INAUGURAL-DISSERTATION

zur
Erlangung der Doktorwürde
der
Naturwissenschaftlich-Mathematischen Gesamtfakultät
der Ruprecht-Karls-Universität
Heidelberg

vorgelegt von

Diplom-Mathematiker Fritz Moritz von Rohrscheidt

aus Mainz

Tag der mündlichen Prüfung:

Bayesian Nonparametric Inference for Queueing Systems

Betreuer: **Prof. Dr. Rainer Dahlhaus**
Dr. Cornelia Wichelhaus

Acknowledgments

I would like to thank my supervisor, Prof. Dr. Dahlhaus, for giving me the opportunity of writing this thesis. Since my undergraduate levels I enjoyed lectures on probability theory and statistics given by him and I am grateful for everything he taught me during the years.

Dr. Cornelia Wichelhaus provided me with the background about the theory of queues necessary for writing this thesis. I savored her support not only with respect to this thesis but also since the time I started my Diploma. I am looking back upon a range of interesting and vivid seminars given by her.

Also, I am very grateful to Prof. Dr. Jan Johannes who strongly supported me during the last year of my time as doctoral student. Numerous discussions on Bayesian statistics led me to improve the thesis considerably. I admire his motivating style of teaching as much as his readiness to guide and help.

The time during which I shared an office in Heidelberg with my dear colleague Mehmet Madensoy was a particular pleasure. We know each other since the first lecture I took in Heidelberg as an undergraduate student and we experienced most highs and lows of our studies together. He is a great discussion partner and we have had interestingly deep conversations ranging from probability theory to politics.

Of course, I want to thank my beloved girlfriend Maike and my entire family for all their aid and encouragement they lend on me during the years.

Finally, I am grateful to the *Deutsche Forschungs Gesellschaft (DFG)* for their support that was accompanied with the work in the Research Training Group (RTG) 1953 "Statistical Modelling of Complex Systems and Processes" as well as to the administration of the *Combined Faculty of Natural Sciences and Mathematics* of the Ruprecht-Karls-Universität Heidelberg.

Zusammenfassung

Die vorliegende Arbeit handelt von statistischer Inferenz für Modelle von Warteschlangen. Dabei bedienen sich die statistischen Ansätze der Bayesianischen Methodik. Die Arbeit gliedert sich in drei Hauptteile, die über das zugrunde liegende Wartschlangenmodell miteinander verbunden sind. Die Gemeinsamkeit aller Teile besteht darin, dass durchweg Warteschlangen in stetiger Zeit betrachtet werden, die von einem Kundenstrom in Form eines homogenen Poisson Prozesses gespeist werden. Außerdem wird durchweg keine endlich-dimensionale, parametrische Form der Bedienzeit-Verteilung angenommen. Letzteres führt zu sogenannter nichtparametrischer Bayes Statistik. Der Unterschied zwischen den einzelnen Teilen ergibt sich zum einen daraus, dass die beiden ersten Teile den Fall eines einzelnen Bedieners behandeln, während der dritte Teil den Fall von unendlich vielen Bedienern abdeckt. Zum anderen unterscheiden sich alle Teile bezüglich der Annahmen an die Beobachtungen des jeweiligen Systems. Im ersten Teil dienen Ankunfts- und Bedienprozess als Beobachtungsgrundlage. Dabei wird statistische Inferenz für die Verteilungen der Wartezeit und der Systemgröße entwickelt. Die weitere Untersuchung dieser Methoden führt neben dem Herausarbeiten ihrer wichtigsten Eigenschaften zu neuen theoretischen Resultaten der Bayes Statistik selbst. Teil zwei ist dem gleichen Warteschlangenmodell gewidmet, nun allerdings bezüglich anderer Beobachtungen. Dabei entfällt die Annahme, dass Ankunfts- und Bedienprozess beobachtbar sind, sodass die einzig mögliche Beobachtung im Abgangsprozess der Kunden besteht. In diesem Rahmen wird Inferenz für die Bedienzeit-Verteilung entwickelt. Dabei liegt das Augenmerk vor allem auf der probabilistischen Struktur des Systems. Die daraus entstehenden theoretischen Resultate über die zugrunde liegende suffiziente Statistik erlauben es anschließend Mischungen solcher Systeme zu betrachten, die das Fundament für Bayesianische Inferenz innerhalb dieses Modellrahmens bilden. Der letzte Teil unterscheidet sich von den beiden vorangegangenen insofern, dass nun die Warteschlange mit unendlich vielen Bedienern betrachtet wird, d.h. es wird angenommen, dass jedem Kunden sein eigener Bediener zugewiesen wird und sich somit keine Warteschlange im eigentlichen Sinn bildet. Das statistische Interesse liegt dabei wiederum auf der Bedienzeit-Verteilung und die Beobachtungen sind von indirekter Natur. Letzteres bedeutet, dass lediglich die Zeiten zu denen Kunden das System betreten bzw. verlassen aufgezeichnet werden. Die Verteilung der daraus resultierenden Rohdaten besitzt einen bekannten Zusammenhang zur Bedienzeit-Verteilung. Dieser Zusammenhang wurde in der Vergangenheit ausgenutzt, um Statistik für die Bedienzeit-Verteilung im frequentistischen Sinne zu betreiben. Allerdings wurden dabei bislang ungeklärte Fragen aufgeworfen, deren Klärung Voraussetzung ist für eine Bayesianische Behandlung des Problems. Da die Rohdaten einen allgemeinen stationären Prozess bilden, besteht das zentrale Problem in einer geeigneten Parametrisierung der stationär-ergodischen Maße, deren Mischung das datengenerierende Maß bildet. Eine solche Parametrisierung wird entwickelt und anschließend benutzt um Bayesianische Inferenz für stationäre Daten zu betreiben.

Abstract

The present thesis deals with statistical inference for queueing models. Thereby, the considered approaches follow the Bayesian methodology. The work divides into three main parts which are related to each other by the underlying queueing model. The similarity lies in the assumption of continuous-time systems which are fed by a homogenous Poisson arrival stream of customers. Moreover, throughout the thesis no finite-dimensional parametric constraint is placed on the distribution of the service times. The latter leads to Bayesian nonparametric statistics. The distinction among the parts arise from the assumed number of servers as well as from the observational setups. The first two parts are about the single server queue while the last part deals with infinitely many servers. In the first part the arrival and service processes are taken as observations. Thereby, the main interest is in inference for the distributions of the waiting times of the customers and the occupation of the system, respectively. Besides the elaboration of the key properties of the statistical methods, their further examinations lead to new results for the theory of Bayesian statistics itself. The second part is also about the same system but a different observational setup is used. This means that the assumption of observations of the arrival and service process is dropped. Instead merely the customer's departure stream is assumed to be observable. Within this setup the main interest is in making inference for the service time distribution. This is done by studying the probabilistic structure of the system in more depth. The emerging theoretical considerations about the sufficient statistic of this inner structure make it possible to think about mixtures of such systems. These mixtures build the basis for the development of further Bayesian nonparametric inference for the service time distribution within this framework. The last part departs from the previous two in such that the queue with infinitely many servers is considered. This means that a separate server is assigned to each customer and no queue builds up. The interest is again in the service time distribution. The observations are indirect, meaning that merely the instants when customers arrive to and depart from the system, respectively, are recorded. The distribution of the emerging raw data has a known relationship to the service time distribution. This relationship was exploited in the past in order to make statistical inference for the service time distribution using the frequentist methodology. However, this raised several questions which have not been answered yet. This thesis provides answers to these questions which are necessary to deal with the issue from a Bayesian perspective. Since the raw data forms a general stationary process, the major problem consists in finding a suitable parametrization of the shift-ergodic measures. That is mainly due to the fact that this enables one to formalize mixtures of ergodic measures which in turn generate the observed data. Such a parametrization is developed and subsequently used for making Bayesian inference for stationary data.

Dedicated to the memory of our dear friend
Christopher Paul Kobrak

Contents

1 Introduction

Much attention has been drawn to the analysis and the statistics of queueing systems. Beside approaches to statistical inference for queueing systems from the classical perspective, since the early 1980s there has taken place remarkable research in what is frequently called Bayesian queues. That means inference for queueing systems from the Bayesian point of view. Bayesian statistical approaches are often feasible and useful in the area of Operations Research since one is able to express prior knowledge about the system which is often present. To cite Dennis Lindley, who was a great advocate of the Bayesian idea on the one hand side and also worked in the field of stochastic processes affecting queueing theory

> *Operational research workers are continually trying to express ideas to management that involve uncertainty: they should do it using the concepts contained therein.*
>
> -Excerpt of the foreword of De Finetti (1974) by Dennis Lindley-

The possibility of incorporating prior knowledge into inferential procedures makes inference more robust especially if the à priori knowledge is good and the amount of available data is small. However, it is by far not the only feature that makes Bayesian statistics attractive to statisticians examining queueing systems. Another one is the richness of designs and statistical modeling which is offered to the statistician. This richness is mainly due to the fact that rather sophisticated methods are utilized than merely replacing parameters by their empirical versions and, thus, often has to come along with deep theoretical considerations. Presumably the most striking advantage is that the Bayesian approach to statistics enables one to epistemically predict the future behavior of the system which is due to the subjectivistic approach to probability which Bayesian statistics build on. More precisely, the interpretation of probability as a personal state of information or uncertainty gives rise to a prior distribution. Loosely speaking, the prior distribution models one's uncertainty by mixing up different rigid states which are themselves represented by probability distributions, the so called likelihoods. One's state of information is changed by observing data in a certain way to be further described. Informally, the prior is updated to the posterior by the data. This updating procedure can by accomplished by different tools. The most popular one is the celebrated Bayes theorem

$$\mathbb{P}(\mathsf{target} \mid \mathsf{data}) = \frac{\mathbb{P}(\mathsf{data} \mid \mathsf{target})\mathbb{P}(\mathsf{target})}{\mathbb{P}(\mathsf{data})}$$

which roughly says that the posterior behaves proportionally to the product of the likelihood and the prior. However, in situations where the interest lies in objects of infinite

dimension, which often appear in queueing theory, classical approaches break down. Then other theoretical methods need to be found in order to define priors on measure spaces that are not representable by finite dimensional parameter spaces and to update them properly to a posterior. This leads to the field of Bayesian nonparametrics. Having access to the entire posterior measure can be more telling than solely regarding one of its specific functionals in form of estimators. Moreover, one can actually write down a posterior predictive distribution

$$\mathbb{P}(\text{future data} \mid \text{observed data}) = \int \mathbb{P}(\text{future data} \mid \text{target}) d\mathbb{P}(\text{target} \mid \text{observed data})$$

which is vacuous in general from the viewpoint of classical frequentistic statistics since the data mostly has to fulfill several independence assumptions. In contrast, the associated assumption in Bayesian statistics is relaxed to conditional independence which, by the celebrated de Finetti theorem, is equivalent to a judgment of the data that has observable character. That is the judgment of exchangeability which basically means that the order of observable data does not affect the update, i.e. the order statistic is judged to be sufficient for future predictions. Certainly this is not true in many situation. However, following the same methodology other useful judgments can be found to express ones belief in the situation to be modeled. From a more theoretical viewpoint all these considerations are dealt with by ergodic theory, which might be regarded as another advantage since ergodic theory has long reaching interrelations with other mathematical theories.

