of the differentiable vector-valued function $f : \mathcal{X} \mapsto \mathbb{R}^n$. The Jacobian (matrix) of f is the function $\nabla f : \mathcal{X} \mapsto \mathbb{R}^{n \times m}$ defined as

$$\nabla f(x) := \begin{bmatrix} \frac{\partial f_1(x)}{\partial x_1} & \dots & \frac{\partial^2 f_1(x)}{x_m} \\ \vdots & & \vdots \\ \frac{\partial f_n(x)}{\partial x_1} & \dots & \frac{\partial f_n(x)}{\partial x_m} \end{bmatrix} \text{ for all } x \in \mathcal{X}. \tag{A.3}$$

Note that the gradient is a *row* vector in our convention. The definition of the Hessian matrix implies that it is symmetric.

Theorem A.2 (Matrix Square Root)

Let $\boldsymbol{A}$ be a real-valued symmetric positive semi-definite (SPSD) $m \times m$ matrix with eigenvalues $\lambda_1, \ldots, \lambda_m$. Then, there exists a unique symmetric positive definite (SPD) $m \times m$ matrix $\boldsymbol{A}^{\frac{1}{2}}$, the MATRIX SQUARE ROOT OF $\boldsymbol{A}$, such that $\boldsymbol{A}^{\frac{1}{2}} \boldsymbol{A}^{\frac{1}{2}} = \boldsymbol{A}$. It can be written as

$$\boldsymbol{A}^{\frac{1}{2}} = \boldsymbol{\Lambda} \boldsymbol{\Gamma} \boldsymbol{\Lambda}^{\top}, \tag{A.4}$$

where $\boldsymbol{\Gamma} = \operatorname{diag}\left(\lambda_1^{\frac{1}{2}}, \ldots, \lambda_m^{\frac{1}{2}}\right)$ and $\boldsymbol{\Lambda}$ is the matrix of eigenvectors of $\boldsymbol{A}$.

Proof Proofs are given in most books on matrix algebra, for example in that of Horn and Johnson [121, Thm. 7.2.6]. □

It is easy to verify that $\left(\boldsymbol{A}^{\frac{1}{2}}\right)^{-1} = \left(\boldsymbol{A}^{-1}\right)^{\frac{1}{2}} =: \boldsymbol{A}^{-\frac{1}{2}}$ and thus that $\boldsymbol{A}^{\frac{1}{2}} \boldsymbol{A}^{-\frac{1}{2}} = \boldsymbol{I}$.

Definition A.3 (Weighted Frobenius Norm)

Let $\boldsymbol{A}$ be a real-valued $m \times n$ matrix, and let $\boldsymbol{W}$ be a real-valued SPD $m \times m$ matrix. The ($\boldsymbol{W}$-)WEIGHTED FROBENIUS NORM of $\boldsymbol{A}$ is

$$\|\boldsymbol{A}\|_{\boldsymbol{W}} := \sqrt{\operatorname{tr}(\boldsymbol{A}^{\top} \boldsymbol{W} \boldsymbol{A})}. \tag{A.5}$$

For $\boldsymbol{W} = \boldsymbol{I}$, the norm reduces to the Frobenius norm and is written as $\|\cdot\|_{\mathrm{F}}$.

For $\boldsymbol{W} = \boldsymbol{I}$, the weighted vector norm reduces to the Euclidean norm. It is easy to show that $\|\boldsymbol{A}\|^2_{\boldsymbol{W}_1 + \boldsymbol{W}_2} = \|\boldsymbol{A}\|^2_{\boldsymbol{W}_1} + \|\boldsymbol{A}\|^2_{\boldsymbol{W}_2}$. If $n = 1$, the matrix $\boldsymbol{A}$ is a column vector. If we denote this vector a, we have $\|a\|^2_{\boldsymbol{W}} = a^{\top} \boldsymbol{W} a$. Any positive definite quadratic form in a can thus be written using the weighted Frobenius norm.

Proposition A.4 (Combination of Quadratic Forms)

Let $a, b, c \in \mathbb{R}^m$ and let $\boldsymbol{A}, \boldsymbol{B}$ be real-valued symmetric $m \times m$ matrices for which the inverse of $\boldsymbol{C} := \boldsymbol{A} + \boldsymbol{B}$ exists, and define $c := \boldsymbol{C}^{-1}(\boldsymbol{A}a + \boldsymbol{B}b)$. Then,

$$\|x - a\|^2_{\boldsymbol{A}} + \|x - b\|^2_{\boldsymbol{B}} = \|x - c\|^2_{\boldsymbol{C}} + \|a - b\|^2_{\boldsymbol{A}\boldsymbol{C}^{-1}\boldsymbol{B}}. \tag{A.6}$$

Proof The proof follows from Def. A.3 and some simple algebra of a function. □

Proposition A.5 (Gradient and Hessian of a Weighted Vector Norm)

Let $m, n \in \mathbb{N}$ and let $\mathcal{X} \subseteq \mathbb{R}^m$. Suppose $f: \mathbb{R}^m \mapsto \mathbb{R}^n$ is twice differentiable and let $\boldsymbol{J}$ be the Jacobian of f. For all $i \in \{1, \ldots, n\}$, let f_i be the i-th component of f and let $\boldsymbol{H}_i$ be its Hessian. Let $\boldsymbol{W}$ be a real-valued SPD $n \times n$-matrix and let w_{ij} be the components of $\boldsymbol{W}^{½}$. Then it holds for all $x \in \mathcal{X}$ that

$$\nabla \|f(x)\|_{\boldsymbol{W}}^2 = 2\boldsymbol{J}^\top(x)\boldsymbol{W} f(x), \text{ and} \tag{A.7}$$

$$\nabla^2 \|f(x)\|_{\boldsymbol{W}}^2 = 2\boldsymbol{J}^\top(x)\boldsymbol{W}\boldsymbol{J}(x) + 2\sum_{i=1}^{m}\left(\sum_{j=1}^{m} w_{ij} f_i\right)\left(\sum_{k=1}^{m} w_{ik}\boldsymbol{H}_k\right). \tag{A.8}$$

Proof The proof is essentially an application of the chain rule, combined with some basic matrix algebra which can be found in Brookes [52] or Petersen and Pedersen [201]. For $\boldsymbol{W} = \boldsymbol{I}$, the proof can be found in Nocedal and Wright [194, Chap. 10]. The general case for $\boldsymbol{W} \neq \boldsymbol{I}$ follows from replacing $f(x)$ by $\boldsymbol{W}^{½} f(x)$. □

B. Selected Topics from Probability Theory and Statistics

This chapter collects some results from probability theory and statistics for the reader's convenience. Unless said otherwise, they can be found in textbooks, together with a rigorous discussion of elementary concepts like probability, random variables and distributions. Standard references are the books of Gut [111], Jaynes [134], Kallenberg [137], and Shao [232].

The symbols $\mathbb{P}[\cdot]$, $\mathbb{E}[\cdot]$, and $\mathbb{C}[\cdot]$ denote the probability, the expectation and the covariance, respectively.

Theorem B.1 (Expectation of a Quadratic Form)

Let $\mathcal{U}$ be an s-dimensional continuous or discrete random variable, and let $\boldsymbol{A} \in \mathbb{R}^{s \times s}$. Then,

$$\mathbb{E}\left[\mathcal{U}^\top \boldsymbol{A}\mathcal{U}\right] = \mathbb{E}[\mathcal{U}]^\top \boldsymbol{A}\,\mathbb{E}[\mathcal{U}] + \operatorname{tr}(\boldsymbol{A}\,\mathbb{C}[\mathcal{U}]). \tag{B.1}$$

Proof A proof is given, for example, by Muirhead [190]. □

This result holds regardless of the distribution of $\mathcal{U}$. In particular, it does not assume that $\mathcal{U}$ is normally distributed. If $\boldsymbol{A}$ is symmetric positive definite (SPD), (B.1) can be rewritten using the identities $\mathcal{U}^\top \boldsymbol{A}\mathcal{U} = \|\mathcal{U}\|_{\boldsymbol{A}}^2$, $\mathbb{E}[\mathcal{U}]^\top \boldsymbol{A}\,\mathbb{E}[\mathcal{U}] = \|\mathbb{E}[\mathcal{U}]\|_{\boldsymbol{A}}^2$, and $\operatorname{tr}(\boldsymbol{A}\,\mathbb{C}[\mathcal{U}]) = \|\mathbb{C}[\mathcal{U}]^{½}\|_{\boldsymbol{A}}^2$ (which requires that $\mathcal{U}$ has full rank) that follow from Def. A.3 and Thm. A.2.

Definition B.2 (Convergence of Random Variables)

Let $\mathcal{U}$ and $(\mathcal{U}_i : i \in \mathbb{N})$ be continuous (discrete) s-dimensional random variables and $\|\cdot\|$ some norm on $\mathbb{R}^s$.

(i) The sequence $\mathcal{U}_1, \mathcal{U}_2, \ldots$ CONVERGES IN PROBABILITY to $\mathcal{U}$, written as $\mathcal{U}_n \xrightarrow{\text{p}} \mathcal{U}$ for $n \to \infty$, iff $\lim_{n\to\infty} \mathbb{P}[\|\mathcal{U}_n - \mathcal{U}\| < \epsilon] = 1$, for all $\epsilon > 0$.

(ii) The sequence $\mathcal{U}_1, \mathcal{U}_2, \ldots$ converges almost surely to $\mathcal{U}$, written as $\mathcal{U}_n \xrightarrow{\text{a.s.}} \mathcal{U}$ for $n \to \infty$, iff $\mathbb{P}\left[\lim_{n\to\infty} \mathcal{U}_n = \mathcal{U}\right] = 1$.

Almost sure convergence implies convergence in probability,

$$\mathcal{U}_n \xrightarrow{\text{a.s.}} \mathcal{U} \Rightarrow \mathcal{U}_n \xrightarrow{\text{p}} \mathcal{U}. \tag{B.2}$$

If $\mathcal{U}_n$ and $\mathcal{U}$ are *discrete* random variables, the reverse is also true. That is, almost sure convergence and convergence in probability are equivalent for discrete random variables,

$$\mathcal{U}_n \xrightarrow{\text{a.s.}} \mathcal{U} \Leftrightarrow \mathcal{U}_n \xrightarrow{\text{p}} \mathcal{U}. \tag{B.3}$$

Theorem B.3 (Continuous Mapping)

Let $\mathcal{U}$ and $(\mathcal{U}_i : i \in \mathbb{N})$ be m-dimensional random variables. Suppose that $f\colon \mathbb{R}^m \mapsto \mathbb{R}^n$ is measurable and continuous almost surely on $\mathbb{R}^n$. Then,

(i) $\mathcal{U}_n \xrightarrow{\text{p}} \mathcal{U} \Rightarrow f(\mathcal{U}_n) \xrightarrow{\text{p}} f(\mathcal{U})$, and

(ii) $\mathcal{U}_n \xrightarrow{\text{a.s.}} \mathcal{U} \Rightarrow f(\mathcal{U}_n) \xrightarrow{\text{a.s.}} f(\mathcal{U})$.

Theorem B.4 (Generalized Slutsky's Theorem)

Let $(\mathcal{V}_n : n \in \mathbb{N})$ be s-dimensional random variables, let $\mathscr{A}$ be a closed and bounded subset of $\mathbb{R}^s$, let $c \in \mathscr{A}$ be some constant, and let f be a continuous function defined on the domain $\mathscr{A}$. Furthermore, let $(\mathcal{U}_n : n \in \mathbb{N})$ be r-dimensional random variables and let $(f_n : n \in \mathbb{N})$ be functions defined on the domain $\mathbb{R}^r \times \mathbb{R}^s$. Suppose that

$$\mathcal{V}_n \xrightarrow{\text{a.s.}} c \text{ and } f_n(\mathcal{U}_n, v) \xrightarrow{\text{p}} f(v) \text{ uniformly in } v \in \mathscr{A}, \tag{B.4}$$

for $n \to \infty$. Then,

$$f_n(\mathcal{U}_n, \mathcal{V}_n) \xrightarrow{\text{a.s.}} f(c), \text{ for } n \to \infty. \tag{B.5}$$

If the $\mathcal{V}_n$ converge only in probability to c, then also $f_n(\mathcal{U}_n, \mathcal{V}_n)$ converges only in probability.

Proof A proof is given, for example, by Bierens [30, Thms. 6.12 and 6.15]. □

An important special case is that $\mathcal{U}_n$ are almost surely constant. Then, they effectively behave like fixed numbers and can be absorbed in the functions f_n. The theorem then essentially states that if

$$\mathcal{V}_n \xrightarrow{\text{a.s.}} c \text{ and } f_n \xrightarrow{\text{p}} f \text{ uniformly}, \text{ for } n \to \infty, \tag{B.6}$$

then

$$f_n(\mathcal{V}_n) \xrightarrow{\text{a.s.}} f(c), \text{ for } n \to \infty. \tag{B.7}$$

Definition B.5 (Laws Of Large Numbers)

Let $(\mathcal{U}_i : i \in \mathbb{N})$ be s-dimensional random variables with finite expectation.

(i) The sequence $\mathcal{U}_1, \mathcal{U}_2, \ldots$ obeys the WEAK LAW OF LARGE NUMBERS, iff

$$\tfrac{1}{n} \sum_{i=1}^{n} (\mathcal{U}_i - \mathbb{E}[\mathcal{U}_i]) \xrightarrow{\text{p}} 0, \text{ for } n \to \infty. \tag{B.8}$$

(ii) The sequence $\mathcal{U}_1, \mathcal{U}_2, \ldots$ obeys the STRONG LAW OF LARGE NUMBERS, iff the convergence in (B.8) is almost surely.

Note that it is neither assumed that the random variables are independent not that they are identically distributed. The strong law implies the weak law due to (B.2). For discrete random variables, both laws are equivalent according to (B.3).

Bernoulli [29] was the first to proof what was later termed a "law of large numbers" by Poisson [204]. The possibly first complete proof of a law of large number for arbitrary random variables was given by Khinchin [141]. Since then, various proofs have been given under different sets of assumptions. An short and elementary proof of the strong law of large numbers for independently and identically distributed (IID) random variables is given by Etemadi [89, Thm. 1]. Hu, Rosalsky, and Volodin [124] and Kuczmaszewska [154] provide fairly general sufficient conditions that allow the random variables to be statistically dependent. The following conditions suffice for this thesis.

Theorem B.6 (Sufficient Conditions for Laws of Large Numbers)

Consider Def. B.5 and assume all random variables of the sequence $(\mathcal{U}_i : i \in \mathbb{N})$ have finite expectations.

(i) The sequence obeys the weak law of large numbers if there is a constant $p \in [1, 2]$ such that

$$\lim_{n \to \infty} \frac{1}{n^p} \sum_{i=1}^{n} \mathbb{E}\left[|\mathcal{U}_i|^p \right] = 0. \tag{B.9}$$

(ii) The sequence obeys the strong law of large numbers if there is a constant $p \in [1, 2]$ such that

$$\sum_{i=1}^{\infty} \frac{1}{i^p} \mathbb{E}\left[|\mathcal{U}_i|^p \right] < \infty. \tag{B.10}$$

Theorem B.7 (Law of the Iterated Logarithm)

Let $(\mathcal{U}_n : n \in \mathbb{N})$ be iid $\mathbb{R}$-valued random variables with zero mean and variance one. For all $n \in \mathbb{N}$, define $\mathcal{V}_n := \sum_{i=1}^{n} \mathcal{U}_i$. Then, it holds almost surely that

$$\limsup_{n \to \infty} \frac{\mathcal{V}_n}{\sqrt{n \ln \ln n}} = \sqrt{2}. \tag{B.11}$$

Proof The first proof was given by Khintchine [142]. □

Various generalizations of this theorem are available for random variables that are not iid.

Definition B.8 (PDF of a Normal Distribution)

The probability density function (pdf) of a s-dimensional normal distribution with expectation $\mu \in \mathbb{R}^s$ and symmetric positive semi-definite (spsd) covariance matrix $\boldsymbol{C} \in \mathbb{R}^{s \times s}$ is the real-valued non-negative function $\phi_s(\cdot\,|\,\mu, \boldsymbol{C})$ defined for all $u \in \mathbb{R}^m$ as

$$\phi_s(u\,|\,\mu, \boldsymbol{C}) := (2\pi)^{-\frac{s}{2}} \det(\boldsymbol{C})^{-½} \exp\Big(-\tfrac{1}{2}\|u-\mu\|^2_{\boldsymbol{C}^{-1}}\Big) \tag{B.12a}$$

$$= \exp\Big(-\tfrac{1}{2}\Big(\|u-\mu\|^2_{\boldsymbol{C}^{-1}} + \ln\det\boldsymbol{C} + n\ln(2\pi)\Big)\Big). \tag{B.12b}$$

We write $\mathcal{U} \sim \mathcal{N}_s(\mu, \boldsymbol{C})$ to express that the continuous random variable $\mathcal{U}$ is subject to an s-dimensional normal distribution with expectation μ and covariance $\boldsymbol{C}$. The subscript s is omitted if the dimension is clear from the context. We refer to $\mathcal{N}(0, \boldsymbol{I})$ as STANDARD NORMAL DISTRIBUTION.

The normal PDF is symmetric in the first two arguments,

$$\phi_s(u\,|\,\mu, \boldsymbol{C}) = \phi_s(\mu\,|\,u, \boldsymbol{C}), \tag{B.13}$$

for all $u, \mu \in \mathbb{R}^s$. Let $\mathcal{U} \sim \mathcal{N}(\mu, \boldsymbol{C})$ and let $\mathcal{V} \sim \mathcal{N}(\nu, \boldsymbol{D})$. The set of normal distributions is closed under affine-linear transformation,

$$\boldsymbol{A}\mathcal{U} + \boldsymbol{B}\mathcal{V} + c \sim \mathcal{N}\big(\boldsymbol{A}\mu + \boldsymbol{B}\nu + c, \boldsymbol{A}\boldsymbol{C}\boldsymbol{A}^\top + \boldsymbol{B}\boldsymbol{D}\boldsymbol{B}^\top\big) \tag{B.14}$$

for all $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{r\times s}$ and all $c \in \mathbb{R}^r$. Furthermore, the integral over a product of two normal PDFS is again a normal PDF,

$$\int_{\mathbb{R}^s} \phi_s(u\,|\,\mu, \boldsymbol{C})\phi_s(u\,|\,\nu, \boldsymbol{D})\,\mathrm{d}u = \phi(\mu\,|\,\nu, \boldsymbol{C}+\boldsymbol{D}). \tag{B.15}$$

Proofs for are given, for example, in the notes of Ahrendt [1], Larsen [164], and Roweis [218]. More details about the normal distribution are assembled in the books of Johnson, Kotz, and Balakrishnan [135, 136] and Kotz, Balakrishnan, and Johnson [150].

C. Essential Concepts of Information Theory

Information theory deals with the quantification of information. It gives a rigorous meaning to the notion of "information" and clarifies its relation to "data" and "uncertainty". It turns out that one requires (at least) three quantities to capture the concept of information, namely entropy, Kullback-Leibler distance (KLD) and mutual information. We introduce and interpret these quantities here and clarify their relations. We limit our considerations to those results required in this thesis. For more details we refer to the standard works Cover and Thomas [73] and Gray [109].

Information theory originates from the communication-theoretic papers of Shannon [229] and Shannon [230] dealing with transmission, compression and storage of data and the information contained in it. Since then, information theory has rapidly grown into an independent area of research with applications in many fields that have to process data. The original publications have been reprinted in the book of Shannon and Weaver [231], a brief summary of central concepts is given by McMillan [184].

Introduction

Assume that $\mathcal{U}, \mathcal{U}_1, \dots, \mathcal{U}_n$ and $\mathcal{V}$ are continuous random variables with a joint distribution. Let $p(u, v)$ be the joint probability density function (PDF) of $\mathcal{U}$ and $\mathcal{V}$, $p(u \,|\, v)$ be the conditional PDF of $\mathcal{U}$ for given $\mathcal{V}$, $p(v \,|\, u)$ be the conditional PDF of $\mathcal{V}$ for given $\mathcal{U}$, $p(u)$ be the marginal PDF of $\mathcal{U}$, and $p(v)$ be the marginal PDF of $\mathcal{V}$. The same notation is used likewise for the PDFs related to the pair $(\mathcal{U}_2, \mathcal{V})$, for all $2 \in \{1, \dots, n\}$. Note that in this conveniently "overloaded" notation, different PDFs are distinguished through their arguments.

For the sake of brevity the considerations in this chapter are restricted to *continuous* random variables. The counterparts of the following definitions for *discrete* random variables can be obtained by replacing the PDFs by probability mass functions (PMFs) and the integrals by sums. Unless said otherwise, the stated theorems also hold for discrete random variables. In fact, the whole information theory can be formulated in terms of probability measures, which may neither correspond to discrete nor to continuous distributions, see, for example, Gray [109].

Unless said otherwise, all integrals in the following are understood over the respective whole domains.

Entropy

Entropy measures the amount of uncertainty or the amount of unpredictability associated with a random variable. The information-theoretic entropy that we deal with here was introduced by Shannon [229]. It is closely related in form and properties to the thermodynamic entropy of Boltzmann and Gibbs from statistical physics.

Definition C.1 (Entropy)

Let $\mathcal{U}$ and $\mathcal{V}$ be jointly distributed random variables, either both discrete or both continuous. The ENTROPY of $\mathcal{U}$ is the real-valued quantity

$$\mathbb{H}[\mathcal{U}] := -\mathbb{E}\left[\ln p_{\mathcal{U}}(\mathcal{U})\right] \tag{C.1}$$

the (conditional) entropy of $\mathcal{U}$ for given $\mathcal{V}$ is

$$\mathbb{H}[\mathcal{U}\,|\,\mathcal{V}] := -\mathbb{E}\left[\ln p_{\mathcal{U}|\mathcal{V}}(\mathcal{U}\,|\,\mathcal{V})\right], \tag{C.2}$$

and the (joint) entropy of $\mathcal{U}$ and $\mathcal{V}$ is

$$\mathbb{H}[\mathcal{U},\mathcal{V}] := -\mathbb{E}\left[\ln p_{\mathcal{U},\mathcal{V}}(\mathcal{U},\mathcal{V})\right], \tag{C.3}$$

supposed the expectations exist.

For discrete random variables,

$$\mathbb{H}[\mathcal{U}] = -\sum_{u} p_{\mathcal{U}}(u) \ln p_{\mathcal{U}}(u), \tag{C.4}$$

$$\mathbb{H}[\mathcal{U}\,|\,\mathcal{V}] = -\sum_{u,v} p_{\mathcal{U}|\mathcal{V}}(u,v) \ln p_{\mathcal{U}|\mathcal{V}}(u\,|\,v) \tag{C.5}$$

$$= -\sum_{v} p_{\mathcal{V}}(v) \sum_{u} p_{\mathcal{U}|\mathcal{V}}(u\,|\,v) \ln p_{\mathcal{U}|\mathcal{V}}(u\,|\,v), \text{ and} \tag{C.6}$$

$$\mathbb{H}[\mathcal{U},\mathcal{V}] = -\sum_{u,v} p_{\mathcal{U},\mathcal{V}}(u,v)\ln p_{\mathcal{U},\mathcal{V}}(u,v), \tag{C.7}$$

where (C.6) follows from substituting $p_{\mathcal{U},\mathcal{V}}(u,v) = p_{\mathcal{U}\,|\,\mathcal{V}}(u\,|\,v)p_{\mathcal{V}}(v)$. For continuous random variables the expressions are analogous, with sums replaced by the corresponding integrals.

The entropy is solely determined by the distributions of the random variables. To emphasize this, we write $\mathbb{H}[p_{\mathcal{U}}]$, $\mathbb{H}\big[p_{\mathcal{U}\,|\,\mathcal{V}}\big]$, and $\mathbb{H}[p_{\mathcal{U},\mathcal{V}}]$ instead of $\mathbb{H}[\mathcal{U}]$, $\mathbb{H}[\mathcal{U}\,|\,\mathcal{V}]$, and $\mathbb{H}[\mathcal{U},\mathcal{V}]$, respectively.

Proposition C.2 (Fundamental Properties of the Entropy)

The entropy has the following basic properties:

(i) $\mathbb{H}[\mathcal{U}\,|\,\mathcal{V}] \leqslant \mathbb{H}[\mathcal{U}]$

(ii) $\mathbb{H}[\mathcal{U},\mathcal{V}] \leqslant \mathbb{H}[\mathcal{U}] + \mathbb{H}[\mathcal{V}]$, with equality iff $\mathcal{U}$ and $\mathcal{V}$ are independent

(iii) $\mathbb{H}[\mathcal{U},\mathcal{V}] = \mathbb{H}[\mathcal{V}] + \mathbb{H}[\mathcal{U}\,|\,\mathcal{V}]$ (chain rule)

If $\mathcal{U}$ is a discrete random variable that takes values in a finite set of q elements, the entropy has the following additional properties:

(iv) $0 \leqslant \mathbb{H}[\mathcal{U}]$, with equality iff $\mathcal{U}$ is almost surely constant

(v) $\mathbb{H}[\mathcal{U}] \leqslant \ln q$, with equality iff $\mathcal{U}$ is uniformly distributed

Proof Proofs can be found in most textbooks on information theory, for example in that of Cover and Thomas [73]. □

Interpretation

Let u' be a realization of a discrete random variable $\mathcal{U}$, and suppose the value u' is unknown. The entropy $\mathbb{H}[\mathcal{U}]$ is the amount of information[1] that is required in average to identify (that is, to gain certainty about the value) u' if only its distribution (in the form of its probability mass function (PMF) or probability density function (PDF) $p_{\mathcal{U}}$) is known.

[1] Since we use the natural logarithm in the definition of the entropy, the information is measures in *nats*. If we used the logarithm to the basis 2, the information would be measured in bits. These relation between these units is 1 nat = 1/ ln 2 bit.

Therefore, *the entropy quantifies the amount of uncertainty of a random variable.* In our sign-convention, larger entropy means larger uncertainty. Accordingly, $\mathbb{H}[\mathcal{U} \mid \mathcal{V}]$ is the uncertainty of $\mathcal{U}$ if $\mathcal{V}$ is known, and $\mathbb{H}[\mathcal{U}, \mathcal{V}]$ is the combined uncertainty of $\mathcal{U}$ and $\mathcal{V}$.

Property (i) tells us that in average, additional information reduces uncertainty. Property (ii) means that the common uncertainty of two random variables is smaller than the sum of their individual uncertainties, unless they are dependent and thus carry information about each other.[2]

The Principle of Maximum Entropy

The principle of maximum entropy is a method for choosing a probability distribution under constraints due to prior knowledge:

> [...] in making inferences on the basis of partial information we must use that probability distribution which has maximum entropy subject to whatever is known. This is the only unbiased assignment we can make; to use any other would amount to arbitrary assumption of information which by hypothesis we do not have. (Jaynes [132, Sec. 2])

The principle was originally proposed by Jaynes [131, 132, 133] as solution to the problem of arbitrariness when choosing prior distributions in Bayesian inference and has since then found wide acceptance for this purpose.

The "partial information" used in the principle must be testable, that is, it must be possible to determine whether it is consistent with a given distribution or not. Typical testable information is the specification of moments of a distribution, e.g. its expectation or its covariance. A distribution that is chosen according to this principle making no claims except those specified by the testable information. In this sense, it is "maximal ignorant" or has "minimal prejudice".