Several scientists followed these advantages in order to create a vivid research area by combining the field of stochastic processes with Bayesian statistics. While the earliest works include inference based on the classical Bayes theorem for exactly solvable queueing systems, more recent works began to transfer the methodology to more general systems using non-parametric Bayesian statistics. It is worth mentioning some of the recent works on Bayesian queues. Two seminal works were given by Armero (1985) and Mcgrath et al. (1987); McGrath and Singpurwalla (1987) who dealt with Markovian queues, i.e. queues with a homogeneous Poisson process as input stream and exponential service times with varying number of service stations. In this parametric and explicitly solvable model, the aim of these works was to infer the traffic intensity, the waiting time of customers and the number of customers in the system under several different observational setups. These works form the base for further works on Bayesian inference for systems with Markovian characters under several generalizations cf. e.g. Armero and Conesa (1998, 2004, 2006) and Ausín et al. (2007). Generalizations in many ways were performed as for example a non-Markovian setting for the inter arrival times, cf. Wiper (1998) and Ausín et al. (2007).
While assuming the arrival stream to be a Poisson process is often well suited, the service times often need to be modeled more flexibly. Hence, it is reasonable to generalize the distribution of the service times. Since the most desirable properties of queueing systems depend on the Poisson process input, this can be done without loosing such. Models with generalizations of the service time distribution can be found in Insua et al. (1998), where Erlang- and hyperexponential distributions are employed to model the service times more flexibly. Ausín et al. (2004) use the well known probabilistic result that the class of phase-type distributions forms a dense set in the space of all probability distributions [cf. Asmussen (2008, p.84)]. They model the service time distribution semi-parametrically, assign prior distributions to its parameters and infer the system using MCMC procedures. However, one is often interested in an even more general approach for modeling the ser-

vice time distribution which leads to a non-parametric Bayesian approach to estimate the unknown service time distribution. The first work in a discrete-time framework concerning this was given by Conti (1999) for a $Geo/G/1$. The basic assumption of this paper was that one is able to observe both, the arrival and service process while its aim was to make inference for the waiting-time distribution. This is a difficult task for at least three reasons. The first is that the process of waiting-times belongs, from a theoretical perspective, to a quite general family and methods for Bayesian statistics in such generality are not known yet. Secondly, the waiting-times are not even observed directly which makes the task of Bayesian inference for such a general process even harder since a hypothetical prior distribution would have to be updated to the posterior by merely observing the arrival and service process. The third reason is that the service time distribution needs to be modeled in an infinite dimensional parameter space since it is not assumed to be of a certain form that can be parametrized in a finite manner. Conti solved this using the celebrated Dirichlet process as a nonparametric prior for the service time distribution and exploiting a functional relationship between inter-arrival, service and waiting times. Thereby, he obtained Bayesian estimators for a suitable functional of the service time distribution and showed its goodness in terms of posterior consistency and a Bernstein-von Mises type result.

The aim of this thesis is to extend these ideas to systems in continuous-time. It splits in three parts. The first part (chapter 3) deals with Bayesian statistics for the $M/G/1$ under the same observational assumptions as in Conti's paper. That means, arrival- and service-times are assumed to be observables and the wish is to make Bayesian inference for the waiting-time distribution and the distribution of the size of the system. Therefor a class of prior distributions is employed which contains the Dirichlet process as a special case, i.e. the neutral-to-right priors. Exploiting a well known functional relationship of the transforms of the distributions of the service and inter-arrival times and the system size, a Bayesian estimator for the latter is obtained as well as posterior consistency and posterior normality. The posterior consistency of the estimator for the particular characteristics requires the posterior consistency of the random expected value of the service time distribution. Since this is a functional of a random distribution function which is not supported by a compact set in general, this is a theoretically deep question. However, an affirmative answer in form of a new result is given under relatively mild conditions.

The second part (chapter 4) of the thesis deals again with the $M/G/1$ queueing system but under a distinct observational setup. Thereby, the entire queueing system is assumed to be a complete black box such that the only thing that can be observed is the departure stream of customers. The interest lies in making inference for the arrival stream and the general service time distribution, respectively. So, the departure stream has to carry information for both of these targets. Since, assuming the system in steady state, the arrival stream and the departure stream are of the same stochastic nature, inference for the arrival stream can be made rather directly by observations of the inter-departures. However, at the same time, the assumption of the system being in steady state disables one to make inference for the service time distribution by merely observing inter-departure times. Hence, an additional observation is introduced, namely the number of customers a departing customer leaves behind in the queue. The subsequent inference procedure for the service time distribution is based on the fact that the marks of the observation process form a Markov chain, the so called embedded Markov chain whose stochastic matrix is

of a particular delta shape. Dealing with this from a subjectivistic viewpoint, a symmetry condition for Markov measures being governed by those stochastic matrices is found that lies in between of exchangeability and partial exchangeability [c.f. Diaconis and Freedman (1980)]. Measures possessing that symmetry are shown to be summarized by a statistic which, in turn, is shown to have S-structure. S-structure is a certain property of statistics that, roughly speaking, ensures their stability with respect to extension of the data. By a classical result of Freedman (1962) one can think of stationary measures being summarized by the particular statistic as mixtures of shift-ergodic measures being summarized by the statistic. The mixing measure is then modeled by exploiting the special structure of the transition matrix. More precisely, it is modeled using a family of Dirichlet processes which basically consists of shifted versions of one particular Dirichlet process. Based on the observation of the marked departure process and theoretical considerations gone ahead, inference procedures for the general service time distribution are obtained as well as results justifying their usage in terms of posterior consistency and posterior normality.

The third part (chapter 5) of the thesis is written in the light of the $M/G/\infty$ system. An infinite amount of servers represents the idea that every customer gets her own service station when arriving at the system. Thus, there is no queue building up at all. However, since the $M/G/\infty$ model is often used for black box type systems, the interest is again in inferring the service time distribution on basis of rather indirect observations. To put it more precisely, the observations are the instants customers arrive to and depart from the system, respectively. However, in many situations it can not be tracked which instants belong to a certain customer. A classical example is that of a motorway with a sensor detecting arriving cars at its ramps and exits. A work of Brown (1970) dealt with the issue of making inference for this setup from a frequentist point of view. Brown discovered that the sequence of differences, i.e. the difference between some departure and the nearest instant of some arrival right before that departure is a stationary and ergodic process. Exploiting the ergodic theorem he obtained an estimator for the cumulative distribution function (c.d.f.) of the sequence of differences. Subsequently, he showed that the c.d.f. of the sequence of differences has a direct link to the c.d.f. of the service time distribution, a fact that allows to transfer inference procedures to the latter. But according to Brown

> [$\cdots$] *we have obtained an estimator for the c.d.f. G,* [*however*] *it is clearly not the best estimator in any sense because we do not use all the information. The problem of finding a best estimator (according to any criterion) is still open.*
>
> -Comments and additions from Brown (1970)-

However, Brown did not further specify what he meant by *all the information*, that might improve inference. It turns out in the present thesis, that exactly this question comes up right at the beginning rather than at the end of a Bayesian's analysis of the $M/G/\infty$ system in the setup sketched above. Since Brown's estimators are based one some empirical version of the c.d.f., this suggests itself that the omitted information is the interplay of the data from the sequence of differences. Regarding the issue from a subjectivistic viewpoint, the difficulty is that one has to express her uncertainty in the shift-ergodic

measures. While Choquet theory guarantees the existence of a proper mixing measure, finding a suitable model for it is a rather hard task. This is mainly due to the overwhelming size of the space of shift ergodic measures and the resulting lack of a suitable parametrization which in turn would enable one to execute the integration with respect to the mixing measure. Since the underlying state space of the stationary data has direct influence on this size, the state space is assumed to be a finite set. From an applied perspective this would mean that one merely measures the length of the elements of the sequence of differences by finitely categorizing the positive real axis, an assumption which doesn't appear too bad to the practitioner. However, from a theoretical viewpoint this assumption has great advantage since it enables the theorist to exploit the theory of Markov chains in order to create a reasonable parameter space. Since this parameter space has to encapsulate ergodic measures of any dependence range, it is not supposed to be actually graspable. As in the case of e.g. fractals, the entire object can not be described in full extent. Instead it is described by an infinite sequence of finite dimensional objects which are belted by some rule of their reciprocal behavior. The finite dimensional objects are taken to be decent multi-dimensional generalizations of stochastic matrices, which I call stochastic tensors. The rule which sticks those tensors together is a family of projections that reduce the dimension of the tensors in a certain way. The family of tensors and the family of mappings are then used to define the parameter space in terms of the inverse limit of the system consisting of these both families. Having clarified the parameter space of the shift-ergodic measures the next step in the direction of Bayesian statistics for stationary data is to model the mixing measure in a useful manner. That is, it should have big support in order to express one's à priori knowledge, it should be mathematically feasible and it should allow for an update mechanism that provides the posterior law in an analytically closed way. This is done by several assumptions on the sampling scheme data is generated from. One such assumption is that the range of dependence is finite but by no means bounded. Another is that, given the order of dependence, the stochastic tensor of the dimension given by this order is sampled due to a certain independence constraint. This means that sectors of the tensor which encode the transition probabilities with respect to different predecessor states are sampled from independent Dirichlet distributions. This specific sampling scheme allows to define a posterior measure in closed form, i.e. to update the distributions appearing in the sampling scheme by observed data. While updating the tensors themselves is a minor problem, updating the distribution on the order of dependence is more sophisticated. Although I conjecture that the defined posterior will center around the true data generating measure when the data size increases, I was unable to obtain an appropriate posterior consistency result yet. Fully answering this question is future work that might require further topological and geometrical arguments. However, there is hope for an affirmative answer since the work of Lijoi et al. (2007) gives a non-constructive Doob-type consistency result for stationary data but without specifying a certain model as it is done in the present thesis. Throughout chapter 5 several examples are given in order to strengthen the grasp of the reader with respect to the core of Bayesian statistics for stationary data.

The present thesis is organized in several chapters. In chapter 2 the theory of continuous-time queueing models is recalled in order to provide the theoretical background which is necessary for the statistical evaluation of those models in the subsequent chapters. Chapter 3 is devoted to the $M/G/1$ model. Thereby the arrival stream as well as the service process are assumed to be observable. The main task is to develop inference for the distributions of the waiting times of the customers and the occupation of the system. The

underlying queueing model in chapter 4 is again the $M/G/1$ system. But the observational setup changes in so far that the system is assumed to be a black box and the only accessible data stems from the departure process. In this framework theoretical examinations from the subjectivist perspective are given which lead to nonparametric Bayesian inference for the non-observable service time distribution G. Chapter 5 departs from the two previous chapters by regarding the $M/G/\infty$ model. Again the system is blackened and the data is assumed to consist of instants of arrivals and departures, respectively. The aim is to make Bayesian inference for G. Since the raw data forms a general stationary process, this aim is accompanied with further deep theoretical considerations. The last chapter provides the conclusions of the thesis as well as outlook to future work. Latter is given with particular respect to chapter 4. So, coarse ideas are given with respect to the possibilities of extending the theory developed in chapter 5 into more general settings.

2 Continuous-Time Queueing Models

Queueing theory is a branch of probability theory which deals with the analysis of stochastic processes formalizing dynamically fluctuating systems of *customers*. Thereby, the term customers can have manifold meanings depending on the situation to be modeled. One can think of actual customers queueing up at a checkout, of information to be transmitted, of jobs needed to be done by a CPU, of items storing in a warehouse, of cars or trains using a joint trail, air planes waiting for a clearance to land or for take off, respectively, and many others. A queuing system mostly consists of a unit serving the customers and a waiting room where customers which can not be processed immediately are stored.

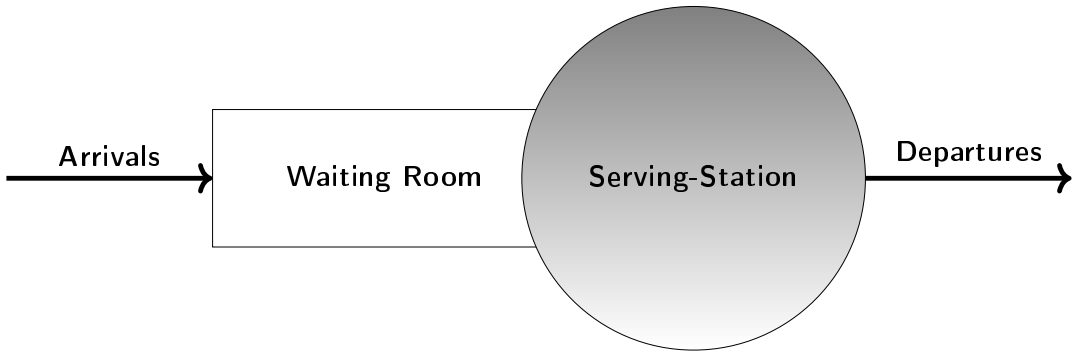


Figure 2.1: Sketch of a queueing system

The first categorization that appears is in the time the system evolves in, i.e. whether it evolves in discrete or in continuous time. Discrete-time systems are commonly used to model situations in which the server handles jobs of fixed size such as e.g. a robot executing some job at a production street or an information system transmitting cells of fixed size. In contrast, continuous-time models are used whenever the server works off jobs which can have arbitrary length. The main difference is that the randomness of the service times in discrete-time systems is encoded in the batch size of the jobs rather than in the time a customer occupies the server. Hence, in discrete-time systems, the service time is most often already determined when the customer enters the system rather than at the beginning of the service, or even during it is executed, respectively.

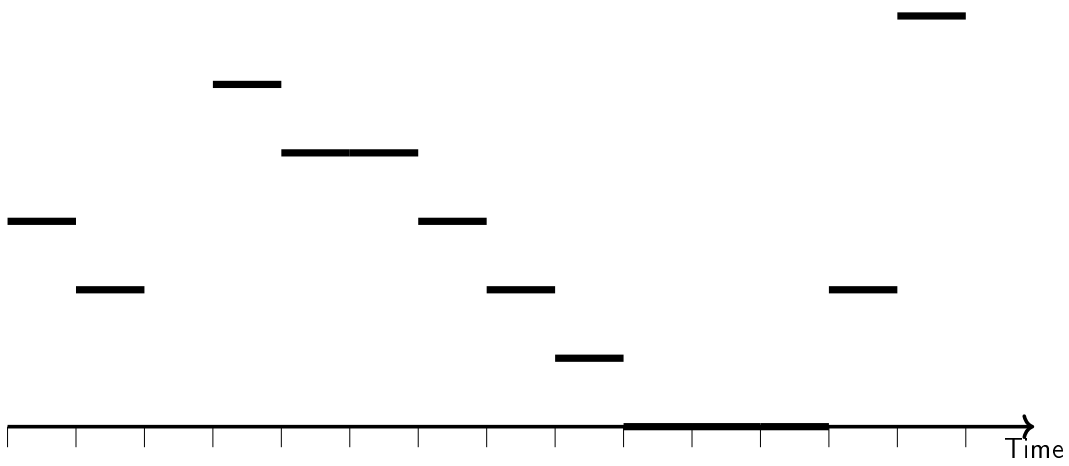


Figure 2.2: Occupation process of a discrete-time queueing system

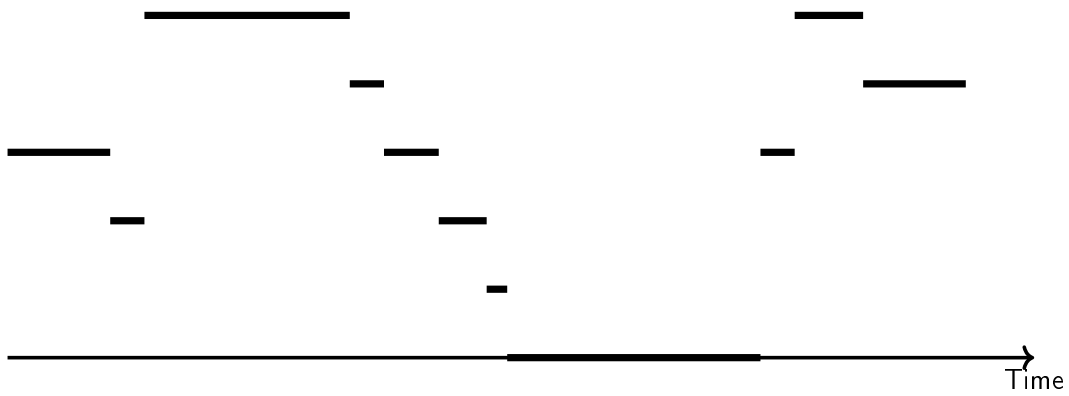


Figure 2.3: Occupation process of a continuous-time queueing system

Once it is clarified whether the system is modeled in discrete or continuous time, the system can be further described. Basically, this is done by the arrival stream of customers and their associated service times. But also factors like the numbers of parallel servers, the size of the the waiting room or the policy of the queue comes into play. Due to the wide-ranging possibilities of setting up such a system, Kendall (1953) introduced a clear-cut nomenclature for them. Following an extended version of this, a queueing system is uniquely determined by a certain tuple commonly denoted by $A/S/c/K/N/D$. Thereby, A describes the arrival process, S the process of service times, c the number of serving stations, K the capacity of the waiting room, N the number of the population of customers and D the queueing policy. Prominent arrival streams are given by $A = Geo$ for the case of discrete-time and $A = M$ for continuous-time, respectively. Thereby, Geo denotes inter-arrival times governed by a geometric distribution while M indicates exponentially distributed inter-arrival times. M stands for *Markovian* in order to depict the memorylessness property of the exponential distribution which the geometric does possess as well. To indicate independent inter-arrival times governed by a general distribution it is common to use GI for discrete-time and just G for continuous-time systems. The notation for the service times is quite similar. The values for c, K and N are self-explanatory. D denotes the queueing policy, i.e. the order customers are processed by the server. Frequently used policies are first-in-first-out (FIFO) which depicts the classical idea of a queue in a supermarket, first-in-last-out (FILO) which is often used in storage

issues, processor-sharing (PS) where the customers receive the same share of the processor capacity, round-robin where customers are served in a simultaneously but alternating manner and many others as e.g. shortest-job-first or priority policies.

The scope of this work is the statistical analysis of the continuous-time $M/G/1/\infty/\infty/\mathsf{FIFO}$ and the $M/G/\infty/\infty/\infty/\mathsf{FIFO}$ (or in abbreviation just $M/G/1$ and $M/G/\infty$) queueing models from a Bayesian perspective and under different schemes of observations. Thereby, the first two parts are devoted to the $M/G/1$ model while the third part deals with the $M/G/\infty$ queue. Studying the $M/G/1$ model from a statistical viewpoint, different objects of interest as well as different levels of difficulty arise depending on which data can actually be observed. While in the first part it is assumed that the scientist has access to the system, i.e. the arrival stream and service process are observable, the basic assumption of the second part is contrary. In the second part, the system is assumed to be a black box such that only the departure stream provides observable data. The last part is about the $M/G/\infty$ model, which models a complete different situation, that is no queue will build up since each customer is assigned her own server. Again, the system is assumed non-accessible and the data merely consists of the instants customers enter and leave the system, respectively. Since this is fairly vague yet, the analysis of the $M/G/\infty$ model will lead to a theoretically deeper study of the foundations of Bayesian statistics.

Before embarking on the main task of this work, namely building up Bayesian models that fit the respective situations best and subsequently analyzing them mathematically, a brief introduction to the $M/G/1$ and $M/G/\infty$ model, respectively, will be given in order to provide the reader with the most important information needed for the subsequent chapters.

2.1 The $M/G/1$ Model

The present subsection is devoted to the $M/G/1$ queueing model. The model is chosen, in Kendall's notation [Kendall (1953)], to be the $M/G/1/\infty/\infty/\mathsf{FIFO}$ system. So, indistinguishable customers arrive consecutively according to a homogeneous Poisson process to the system giving rise to exponentially distributed inter-arrival times $(A_n)_{n\in\mathbb{Z}}$. Note that most of the appreciable properties of the $M/G/1$ system are due to this assumption, see below. Moreover, customers requiring service are served according to a general service time distribution G concentrated on $\mathbb{R}_+$. The service times are assumed to form a sequence of i.i.d. random variables $(S_n)_{n\in\mathbb{Z}}$. The consecutive services are accomplished by one reliable service station in a first-in-first-out manner. Customers who cannot be served immediately are stored in an infinitely large waiting-room and form a queue.

A main characteristic of the $M/G/1$ queue is the traffic coefficient ρ, which serves as an indicator for the load of the system. It is defined as the quotient of the mean service time and the mean inter-arrival time, in symbols $\rho := \frac{\mathbb{E}[S]}{\mathbb{E}[A]}$. Due to exponentially distributed inter-arrival times, this amounts for $M/G/1$ to $\rho = \lambda\mathbb{E}[S] := \lambda\mu$, where λ denotes the arrival intensity and μ the mean service time. The coefficient ρ plays an important role

since it indicates whether the system is stable, i.e. whether the queue attains a stochastic steady state after it has run an infinitely long time (c.f. the definition of the stationary distributions of Markov chains) in the sense that its occupation fluctuates and can even attain arbitrarily large values but it will be cleared in finite time with probability one. Thereby the stable case corresponds to $\rho < 1$, while for $\rho > 1$ the content of the queue will explode almost surely. Analogously as in the theory of power series in complex analysis, $\rho = 1$ is a borderline case which deserves further examination. However, one is usually interested in the stable behavior of the system or the way to it, respectively. The characteristic ρ can be shown to have another nice interpretation; it equals the probability that the system is empty. To see this, it is most efficient to consider the "inner structure" of the system. However, the queue length process cannot be argued to be a Markov process in general. This holds only in introductory cases as $M/M/1$. Generally, it is only a so called semi-Markov process. The main reason is that the queue length process strongly depends on the entire past through the residual service time of the customer just in service. A way out is to take into account this additional information by conditioning on it. The most straight forward way to do so in a practically workable manner is to consider the process of the queue lengths at instances right before customers depart from the system because then the residual service time of the present customer is just known to be naught. This approach yields the so called embedded Markov chain. It is regarded as a discrete-time process being embedded into the continuous-time framework which is governed by a Markov measure of a particular structure. This structure in turn reveals itself in form of the stochastic transition matrix which determines the Markov law. In order to depict it, it is given as

$$\begin{pmatrix} a_0 & a_1 & a_2 & a_3 & a_4 & a_5 & \cdots \\ a_0 & a_1 & a_2 & a_3 & a_4 & a_5 & \cdots \\ 0 & a_0 & a_1 & a_2 & a_3 & a_4 & \cdots \\ 0 & 0 & a_0 & a_1 & a_2 & a_3 & \cdots \\ 0 & 0 & 0 & a_0 & a_1 & a_2 & \cdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix},$$

where a_j denotes the probability that j customers enter the system during a service time. Briefly reflecting the traffic coefficient one can roughly say that the necessary condition for the embedded Markov chain to be "ergodic" (in sense of the theory of Markov chains) is given by $\rho < 1$.

At this point, from a rather statistical point of view, the question may arise whether the distribution of the queue length process observed at departure times is the same as at arbitrary time points. This question has an affirmative answer when the system is stable and the system's input is a homogenous Poisson process, since then the PASTA property and the level-crossing laws appear to hold. **PASTA**, abbreviating *Poisson-arrivals-see-time-averages*, means that the probability to find the stable system in a certain state at an arbitrary time point is equal to the probability that some arriving customer finds the system in this particular state without counting himself. Note that PASTA strongly depends on the lack of anticipation assumption (LAA) of the arrival stream. Roughly speaking, LAA means that future increments of the arrival process are not affected by

the present state $s \in \mathbb{N}_0$ of the queueing system. This assumption is clearly met by the Poisson process governing the arrivals of customers. For more details on LAA and PASTA see Wolff (1982). Thus, from a statistical point of view, if one is interested in making inference for some characteristic of the queueing system $M/G/1$ based on consecutive observations, there is no difference between collecting respective data at instances of arrivals or arbitrarily chosen instances. The second property of an stable M/G/1 system is the so called **level-crossing law** (LC). LC says that the limiting fraction of arriving customers seeing $s \in \mathbb{N}_0$ customers in front equals the limiting fraction of departing customers leaving s customers behind. See Brill (2008) for an exhaustive treatment of level-crossing laws. Thus, in steady state of the queue, we conclude by PASTA and LC that the system size at arbitrary instances equals that at departure times of customers. This will play an essential role in the statistical treatment of the second part. The third property briefly reviewed is perhaps even more unexpected on a first glance. It says that the departure process, i.e. the continuous time stochastic process given by the random departure time points, in equilibrium is a Poisson process with the same intensity rate as the arrival process. This relies on the fact that any stable (in the literature in abuse of notation often called "ergodic") birth-death process is **time-reversible**. Thus, the reversed system evolves as the original $M/G/1$ which in turn implies a Poisson process for the departure stream with the same rate as the arrival process, see e.g. Asmussen (2008, page 115). However, the latter property affects the statistical viewpoint on the system. Whenever the system has reached equilibrium, solely observing the departure stream is completely non-informative with respect to the unknown service time distribution. Hence, in the second part of the thesis additional observations will be required which provide enough information about the services in order to infer on the service time distribution.

For the sake of clarity of what follows, some notation comes in useful. Thereby, it will be assumed that all random variables are defined on a common underlying probability space $(\Omega, \mathcal{A}, \mathbb{P})$.