Kullback-Leibler Distance

The Kullback-Leibler distance (KLD) measures the dissimilarity of two probability distributions. Alternatively, it can be viewed as the amount of additional information required to identify an unknown realization of a random variable (in the sense of the entropy) if one assumes a wrong distribution for this random

[2] This notion is formalized by the mutual information defined below.

variable. The KLD was introduced by Kullback and Leibler [157], its rich set of properties and interpretations were thoroughly discussed by Kullback [158] or its recent reprint Kullback [159]. The KLD plays a key role in statistical inference, summarized, for example, by Eguchi and Copas [87] and by Cover and Thomas [73, Chap. 11].

Definition C.3 (Kullback-Leibler Distance (KLD))

Let $\mathcal{U}, \mathcal{U}', \mathcal{V}, \mathcal{V}'$ and $\mathcal{W}$ be jointly distributed random variables, either all discrete or all continuous. Assume that $\mathcal{U}$ and $\mathcal{V}$ have the same dimension and $\mathcal{U}'$ and $\mathcal{V}'$ have the same dimension.

The KULLBACK-LEIBLER DISTANCE (KLD) from $\mathcal{U}$ to $\mathcal{V}$ is the real-valued quantity

$$\mathbb{D}[\mathcal{U}\|\mathcal{V}] := \mathbb{E}\left[\ln \frac{p_{\mathcal{U}}(\mathcal{U})}{p_{\mathcal{V}}(\mathcal{U})}\right], \tag{C.8}$$

the (conditional) KLD from $\mathcal{U}$ to $\mathcal{V}$ for given $\mathcal{W}$ is

$$\mathbb{D}[\mathcal{U}\,|\,\mathcal{W}\|\mathcal{V}\,|\,\mathcal{W}] := \mathbb{E}\left[\ln \frac{p_{\mathcal{U}|\mathcal{W}}(\mathcal{U}\,|\,\mathcal{W})}{p_{\mathcal{V}|\mathcal{W}}(\mathcal{U}\,|\,\mathcal{W})}\right], \tag{C.9}$$

and the (joint) KLD from $\mathcal{U}$ and $\mathcal{U}'$ to $\mathcal{V}$ and $\mathcal{V}'$ is

$$\mathbb{D}[\mathcal{U},\mathcal{U}'\|\mathcal{V},\mathcal{V}'] := \mathbb{E}\left[\ln \frac{p_{\mathcal{U},\mathcal{U}'}(\mathcal{U},\mathcal{U}')}{p_{\mathcal{V},\mathcal{V}'}(\mathcal{U},\mathcal{U}')}\right], \tag{C.10}$$

supposed the expectations exists.

Kullback and Leibler [157] introduced the quantity $\mathbb{D}[\cdot\|\cdot]$ under the name "discrimination information". According to Kullback [160], not less than nine names are used to refer to it – examples being relative entropy, information gain or Kullback-Leibler divergence – while he still prefers his original term. We nevertheless call it Kullback-Leibler distance, since it seems to be one of the most frequently used terms.

For discrete random variables,

$$\mathbb{D}[\mathcal{U}\|\mathcal{V}] = \sum_{u} p_{\mathcal{U}}(u) \ln \frac{p_{\mathcal{U}}(u)}{p_{\mathcal{V}}(u)} \tag{C.11}$$

$$\mathbb{D}[\mathcal{U}\,|\,\mathcal{W}\|\mathcal{V}\,|\,\mathcal{W}] = \sum_{u,v} p_{\mathcal{U}}(u,v)\ln\frac{p_{\mathcal{U}}(u\,|\,v)}{p_{\mathcal{V}}(u\,|\,v)}, \text{ and} \tag{C.12}$$

$$\mathbb{D}[\mathcal{U},\mathcal{U}'\|\mathcal{V},\mathcal{V}'] = \sum_{u,u'} p_{\mathcal{U},\mathcal{U}'}(u,u')\ln\frac{p_{\mathcal{U},\mathcal{U}'}(u,u')}{p_{\mathcal{V},\mathcal{V}'}(u,u')}. \tag{C.13}$$

The analogous expressions for continuous random variables are obtained by replacing sums by the corresponding integrals.

Like the entropy, the KLD depends only on the distributions of the considered random variables. To emphasize this property, we sometimes write $\mathbb{D}[p_{\mathcal{U}}\|p_{\mathcal{V}}]$ $\mathbb{D}[p_{\mathcal{U},\mathcal{U}'}\|p_{\mathcal{V},\mathcal{V}'}]$, and $\mathbb{D}\big[p_{\mathcal{U}\,|\,\mathcal{W}}\,\big\|\,p_{\mathcal{V}\,|\,\mathcal{W}}\big]$, instead of $\mathbb{D}[\mathcal{U}\|\mathcal{V}]$, $\mathbb{D}[\mathcal{U},\mathcal{U}'\|\mathcal{V},\mathcal{V}']$, and $\mathbb{D}[\mathcal{U}\,|\,\mathcal{W}\|\mathcal{V}\,|\,\mathcal{W}]$, respectively.

Proposition C.4 (Fundamental Properties of the KLD)

The KLD has the following basic properties:

(i) $\mathbb{D}[\mathcal{U}\|\mathcal{V}] \geqslant 0$

(ii) $\mathbb{D}[\mathcal{U}\|\mathcal{V}] = 0$ iff $\mathcal{U}$ and $\mathcal{V}$ have the same distribution.

(iii) If $\mathcal{U}$ and $\mathcal{U}'$ have the same distribution and $\mathcal{V}$ and $\mathcal{V}'$ have the same distribution, then $\mathbb{D}[\mathcal{U}\|\mathcal{V}] = \mathbb{D}[\mathcal{U}'\|\mathcal{V}']$

(iv) If $\mathcal{U}$ and $\mathcal{V}$ have different distributions, then $\mathbb{D}[\mathcal{U}\|\mathcal{V}] \neq \mathbb{D}[\mathcal{V}\|\mathcal{U}]$

(v) If $\mathcal{U}$ and $\mathcal{U}'$ are independent, and $\mathcal{V}$ and $\mathcal{V}'$ are independent, then $\mathbb{D}[\mathcal{U},\mathcal{U}'\|\mathcal{V},\mathcal{V}'] = \mathbb{D}[\mathcal{U}\|\mathcal{U}'] + \mathbb{D}[\mathcal{V}\|\mathcal{V}']$

(vi) $\mathbb{D}[\mathcal{U},\mathcal{W}\|\mathcal{V},\mathcal{W}] = \mathbb{D}[\mathcal{U}\|\mathcal{V}] + \mathbb{D}[\mathcal{U}\,|\,\mathcal{W}\|\mathcal{V}\,|\,\mathcal{W}]$ (chain rule)

(vii) $\mathbb{D}[\lambda p_1 + (1-\lambda)\tilde{p}_1\|\lambda p_2 + (1-\lambda)\tilde{p}_2] \leqslant \lambda\,\mathbb{D}[p_1\|p_2] + (1-\lambda)\,\mathbb{D}[\tilde{p}_1\|\tilde{p}_2]$, for all $\lambda \in [0,1]$ (convexity)

Proof For a proof of the convexity property (vii), see Kullback [159, Cor. 3.1 and subsequent examples in Sec. 2.3]. Proofs of the other properties can be found in the book of Cover and Thomas [73]. □

Property (v) immediately generalizes to arbitrary numbers of independent variables, and property (vii) generalizes to any convex function on the domain of probability mass functions (PMFs)/probability density functions (PDFs).

Interpretation

The KLD quantifies the dissimilarity of two probability distributions: the larger its value, the more dissimilar the distributions. We present three among the many arguments for this claim.

First, the KLD is a premetric or pseudosemimetric due to properties properties (i) and (ii). As such, it gives rise to a topology on the domain of PMFs/PDFs, that it, it defines a notion of "closeness" between distributions. Note, however, the KLD is *not* a metric, as it is neither symmetric nor fulfills the triangle inequality. For details, see the book of Buldygin and Kozachenko [54].

Second, the KLD quantifies the *increase in uncertainty* if $p_{\mathcal{V}}$ is used to approximate $p_{\mathcal{U}}$. Let u' be an unknown realization of the random variable $\mathcal{U}$ with PDF $p_{\mathcal{U}}$. If $p_{\mathcal{U}}$ is known, the information required to identify u' is in average $\mathbb{H}[p_{\mathcal{U}}]$. Now suppose u' is considered as a realization of $\mathcal{V}$ with PDF $p_{\mathcal{V}}$, for example because $p_{\mathcal{U}}$ is unknown and $p_{\mathcal{V}}$ is used as approximation, or simply by mistake. Then, the information required in average to identify u' is

$$\mathbb{H}[p_{\mathcal{U}}] + \mathbb{D}[p_{\mathcal{U}} \| p_{\mathcal{V}}] = \mathbb{H}[\mathcal{U}] + \mathbb{D}[\mathcal{U} \| \mathcal{V}], \tag{C.14}$$

see Cover and Thomas [73, Thm. 5.4.3]. That is, the KLD $\mathbb{D}[\mathcal{U} \| \mathcal{V}]$ is the *additional* information[3] that is in average required to identify u'. In this sense, the KLD measures the loss of information or the increase in uncertainty if $\mathcal{V}$ is used to approximate $\mathcal{U}$, from which the data actually originates. The larger the increase in uncertainty, the more dissimilar is $p_{\mathcal{V}}$ to $p_{\mathcal{U}}$.

Third, the larger $\mathbb{D}[p_{\mathcal{U}} \| p_{\mathcal{V}}]$, the easier to distinguish the distributions $p_{\mathcal{U}}$ and $p_{\mathcal{V}}$ *empirically*, that is, based on data, with a likelihood ratio test, as described in Sec. 4.2.2.

Mutual Information

The mutual information measures the information that two random variables carry about each other. It is the reduction in uncertainty about one random variable due to knowledge about another random variable. In this sense, the mutual information quantifies the amount of dependency between random variables and can be considered as an extended concept of correlation.

[3] in nats

Definition C.5 (Mutual Information)

Let $\mathcal{U}$ and $\mathcal{V}$ be jointly distributed random variables, either both discrete or both continuous. The MUTUAL INFORMATION of $\mathcal{U}$ and $\mathcal{V}$ is

$$\mathbb{I}[\mathcal{U}\|\mathcal{V}] := \mathbb{D}[p_{\mathcal{U},\mathcal{V}}\|p_{\mathcal{U}}p_{\mathcal{V}}], \tag{C.15}$$

supposed that the Kullback-Leibler distance (KLD) exists.

The mutual information is the KLD between the joint probability density function (PDF) of $\mathcal{U}$ and $\mathcal{V}$ and the product of their respective marginal PDFs. Recall that $p_{\mathcal{U}|\mathcal{V}}(u,v) = p_{\mathcal{U}}(u)p_{\mathcal{V}}(v)$ if and only if $\mathcal{U}$ and $\mathcal{V}$ are independent. Therefore, already the definition of the mutual information suggests that it is a measure for the dependency of random variables. This interpretation is supported by the following properties.

Proposition C.6 (Fundamental Properties of Mutual Information)

(i) $\mathbb{I}[\mathcal{U}\|\mathcal{V}] = \mathbb{I}[\mathcal{V}\|\mathcal{U}]$

(ii) $\mathbb{I}[\mathcal{U}\|\mathcal{V}] \geqslant 0$

(iii) $\mathbb{I}[\mathcal{U}\|\mathcal{V}] = 0 \Leftrightarrow \mathcal{U}$ and $\mathcal{V}$ are independent

Proof Proofs are available in most textbooks, for example in Cover and Thomas [73, Chaps. 2 and 8]. An early proof can be found in the article of Lindley [176, Thm. 1]. □

Interpretation and Relation to Entropy and KLD

The mutual information measures the average amount of information that a random variable carries about another random variable and vice versa. The mutual information is best understood by its relation to entropy and KLD, summarized in the next theorem. Its relationship with entropy is also visualized in the Venn diagram in Fig. C.1 on the next page.

Proposition C.7 (Relations between Entropy, KLD and Mutual Information)

$$\mathbb{I}[\mathcal{U}\|\mathcal{V}] = \mathbb{H}[\mathcal{U}] - \mathbb{H}[\mathcal{U}|\mathcal{V}] = \mathbb{H}[\mathcal{V}] - \mathbb{H}[\mathcal{V}|\mathcal{U}] \tag{C.16}$$

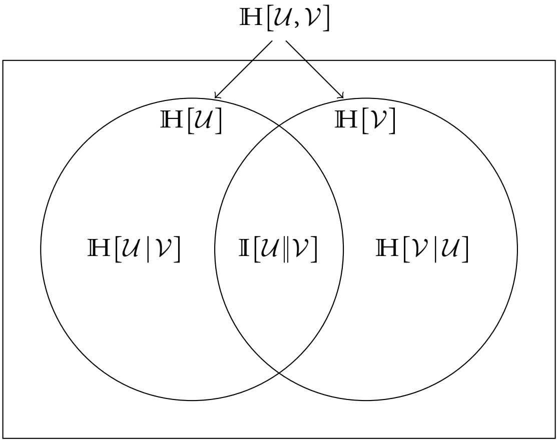
$$\mathbb{I}[\mathcal{U}\|\mathcal{V}] = \mathbb{H}[\mathcal{U}] + \mathbb{H}[\mathcal{V}] - \mathbb{H}[\mathcal{U},\mathcal{V}] \tag{C.17}$$

$$\mathbb{I}[\mathcal{U}\|\mathcal{U}] = \mathbb{H}[\mathcal{U}] \tag{C.18}$$

$$\mathbb{I}[\mathcal{U}\|\mathcal{V}] = \mathbb{D}[\mathcal{U}\,|\,\mathcal{V}\|\mathcal{U}] = \mathbb{D}[\mathcal{V}\,|\,\mathcal{U}\|\mathcal{V}] \tag{C.19}$$

Proof Follows directly from the definitions. See, for example, Cover and Thomas [73, Secs. 2.4 and 8.5]. The first equality in (C.19) follows from $\mathbb{I}[\mathcal{U}\|\mathcal{V}] \overset{(a)}{=} \mathbb{D}[p_{\mathcal{U},\mathcal{V}}\|p_{\mathcal{U}}p_{\mathcal{V}}] \overset{(b)}{=} \mathbb{D}[p_{\mathcal{U}}\|p_{\mathcal{U}}] + \mathbb{D}\big[p_{\mathcal{U}\,|\,\mathcal{V}}\,\|p_{\mathcal{U}}\big] \overset{(c)}{=} \mathbb{D}[\mathcal{U}\,|\,\mathcal{V}\|\mathcal{U}]$, where the equalities are due to (a) (C.15), (b) property (vi) of Prop. C.4, and (c) property (ii) of Prop. C.4. The second equality in (C.19) is a consequence of the symmetry of $\mathbb{I}[\mathcal{U}\|\mathcal{V}]$. □

Figure C.1.: Relationships between entropy and mutual information.