- $Q_t \equiv$ queue length at time $t \in \mathbb{R}$
- $N_t \equiv$ number of customers in the entire system at time $t \in \mathbb{R}$, i.e. $N_t = Q_t + 1$ in case $Q_t > 0$ for $t \in \mathbb{R}$
- $W_n \equiv$ waiting time of customer $n \in \mathbb{Z}$
- $D_n \equiv$ sojourn time in the system of customer $n \in \mathbb{Z}$
- $A_n \equiv$ time between arrival of n-th and $(n-1)$-st customer $n \in \mathbb{Z}$
- $S_n \equiv$ service time of customer $n \in \mathbb{Z}$
- $T_n \equiv$ time point of the departure of the n-th customer
- $D_n = T_{n+1} - T_n \equiv$ time between departure of n-th and the $(n+1)$-th customer
- $A_{S_n} \equiv$ number of customers entering the system during the service of customer $n \in \mathbb{Z}$.

Then, the following statements are well known facts in queueing theory.

$$\mathbb{E}[N] = \rho + \frac{\rho^2 + \lambda \mathsf{Var}[S]}{2(1-\rho)}.$$

This formula is known as the Pollaczek-Khinchin mean formula and relates the mean of the random variable N to that of the random variables A, S and to the variance of S, $\mathsf{Var}[S]$. However, this formula only relates parameters of aforementioned distributions to each other instead of the entire distributions. However, in Bayesian statistics one is typically interested in the entire distribution of a parameter one is uncertain in. A functional relationship is given by the Pollaczek-Khinchine transform formula. To state it, define the following objects. Let $n(z) = \sum_{k=0}^{\infty} z^k \mathbb{P}(N = k)$ be the probability generating function (p.g.f.) of the distribution of N and $g(z) = \int_0^\infty e^{-zs} dG(s)$ the Laplace-Stieltjes transform (LST) of the service time distribution G. Then the following functional relationship can be obtained, which relates the distributions of the system size, the queue length and the waiting times, respectively, to the distribution of the inter-arrivals and the service times.

$$n(z) = g(\lambda(1-z)) \frac{(1-z)(1-\rho)}{g(\lambda(1-z)) - z}, \qquad z \in [0,1].$$

Manipulations of this formula give

$$q(z) = \frac{(1-z)(1-\rho)}{g(\lambda(1-z)) - z}, \qquad z \in [0,1], \qquad \text{and} \qquad w(z) = \frac{z(1-\rho)}{z - \lambda(1-g(z))}, \qquad z \in \mathbb{R}_+,$$

where $q(z)$ denotes the p.g.f. of the queue-length distribution and $w(z)$ the LST of the waiting-time distribution. The latter two quantities are of special statistical interest in chapter 3 since they provide the essential information about the development of the queue.

Moreover, by time-reversibility, one has $\mathcal{L}[A] = \mathcal{L}[D] = \mathcal{E}[\lambda]$, where $\mathcal{L}$ stands for the law of a random quantity and $\mathcal{E}(\lambda)$ for the exponential distribution with rate $\lambda > 0$.

Turning to the inner structure of the system, recall that the process $\{N_t\}_{t\in\mathbb{R}}$ was said to be a semi-Markov Process in general. A mathematical rigorous definition of that is given now. Therefor, let $\{N(t)\}_{t\in\mathbb{R}}$ denote the stochastic process with state space $\mathbb{N}_0$ that describes the number of customers in the system at time t. It is plain that in general $N(t)$ is not a Markov process. The only situation in which $N(t)$ can be considered Markovian is when $G = \mathcal{E}$, which is due to the memorylessness property of the exponential distribution. However, a "sub-process" of N can be found that is a Markov chain and thus makes N a semi-Markov process. Call $N(t)$ a semi-Markov process with state space $\mathbb{N}_0$ according to the following construction. Assume there is a stochastic kernel $\kappa : \mathbb{N}_0 \times (\mathfrak{B}_{\mathbb{N}_0} \otimes \mathfrak{B}_{\mathbb{R}}) \to [0,1]$ and a discrete-time stochastic process $(\tau_n)_{n\in\mathbb{Z}}$ such that $\mathbb{P}$-a.s. $\tau_n < \tau_{n+1}$ and $N(t) = c_{n+1}$, for all $t \in [\tau_n, \tau_{n+1})$. Define a two-component discrete-time stochastic process Y by $Y_n := (c_n, \tau_n - \tau_{n-1})$. Then, $N(t)$ is a semi-Markov process with kernel κ if $\mathbb{P}(Y_{n+1} = y_{n+1} | Y_i; i \le n) = \mathbb{P}(Y_{n+1} = y_{n+1} | Y_n) = \kappa(c_n, y_{n+1})$. Although it will hardly affect the present thesis, it is worth mentioning that Epifani et al. (2002) deal with semi-Markov processes from a subjectivist viewpoint by obtaining de Finetti-style theorems exploiting among others theory from Diaconis and Freedman (1980). These results give rise to think about mixtures of laws governing a semi-Markov process as laws fulfill-

ing a certain kind of symmetry or invariance condition, respectively. They are useful for Bayesian statistics since they form its basis by providing the existence of a prior measure.

In order to move on, notice that the definition of a semi-Markov process fits well the process N considered here since the system size is constant on intervals where no departure and no arrival occurs. The holding times τ_n then are seen to be the minimum of the residual service times of the customer currently being served (if there is some) and the time elapsing until the next arrival of a new customer. Although, due to the memorylessness property of the exponential distribution, the time until the next arrival is an exponential distribution, the description of the entire holding time distribution is difficult. However, it should become easier if we omit to take into account the residual service time. That means we observe the system at instances at which a customer departs from the system, T_n. Indeed, it turns out that the process $\{N(T_n), T_n\}_{n\in\mathbb{Z}}$ is a Markov chain. To see this, note that

$$\begin{aligned} N(T_{n+1}) &= N(T_n) + A_{S_{n+1}} - \delta_{N(T_n)}(\{0\}), \\ T_{n+1} &= [1 - \delta_{N(T_n)}(\{0\})](T_n + S_{n+1}) + \delta_{N(T_n)}(\{0\})(T_n + S_{n+1} + I_{n+1}), \end{aligned}$$

where $I_{n+1} \sim \mathcal{E}(\lambda)$ reflects the remainder of the inter-arrival time between the n-th and $(n+1)$-st customer, which is independent of $\{(N(T_i), T_i) : i < n\}$ as well as S_{n+1} is by assumption. If one focuses solely on the Markov chain $\{N(T_n)\}_n$, above equation gives rise to the stochastic matrix governing the chain $N := \{N(T_n)\}_n$. Let $M = (m_{ij})_{i,j\in\mathbb{N}_0} \in \mathfrak{S} \subset M(\infty, \mathbb{R})$ denote the infinite matrix consisting of all probabilities of transitions of the chain N from state i to state j, where $\mathfrak{S}$ denotes the space of all infinite stochastic matrices. Then, the probability of having a transition from i to j is given through

$$m_{ij} = \mathbb{P}\left(A_S = j - i + [1 - \delta_i(\{0\})]\right),$$

which yields the following form of M, c.f. page 9.

$$M = \begin{pmatrix} a_0 & a_1 & a_2 & a_3 & a_4 & a_5 & \cdots \\ a_0 & a_1 & a_2 & a_3 & a_4 & a_5 & \cdots \\ 0 & a_0 & a_1 & a_2 & a_3 & a_4 & \cdots \\ 0 & 0 & a_0 & a_1 & a_2 & a_3 & \cdots \\ 0 & 0 & 0 & a_0 & a_1 & a_2 & \cdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix},$$

where for $i \in \mathbb{N}_0$ $a_i = \mathbb{P}(A_S = i)$.

Note that this stochastic matrix is a member of a larger family of stochastic matrices, so called Δ-matrices which were introduced by Abolnikov and Dukhovny (1991). To be more precise, M is a positive homogenous $\Delta_{1,1}$, where the first 1 indicates that $p_{ij} = 0$ for all $i - j > 1$, the second that this holds for all rows $i > 1$, homogeneity that $m_{ij} = a_{j-i+1}$ for all $i > 1$ and positivity that $a_{j-i+1} > 0$. Notice that positivity is given since the number of customers who enter the system during a service time is not bounded, even if its probability might decay at a fast rate. Abolnikov and Dukhovny (1991) further study

the "ergodicity" of Markov chains governed by some Δ-matrix and obtain necessary and sufficient conditions for that in terms of the p.g.f. of the discrete distribution $(a_i)_{i\in\mathbb{N}_0}$. For the here considered stochastic matrix M, it is easy to see that it is irreducible and aperiodic since

$$M^k = \begin{pmatrix} x & x & x & x & x & x & \cdots \\ x & x & x & x & x & x & \cdots \\ x & x & x & x & x & x & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots \\ x & x & x & x & x & x & \cdots \\ 0 & x & x & x & x & x & \cdots \\ 0 & 0 & x & x & x & x & \cdots \\ 0 & 0 & 0 & x & x & x & \cdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix} \left.\begin{matrix} \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{matrix}\right\} (k-1)\text{-times}$$

where an entry x means an entry strictly greater than zero. Moreover, Theorem 3.4. in Abolnikov and Dukhovny (1991) shows that the Markov chain N governed by above M is positive recurrent if and only if $a'(1) < 1$. Since $a'(1) = \lambda\mathbb{E}[S] = \mathbb{E}[S]/\mathbb{E}[I] = \rho$, see the explicite form of $a(z)$ below, this requirement becomes the common condition in queueing theory for stability of the $M/G/1$ system. It is well known from the theory of Markov chains that for M of above form with $a'(1) < 1$ there exists a unique M-invariant distribution $p \in \mathcal{P}(\mathbb{N}_0)$, i.e. $p \in c_0^{(+)}$ where $c_0^{(+)}$ denotes the space of null sequences with all projections being strictly positive and such that $p = pM$, meaning $p_j = \sum_{i=0}^{\infty} p_i m_{ij}$. Letting $\pi(\cdot)$ denote the p.g.f. of p, a result by Harris (1967) yields

$$\pi(z) = a(z)\frac{(1-z)(1-a'(1))}{a(z)-z}.$$

This formula relates the p.g.f. of the stationary distribution to that of the distribution of A_S in an explicit form. Put another way, the mapping $M \mapsto p$ that maps the stochastic matrix onto its associated invariant distribution can be explicitly obtained here. This strongly depends on the particular shape of the stochastic matrix and in general this mapping can not be given explicitly even if it is known from probability theory that it is well defined and injective.

Manipulating the distribution of A_S or its p.g.f., respectively, further functional relationships can be obtained.

$$\mathbb{P}(A_S = k) = \int_0^\infty \mathbb{P}(A_S = k|S \le t)G(dt) = \frac{1}{k!}\int_0^\infty e^{-\lambda t}(\lambda t)^k G(dt),$$

which in turn yields

$$a(z) := \sum_{k=0}^{\infty} z^k \mathbb{P}(A_S = k) = \int_0^{\infty} e^{-\lambda t} \left[\sum_{k=0}^{\infty} \frac{(\lambda t)^k}{k!} \right] G(dt) = \int_0^{\infty} e^{-\lambda[1-z]t} G(dt) =: g(\lambda[1-z]),$$

using the monotone convergence theorem. Hence, a functional of the service time distribution G, $g(\cdot)$ is obtained in terms of the distributions of A_S and D . This functional relationship will be the starting point for the inferential analysis in part 2 (chapter 4).

2.2 The $M/G/\infty$ Model

The third part (chapter 5) of the Bayesian statistical analysis of continuous-time queueing systems will be devoted to the $M/G/\infty$ model. The considered observation scheme will be as in Brown (1970), i.e. only the instants of arrivals and departures of customers, respectively, are recorded. Due to that observation scheme foundational analysis of the $M/G/\infty$ queueing model will play a rather minor role. Nevertheless, for the sake of completion, the foundations of $M/G/\infty$ are briefly reviewed. The $M/G/\infty/\bullet/\infty/\bullet$ system can be seen as a generalization of the $M/M/c/\infty/\infty/\mathsf{FIFO}$ queue for two reasons. Firstly the service time distribution G generalizes the exponential distribution in $M/M/c$ and secondly $M/M/\infty$ can be regarded as emerging from $M/M/c$ by taking the limit $c \to \infty$. For that reason, it is natural to start the study with the $M/M/\infty$ system. The $M/M/\infty$ system can be described as a birth-death process using a so called transition intensity matrix Q [c.f. Asmussen (2008, section III.2 ff.)] which indicates the instantaneous rates of change of the content of the system. Letting λ be the parameter of the Poisson arrival process and $1/\mu$ the mean of the exponential service time distribution, Q is given through

$$Q = \begin{pmatrix} -\lambda & \lambda & 0 & 0 & 0 & 0 & \cdots \\ \mu & -(\lambda+\mu) & \lambda & 0 & 0 & 0 & \cdots \\ 0 & 2\mu & -(\lambda+2\mu) & \lambda & 0 & 0 & \cdots \\ 0 & 0 & 3\mu & -(\lambda+3\mu) & \lambda & 0 & \cdots \\ 0 & 0 & 0 & 4\mu & -(\lambda+4\mu) & \lambda & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

Based on the theory of birth-death processes several characteristics of the $M/M/\infty$ system can be developed. However, since no queue builds up at all, different characteristics as for the $M/G/1$ system are of interest. These are for instance the transient behavior of the system as well as the stationary occupation probability. But also characteristics like the busy-time period, or more generally the congestion-time period which amounts to the time the system spends in states exceeding a certain fixed number, and the distribution of the maximum of the system's occupation over a finite time horizon can be obtained. Again, $\rho = \frac{\lambda}{\mu} < 1$ is necessary and sufficient for the system to reach equilibrium. The transient analysis for $M/M/\infty$ is e.g. given in Gross et al. (2008, section 2.11). The

probability that the occupation at time $t > 0$ equals a non-negative integer n, given that the system was empty at $t = 0$ is given by

$$p_{0,n}^{t} = \frac{1}{n!} \exp\left[-\rho\left(1 - e^{-t\mu}\right)\right]\left[\rho\left(1 - e^{-t\mu}\right)\right]^{n},$$

which yields the mean

$$\mathbb{E}\left[N_t | N_0 = 0\right] = \rho\left(1 - e^{-t\mu}\right).$$

Furthermore, taking the limit $t \to \infty$ yields the stationary distribution of the occupation of the system which equals a Poisson distribution with intensity ρ, i.e.

$$p_{0\overset{t}{\to}n} \overset{t\to\infty}{\longrightarrow} \pi(n),$$

where

$$\pi(n) = \frac{\rho^n e^{-\rho}}{n!}.$$

Moreover, Newell (1966) has shown that this probability is just the same for the $M/G/\infty$ system, where $\rho = \lambda\mathbb{E}[S]$. More generally, a transient analysis of $M/G/\infty$ can be found in Kulkarni (2009, chapter 8).

The mean congestion period for $h \in \mathbb{N}_0$ can be argued to equal

$$\mathbb{E}\left[N(t+h) - N(t) > h\right] = \lambda^{-1} \sum_{j=h+1}^{\infty} \frac{h!}{j!} \rho^{j-h},$$

c.f. Guillemin et al. (1996). Moreover, by Guillemin and Simonian (1995), the Laplace transform of the congestion distribution is shown to relate to Kummer's function in a certain way. Using $h = 0$ delivers the respective terms for the busy-time period.

The distribution of the maximum of the system's content over a finite time horizon is more complicated to examine. However, the study was given by Morrison et al. (1987) who obtained some relationship of the maximum distribution and Poisson-Charlier polynomials. For reasons of presentation, the details are omitted here and the interested reader is referred to Morrison et al. (1987).

Another interesting theoretical result for $M/M/\infty$ was given by Knessl and Yang (2001) who showed that some rescaled version of the process describing the occupation of the system converges to an Ornstein-Uhlenbeck process when the traffic increases. In symbols that means

$$\mathcal{L}\left[\rho^{-1/2}(N - \rho)\right] \overset{\rho\to\infty}{\longrightarrow} \mathcal{L}(Y),$$

where Y denotes a stochastic process evolving according to the SPDE $dY_t = -Y\,dt + \sqrt{2}dW_t$, where W is a standard Brownian motion. By that result, $M/M/\infty$ systems under heavy traffic can be approximated by processes which in turn can be studied from a Itō calculus perspective.

2.3 Further Reading

For supplementary material about queueing theory see e.g. Kleinrock (1976); Medhi (2002); Haigh (2004); Gross et al. (2008); Asmussen (2008); Kulkarni (2009); Nelson (2013) as well as references therein.

3 Inference for $M/G/1$ based on Observations of the Arrival Stream and the Service Process

3.1 Introduction

The present chapter is devoted to the problem of making Bayesian nonparametric inference for the $M/G/1$ system's characteristics as e.g. the customers waiting time distribution or the occupation of the system, respectively. It can be regarded as an extension of Conti (1999) to the continuous-time framework. The work of Conti (1999) was about the $Geo/G/1$ system which can be seen to be the discrete-time analog of the continuous-time $M/G/1$. The major issue in Conti (1999) as well as in this chapter is that the objects one wants to infer are stochastic processes that are of quite general appearance. More precisely, the waiting times of customers and the system size are both general stationary processes. It should be stressed that they do not possess a finite range of the dependency. A fact that raises the difficulty for statistics from a Bayesian perspective especially if the processes can be observed directly. See also chapter 5 of this thesis for more details on that. An additional difficulty emerges when the processes are even not observable directly and the only observations consist in the input of the system in form of the arrival stream of customers and the service process. This situation is dealt with in Conti's paper in a discrete-time framework and the ideas will be transferred to the continuous-time setting in this chapter.

Before embarking into the theory, Conti's paper is briefly reviewed. The $Geo/G/1$ queueing system is frequently used to model communication systems where information encoded in packages of fixed size is transmitted by a serving unit. The server is assumed to be able to transmit one package during one time slot. Hence, the randomness of the service times is rather given in form of random batch sizes, i.e. marks of the point process governing the arrival stream. Employing a Dirichlet process prior for the distribution of the magnitude of these marks, Conti obtains estimators for various characteristics of the queue exploiting a well known functional relationship [c.f. Grübel and Pitts (1992)] of the inter-arrival and the service time distribution with the waiting time distribution. Since the target of Conti's work was to infer the waiting time distribution, he took this rather indirect approach mainly because assigning a prior distribution to the waiting time distribution and updating it by data consisting of the marked arrival stream is an infeasible task. Subsequently, he obtained large sample properties for the estimators which establish their justification. One main result, which is always important in Bayesian statistics, was uniform posterior consistency which states that the posterior law centers around the

true data generating probability measure when the data size increases. Hence, posterior consistency ensures that the statistical procedure is not misleading. Another result consisted in a Bernstein-von Mises type result which forms a Bayesian analog to the central limit theorem and gives an idea how the centering of the posterior takes place. This can be interesting for an user applying the estimators since it enables one to sample from a distribution which is "close" to the posterior law whenever the sample size is "large". For the Bernstein-von Mises theorem results in Freedman (1963, section 4) were employed which clarify the behavior of the posterior in the discrete case and under the promises that the true distribution is supported by a finite set.

The aim of the present chapter is to take a similar way of Bayesian statistical inference for the continuous-time $M/G/1$ system which is the more appropriate model in many situations where time-continuity is more reasonable. For example one could think of customers arriving at a cashier in a supermarket, cars arriving at a traffic jam, goods arriving at a storing center and many others. Therefor a continuous-time analog for the functional relationship of observables and objects of interest is used which is known as the functional Pollaczek-Khinchine formula in honor of Felix Pollaczek and Aleksandr Khinchine who derived the steady state behaviour of the $M/G/1$ system in the 1930s, c.f. chapter 2. Thereby, the philosophy is analog to that of Conti (1999), meaning that the main target will be the nonparametric estimation of the waiting time distribution on basis of observations of the arrival stream and the service times. However, since continuous-time is alleged a richer class of prior distributions is used which allows to express prior knowledge more flexibly. The combination of this larger class of prior distributions together with the assumption of continuous-time leads to more intricate proofs of results related to the discrete-time analog. As usual, the Bayesian approach is appreciable if the prior knowledge is good and only few data are available. However, typically large data samples are accessible Therefore, large sample properties as posterior consistency and posterior normality are also investigated.

The chapter is organized the following way. Section 3.2 is devoted to the assignment of prior distributions to the random distributions governing the observables. Therefor, some facts of Bayesian statistics are reviewed. This brief survey includes the probabilistic background as well as the family of prior distributions which is used to model the general random c.d.f. of the service times. Furthermore, suitable estimators are defined on basis of the expressions recalled in the first chapter. Subsequently, frequentist validations of the suggested estimators is given in section 3.3 and section 3.4. Section 3.3 deals with the concentration of the posterior law of the random quantities in form of posterior consistency results. Since the estimators depend on the mean of the service time distribution through the traffic coefficient, posterior consistency of this mean needs to be shown which leads to a new result. In section 3.4 it is examined how the concentrations of the respective posterior laws take place. This typically leads to Bernstein-von Mises type results which are obtained for the present setup.

3.2 Prior Assignments and Estimators

All the transforms $n(z)$, $q(z)$ and $w(z)$ introduced in chapter 2 depend on the arrival rate λ and on the LST of the service time distribution $g(z)$. These values are typically assumed to be unknown and thus need to be inferred. For this inference, we choose a Bayesian approach which is further introduced in the present section. Since it does not make sense to a subjectivist statistician to talk about "fixed but unknown" parameters [see De Finetti (1974)],
another philosophical concept is used. That is one interprets the "fixed but unknown" parameter as a random quantity itself. This approach leads to the concept of exchangeability which provides a *meaningful, observable character* of the data. This approach is briefly reviewed. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be an abstract probability space. Call an infinite sequence of random variables $(X_i)_{i=1}^{\infty}$ with $X_i : \Omega \to \mathbb{R}, i \in \mathbb{N}$, exchangeable if for any $m \in \mathbb{N}$ and any permutation π of m elements it holds that

$$\mathcal{L}[X_1, \dots, X_m] = \mathcal{L}\left[X_{\pi(1)}, \dots, X_{\pi(m)}\right], \tag{E}$$

where $\mathcal{L}$ denotes the joint law of the respective random object.

Now, assume that we observe n inter-arrival times $A_1^n = (A_1, \dots, A_n)$ between the first $(n+1)$ consecutive customers as well as the service times $S_1^n = (S_1, \dots, S_n)$ of the first n customers. The data A_1^n and S_1^n are assumed to be the first n projections of two independent sequences of exchangeable random variables A_1^{∞} and S_1^{∞} which, in turn, are assumed to be independent of each other, i.e. $\mathcal{L}(S_i; i \in I | A_j; j \in J) = \mathcal{L}(S_i; i \in I)$ for all finite subsets $I, J \subset \mathbb{N}$.

We now turn to de Finetti's theorem for Polish spaces [Hewitt and Savage (1955)]. Let $\mathcal{P}(\mathcal{S})$ denote the space of all probability measures on some Polish space $\mathcal{S}$ and consider in particular $\mathcal{P}(\mathbb{R})$ equipped with the topology of weak convergence of measures. This leads to a measurable space $(\mathcal{P}(\mathbb{R}), \mathfrak{B}_{\mathcal{P}(\mathbb{R})})$ which is itself Polish [e.g. Kechris (1995)]. By the de Finetti theorem for exchangeables, it holds for all $n \in \mathbb{N}$ and for all measurable subsets $A_i \subset \mathbb{R}$, $i = 1, \dots, n$ that $X_1^{\infty} = (X_1, X_2 \dots)$ is exchangeable if and only if there is a unique mixing measure $\nu \in \mathcal{P}(\mathcal{P}(\mathbb{R}))$ such that

$$\mathbb{P}(X_i \in A_i; i = 1, \dots, n) = \int_{\mathcal{P}(\mathbb{R})} \prod_{i=1}^{n} P(X_i \in A_i) \nu(dP).$$

The right-hand side reflects the equivalent property of the sequence X_1^{∞} being exchangeable the following way. The data X_1^{∞} are conditionally i.i.d. given some probability measure $P \in \mathcal{P}(\mathbb{R})$, in symbols write $X_1^{\infty} | P \sim \bigotimes_{\mathbb{N}} P$. The probability measure P itself is random and distributed according to the mixing measure ν which is called the prior distribution from a Bayesian statistical point of view. Moreover, notice that unconditionally the data X_1^{∞} are in general not independent. Indeed, from a result from Kingman (1978), it is seen that in general exchangeable data are positively correlated, a fact that makes Bayesian statistics a theory of statistical prediction and thus convenient for other theories as e.g. machine learning.