$\mathbb{H}[\mathcal{U},\mathcal{V}]$
$\mathbb{H}[\mathcal{U}]$
$\mathbb{H}[\mathcal{V}]$
$\mathbb{H}[\mathcal{U}\,|\,\mathcal{V}]$
$\mathbb{I}[\mathcal{U}\|\mathcal{V}]$
$\mathbb{H}[\mathcal{V}\,|\,\mathcal{U}]$

Special Cases for Normal Distributions

We consider the introduced information-theoretic quantities for the frequently required special case of continuous random variables with a normal distribution (see Def. B.8).

Proposition C.8 (Entropy of a Normal Distribution)

If $\mathcal{U} \sim \mathcal{N}_r(\mu, \boldsymbol{C})$, then $\mathbb{H}[\mathcal{U}] = \frac{1}{2}\ln\det(\boldsymbol{C}) + \frac{1}{2}(\ln(2\pi)+1)r$.

Proof Proofs can be found in several textbooks, for example in that of Cover and Thomas [73, Thm. 8.4.1]. □

It is important to realize that *the entropy of a normal distribution is independent of its mean.*

Proposition C.9 (Maximum-Entropy Property of the Normal Distribution)

If $\mathcal{U}$ is a r-dimensional continuous random variable, then

$$\mathbb{H}[\mathcal{U}] \leqslant \frac{1}{2}\ln\det(\mathbb{C}[\mathcal{U}]) + \frac{1}{2}(\ln(2\pi)+1)r, \tag{C.20}$$

with equality iff $\mathcal{U} \sim \mathcal{N}_r(\cdot, \mathbb{C}[\mathcal{U}])$.

Proof See, for example, the book of Cover and Thomas [73, Thm. 8.4.1]. □

The right-hand side of the inequality in (C.20) is the entropy of a normal distribution with the same covariance as $\mathcal{U}$. Therefore, the theorem tells us that *the normal distribution has the largest entropy among all distributions with the same covariance.*

Theorem C.10 (KLD between Normal Distributions)

If $\mathcal{U}_1 \sim \mathcal{N}_r(\mu_1, \boldsymbol{C}_1)$ and $\mathcal{U}_2 \sim \mathcal{N}_r(\mu_2, \boldsymbol{C}_2)$, and $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$ have full rank, then

$$\mathbb{D}[\mathcal{U}_1 \| \mathcal{U}_2] = \frac{1}{2}\Big(\|\mu_1 - \mu_2\|^2_{\boldsymbol{C}_2^{-1}} + \operatorname{tr}\big(\boldsymbol{C}_1\boldsymbol{C}_2^{-1}\big) - \ln\det\big(\boldsymbol{C}_1\boldsymbol{C}_2^{-1}\big) - r\Big), \tag{C.21}$$

Proof By definition, $\mathbb{D}[\mathcal{U}_1 \| \mathcal{U}_2] = \mathbb{E}\left[\ln \frac{p_1(\mathcal{U}_1)}{p_2(\mathcal{U}_1)}\right]$, where p_1 and p_2 are the probability density functions (pdfs) of $\mathcal{U}_1$ and $\mathcal{U}_2$, respectively. By substituting the pdfs of a normal distribution (B.12) we get

$$\ln\frac{p_1(\mathcal{U}_1)}{p_2(\mathcal{U}_1)} = \frac{1}{2}\Big(\|\mathcal{U}_1 - \mu_2\|^2_{\boldsymbol{C}_2^{-1}} - \|\mathcal{U}_1 - \mu_1\|^2_{\boldsymbol{C}_1^{-1}} - \ln\det\big(\boldsymbol{C}_1\boldsymbol{C}_2^{-1}\big)\Big). \tag{C.22}$$

We calculate the expectation of (C.22) term by term. The expectation of the first term is

$$\begin{aligned}
\mathbb{E}\left[\|\mathcal{U}_1 - \mu_2\|^2_{\boldsymbol{C}_2^{-1}}\right] &= \|\mathbb{E}\left[\mathcal{U}_1 - \mu_2\right]\|^2_{\boldsymbol{C}_2^{-1}} + \operatorname{tr}\left(\mathbb{C}\left[\mathcal{U}_1 - \mu_2\right]\boldsymbol{C}_2^{-1}\right) && \text{(C.23)}\\
&= \|\mathbb{E}\left[\mathcal{U}_1\right] - \mu_2\|^2_{\boldsymbol{C}_2^{-1}} + \operatorname{tr}\left(\mathbb{C}\left[\mathcal{U}_1\right]\boldsymbol{C}_2^{-1}\right) && \text{(C.24)}\\
&= \|\mu_1 - \mu_2\|^2_{\boldsymbol{C}_2^{-1}} \qquad + \operatorname{tr}\left(\boldsymbol{C}_1\boldsymbol{C}_2^{-1}\right), && \text{(C.25)}
\end{aligned}$$

where we used Thm. B.1 for the first equality. Analogously, the expectation of the second term in (C.22) is

$$\begin{aligned}
\mathbb{E}\left[\|\mathcal{U}_1 - \mu_1\|^2_{\boldsymbol{C}_1^{-1}}\right] &= \operatorname{tr}\left(\mathbb{C}\left[\mathcal{U}_1 - \mu_1\right]\boldsymbol{C}_1^{-1}\right) + \|\mathbb{E}\left[\mathcal{U}_1 - \mu_1\right]\|^2_{\boldsymbol{C}_1^{-1}} && \text{(C.26)}\\
&= \operatorname{tr}\left(\mathbb{C}\left[\mathcal{U}_1\right]\boldsymbol{C}_1^{-1}\right) \qquad + \|\mathbb{E}\left[\mathcal{U}_1\right] - \mu_1\|^2_{\boldsymbol{C}_1^{-1}} && \text{(C.27)}\\
&= \operatorname{tr}\left(\boldsymbol{C}_1\boldsymbol{C}_1^{-1}\right) \qquad\qquad = r. && \text{(C.28)}
\end{aligned}$$

Since the third term in (C.22) is independent of $\mathcal{U}_1$, $\mathbb{E}\left[\ln\det\left(\boldsymbol{C}_1\boldsymbol{C}_2^{-1}\right)\right] = \ln\det\left(\boldsymbol{C}_1\boldsymbol{C}_2^{-1}\right)$. By taking together these results we obtain (C.21). □

For the special case $\boldsymbol{C}_1 = \boldsymbol{C}_2 = \boldsymbol{C}$, the last three terms in (C.21) add up to zero so that $\mathbb{D}[\mathcal{U}_1\|\mathcal{U}_2] = \frac{1}{2}\|\mu_1 - \mu_2\|^2_{\boldsymbol{C}^{-1}}$.

Bibliography

[1] Peter Ahrendt. *The Multivariate Gaussian Probability Distribution*. Tech. rep. IMM, Technical University of Denmark, Jan. 2005.

[2] Hirotugu Akaike. “Information Theory as an Extension of the Maximum Likelihood Principle.” In: *Proceedings of the Second International Symposium on Information Theory*. Ed. by B. N. Petrov and F. Csaki. Budapest, Hungary, 1973, pp. 267–281.

[3] Hirotugu Akaike. “On Entropy Maximization Principle.” In: *Applications of Statistics: Proceedings of the Symposium Held at Wright State University, Dayton, Ohio, 14-18 June 1976*. Ed. by Paruchuri R. Krishnaiah. Elsevier Science & Technology Books, 1977.

[4] Giacomo Aletti, Caterina May, and Chiara Tommasi. “A Convergent Algorithm for Finding KL-optimum Designs and Related Properties.” In: *mODa 10 – Advances in Model-Oriented Design and Analysis*. Ed. by Dariusz Uciński, Maciej Patan, and Anthony C. Atkinson. Contributions to Statistics. Springer, 2013, pp. 1–9. ISBN: 9783319002187. DOI: 10.1007/978-3-319-00218-7.

[5] Giacomo Aletti, Caterina May, and Chiara Tommasi. “KL-optimum designs: theoretical properties and practical computation.” In: *Statistics and Computing* (Sept. 2014). DOI: 10.1007/s11222-014-9515-8. URL: http://dx.doi.org/10.1007/s11222-014-9515-8.

[6] S. M. Ali and S. D. Silvey. “A General Class of Coefficients of Divergence of One Distribution from Another.” In: *Journal of the Royal Statistical Society. Series B (Methodological)* 28.1 (1966), pp. 131–142. URL: http://www.jstor.org/stable/2984279.

[7] Tomohiro Ando. *Bayesian Model Selection and Statistical Modeling*. Statistics: Textbooks and Monographs. Chapman and Hall/CRC, May 2010. ISBN: 9781439836149.

[8] Sanjeev Arora and Boaz Bara. *Computational Complexity: A Modern Approach*. Cambridge University Press, 2009. ISBN: 9780521424264.

[9] S. P. Asprey and S. Macchietto. "Designing robust optimal dynamic experiments." In: *Journal of Process Control* 12 (2002), pp. 545–556.

[10] S. P. Asprey and S. Macchietto. "Statistical tools for optimal dynamic model building." In: *Computers and Chemical Engineering* 24 (2000), pp. 1261–1267.

[11] A. C. Atkinson and A. N. Donev. *Optimum Experimental Designs*. Ed. by Anthony C. Atkinson, J. B. Copas, D. A. Pierce, Mark J. Schervish, and D. M. Titterington. Oxford Statistical Sciences Series 8. Oxford: Claredon Press, 1992. isbn: 978-0198522546.

[12] Anthony C. Atkinson. "A comparison of two criteria for the design of experiments for discriminating between models." In: *Technometrics* 23 (1981), 3ff.

[13] Anthony C. Atkinson. "DT-optimum designs for model discrimination and parameter estimation." In: *Journal of Statistical Planning and Inference* 138 (2008), pp. 56–64.

[14] Anthony C. Atkinson. "Optimum experimental designs for parameter estimation and for discrimination between models in the presence of prior information." In: *Model Oriented Data-Analysis – A Survey of Recent Methods. Proceedings of the 2nd IIASA-workshop in St. Kyrik, Bulgaria, May 28 – June 1, 1990*. Ed. by V. V. Fedorov, W. G. Müller, and I. N. Vuchkov. Contributions to Statistics. Heidelberg, Germany: Physica Verlag, 1990, pp. 3–30. isbn: 3-7908-0624-2.

[15] Anthony C. Atkinson. "Planning experiments for model testing and discrimination." In: *Mathematische Operationsforschung und Statistik* 6 (Jan. 1975), pp. 253–267. doi: 10.1080/02331937508842248.

[16] Anthony C. Atkinson. "Posterior probabilities for choosing a regression model." In: *Biometrika* 65 (1978), pp. 39–48.

[17] Anthony C. Atkinson and R. A. Bailey. "One hundred years of the design of experiments on and off the pages of Biometrika." In: *Biometrika* 88 (2001), pp. 53–97.

[18] Anthony C. Atkinson, Barbara Bogacka, and Mariusz B. Bogacki. "D- and T-optimum designs for the kinetics of a reversible chemical reaction." In: *Chemometrics and Intelligent Laboratory Systems* 43 (1998), pp. 185–198.

[19] Anthony C. Atkinson and David R. Cox. “Planning experiments for discriminating between models.” In: *Journal of the Royal Statistical Society* B36 (1974), pp. 321–348. URL: http://www.jstor.org/stable/298492.

[20] Anthony C. Atkinson and V. Fedorov. “Optimal design: Experiments for discriminating between several models.” In: *Biometrika* 62 (1975), pp. 289–303.

[21] Anthony C. Atkinson and V. Fedorov. “The design of experiments for discriminating between two rival models.” In: *Biometrika* 62 (1975), pp. 57–60.

[22] Adriano Azevedo-Filho and Ross D. Shachter. “Laplace’s method approximations for probabilistic inference in belief networks with continuous variables.” In: *UAI’94 Proceedings of the Tenth international conference on Uncertainty in artificial intelligence.* Ed. by Ramon Lopez De Mantaras and David Poole. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1994, pp. 28–36. ISBN: 1558603328.

[23] Yonathan Bard. *Nonlinear Parameter Estimation.* Academic Press, 1974.

[24] André Bardow, Ernesto Kriesten, Mihai Adrian Voda, Federico Casanova, Bernhard Blümich, and Wolfgang Marquardt. “Prediction of multicomponent mutual diffusion in liquids: Model discrimination using NMR data.” In: *Fluid Phase Equilibria* 278 (2009), pp. 27–35.

[25] M. Bayes. “An Essay towards solving a Problem in the Doctrine of Chances.” In: *Philosophical Transactions of the Royal Society of London* 53 (1763), pp. 370–418.

[26] Robert H. Berk. “Consistency a Posteriori.” In: *Ann. Math. Statist.* 41.3 (1970), pp. 894–906. DOI: 10.1214/aoms/1177696967. URL: http://projecteuclid.org/euclid.aoms/1177696967.

[27] Robert H. Berk. “Limiting Behavior of Posterior Distributions when the Model is Incorrect.” In: *Ann. Math. Statist.* 37.1 (1966), pp. 51–58. DOI: 10.1214/aoms/1177699597. URL: http://projecteuclid.org/euclid.aoms/1177699597.

[28] José M. Bernardo and Adrian F. M. Smith. *Bayesian Theory.* Wiley Series in Probability & Mathematical Statistics. John Wiley & Sons, 1994. ISBN: 0471924164.

[29] Jakob Bernoulli. *Ars Conjectandi: Usum & Applicationem Praecedentis Doctrinae in Civilibus, Moralibus & Oeconomicis*. Translated into English by Oscar Sheynin. Basel: Thurneysen Brothers, 1713. Chap. 4.

[30] Herman J. Bierens. *Introduction to the Mathematical and Statistical Foundations of Econometrics*. Themes in Modern Econometrics. Cambridge University Press, 2004. ISBN: 9780511079658 (eBook), 9780521834315 (hardback), 9780521542241 (paperpback). URL: www.cambridge.org/9780521834315.

[31] Patrick Bilingsley. *Convergence of Probability Measures*. 2nd ed. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., 1999. ISBN: 0471197459.

[32] David Blackwell. "Comparison of Experiments." In: *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability. July 31-August 12, 1950. Statistical Laboratory of the University of California, Berkeley.* Ed. by Jerzy Neyman. Berkeley, California, USA: University of California Press, 1950, pp. 93–102.

[33] David Blackwell. "Equivalent Comparisons of Experiments." In: *he Annals of Mathematical Statistics* 24.2 (June 1953), pp. 265–272. URL: http://www.jstor.org/stable/2236332.

[34] William J. Blot and Duane A. Meeter. "Sequential Experimental Design Procedures." In: *Journal of the American Statistical Association* 68.343 (Sept. 1973), pp. 586–593. URL: http://www.jstor.org/stable/2284782.

[35] H. G. Bock. *Randwertproblemmethoden zur Parameteridentifizierung in Systemen nichtlinearer Differentialgleichungen*. Vol. 183. Bonner Mathematische Schriften. Bonn: Universität Bonn, 1987. URL: http://www.iwr.uni-heidelberg.de/groups/agbock/FILES/Bock1987.pdf.

[36] H. G. Bock, E. Kostina, and J. P. Schlöder. "Numerical Methods for Parameter Estimation in Nonlinear Differential Algebraic Equations." In: *GAMM Mitteilungen* 30/2 (2007), pp. 352–375.

[37] P. T. Boggs and J. W. Tolle. "Sequential Quadratic Programming." In: *Acta Numerica* 4 (1995), pp. 1–51.

[38] Roger Bowden. "The Theory of Parametric Identification." In: *Econometrica* 41.6 (Nov. 1973), pp. 1069–1074.

[39] G. E. P. Box and T. L. Henson. *MODEL FITTING AND DISCRIMINATION*. Tech. rep. 211. Madison, Wisconsin, USA: University of Wisconsin, July 1969. URL: http://www.stat.wisc.edu/node/635.

[40] G. E. P. Box and T. L. Henson. "Some aspects of Mathematical Modeling in Chemical Engineering." In: *Proceedings of the Inaugural Conference of the Scientific Computation Centre and the Institute of Statistical Studies and Research*. Cairo, Egypt: Cairo University Press, 1970, pp. 548–570.

[41] G. E. P. Box and William G. Hunter. "The Experimental Study of Physical Mechanisms." In: *Technometrics* 7.1 (1965), pp. 23–42.

[42] G. Box and W. Hill. "Discrimination among mechanistic models." In: *Technometrics* 9.1 (1967), pp. 57–71. URL: http://www.jstor.org/stable/1266318.

[43] George E. P. Box. "Science and Statistics." In: *Journal of the American Statistical Association* 71.356 (Dec. 1976), pp. 791–799.

[44] George E. P. Box and Norman R. Draper. *Empirical Model-Building and Response Surfaces*. Wiley Series in Probability and Statistics. New York, NY, USA: John Wiley & Sons, Inc., Jan. 1987. ISBN: 978-0471810339.

[45] George E. P. Box and George C. Tiao. *Bayesian inference in statistical analysis*. Addison-Wesley series in behavioral science: quantitative methods. Reading, Massachusetts: Addison-Wesley, 1973. ISBN: 978-0-201-00622-3.

[46] M. J. Box. *A note on the design of experiments for model discrimination*. Research Note 68|22. I.C.I. Central Instruments Research Laboratory, 1968.

[47] Hamparsum Bozdogan. "Akaike's Information Criterion and Recent Developments in Information Complexity." In: *Journal of Mathematical Psychology* 44 (2000), pp. 62–91. DOI: 10.1006/jmps.1999.1277.

[48] Hamparsum Bozdogan. "Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions." In: *Psychometrika* 52.3 (Sept. 1987), pp. 345–370.

[49] Eric Braaten and George Weller. "An Improved Low-Discrepancy Sequence for Multidimensional Quasi-Monte Carlo Integration." In: *Journal of Computational Physics* 33 (1979), pp. 249–258.

[50] Russell N. Bradt and Samuel Karlin. “On the Design and Comparison of Certain Dichotomous Experiments.” In: *The Annals of Mathematical Statistics* 27.2 (June 1956), pp. 390–409. URL: http://www.jstor.org/stable/2237000.

[51] Paul Bratley and Bennett L. Fox. “Algorithm 659: Implementing Sobol’s quasirandom sequence generator.” In: *Journal ACM Transactions on Mathematical Software (TOMS)* 14.1 (Mar. 1988), pp. 88–100.

[52] M. Brookes. *The Matrix Reference Manual.* 2011. URL: http://www.ee.imperial.ac.uk/hp/staff/dmb/matrix/intro.html.

[53] C. G. Broyden. “The convergence of a class of double–rank minimization algorithms.” In: *Journal of the Institute of Mathematics and its Applications* 6 (1970), pp. 76–90. DOI: 10.1093/imamat/6.1.76. URL: http://imamat.oxfordjournals.org/cgi/content/abstract/6/1/76.

[54] Valerii Vladimirovich Buldygin and Yu.V. Kozachenko. *Metric Characterization of Random Variables and Random Processes.* Vol. 188. Translations of Mathematical Monographs. American Mathematical Society, 2000. ISBN: 978-0821805336.

[55] Olaf Bunke and Xavier Milhaus. “Asymptotic Behavior of Bayes Estimates under Possibly Incorrect Models.” In: *The Annals of Statistics* 26.2 (1998), pp. 617–644.

[56] Samuel Burer and Dieter Vandenbussche. “A finite branch-and-bound algorithm for nonconvex quadratic programming via semidefinite relaxations.” In: *Mathematical Programming* 113.2 (June 2008), pp. 259–282.

[57] Samuel Burer and Dieter Vandenbussche. “Globally solving box-constrained nonconvex quadratic programs with semidefinite-based finite branch-and-bound.” In: *Computational Optimization and Applications* 43.2 (June 2009), pp. 181–195.

[58] Jose F. Burguete, A. Ronald Gallant, and Geraldo Souza. “On unification of the asymptotic theory of nonlinear econometric models.” In: *Econometric Reviews* 1.2 (1982), pp. 151–190. DOI: 10.1080/07311768208800012.

[59] Annette Burke. “Discriminating between the Terminal and Penultimate Models Using Designed Experiments: An Overview.” In: *Industrial and Engineering Chemistry Research* 36.4 (1997), pp. 1016–1035.

[60] R. W. Butler, P. L. Davies, and M. Jhun. "Asymptotics for the Minimum Covariance Determinant Estimator." In: *The Annals of Statistics* 21.3 (Sept. 1993), pp. 1385–1400. URL: http://www.jstor.org/stable/2242201.

[61] G. Buzzi-Ferraris, P. Forzatti, G. Emig, and H. Hofmann. "Sequential experimental design for model discrimination in the case of multiple responses." In: *Chemical Engineering Science* 39.1 (1984), pp. 81–85.

[62] Guido Buzzi-Ferraris. "Some Observations on the Paper 'Optimal Experimental Design for Discriminating Numerous Model Candidates: The AWDC Criterion'." In: *Industrial and Engineering Chemistry Research* 49.19 (2010), pp. 9561–9562. DOI: 10.1021/ie100373t.

[63] Guido Buzzi-Ferraris and Pio Forzatti. "A new sequential experimental design procedure for discriminating among rival models." In: *Chemical Engineering Science* 38.2 (1983), pp. 225–232.

[64] Guido Buzzi-Ferraris, Pio Forzatti, and Paolo Canu. "An improved version of a sequential design criterion discriminating among rival multiresponse models." In: *Chemical Engineering Science* 45.2 (1990), pp. 477–481.

[65] LL. D. Campbell and William Garnett. *The Life of James Clerk Maxwell with a selection from his correspondence and occasional writings and a sketch of his contibutions to science.* Digitally preserved version by James C. Rautio, 2nd edition, 1999. London: Macmillan and Co., 1882.

[66] Kathryn Chaloner and Isabella Verdinelli. "Bayesian Experimental Design: A Review." In: *Statistical Science* 10.3 (Aug. 1995), pp. 273–304. URL: http://www.jstor.org/stable/2246015.

[67] Bing H. Chen and Steven P. Asprey. "On the Design of Optimally Informative Dynamic Experiments for Model Discrimination in Multiresponse Nonlinear Situations." In: *Industrial and Engineering Chemistry Research* 42 (2003), pp. 1379–1390.

[68] Hongmei Chi and Edward L. Jones. "Generating parallel quasirandom sequences via randomization." In: *Journal of Parallel and Distributed Computation* 67 (2007), pp. 876–881. DOI: 10.1016/j.jpdc.2007.04.004.

[69] Christine Choirat and Raffaello Seri. "Estimation in Discrete Parameter Models." In: *Statistical Science* 27.2 (2012), pp. 278–293. DOI: 10.1214/11-STS371.

[70] Gregory C. Chow. "Maximum-likelihood estimation of misspecified models." In: *Economic Modelling* 1.2 (Apr. 1984), pp. 134–138. doi: 10.1016/0264-9993(84)90001-4.

[71] Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. "Power-Law Distributions in Empirical Data." In: *SIAM Review* 51.4 (2009), pp. 661–703. doi: 10.1137/070710111.

[72] M. J. Cooney and K. A. McDonald. "Optimal dynamic experiments for bioreactor model discrimination." In: *Applied Microbiology and Biotechnology* 43 (1995), pp. 826–837.

[73] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. 2nd ed. Wiley Series in Telecommunications and Signal Processing. Hoboken, New Jersey, USA: John Wiley & Sons, Inc., Sept. 2006. isbn: 978-0-471-24195-9.

[74] David R. Cox and Nancy Reid. *The theory of the design of experiments*. Monographs on statistics and applied probability. Boca Raton, Fla.: Chapman & Hall/CRC, 2000. isbn: 978-1-58488-195-7.

[75] Harald Cramér. *Mathematical Methods of Statistics*. 2nd ed. Princeton, New Jersey, USA: Princeton University Press, 1946. isbn: 978-0691005478.

[76] I. Csiszár. "Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten." In: *Publications of the Mathematical Institute of the Hungarian Academy of Sciences* 8 (1963), pp. 95–108.

[77] A. P. Dawid. "Probability Forecasting." In: *Plackett Family of Distribution to Regression, Wrong*. Ed. by Samuel Kotz, Norman Lloyd Johnson, and Campbell B. Read. Vol. 7. Encyclopedia of Statistical Sciences. Wiley-Interscience, 1986, pp. 210–218. isbn: 9780471055556.

[78] M. H. DeGroot. "Uncertainty, Information and Sequential Experiments." In: *Annals of Mathematical Statistics* 33.2 (June 1962), pp. 404–419. url: http://www.jstor.org/stable/2237520.

[79] Holger Dette and Stefanie Titoff. "Optimal Discrimination Designs." In: *The Annals of Statistics* 37.4 (2009), pp. 2056–2082. doi: 10.1214/08-AOS635.

[80] A. Dieses. “Numerische Verfahren zur Diskriminierung nichtlinearer Modelle für dynamische chemische Prozesse.” Diploma thesis. Universität Heidelberg, 1997.

[81] Brecht M. R. Donckels, Dirk J. W. De Pauw, Bernard De Baets, Jo Maertens, and Peter A. Vanrolleghem. “An anticipatory approach to optimal experimental design for model discrimination.” In: *Chemometrics and Intelligent Laboratory Systems* 95 (Aug. 2009), pp. 53–63.

[82] Brecht M.R. Donckels, Dirk J.W. De Pauw, Peter A. Vanrolleghem, and Bernard De Baets. “Performance assessment of the anticipatory approach to optimal experimental design for model discrimination.” In: *Chemometrics and Intelligent Laboratory Systems* 110 (2012), pp. 20–31.

[83] J. L. Doob. “Applications of the theory of martingales.” In: *Colloques Internationaux du Centre National de la Recherche Scientifique* 13 (1949), pp. 22–28.

[84] J. L. Doob. “Probability and Statistics.” In: *Transactions of the American Mathematical Society* 36 (1934), pp. 759–775. DOI: 10.1090/S0002-9947-1934-1501765-1.

[85] David Draper. “Assessment and Propagation od Model Uncertainty.” In: *Journal of the Royal Statistical Society, Series B (Methodological)* 57.1 (1995), pp. 45–97. URL: http://www.jstor.org/stable/2346087.

[86] Anthony W. F. Edwards. *Likelihood*. The Johns Hopkins University Press, Oct. 1992. ISBN: 978-0801844430.

[87] Shinto Eguchi and John Copas. “Interpreting Kullback-Leibler divergence with the Neyman-Pearson lemma.” In: *Journal of Multivariate Analysis* 97 (2006), pp. 2034–2040.