For actual applications, however, the prior ν and the integration in de Finetti's theorem becomes infeasible in general. This is mainly due to the fact that one has merely a nonconstructive proof of the existence of ν. Another difficulty is that the set $\mathcal{P}(\mathbb{R})$ is quite large. One has mainly two ways to choose to circumvent this problem in applications. The first is to shrink the support of ν to a reasonable subset of $\mathcal{P}(\mathbb{R})$ by an additional *symmetry constraint* on the sequence of random variables, thus, to put additional information on the data, to boil down above integration to a finite dimensional parameter space. The second is to model ν nonparametrically by sophisticated probabilistic tools. We use the first approach here to get a prior distribution for the random arrival rate $\lambda : \Omega \to \mathbb{R}_+$ and the second to get a prior distribution for the random service time distribution $G : \Omega \to \mathcal{P}(\mathbb{R}_+)$.

3.2.1 Arrival Rate

In case of the inter-arrivals an additional symmetry assumption on the law of the infinite exchangeable sequence A_1^∞ will give rise to a mixing measure which is supported by all exponential distributions. This additional symmetry condition is stated as follows. Let $n \in \mathbb{N}$ and for $i = 1, ..., n$, $B_i \in \mathfrak{B}_{\mathbb{R}_+}$, the Borel sigma-field on $\mathbb{R}_+$. Furthermore, let $c_i \in \mathbb{R}$ be real constants such that $\sum_{i=1}^n c_i = 0$ and $C_i = c_i + B_i = \{c_i + r : r \in B_i\} \subset \mathbb{R}_+$, $i = 1, \ldots, n$. The law of the exchangeable sequence A_1^∞ fulfills the symmetry condition

$$\mathbb{P}(A_i \in B_i; i = 1, ..., n) = \mathbb{P}(A_i \in C_i; i = 1, ..., n), \tag{S}$$

for all n, B_i, C_i as above if and only if it is a mixture of exponential distributions [Diaconis and Ylvisaker (1985)], in symbols

$$\begin{aligned}\mathbb{P}(A_i \in B_i; i = 1, ..., n) &= \int_{\mathcal{E}} \prod_{i=1}^n P(A_i \in B_i)\nu(dP) \\ &= \int_{\mathbb{R}_+} \prod_{i=1}^n \int_{B_i} \lambda e^{-\lambda x_i} dx_i \tilde{\nu}(d\lambda),\end{aligned}$$

where $\mathcal{E}$ denotes the space of all exponential distributions on $\mathbb{R}_+$ and $\tilde{\nu}$ denotes the push-forward measure of ν along the natural parametrization$\tilde{}: \mathcal{E} \to \mathbb{R}_+; P \mapsto \lambda$ of the exponential distributions. Moreover, if in above situation additionally it holds $\mathbb{E}[A_2|A_1] = \alpha A_1 + \beta$ for real constants $\alpha, \beta > 0$, then the mixing measure $\tilde{\nu}$ can be shown to be a Gamma distribution.

Since the $M/G/1$ queueing model implies a mixture of exponential distributions for the joint law of the inter-arrivals, from a Bayesian point of view, we assume A_1^∞ to meet constraints (E), (S) and that $\mathbb{E}[A_2|A_1] = \alpha A_1 + \beta$ holds as well (note that by exchangeability, this extends to all random inter-arrival times). The latter directly leads to a conjugate prior for the arrival rate. To be more precise, if we assume λ to be a random variable distributed according to a Gamma distribution with hyper parameters $a, b > 0$, then the posterior distribution given the data X_1^n is a Gamma distribution as well with updated hyper-parameters $(a + n, b + \sum_{i=1}^n A_i)$, i.e. the family of gamma distributions is closed with

respect to exponential sampling. In summary, we assume the following sampling scheme

$$\begin{aligned}\lambda|(a,b) &\sim \Gamma(a,b)\\ A_1^\infty|\lambda &\sim \bigotimes_{\mathbb{N}} \mathcal{E}(\lambda),\end{aligned}$$

which leads, through Bayes theorem, to the posterior distribution, the Bayes estimate for squared error loss and the posterior predictive distribution density, respectively, given by

$$\begin{aligned}\lambda|(a,b), A_1^n &\sim \Gamma\left(a+n, b+\sum_{i=1}^n A_i\right)\\ \mathbb{E}_\Gamma[\lambda|A_1^n,(a,b)] &= \frac{a+n}{b+\sum_{i=1}^n A_i}\\ f\left(a_{n+1}|A_1^n,(a,b)\right) &= \frac{(a+n)\left(b+\sum_{i=1}^n A_i\right)^{a+n}}{\left(b+\sum_{i=1}^n A_i+a_{n+1}\right)^{a+n+1}}.\end{aligned}$$

The latter leads to the predictive value for the next observation

$$\mathbb{E}[A_{n+1}|A_1^n,(a,b)] = \frac{1}{a+n-1}\sum_{i=1}^n A_i + \frac{b}{a+n-1}.$$

Note that the latter equation again reflects the learning process which does not exist in the frequentistic approach in such an explicit form and that for $n = 1$ it is given by $\mathbb{E}[A_2|A_1,(a,b)] = 1/aA_1 + b/a$.

3.2.2 Service Time Distribution

Since the $M/G/1$ model does not imply a parametric mixture for the service time random variables as for the inter-arrivals, assigning a suitable prior is a more difficult task. Not being able to shrink the support of the mixing measure ν in the de Finetti theorem to a finite-dimensional set, we have to choose a prior that supports most of $\mathcal{P}(\mathbb{R}_+)$. The common way is to parametrize $\mathcal{P}(\mathbb{R}_+)$ by a reasonable dense subset. This is taken to be the set of all discrete distributions on $\mathbb{R}_+$, $\mathcal{P}_d(\mathbb{R}_+) = \{P \in \mathcal{P}(\mathbb{R}_+) : P(\cdot) = \sum_{i=1}^\infty w_i\delta_{x_i}(\cdot); w_i \in [0,1], \sum_{i=1}^\infty w_i = 1, x_i \in \mathbb{R}_+; \forall i \in \mathbb{N}\}$. It is well-known that the most famous non-parametric prior in Bayesian statistics, namely the Dirichlet process prior obtained in Ferguson (1973), samples discrete probability measures with probability one. Moreover, it is known that it has full weak support of all of $\mathcal{P}(\mathbb{R}_+)$ if its base measure has support all of $\mathbb{R}_+$. However, here we will use a slightly more general family of prior distributions that enables us to model more flexibly prior beliefs of the true data generating distribution G_0. This larger family will be a subclass of so called neutral to the right prior processes, namely the beta-Stacy processes. Although these priors sample discrete probabilities with probability one, too, an analogue of the result concerning the weak support is known as well. The class of neutral to the right priors is now briefly introduced and its most important properties will be stated.

Neutral to the right Priors

Let $\mathcal{F}(\mathbb{R}_+)$ denote the space of all cumulative distribution functions (c.d.f.) on $\mathbb{R}_+$. Then, a random distribution function $F \in \Omega^{\mathcal{F}(\mathbb{R}_+)}$ is said to be neutral to the right (NTR) if for each $k > 1$ and $0 < t_1 < t_2 < ... < t_k$ the normalized increments

$$F(t_1), \frac{F(t_2) - F(t_1)}{1 - F(t_1)}, ..., \frac{F(t_k) - F(t_{k-1})}{1 - F(t_{k-1})}$$

are independent assuming $F(t) < 1, \forall t \in \mathbb{R}_+$. That is, for all $i \in \mathbb{N}$, $\bar{F}(t_i)/\bar{F}(t_{i-1})$ is independent of the sigma-field generated by F up to time t_{i-1}, $\sigma(\{F(t) : t < t_{i-1})$, where $\bar{F}(\cdot) := 1 - F(\cdot)$ is the survival function associated to F. This essentially asserts that the proportion of mass that F assigns to (t_i, ∞) with respect to (t_{i-1}, ∞) does not depend on how F behaves left of t_{i-1}. This property coined the name neutral to the right. Doksum (1974) has shown that $F(\cdot) \in \Omega^{\mathcal{F}}$ is NTR if and only if $\mathcal{L}(F(\cdot)) = \mathcal{L}(1 - \exp[-A(\cdot)])$ for some independent increment process $A(\cdot)$ which is almost surely non-decreasing, right continuous and such that $\lim_{t\to-\infty} A(t) = 0$ and $\lim_{t\to\infty} A(t) = \infty$. Such objects are called increasing additive processes, see e.g. Sato (1999). For more details on the construction of NTR priors see e.g. Phadia (2015). Since independent increment processes are well understood, the definition of NTR priors leads to a rich class of non-parametric priors which are analytically tractable. Another nice feature of NTR priors is that this family is conjugate with respect to (possibly right-censored) exchangeable data. A fact that makes NTR priors appreciable in statistical survival analysis. Notice, that a Dirichlet process prior updated by right censored data is not longer Dirichlet but can be shown to be structurally conjugate considered as a member of NTR.

The next proposition makes a statement about the weak support of a NTR prior. Recall that the topological support of a measure is the smallest closed set with full measure. In the following theorem $\mathcal{F}(\mathbb{R}_+)$ is identified with the space of all probability measures $\mathcal{P}(\mathbb{R}_+)$ which is equipped with the weak topology.

Proposition 3.2.1. *Dey et al. (2003)*
Let F be a random distribution function which is governed by a NTR prior $\Pi \in \mathcal{P}(\mathcal{F})$, i.e. $F \sim \Pi$, and let $A(\cdot) = -\log(1 - F(\cdot))$ be the corresponding positive increasing additive process with Lévy measure L. Then, Π has full support if L has full support. The assertion remains true if $\mathcal{F}(\mathbb{R}_+)$ is equipped with the sup-norm.

Proof. It has to be shown that all weak neighborhoods of any probability measure have positive Π-mass. Since continuous distributions are dense in $\mathcal{P}(\mathbb{R}_+)$ with respect to the weak topology, it suffices to show the assertion for all weak neighborhoods of continuous distributions. Now, choose some random distribution function F_0 and some $\epsilon > 0$ and consider $U_\epsilon := \left\{F : \sup_{0\le t<\infty} |F(t) - F_0(t)| < \epsilon\right\}$. Then [e.g. Ghosh and Ramamoorthi (2003))] $W \subseteq U_\epsilon$ for some weak neighborhood W of F_0. Moreover, there is $\delta > 0$, $m \in \mathbb{N}$ and $0 < t_1 < t_2 < \cdots < t_m$ such that $\{F : |F(t_i, t_{i+1}] - F_0(t_i, t_{i+1}]| < \delta, i = 1, \ldots, m\} \subseteq W$. Hence, it is enough to show that sup-neighborhoods restricted to compact sets possess positive prior probability. By the homeomorphism $\phi : F(\cdot) \mapsto -\log[1 - F(\cdot)]$ the problem can be translated to the analogous one for the corresponding Lévy process. That is to show that $\mathcal{L}(A)$ gives positive probability to sets of the form $C := \left\{A : \sup_{0\le t\le r} |A(t) - \phi^{-1}F_0(t)| < \gamma\right\}$ for

fixed $r \in \mathbb{Q}$. Now, take a partition $\rho = \cup_{i=1}^{k}(a_i, a_{i+1}]$ of $(0, r]$ such that $\sup_{1\le j\le k} \phi^{-1}F_0(a_j, a_{j+1}] < \gamma$ and define

$$B_i := (a_i, a_{i+1}] \times (\phi^{-1}F_0(a_i, a_{i+1}] - \gamma/k, \phi^{-1}F_0(a_i, a_{i+1}] + \gamma/k),$$
$$B := \cap_{i=1}^{k} \{A : \#\{(t, A(\{t\}) \in B_i\} = 1\}.$$