[88] D. Espie and S. Macchietto. “The Optimal Design of Dynamic Experiments.” In: *AIChE Journal* 35.2 (Feb. 1989), pp. 223–229.

[89] Nasrollah Etemadi. “An elementary proof of the strong law of large numbers.” In: *Probability Theory and Related Fields* 55 (1981), pp. 119–122.

[90] Henri Faure and Christiane Lemieux. “Generalized Halton Sequences in 2008: A Comparative Study.” In: *ACM Transactions on Modeling and Computer Simulation* 19.4 (Oct. 2009), 15:1–15:31. DOI: 10.1145/1596519.1596520.

[91] Valerii V. Fedorov. "Asymptotically optimal designs of experiments for discriminating two rival regression models." In: *Theory Prob. Applic.* 16 (1971), pp. 561–562.

[92] Valerii V. Fedorov. "[Follow-Up Designs to Resolve Confounding in Multifactor Experiments]: Discussion." In: *Technometrics* 38.4 (Nov. 1996), pp. 321–322.

[93] Valerii V. Fedorov. "Optimal experimental designs for discriminating two rival regression models." In: *A survey of statistical design and linear models*. Ed. by Jagdish N. Srivastava. North-Holland Publishing / American Elsevier, 1975, pp. 155–164.

[94] Valerii V. Fedorov. "Some Extremal Problems in Designing Discriminating Experiments." In: *Communications in Statistics A (Theory and Methods)* 7.14 (1978), pp. 1339–1345.

[95] Valerii V. Fedorov. *Theory of optimal experiments*. Probability and Mathematical Statistics. Originally published in Russian under the title "TEORIYA OPTIMAL'NOGO EKSPERIMENTA" by Izdatel'stvo Moskovskogo Universieteta, 1969. New York and London: Academic Press, 1972.

[96] Valerii V. Fedorov and Peter Hackl. *Model-Oriented Design of Experiments*. Vol. 125. Lecture Notes in Statistics. Springer, 1997. ISBN: 978-0-387-98215-1.

[97] Valerii V. Fedorov and V. Khabarov. "Duality of optimal designs for model discrimination and parameter estimation." In: *Biometrika* 73.1 (1986), pp. 183–190.

[98] Valerii V. Fedorov and M. Malyutov. "Optimal designs in regression problems." In: *Math. Operationsforschung and Statistik* 3 (1972), pp. 281–308.

[99] Valerii Fedorov. "Optimal experimental design." In: *Wiley Interdisciplinary Reviews: Computational Statistics* 2.5 (2010), pp. 581–589. ISSN: 1939-0068. DOI: 10.1002/wics.100. URL: http://onlinelibrary.wiley.com/doi/10.1002/wics.100/pdf.

[100] R. Fletcher. "A new approach to variable metric algorithms." In: *Computer Journal* 13 (1970), pp. 317–322. DOI: 10.1093/comjnl/13.3.317. URL: http://comjnl.oxfordjournals.org/cgi/content/abstract/13/3/317.

[101] Thomas B. Fomby and R. Carter Hill, eds. *Maximum Likelihood Estimation of Misspecified Models: Twenty Years Later.* Advances in Econometrics 17. Emerald Group Publishing Limited, Dec. 2003. ISBN: 9780762310753.

[102] Wolfgang Förstner and Boudewijn Moonen. "A Metric for Covariance Matrices." In: *Quo vadis geodesia…? Festschrift for Erik W. Grafarend on the occasion of his 60th birthday.* Ed. by Friedhelm Krumm and Volker S. Schwarze. Technical Reports of the Department of Geodesy and Geoinformatics 1999.6. ISSN 0933-2839. Stuttgart University, Oct. 1999. Chap. 12, pp. 113–128.

[103] Gaia Franceschini and Sandro Macchietto. "Model-based design of experiments for parameter precision: State of the art." In: *Chemical Engineering Science* 63 (2008), pp. 4846–4872. DOI: 10.1016/j.ces.2007.11.034.

[104] A. Ronald Gallant and Alberto Holly. "Statistical Inference in an Implicit, Nonlinear, Simultaneous Equation Model in the Context of Maximum Likelihood Estimation." In: *Econometrica* 48.3 (Apr. 1980), pp. 697–720.

[105] M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness.* New York: W. H. Freeman, 1979.

[106] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian data analysis.* 2nd ed. Texts in statistical science. Chapman & Hall/CRC, 2004. ISBN: 9781584883883.

[107] D. Goldfarb. "A family of variable metric updates derived by variational means." In: *Mathematics of Computation* 24 (1970), pp. 23–26. URL: http://www.jstor.org/pss/2004873.

[108] Nick Gould, Dominique Orban, and Philippe Toint. "Numerical methods for large-scale nonlinear optimization." In: *Acta Numerica* 14 (May 2005), pp. 299–361. URL: http://journals.cambridge.org/abstract_S0962492904000248.

[109] Robert M. Gray. *Entropy and Information Theory.* Fourth revised online edition, 2009. Springer, July 1991. URL: http://ee.stanford.edu/~gray/it.html.

[110] Pushpa L. Gupta and R. D. Gupta. "Sample size determination in estimating a covariance matrix." In: *Computational Statistics & Data Analysis* 5.3 (Aug. 1987), pp. 185–192. DOI: 10.1016/0167-9473(87)90014-4.

[111] Allan Gut. *Probability: A Graduate Course.* 2nd ed. Springer Texts in Statistics. Springer, 2013. ISBN: 0387228330.

[112] J. H. Halton. "Algorithm 247: Radical-inverse quasi-random point sequence." In: *Communications of the ACM* 7.12 (Dec. 1964), pp. 701–702.

[113] J. H. Halton. "On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals." In: *Numerische Mathematik* 2.1 (Dec. 1960), pp. 84–90. DOI: 10.1007/BF01386213. URL: http://dx.doi.org/10.1007/BF01386213.

[114] Kathrin Hatz. "Efficient Numerical Methods for Hierarchical Dynamic Optimization with Application to Cerebral Palsy Gait Modeling." PhD thesis. Heidelberg University, 2014.

[115] John R. Hershey and Peder A. Olsen. "Approximating the Kullback Leibler Divergence Between Gaussian Mixture Models." In: *Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2007. ICASSP 2007.* Apr. 2007, pp. IV–317–IV–320. DOI: 10.1109/ICASSP.2007.366913.

[116] P. D. H. Hill. "A Review of Experimental Design Procedures for Regression Model Discrimination." In: *Technometrics* 20 (Feb. 1978), pp. 15–21. URL: http://www.jstor.org/stable/1268155.

[117] William J. Hill and William G. Hunter. "A Note on Designs for Model Discrimination: Variance Unknown Case." In: *Technometrics* 11.2 (May 1969), pp. 396–400. URL: http://www.jstor.org/stable/1267271.

[118] William J. Hill and William G. Hunter. *Design of Experiments for Model Discrimination in Multiresponse Situations.* Tech. rep. 65. Madison, Wisconsin: University of Wisconsin, Feb. 1966.

[119] Jennifer A. Hoeting, David Madigan, Adrian E. Raftery, and Chris T. Volinsky. "Bayesian Model Averaging: A Tutorial." In: *Statistical Science* 14.4 (Dec. 1999). With comments by M. Clyde, David Draper and E. I. George, and a rejoinder by the authors, pp. 382–417. DOI: doi:10.1214/ss/1009212519. URL: http://projecteuclid.org/euclid.ss/1009212519.

[120] Christian Hoffmann. "Numerical Methods for the Discrimination of DAE Models with Applications in Enzyme Kinetics." Diplomarbeit. Heidelberg University, Dec. 2005.

[121] Roger A. Horn and Charles R. Johnson. *Matrix Analysis.* Cambridge University Press, 2007. ISBN: 9780532305860.

[122] R. Horst, Panos M. Pardalos, and Nguyen Van Thoai. *Introduction to Global Optimization*. Vol. 48. Nonconvex Optimization and Its Applications. Springer, 2000. ISBN: 9780792365747.

[123] Reiner Horst and Panos M. Pardalos, eds. *Handbook of Global Optimization*. Vol. 2. Nonconcex Optimization and Its Applications. Kluwer Academic Publishers, 1995. ISBN: 0792331206.

[124] Tien-Chung Hu, Andrew Rosalsky, and Andrei Volodin. "On convergence properties of sums of dependent random variables under second moment and covariance restrictions." In: *Statistics & Probability Letters* 78.14 (Oct. 2008), pp. 1999–2005. DOI: 10.1016/j.spl.2008.01.073.

[125] Marco F. Huber, Tim Bailey, Hugh Durrant-Whyte, and Uwe D. Hanebeck. "On Entropy Approximation for Gaussian Mixture Random Vectors." In: *Proceedings of the IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems, 2008. MFI 2008.* 2008, pp. 181–188.

[126] Peter J. Huber. "The behavior of maximum likelihood estimates under nonstandard conditions." In: *Proceeding of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Vol. 1. University of California Press, 1967, pp. 221–233. URL: http://projecteuclid.org/euclid.bsmsp/1200512988.

[127] Mia Hubert and Michiel Debruyne. "Minimum covariance determinant." In: *Wiley Interdisciplinary Reviews: Computational Statistics* 2 (Feb. 2009), pp. 36–43. DOI: 10.1002/wics.61.

[128] Mia Hubert, Peter J. Rousseeuw, and Stefan Van Aelst. "High-Breakdown Robust Multivariate Methods." In: *Statistical Science* 23.1 (2008), pp. 92–119. DOI: 10.1214/088342307000000087.

[129] William. G. Hunter and Albey M. Reiner. "Methods for Discriminating Between Two Rival Models." In: *Technometrics* 7.3 (Aug. 1965), pp. 307–323. URL: http://www.jstor.org/stable/1266591.

[130] E. T. Jaynes. "Bayesian Methods: General Background." In: *Maximum Entropy and Bayesian Methods in Applied Statistics*. Ed. by J. H. Justice. Cambridge University Press, 1985, pp. 1–25.

[131] Edwin Thompson Jaynes. "Information Theory and Statistical Mechanics II." In: *Physical Review* 108.2 (Oct. 1957), pp. 171–190. DOI: 10.1103/PhysRev.108.171.

[132] Edwin Thompson Jaynes. “Information Theory and Statistical Mechanics.” In: *Physical Review* 106.4 (May 1957), pp. 620–630. DOI: 10.1103/PhysRev.106.620.

[133] Edwin Thompson Jaynes. “Prior Probabilities.” In: *IEEE Transactions on Systems Science and Cybernetics* 4.3 (Sept. 1968), pp. 227–241. DOI: 10.1109/TSSC.1968.300117.

[134] Edwin Thompson Jaynes. *Probability theory: the Logic of Science*. 8th ed. Ed. by G. Larry Bretthorst. Cambridge University Press, 2011. ISBN: 978-0-521-59271-0.

[135] Norman L. Johnson, Samuel Kotz, and N. Balakrishnan. *Continuous Univariate Distributions. Volume 1*. 2nd ed. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc., 1994. ISBN: 9780471584957.

[136] Norman L. Johnson, Samuel Kotz, and N. Balakrishnan. *Continuous Univariate Distributions. Volume 2*. 2nd ed. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc., 1995. ISBN: 9780471584940.

[137] Olav Kallenberg. *Foundations of Modern Probability*. 2nd ed. Probability and Its Applications. Springer, Jan. 2002. ISBN: 0387953132.

[138] W. Karush. “Minima of functions of several variables with inequalities as side conditions.” MA thesis. Department of Mathematics, University of Chicago, 1939.

[139] Robert. E. Kass and Adrian E. Raftery. “Bayes Factors.” In: *Journal of the American Statistical Association* 90.430 (June 1995), pp. 773–795. URL: http://www.jstor.org/stable/2291091.

[140] Robert E. Kass, Luke Tierney, and Joseph B. Kadane. “The Validity of Posterior Expansions based on Laplace’s Method.” In: *Essays in Honor of George Barnard*. Ed. by S. Geisser, J. S. Hodges, S. J. Press, and A. Zellner. Bayesian and Likelihood Methods in Statistics and Economics. North-Holland Publishing, 1990.

[141] A. Khinchin. “Sur la loi des grands nombres.” In: *Comptes rendus de l’Académie des Sciences* 189 (1929), pp. 477–479.

[142] Aleksandr Khintchine. “Über einen Satz der Wahrscheinlichkeitsrechnung.” In: *Fundamenta Mathematicae* 6.1 (1924), pp. 9–20. URL: http://eudml.org/doc/214283.

[143] J. Kiefer and J. Wolfowitz. "Optimum Designs in Regression Problems." In: *The Annals of Mathematical Statistics* 30.2 (June 1959), pp. 271–294. URL: http://www.jstor.org/stable/2237082.

[144] B. J. K. Kleijn and A. W. van der Vaart. "The Bernstein-von-Mises theorem under misspecification." In: *Electronic Journal of Statistics* 6 (2012), pp. 354–381. DOI: 10.1214/12-EJS675.

[145] Bastiaan Jan Korneel Kleijn. "Bayesian Asymptotics Under Misspecification." PhD thesis. Faculteit der Exacte Wetenschappen, Afdeling Wiskunde, Vrije Universiteit Amsterdam, 2003.

[146] Ladislav Kocis and William J. Whiten. "Computational Investigations of Low-Discrepancy Sequences." In: *ACM Transactions on Mathematical Software* 23.2 (June 1997), pp. 266–294.

[147] S. Körkel. "Numerische Methoden für Optimale Versuchsplanungsprobleme bei nichtlinearen DAE-Modellen." PhD thesis. Heidelberg: Universität Heidelberg, 2002. URL: http://www.koerkel.de.

[148] S. Körkel, E. Kostina, H. G. Bock, and J. P. Schlöder. "Numerical Methods for Optimal Control Problems in Design of Robust Optimal Experiments for Nonlinear Dynamic Processes." In: *Optimization Methods and Software* 19 (2004), pp. 327–338.