It follows that $B \subseteq C$ and, since L was assumed to have full support,

$$\phi^{-1}\Pi(C) \ge \phi^{-1}\Pi(B) = \prod_{i=1}^{k} L(B_i)e^{-L(B_i)} > 0.$$

□

Beta Processes and Beta-Stacy Processes

For our purposes, we choose a NTR prior with corresponding process $Y(\cdot)$ being driven by a certain class of Lévy measures. These were studied by Hjort (1990) and Walker and Muliere (1997). Hjort studied beta-processes from a survival analysis viewpoint and therefor elicited a non-parametric prior for the cumulative hazard function (c.h.f.) given for $F \in \mathcal{F}(\mathbb{R}_+)$ by

$$H(t) = \int_0^t \frac{dF(s)}{\bar{F}(s)},\ t > 0.$$

Walker and Muliere (1997) give the definition of the analog of the beta-process as a prior for the c.d.f. directly as follows. F is said to be distributed according to a beta-Stacy process with parameters $(c(\cdot), H(\cdot)) \in \mathbb{R}_+^{\mathbb{R}_+} \times \mathcal{F}(\mathbb{R}_+)$ (for short $F \sim BS(c, H)$) if for all $t \ge 0$ the corresponding process $Y(\cdot)$ fulfilling $F(\cdot) = 1 - \exp[-Y(\cdot)]$ has Lévy measure

$$dL_t(x) = \frac{dx}{1 - e^{-x}} \int_0^t e^{-xc(s)(1-H(s))} c(s) dH_c(s),$$

for all $x > 0$, where $H_c(t) = H(t) - \sum_{k:t_k<t} H(t_k)$ is the continuous part of H with t_i as the fixed points of discontinuity of H. Since we work in a continuous-time framework, we will always choose H to be continuous. However, discontinuities appear in the Lévy measure governing the posterior law. Note that $\mathbb{E}_{BS}[F(\cdot)] = H(\cdot)$ is the prior guess on the c.d.f. F and the function $c(\cdot)$ acts like a tuning parameter affecting the magnitude of the increments and is sometimes interpreted as the "flexible belief in the prior guess". For the sake of clarity, note that the Dirichlet process with finite measure α as parameter admits a similar representation with Lévy-measure

$$dD_t(x) = \frac{dx}{1 - e^{-x}} \int_0^t e^{-xc(1-\bar{\alpha}(s))} c\bar{\alpha}(ds),$$

where $c = \alpha(\mathbb{R})$ and $\bar{\alpha}(\cdot)$ denotes the c.d.f. corresponding to the probability measure $\alpha(\cdot)/c$. In the case of the Dirichlet process it is well-known that the prior guess on the random probability measure equals $\alpha(\cdot)/c$ and c itself is often interpreted as the strength

of belief in the prior guess. Thus, a Dirichlet process is a beta-Stacy process whose parameters are determined by α alone.

As already mentioned, the beta-Stacy process is a parametrically conjugate prior, meaning that the posterior law of F given exchangeable possibly right-censored data is a beta-Stacy process as well. To be more precise, let S_1^n be the first n projections of infinite exchangeable data S_1^∞. Let S_1^∞ be conditional i.i.d. given the c.d.f. F and let $F \sim BS(c, H)$.

Furthermore, define $M_n(t) = \sum_{i=1}^{n} 1_{[t,\infty)}(S_i)$ and $N_n(t) = \sum_{i=1}^{n} 1_{[0,t]}(S_i)$. Then it holds [Walker and Muliere (1997, Theorem 4)] that $F|S_1^n \sim BS(c_n^*, H_n^*)$, where $c_n^*(\cdot)$ and $H_n^*(\cdot)$ are given by

$$H_n^*(t) = 1 - \prod_{s\in[0,t]} \left(1 - \frac{c(s)dH(s) + dN_n(s)}{c(s)\bar{H}(s) + M_n(s)}\right),$$
$$c_n^*(t) = \frac{c(t)\bar{H}(t) + M_n(t) - N_n(t)}{\bar{H}_n^*(t)}.$$

Thereby, $H_n^*(t)$ is defined by means of the product integral, see Gill and Johansen (1990). Note, that the posterior process possesses fixed points of discontinuity at the observations and that the posterior guess on F, i.e. the Bayes estimate with respect to squared error loss is given by $\hat{F}_n(\cdot) = H_n^*(\cdot)$.

Having a look at the Pollazcek-Khinchine transform formulas in chapter 2, one may ask if they are well defined for almost all λ and F drawn from their respective prior and posterior distributions. While the posterior of the mean inter-arrival times is straightforward by definition, the mean and the second moment of the random c.d.f. describing the general service times deserves more attention. The existence of functionals of c.d.f.'s drawn according to a beta-Stacy process was studied in Epifani et al. (2003). Relating $H(\cdot)$ and $c(\cdot)$ to the existence of a certain functional, they obtained sufficient conditions for moments of order m to exist [equation (10) in their article]. We will assume throughout that this condition holds at least for the moment of second order. Moreover, note that they obtained an explicit formula for the prior moments [equation (11)] as well as for the posterior moments [equation (13)].

In summary, we assume the following sample scheme for the service times in the $M/G/1$ system.

$$G|(c, H) \sim BS(c, H)$$
$$S_1^\infty|G \sim \bigotimes_{\mathbb{N}} G.$$

This leads to the posterior distribution

$$G|(c, H), S_1^n \sim BS(c_n^*, H_n^*).$$

Notice again that this updating continues to hold for right-censored observations. Hence, the model can be enlarged in that one does not have to keep exact track of the customers and is still able to make reasonable inference even if e.g. it is solely known that the service of customers has exceeded a certain threshold.

3.2.3 Estimators for Queueing Characteristics

In this subsection, we study the Bayes estimators, i.e. the posterior means, of several characteristics of the $M/G/1$ system. Not all posterior laws of each single characteristic are obtainable explicitly. Hence, for those characteristics whose posterior laws are not obtainable in closed form, we define suitable estimators by replacing appropriate estimators for the corresponding values.

Assume for the general service time distribution $G \sim BS(c, H)$ and let

$$\begin{aligned}\hat{G}_n &= \mathbb{E}_{BS}\left[G|S_1^n\right],\\ \hat{\mu}_n &= \mathbb{E}_{BS}\left[\int_0^\infty t dG(t)|S_1^n\right] \text{ and}\\ \hat{\lambda}_n &= \mathbb{E}_{\Gamma}\left[\lambda|A_1^n\right]\end{aligned}$$

be the Bayes estimators with respect to squared error loss. Note that with $M_n(s) = \sum_{j=1}^n \delta_{X_j}[s,\infty)$, as in Epifani et al. (2003), $\hat{\mu}_n$ can be given as

$$\hat{\mu}_n = \int_0^\infty \exp\left[-\int_0^t \frac{\alpha(ds)}{\beta(s)+M_n(s)}\right]\exp\left[-\int_0^t \frac{\beta(s)+M_n(s)-1}{\beta(s)+M_n(s)}N_n(ds)\right]dt,$$

where if H is continuous [see Phadia (2015)]

- $\beta(t) = c(t)[1-H(t)]$
- $\alpha(t) = \int_0^t c(s)dH(s)$.

Observe that $\hat{\rho}_n = \mathbb{E}_{BS\otimes\Gamma}\left[\lambda\mu|S_1^n, A_1^n\right] = \hat{\mu}_n\hat{\lambda}_n$ by above independence assumption. However, in Bayesian statistics, one is interested in the entire posterior law rather than merely in a certain functional. The posterior law of ρ is obtainable explicitly in the following form

$$\begin{aligned}&P_{BS\otimes\Gamma}(\rho \le t|A_1^n, S_1^n) = P_{BS\otimes\Gamma}(\mu\lambda \le t|A_1^n, S_1^n)\\ &= \int_0^\infty P_{BS\otimes\Gamma}(\mu \le t/\lambda|A_1^n, S_1^n, \lambda)P_{BS\otimes\Gamma}(\lambda|A_1^n, S_1^n)d\lambda\\ &= \int_0^\infty P_{BS}(\mu \le t/\lambda|S_1^n, \lambda)P_{\Gamma}(\lambda|A_1^n)d\lambda\\ &= \frac{(b+\sum_{i=1}^n A_i)^{a+n}}{\Gamma(a+n)}\int_0^\infty P_{BS}(\mu \le t/\lambda|S_1^n, \lambda)\lambda^{a+n-1}e^{-\lambda(b+\sum_{i=1}^n A_i)}d\lambda.\end{aligned}$$

Note that $P_{BS}(\mu \le t/\lambda|S_1^n, \lambda)$ can be stated more explicitly in form of a density by means of Proposition 4 in Regazzini et al. (2003). This explicit form is omitted here due to

technical reasons and left to the interested reader. If one observes the queueing system and is, as is usually the case, not aware of the condition $\rho_0 < 1$, one can at least obtain the posterior probability that the system is stable $P_{BS\otimes\Gamma}(\rho < 1 | A_1^n, S_1^n)$.

We close this section by defining estimators for:

- mean service time: $\hat{\mu}_n$
- traffic intensity: $\hat{\rho}_n = \hat{\lambda}_n \hat{\mu}_n$
- LST of service-time distribution: $g_n^*(z) = \int_0^\infty e^{-zt} d\hat{G}_n(t)$
- LST of waiting-time distribution: $w_n^*(z) = \frac{z(1-\hat{\rho}_n)}{z-\hat{\lambda}_n(1-g_n^*(z))}$
- pgf of queue-size distribution: $q_n^*(z) = \frac{(1-\hat{\rho}_n)(1-z)}{g_n^*(\hat{\lambda}_n(1-z))-z}$
- system-size distribution: $n_n^*(z) = q_n^*(z) g_n^*(\hat{\lambda}_n(1-z))$.

Notice that for the random arrival- and service rate and the random distribution function of the service times the natural Bayes estimator is used, i.e. the minimizer with respect to squared error loss. They are analytically tractable and obtainable in closed form. For the remaining queueing characteristics obvious plug-in estimators are used. The reason therefor is that a closed form of the push-forward law under the mapping

$$(\lambda, G(\cdot)) \mapsto f(\cdot),$$

where $f \in \{n, g, q\}$, of the prior $\Pi_{BS\otimes\Gamma}$ is not easy to obtain. Neither is known how to update such a pushed prior by exchangeable data $(S, A)_1^n$, which we actually have access to, in a natural Bayesian way.

The goodness of these estimates in a rigorous mathematical sense is established in the next sections.

3.3 Posterior Consistency Results

Posterior consistency provides a tool for validation of Bayesian procedures. Roughly speaking, it is defined to be the property of the posterior law to center around the true "parameter" when the number of observed data increases. As a consequence of posterior consistency two different priors will asymptotically lead to the same prediction, a fact often called merging of prior opinions. So, prior information is consecutively washed away by the new state of information provided by the data. Posterior consistency is the most desired property of Bayesian procedures since it states that one can recover the true measure from the data. For examples of inconsistent Bayes procedures see e.g. Diaconis and Freedman (1986) and Kim and Lee (2001) and references therein.

We will have to deal with two different posterior consistency issues, a parametric one for the posterior of the arrival rate and a non-parametric one for the service time distribution and values depending on it. The first was considered by Doob (1949) in a rather general way by the use of martingale theory. It clarifies that, under very weak constraints (the state- and the parameter spaces are assumed to be Polish spaces and the likelihood is identifiable), there is a subset of the parameter space with full prior mass such that the sequence of posterior laws is consistent at any parameter of that subset that is taken as the true one. However, one problem is that in high-dimensional parameter spaces the set with full prior mass might become topologically small. Therefore, especially nonparametric Bayesian statistical problems deserve a deeper study of posterior consistency since the set of possible likelihoods can not longer be assumed to be parametrized finitely.

First of all, the definition of posterior consistency in a rather general framework is given as it can be found e.g. in Schervish (1995). Notice that for any Polish state space $\mathcal{S}$ the space of all probability measures, $\mathcal{P}(\mathcal{S})$, can be equipped with the topology induced by weak convergence which has neighborhood bases for $P \in \mathcal{P}(\mathcal{S})$ given by the collection of sets of the form $U_{P;\epsilon} = \{Q \in \mathcal{P}(\mathbb{R}) : |\int f_i dP - \int f_i dQ| < \epsilon, f_i \in C_b(\mathbb{R}), \text{ for all } i = 1, .., k\}$. This topology makes $\mathcal{P}(\mathcal{S})$ itself a Polish space [see Kechris (1995)] and the topology can be metrized e.g. by the Prohorov metric. The Borel σ-field, $\mathfrak{B}(\mathcal{P})$, with respect to the weak topology serves as a natural measure-theoretical structure to turn $\mathcal{P}(\mathcal{S})$ into a measurable space $(\mathcal{P}(\mathcal{S}), \mathfrak{B}(\mathcal{P}))$. $\mathfrak{B}(\mathcal{P})$ turns out to be the smallest σ-field on $\mathcal{P}(\mathcal{S})$ that makes the mappings $P \mapsto P(A)$ $(\mathfrak{B}(\mathcal{P}), \mathfrak{B}([0,1]))$-measurable for all $A \in \mathfrak{B}(\mathcal{S})$. In the sequel Π, occasionally with an appropriate index, shall denote a prior distribution and Π_n the corresponding posterior distribution after having seen the first n projections of the exchangeable data. A distribution (or a value that parametrizes a distribution) indexed by a naught will always denote the true data generating measure.