[149] Ekaterina Kostina. "Robust Parameter Estimation in Dynamic Systems." In: *Optimization and Engineering* 5.4 (Dec. 2004), pp. 461–484.

[150] Samuel Kotz, N. Balakrishnan, and Norman L. Johnson. *Continuous Multivariate Distributions. Volume 1: Models and Applications*. 2nd ed. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc., 2000. ISBN: 9780471183877.

[151] M. K. Kozlov, S. P. Tarasov, and L. G. Khachiyan. "The polynomial solvability of convex quadratic programming." In: *USSR Computational Mathematics and Mathematical Physics* 20.5 (1980), pp. 223–228.

[152] Clemens Kreutz and Jens Timmer. "Systems biology: experimental design." In: *The FEBS Journal* 276 (2009), pp. 923–942.

[153] Bartosz Kuczewski. "Computational Aspects of Discrimination between Models of Dynamic Systems." PhD thesis. 65-246 Zielona Góra, Poland: University of Zielona Góra, 2006. ISBN: 83-7481-030-0.

[154] Anna Kuczmaszewska. “The strong law of large numbers for dependent random variables.” In: *Statistics & Probability Letters* 73.3 (July 2005), pp. 305–315. DOI: 10.1016/j.spl.2005.04.005.

[155] Peter Kühl, Moritz Diehl, Tom Kraus, Johannes P. Schlöder J.der, and Hans Georg Bock. “A real-time algorithm for moving horizon state and parameter estimation.” In: *Computers and Chemical Engineering* 35 (2011), pp. 71–83. URL: 10.1016/j.compchemeng.2010.07.012.

[156] H. W. Kuhn and A. W. Tucker. “Nonlinear programming.” In: *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*. Ed. by J. Neyman. Berkeley: University of California Press, 1951.

[157] S. Kullback and R. A. Leibler. “On Information and Sufficiency.” In: *The Annals of Mathematical Statistics* 22.1 (Mar. 1951), pp. 79–86. URL: http://www.jstor.org/stable/2236703.

[158] Solomon Kullback. *Information Theory and Statistics*. Wiley publications in statistics. New York: John Wiley & Sons, Inc., 1959.

[159] Solomon Kullback. *Information Theory and Statistics*. Dover Books on Mathematics. Corrected and extended republication of the book of Kullback [158]. Dover Publications Inc., 1997. ISBN: 0486696847.

[160] Solomon Kullback. “Letter to the Editor: The Kullback-Leibler distance.” In: *The American Statistician* 41 (1987), pp. 340–341. URL: http://www.jstor.org/stable/2684769.

[161] Serge Lang. *Fundamentals of Differential Geometry*. Vol. 191. Graduate Texts in Mathematics. Springer, 1999. ISBN: 9780387985930 (Hardcover) 9781461205418 (eBook).

[162] Pierre Simon Laplace. “Memoir on the Probability of the Causes of Events.” In: *Statistical Science* 1.3 (Aug. 1986). Translated from the original French by S. M. Stiegler, University of Chicago. Originally published as “Mémoire sur la probabilité des causes par les évènements,” par M. de la Place, Professeur á l’Ècole royal Militaire, in *Mémoires des Mathématique et de Physique, Presentés à l’Académie Royale des Sciences, par divers Savanas & lûs dans ses Assemblées, Tome Sixieme* (1774) 621–656., pp. 364–378.

[163] Pierre Simon Laplace. *Théorie analytique des probabilités*. Paris: Ve. Courcier, 1814.

[164] Jan Larsen. *Gaussian Integrals.* Tech. rep. Version 2. Intelligent Signal Processing Group, Informatics and Mathematical Modelling, Technical University of Denmark, Jan. 2005.

[165] Lucien M. Le Cam. *Les Propriétés Asymptotiques Des Solutions De Bayes.* 7. Publications de l'institut de statistique de l'universite de Paris, 1958, pp. 17–35.

[166] Lucien M. Le Cam. *On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates.* Vol. 1. University of California publications in statistics 11. University of California Press, 1953, pp. 227–329.

[167] Pierre L'Ecuyer. "Good Parameters And Implementations For Combined Multiple Recursive Random Number Generators." In: *Operations Research* 47.1 (Feb. 1999), pp. 159–164. DOI: 10.1287/opre.47.1.159.

[168] Juhee Lee and Steven N. MacEachern. "Consistency of Bayes estimators without the assumption that the model is correct." In: *Journal of Statistical Planning and Inference* 141 (2011), pp. 748–757. DOI: 10.1016/j.jspi.2010.07.022.

[169] Peter M. Lee. *Bayesian Statistics: An Introduction.* 4th ed. Wiley, 2012. ISBN: 978-1-118-33257-3;

[170] E. L. Lehmann and George Casella. *Theory of Point Estimation.* 2nd ed. Springer Texts in Statistics. Springer, 1998. ISBN: 0387985026.

[171] E. L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses.* Springer Texts in Statistics. Spr, 2005. ISBN: 0378988645.

[172] Simon Lenz. "Impulsive Hybrid Discrete-Continuous Delay Differential Equations." PhD thesis. Heidelberg University, 2014. URL: http://www.ub.uni-heidelberg.de/archiv/17117.

[173] A. C. Ponce de Leon and A. C. Atkinson. "Optimum experimental design for discriminating between two rival models in the presence of prior information." In: *Biometrika* 78.3 (1991), pp. 601–608.

[174] Antonio Carlos Monteiro Ponce de Leon. "Optimum Experimental Design for Model Discrimination and Generalized Linear Models." PhD thesis. London: London School of Economics, Political Sciences, Department of Statistical, and Mathematical Sciences, June 1993.

[175] Friedrich Liese and Igor Vajda. “On Divergences and Informations in Statistics and Information Theory.” In: *IEEE Transactions on Information Theory* 52.10 (Oct. 2006), pp. 4394–4412.

[176] D. V. Lindley. “On a Measure of the Information Provided by an Experiment.” In: *The Annals of Mathematical Statistics* 27.4 (Dec. 1956), pp. 986–1005. URL: http://www.jstor.org/stable/2237191.

[177] J. López-Fidalgo, C. Tommasi, and P. C. Trandafir. “An optimal experimental design criterion for discriminating between non-normal models.” In: *Journal of the Royal Statistical Society, Series B (Methodological)* 69.2 (2007), pp. 231–242. URL: http://www3.interscience.wiley.com/cgi-bin/fulltext/118490758/PDFSTART.

[178] J. N. Lyness and C. B. Moler. “Numerical Differentiation of Analytic Functions.” In: *SIAM Journal on Numerical Analysis* 4 (1967), pp. 202–210.

[179] David Madigan and Adrian E. Raftery. “Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam’s Window.” In: *Journal of the American Statistical Association* 89.428 (Dec. 1994), pp. 1535–1546. DOI: 10.1080/01621459.1994.10476894.

[180] George Marsaglia and Wai Wan Tsang. “The Ziggurat Method for Generating Random Variables.” In: *Journal of Statistical Software* 5.8 (2000), pp. 1–7. DOI: 10.18637/jss.v005.i08. URL: http://www.jstatsoft.org/article/view/v005i08.

[181] J. R. R. A. Martins, P. Sturdza, and J. J. Alonso. “The Complex-Step Derivative Approximation.” In: *ACM Transactions on Mathematical Software* 29.3 (Sept. 2003), pp. 245–262. DOI: 10.1145/838250.838251.

[182] Jiří Matoušek. “On the L_2-Discrepancy for Anchored Boxes.” In: *Journal of Complexity* 14 (1998), pp. 527–556.

[183] Caterina May and Chiara Tommasi. “Model Selection and Parameter Estimation in Non-Linear Nested Models: a Sequential Generalized DKL-Optimum Design.” In: *Statistica Sinica* 24.1 (2014), pp. 63–82.

[184] McMillan. “The basic theorems of information theory.” In: *Annals of Mathematical Statistics* 24.2 (1953), pp. 196–219. DOI: 10.1214/aoms/1177729028. URL: http://projecteuclid.org/euclid.aoms/1177729028.

[185] Duane Meeter, Walter Pirie, and William Blot. "A Comparison of Two Model-Discrimination Criteria." In: *Technometrics* 12.3 (1970), pp. 457–470. URL: http://www.jstor.org/stable/1267196.

[186] Claas Michalik, Maxim Stuckert, and Wolfgang Marquardt. "Optimal Experimental Design for Discriminating Numerous Model Candidates: The AWDC Criterion." In: *Industrial and Engineering Chemistry Research* 49.2 (2010), pp. 913–919. DOI: 10.1021/ie900903u. URL: http://pubs.acs.org/doi/abs/10.1021/ie900903u.

[187] Maher Moakher and Philipp G. Batchelor. "Symmetric Positive-Definite Matrices: From Geometry to Applications and Visualization." In: *Visualization and Processing of Tensor Fields*. Ed. by Joachim Weickert and Hans Hagen. Mathematics and Visualization. Springer, 2006. Chap. 17, pp. 285–298. ISBN: 9783540250326 (Print), 9783540312727 (Online). DOI: 10.1007/3-540-31272-2_17.

[188] Ilya Molchanov and Sergei Zuyev. "Steepest Descent Algorithms in Space of Measures." In: *Statistics and Computing* 12 (Dec. 2002), pp. 115–123. DOI: 10.1023/A:1014878317736.

[189] William J. Morokoff and Russell E. Catflisch. "Quasi-Random Sequences and Their Discrepancies." In: *SIAM Journal on Scientific Computing* 15.6 (Nov. 1994), pp. 1251–1279.

[190] Robb J. Muirhead. *Aspects of Multivariate Statistical Theory*. 2nd ed. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Sept. 2005. ISBN: 978-0471769859.

[191] K. G. Murty. "Some NP-complete problems in quadratic and nonlinear programming." In: *Mathematical Programming* 39 (1987), pp. 117–129.

[192] Harald Niederreiter. "Quasi-Monte Carlo Methods and Pseudo-Random Numbers." In: *Bulletin of the Americal Mathematical Society* 84.6 (Nov. 1978), pp. 957–1041.

[193] Harald Niederreiter. *Random Number Generation and Quasi-Monte Carlo Methods*. CBMS-NSF Regional Conference Series in Applied Mathematics 63. Society for Industrial and Applied Mathematics, June 1992. ISBN: 9780898712957.

[194] Jorge Nocedal and Stephen G. Wright. *Numerical Optimization*. 2nd ed. Springer Series in Operations Research and Financial Engineering. Springer, 2006. ISBN: 9780387303031 (Hardcover), 9780387400655 (eBook).

[195] Anthony O'Hagan and Jon Forster. *Bayesian Inference*. 2nd ed. Vol. 2B. Kendalls Advanced Theory of Statistic. Wiley-Blackwell, June 2009. ISBN: 9780470685693.

[196] Art B. Owen. "Monte Carlo Variance of Scrambled Net Quadrature." In: *SIAM Journal on Numerical Analysis* 34.5 (Oct. 1997), pp. 1884–1910. URL: http://www.siam.org/journals/sinum/34-5/27746.html.

[197] P. M. Pardalos and G. Schnitger. "Checking local optimality in constrained quadratic programming is NP-hard." In: *Operational Research Letters* 7.1 (Feb. 1988), pp. 33–35.

[198] Panos M. Pardalos. "Global optimization algorithms for linearly constrained indefinite quadratic problems." In: *Computers & Mathematics with Applications* 21.6–7 (1991), pp. 87–97.

[199] Panos M. Pardalos and Stephen A. Vavasis. "Quadratic Programming with One Negative Eigenvalue Is NP-Hard." In: *Journal of Global Optimization* 1.1 (1991), pp. 15–22. DOI: 10.1007/BF00120662.

[200] Yudi Pawitan. *In All Likelihood*. Oxford University Press, 2001. ISBN: 0198507658.

[201] Kaare Brandt Petersen and Michael Syskind Pedersen. *The Matrix Cookbook*. Tech. rep. Nov. 2012. URL: http://matrixcookbook.com.

[202] Hai V. Pham and Frank T.-C. Tsai. "Optimal observation network design for conceptual model discrimination and uncertainty reduction." In: *Water Resources Research* 52.2 (Feb. 2016), pp. 1245–1264. DOI: 10.1002/2015WR017474.

[203] A. N. Philippou and G. G. Roussas. "Asymptotic Normality of the Maximum Likelihood Estimate in the Independent not Identically Distributed Case." In: *Annals of the Institute of Statistical Mathematics* 27.1 (1975), pp. 45–55. DOI: 10.1007/BF02504623.

[204] S. D. Poisson. *Recherche sur la Probabilité des jugements en matière criminelle et en matière civile, précédées des règles générales du calcul des probabilitiés*. Paris, France: Bachelier, 1837.

[205] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes: The Art of Scientific Computing*. 3rd ed. Cambridge University Press, 2007. ISBN: 9780521880688.

[206] Friedrich Pukelsheim. *Optimal Design of Experiments*. John Wiley & Sons, Inc., 1993. ISBN: 0898716047.

[207] Friedrich Pukelsheim and Sabine Rieder. "Efficient rounding of approximate designs." In: *Biometrika* 79.4 (1992), pp. 763–770.

[208] Adrian E. Raftery. "Bayesian Model Selection in Social Research." In: *Sociological Methodology* 25 (1995), pp. 111–163. DOI: 10.2307/271063.

[209] Adrian E. Raftery, David Madigan, and Chris T. Volinsky. "Accounting for Model Uncertainty in Survival Analysis Improves Predictive Performance." In: *Bayesian Statistics 5. Proceedings of the Fifth Valencia International Meeting, June 5-9, 1994.* Ed. by J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith. Oxford University Press, May 1996. ISBN: 9780198523567.

[210] Calyampudi Radakrishna Rao. "Information and the accuracy attainable in the estimation of statistical parameters." In: *Bulletin of the Calcutta Mathematical Society* 37 (1945), pp. 81–89.

[211] Nancy Reid. "Likelihood inference." In: *Wiley Interdisciplinary Reviews: Computational Statistics* 2.5 (Sept. 2010), pp. 517–525. ISSN: 1939-0068. DOI: 10.1002/wics.110.

[212] Alfréd Rényi. "On Measures of Entropy and Information." In: *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability.* Vol. 1. Berkeley: University of California Press, 1691, pp. 547–561.

[213] Christian P. Robert. *The Bayesian Choice - A Decision-Theoretic Motivation.* Springer Texts in Statistics. Springer, 1994.

[214] Thomas J. Rothenberg. "Identification in Parametric Models." In: *Econometric* 39.3 (May 1971), pp. 577–591. URL: http://www.jstor.org/stable/1913267.

[215] Peter J. Rousseeuw. "Least Median of Squares Regression." In: *Journal of the American Statistical Association* 79.388 (Dec. 1984), pp. 871–880. URL: http://www.jstor.org/stable/2288718.

[216] Peter J. Rousseeuw and Katrien Van Driessen. "A Fast Algorithm for the Minimum Covariance Determinant Estimator." In: *Technometrics* 41.3 (Aug. 1999), pp. 212–223. DOI: 10.1080/00401706.1999.10485670.

[217] Peter Rousseeuw. "Multivariate Estimation with High Breakdown Point." In: *Mathematical Statistics and Applications.* Ed. by W. Grossmann, G. C. Pflug, I. Vincze, and W. Wertz. Dordrecht, Netherlands: Reidel Publishing Company, 1985, pp. 283–297. ISBN: 9789027720887.

[218] Sam Roweis. *gaussian identities*. Tech. rep. New York University, July 1999.

[219] Sartaj Sahni. "Computationally Related Problems." In: *SIAM Journal on Computing* 3.4 (Dec. 1974), pp. 262–279. doi: 10.1137/0203021. url: http://dx.doi.org/10.1137/0203021.

[220] Takamitsu Sawa. "Information Criteria for Discriminating Among Alternative Regression Models." In: *Econometrica* 46.6 (1978), pp. 1273–1291. url: http://www.jstor.org/stable/1913828.

[221] Spencer D. Schaber, Stephen C. Born, Klavs F. Jensen, and Paul I. Barton. "Design, Execution, and Analysis of Time-Varying Experiments for Model Discrimination and Parameter Estimation in Microreactors." In: *Organis Process Research & Development* 18 (Sept. 2014), pp. 1461–1467. doi: 10.1021/op500179r.

[222] Christoph Schlier. "On scrambled Halton sequences." In: *Applied Numerical Mathematics* 58 (Sept. 2008), pp. 1467–1478. doi: 10.1016/j.apnum.2007.09.001.

[223] J. P. Schlöder. *Numerische Methoden zur Behandlung hochdimensionaler Aufgaben der Parameteridentifizierung.* Vol. 187. Bonner Mathematische Schriften. Bonn: Universität Bonn, 1988.

[224] Marcio Schwaab, José Luiz Monteiro, and José Carlos Pinto. "Sequential experimental design for model discrimination taking into account the posterior covariance matrix of differences between model predictions." In: *Chemical Engineering Science* 63 (Feb. 2008), pp. 2408–2419.

[225] Marcio Schwaab, Fabrício M. Silva, Christian A. Queipo, Amaro G. Barreto Jr., Márcio Nele, and José Carlos Pinto. "A new approach for sequential experimental design for model discrimination." In: *Chemical Engineering Science* 61 (Apr. 2006), pp. 5791–5806.

[226] Loraine Schwartz. "On Bayes Procedures." In: *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 4 (1965), pp. 10–26.

[227] Loraine Schwartz. "On Consistency of Bayes Procedures." In: *Proceedings of the National Academy of Sciences* 52 (1964), pp. 46–49.

[228] D. F. Shanno. "Conditioning of Quasi–Newton methods for function minimization." In: *Mathematics of Computation* 24.111 (July 1970), pp. 647–656. url: http://www.jstor.org/pss/2004840.

[229] C. E. Shannon. "A Mathematical Theory of Communication." In: *The Bell System Technical Journal* 27 (July 1948), pp. 379–423, 623–656.

[230] Claude E. Shannon. "Communication in the Presence of Noise." In: *Proceedings of the IRE* 37.1 (Jan. 1949), pp. 10–21.

[231] Claude E. Shannon and Warren Weaver. *The Mathematical Theory of Communication.* University of Illinois Press, 1949. ISBN: 9780252725487.

[232] Jun Shao. *Mathematical Statistics.* 2nd ed. Springer texts in statistics. Springer, 2003. ISBN: 0387953825.

[233] Chor-Yiu Sin and Halbert White. "Information criteria for selecting possibly misspecified parametric models." In: *Journal of Econometrics* 71.1-2 (Apr. 1996), pp. 207–225.

[234] Dominik Skanda and Dirk Lebiedz. "An optimal experimental design approach to model discrimination in dynamic biochemical systems." In: *Bioinformatics* 26.7 (2010), pp. 939–945. DOI: 10.1093/bioinformatics/btq074. URL: http://bioinformatics.oxfordjournals.org/cgi/reprint/btq074v1.pdf.

[235] Il'ya Meerovich Sobol. "On the distribution of points in a cube and the approximate evaluation of integrals." In: *USSR Computational Mathematics and Mathematical Physics* 7.4 (1967), pp. 86–112. DOI: 10.1016/0041-5553(67)90144-9.

[236] Geraldo Souza. "Statistical inference in nonlinear models: a pseudo likelihood approach." PhD thesis. Rayleigh, NC, USA: North Caroline State University, 1979.

[237] I. Stamati, F. Logist, S. Akkermans, E. Noriega Fernández, and J. Van Impe. "On the effect of sampling rate and experimental noise in the discrimination between microbial growth models in the suboptimal temperature range." In: *Computers and Chemical Engineering* 85 (2016), pp. 84–93. URL: http://dx.doi.org/10.1016/j.compchemeng.2015.10.005.

[238] David M. Steinberg and William G. Hunter. "Experimental Design: Review and Comment." In: *Technometrics* 26.2 (May 1984), pp. 71–97. URL: http://www.jstor.org/stable/1268097.

[239] W. E. Stewart, Y. Shon, and G. E. P. Box. "Discrimination and Goodness of Fit of Multiresponse Mechanistic Models." In: *AIChE Journal* 44.6 (June 1998), pp. 1404–1412.

[240] Warren E. Stewart, Thomas L. Henson, and George E. P. Box. "Model Discrimination and Criticism with Single-Response Data." In: *AIChE Journal* 42.11 (Nov. 1996), pp. 3055–3062.

[241] M. Stone. "Application of a Measure of Information to the Design and Comparison of Regression Experiments." In: *The Annals of Mathematical Statistics* 30.1 (1959), pp. 55–70. URL: http://www.jstor.org/stable/2237120.

[242] T. J. Sweeting. "Uniform Asymptotic Normality of the Maximum Likelihood Estimator." In: *The Annals of Statistics* 8.6 (Nov. 1980), pp. 1375–1381. URL: http://www.jstor.org/stable/2240949.

[243] C. Tommasi and J. López-Fidalgo. "Bayesian optimum designs for discriminating between models with any distribution." In: *Computational Statistics & Data Analysis* 54 (2010), pp. 143–150. DOI: 10.1016/j.csda.2009.07.022.

[244] Chiara Tommasi. "Optimal designs for discriminating among several non-Normal models." In: *mODa 8 – Advances in Model-Oriented Design and Analysis*. Ed. by Jesus López-Fidalgo, Juan Manuel Rodríguez-Díaz, and Ben Torsney. Contributions to Statistics. Heidelberg, Germany: Physica Verlag, 2007, pp. 213–220.

[245] Chiara Tommasi and Jesús López-Fidalgo. *Bayesian optimal designs for discriminating between non-Normal models*. UNIMI - Research Papers in Economics, Business, and Statistics 1055. Universitá degli Studi di Milano, May 2007. URL: http://ideas.repec.org/p/bep/unimip/1055.html.

[246] Chiara Tommasi, Maria Teresa Santos-Martín, and Juan Manuel Rodríguez-Díaz. "Discrimination Between Random and Fixed Effect Logistic Regression Models." In: *mODa 9 – Advances in Model-Oriented Design and Analysis*. Ed. by Alessandra Giovagnoli, Anthony C. Atkinson, and Bernard Torsney. Contributions to Statistics 2. Berlin, Germany: Springer, 2010. ISBN: 978-3-7908-2409-4. DOI: 10.1007/978-3-7908-2410-0.

[247] Lloyd N. Trefethen. *The Definition of Numerical Analysis*. Tech. rep. TR 92-1304. Cornell University, Sept. 1992.

[248] Dariusz Uciński and Barbara Bogacka. "A constrained optimum experimental design problem for model discrimination with a continuously varying factor." In: *Journal of Statistical Planning and Inference* 137 (Apr. 2007), pp. 4048–4065.

[249] Dariusz Uciński and Barbara Bogacka. “A functional experimental design factor for optimum discrimination between multiresponse models.” In: *Proceedings of the 5th St. Petersburg Workshop on Simulation.* Ed. by S. M. Ermakov, V. B. Melas, and A. N. Pepelyshev. St. Petersburg, Russia: St. Petersburg University Press, 2005, pp. 703–708.

[250] Dariusz Uciński and Barbara Bogacka. “Construction of T-optimum designs for multiresponse dynamic models.” In: *COMPSTAT – Proceedings in Computational Statistics.* Ed. by Wolfgang Härdle and Bernd Rönz. Vol. 15. Berlin: Physica Verlag, 2002.

[251] Dariusz Uciński and Barbara Bogacka. “T-Optimum Designs for Discrimination between two Multiresponse Dynamic Models.” In: *Journal of the Royal Statistical Society, Series B (Methodological)* 67.1 (2005), pp. 3–18. URL: http://www.jstor.org/stable/3647597.

[252] Dariusz Uciński and Barbara Bogacka. “T-Optimum Designs for Multiresponse Dynamic Heteroscedastic Models.” In: *mODa 7 – Advances in Model-Oriented Design and Analysis. Proc. 7-th Int. Workshop in Model-Oriented Design and Analysis 2004.* Ed. by A. Di Bucchianico, H. Läuter, and H. P. Wynn. Heeze, the Netherlands: Physica Verlag, Heidelberg, 2004, pp. 191–199.

[253] Johannes Gualtherus van der Corput. “Verteilungsfunktionen.” In: *Proceedings of the Nederlandse Akademie van Wetenschappen* 38 (1935), pp. 813–821.

[254] Stephen A. Vavasis. “Approximation algorithms for indefinite quadratic programming.” In: *Mathematical Programming* 57.1 (May 1992), pp. 279–311.

[255] Stephen A. Vavasis. *Nonlinear Optimization: Complexity Issues.* Vol. 8. International series of monographs on computer science. Oxford University Press, 1991. ISBN: 0195072081.

[256] Stephen A. Vavasis. “Quadratic Programming is in NP.” In: *Information Processing Letters* 39.2 (Oct. 1990), pp. 73–77.

[257] Sabine Verboven and Mia Hubert. “LIBRA: a MATLAB Library for Robust Analysis.” In: *Chemometrics and Intelligent Laboratory Systems* 75 (June 2004), pp. 127–136.

[258] Sabine Verboven and Mia Hubert. “MATLAB library LIBRA.” In: *Wiley Interdisciplinary Reviews: Computational Statistics* 2.4 (Aug. 2010), pp. 509–515. DOI: 10.1002/wics.96.

[259] Udo von Toussaint. "Bayesian inference in physics." In: *Reviews of Modern Physics* 83 (Sept. 2011), pp. 943–999. doi: 10.1103/RevModPhys.83.943. url: http://link.aps.org/doi/10.1103/RevModPhys.83.943.

[260] Quang H. Vuong. *Cramér-Rao Bounds for Misspecified Models.* Working Papers 652. Division of the Humanities and Social Sciences, 228-77, Caltech, Pasadena CA 91125: California Institute of Technology, Division of the Humanities and Social Sciences, 1986. url: http://EconPapers.repec.org/RePEc:clt:sswopa:652.

[261] Quang H. Vuong. "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses." In: *Econometrica* 57.2 (Mar. 1989), pp. 307–333. url: http://www.jstor.org/stable/1912557.

[262] Abraham Wald. "Note on the Consistency of the Maximum Likelihood Estimate." In: *Annals of Mathematical Statistics* 20.4 (1949), pp. 595–601. doi: 10.1214/aoms/1177729952. url: http://projecteuclid.org/euclid.aoms/1177729952.

[263] Éric Walter and Luc Pronzato. *Identification of Parametric Models.* Communications and Control Engineering. Springer, 1997. isbn: 3540761195.

[264] Dirk Werner. *Funktionalanalysis.* 4th ed. Springer, 2002. isbn: 9783540435860.

[265] Halbert White. *Estimation, Inference and Specification Analysis.* Econometric Society Monographs 22. Cambridge University Press, 1994. isbn: 0521252806.

[266] Halbert White. "Maximum Likelihood Estimation of Misspecified Models: Corrigendum." In: *Econometrica* 51.2 (Mar. 1983), p. 513. url: http://www.jstor.org/stable/1912004.

[267] Halbert White. "Maximum Likelihood Estimation of Misspecified Models." In: *Econometrica* 50.1 (Jan. 1982), pp. 1–25. url: http://www.jstor.org/stable/1912526.

[268] R. B. Wilson. "A simplicial algorithm for concave programming." PhD thesis. Harvard University, 1963.

[269] Wenling Zhang, Michael Binns, Constantinos Theodoropoulos, Jin-Kuk Kim, and Robin Smith. "Model Building Methodology for Complex Reaction Systems." In: *Industrial & Engineering Chemistry Research* 54 (2015), pp. 4603–4615. doi: 10.1021/ie504343d.