Definition 3.3.1. *Let $\Pi \in \mathcal{P}(\mathcal{P}(\mathbb{R}))$ be a prior distribution, i.e. the distribution of some random probability measure $P \in \mathcal{P}(\mathbb{R})$ and let X_1^∞ be a sequence of exchangeable data which is conditionally i.i.d. given P. Moreover, let $P_0 \in \mathcal{P}(\mathbb{R})$ be the true data generating distribution and $(\Pi_n)_{n\geq 0} = (\Pi(\cdot|X_1^n))_{n\geq 0}$ be the sequence of posterior laws. Then call $(\Pi_n)_{n\geq 0}$ weakly consistent at P_0 if for all weak neighborhoods $U_{P_0,\epsilon}$ of P_0 it holds*

$$\Pi_n(U_{P_0,\epsilon}) \overset{n\to\infty}{\longrightarrow} 1,$$

for P_0^∞-almost all data sequences X_1^∞.

Here, P_0^∞ denotes the true joint law governing the sequence X_1^∞.

Needless to say that in the case of the arrival rate λ things become easier because, due to the additional judgment concerning the symmetry of the distribution of the exchangeable sequence of inter-arrival times, one can reduce the problem to that of parametric consistency that is rather easy to handle by conjugacy.

Proposition 3.3.2. *Let Π_Γ stand for the Gamma prior of the random arrival rate λ and $(\Pi_{\Gamma;n}(\cdot))_{n\geq 1} = (\Pi_\Gamma(\cdot|A_1^n))_{n\geq 1}$ be the sequence of posterior laws. Then, for all $\epsilon > 0$ and for $P_{\lambda_0}^\infty$ almost all sequences A_1^∞ it holds*

$$\Pi_{\Gamma;n}\left(|\lambda - \lambda_0| < \epsilon\right) \overset{n\to\infty}{\longrightarrow} 1.$$

Proof. Taking into account that the gamma prior is conjugate for exponentially distributed data, the posterior density can be shown to be that of a $\Gamma(a+n, b+\sum_{i=1}^{n} A_i)$ distribution. By well known properties of the Gamma distribution one has $\mathbb{E}_{\Gamma;n}[\lambda] = (a+n)(b+\sum_{i=1}^{n} A_i)^{-1}$ and $\mathsf{Var}_{\Gamma;n}[\lambda] = (a+n)\,(b+\sum_{i=1}^{n} A_i)^{-2}$. Applying the continuous mapping theorem the assertion follows from the strong law of large numbers and the Markov inequality. □

Let us now turn to the nonparametric part concerning the random c.d.f. G. The following two results are due to Dey et al. (2003) and give the consistency of the Bayes estimator and the random c.d.f. G.

Lemma 3.3.3 (Dey et al. (2003))**.** *Let $G_0 \in \mathcal{F}(\mathbb{R}_+)$ be the true c.d.f. of the service times. Furthermore suppose that G_0 is continuous and that $G_0(t) < 1$ for all $t \in \mathbb{R}_+$. Let $G \sim BS(c, H)$ and $\Pi_{BS;n}$ denote the posterior distribution of G. Then, for G_0^∞-almost all sequences of data S_1^∞ and all $t \in \mathbb{R}_+$ it holds*

$$\mathbb{E}_{BS;n}[G(t)] \xrightarrow{n\to\infty} G_0(t).$$

Next, the posterior consistency of the random service time distribution G induced by a beta-Stacy prior will be stated. Since the service time distribution function $G(\cdot)$ is seen to be a random function rather than a random variable, we investigate deviations in the sup-norm.

Theorem 3.3.4 (Dey et al. (2003))**.** *Let $G_0 \in \mathcal{F}(\mathbb{R}_+)$ be the true c.d.f. of the service times. Again, suppose that G_0 is continuous and that $G_0(t) < 1$ for all $t \in \mathbb{R}_+$. Let $G \sim BS(c, H)$ and $\Pi_{BS;n}$ denote the posterior distribution of G. Then, for all $\epsilon > 0$ it holds*

$$\Pi_{BS;n}\left(\sup_{0\le t<\infty} |G(t) - G_0(t)| < \epsilon\right) \xrightarrow{n\to\infty} 1$$

for G_0^∞-almost all sequences S_1^∞.

Remark 3.3.5. *We stress some peculiarity of a certain class of neutral to the right measures. For a full formal treatment see Dey et al. (2003). As already mentioned previously, neighborhoods w.r.t. the sup-norm contain some weak neighborhoods. These, in turn, are given as finite intersections of sets of the form $\{G : |G(t) - G_0(t)| < \gamma\}$. Thus, it is enough to show that the posterior mass of such sets converges to one. But since the Lévy measure that corresponds to the beta-Stacy prior is of the form $L(dt, ds) = a(t, s)dsK(dt)$ for suitable a and K, the convergence of the expected value of the posterior law to the true c.d.f. already ensures that the posterior variance vanishes with increasing data size.*

As an immediate consequence one has the uniform consistency of the Bayes estimator.

Corollary 3.3.6. *Let the conditions of Lemma* 3.3.3 *be fulfilled. Then the Bayes estimate $\mathbb{E}_{BS;n}[G(\cdot)]$ of the service time distribution is uniformly consistent at the true continuous service time distribution $G_0(\cdot)$, that is*

$$\sup_{0\le t<\infty} \left|\mathbb{E}_{BS;n}[G(t)] - G_0(t)\right| \xrightarrow{n\to\infty} 0,$$

for G_0^∞-almost all sequences S_1^∞.

Proof. By above theorem one has

$$\begin{aligned}
&\sup_{0\le t<\infty}\left|\mathbb{E}_{BS;n}[G(t)]-G_0(t)\right| = \sup_{0\le t<\infty}\left|\int_{\mathcal{F}}(G(t)-G_0(t))\,\Pi_{BS;n}(dG)\right| \\
&\le \int_{\{G:\sup|G(t)-G_0(t)|\ge\epsilon\}}\sup_{0\le t<\infty}|G(t)-G_0(t)|\,\Pi_{BS;n}(dG) \\
&\qquad\qquad + \int_{\{G:\sup|G(t)-G_0(t)|<\epsilon\}}\sup_{0\le t<\infty}|G(t)-G_0(t)|\,\Pi_{BS;n}(dG) \\
&< \Pi_{BS;n}\left(\sup_{0\le t<\infty}|G(t)-G_0(t)|\ge\epsilon\right)+\epsilon \overset{n\to\infty}{\longrightarrow} \epsilon.
\end{aligned}$$

Since $\epsilon > 0$ can be chosen arbitrarily small, the proof is completed and the assertion follows. □

The next lemma establishes the uniform consistency on $\mathbb{R}_+$ of the service time LST in posterior law.

Proposition 3.3.7. *Let $g(z) = \int e^{-sz}dG(s)$ denote the LST of the random service time distribution G possessing a beta-Stacy process prior and $g_0(z) = \int e^{-sz}dG_0(s)$ the LST of the corresponding true data generating distribution. Then, if the constraints of Theorem* 3.3.4 *are fulfilled, it holds*

$$\Pi_{BS;n}\left(\sup_{0\le z<\infty}|g(z)-g_0(z)|<\epsilon\right)\overset{n\to\infty}{\longrightarrow}1,$$

for all $\epsilon > 0$ and for G_0^∞-almost all data sequences S_1^∞.

Proof. The assertion of the proposition follows from Theorem 3.3.4 and the continuous mapping theorem applied to the mapping $G(\cdot)\mapsto g(\cdot)$ which is continuous w.r.t. the sup-norm. Indeed, take a $\delta > 0$ and let $U_{G_0,\delta}$ be a uniform δ-neighborhood of the true service time distribution and let $G \in U_{G_0,\delta}$ with corresponding LST g. Then, by integration by parts of the Riemann-Stieltjes integral it holds that

$$\begin{aligned}
&\sup_{0\le z<\infty}|g(z)-g_0(z)| = \sup_{0\le z<\infty}\left|\int_0^\infty e^{-sz}[G-G_0](ds)\right| \\
&= \sup_{0\le z<\infty}\left|[e^{-sz}[G(s)-G_0(s)]_{|s=\infty}-[e^{-sz}[G(s)-G_0(s)]_{|s=0}-\int_0^\infty[G(s)-G_0(s)]de^{-zs}\right| \\
&= \sup_{0\le z<\infty}\left|\int_0^\infty[G(s)-G_0(s)]de^{-zs}\right| = \sup_{0\le z<\infty}\left|\int_0^\infty[G(s)-G_0(s)]ze^{-zs}ds\right| \\
&\le \sup_{0\le z<\infty}\int_0^\infty ze^{-zs}|G(s)-G_0(s)|\,ds \\
&< \delta\sup_{0\le z<\infty}\int_0^\infty ze^{-zs}ds = \delta,
\end{aligned}$$

which completes the proof and shows the claim. □

Next, the consistency of the estimator $g_n^*(z)$ of the LST of the service time distribution is studied. On basis of the previous result, the following lemma establishes the uniform

consistency on $\mathbb{R}_+$ of the estimator $g_n^*(z)$.

Lemma 3.3.8. *Let $g_n^*(\cdot), g_0(\cdot)$ as above. Then, under the constraints of Theorem 3.3.4, it holds*

$$\sup_{0\le z<\infty} |g_n^*(z) - g_0(z)| \overset{n\to\infty}{\longrightarrow} 0,$$

for G_0^∞-almost all sequences of data S_1^∞.

Proof. Let $\epsilon > 0$ be arbitrary and let $\hat{G}_n(\cdot) := \mathbb{E}_{BS;n}[G(\cdot)]$. Then by properties of the Riemann-Stieltjes integral, one has

$$\begin{aligned}
&\sup_{0\le z<\infty} |g_n^*(z) - g_0(z)| \\
&= \sup_{0\le z<\infty} \left| \int_0^\infty e^{-zs} d\hat{G}_n(s) - \int_0^\infty e^{-zs} dG_0(s) \right| \\
&= \sup_{0\le z<\infty} \Big| \left[e^{-zs}\hat{G}_n(s)\right]_{|s=\infty} - \left[e^{-zs}\hat{G}_n(s)\right]_{|s=0} - \int_0^\infty \hat{G}_n(s) de^{-zs} \\
&\qquad\qquad - \left[e^{-zs}G_0(s)\right]_{|s=\infty} + \left[e^{-zs}G_0(s)\right]_{|s=0} + \int_0^\infty G_0(s) de^{-zs} \Big| \\
&= \sup_{0\le z<\infty} \left| \int_0^\infty \hat{G}_n(s) d(e^{-sz}) - \int_0^\infty G_0(s) d(e^{-sz}) \right| \\
&= \sup_{0\le z<\infty} \left| \int_0^\infty [\hat{G}_n(s) - G_0(s)] z e^{-zs} ds \right| \\
&< \sup_{0\le s<\infty} \left| \hat{G}_n(s) - G_0(s) \right|.
\end{aligned}$$

Hence, the assertion follows from Corollary 3.3.6 and the proof is completed. □

Since the Bayesian estimate of the random mean of G, i.e. $\mathbb{E}_{BS;n}\left[\int_0^\infty t dG(t)\right]$, is used to define estimators of several queueing characteristics, its posterior consistency is examined next. For the most prominent prior process, namely the Dirichlet process a rather general result is known. This is reviewed briefly. So, let $P \sim \mathcal{D}_\alpha$ be a random probability measure that is distributed according to a Dirichlet prior with finite measure α as parameter. Let $f : \mathcal{S} \to \mathbb{R}$ be measurable. Then [c.f. Feigin and Tweedie (1989)] $\int |f| \, d\alpha < \infty \Rightarrow \int |f| \, dP < \infty$ with $\mathcal{D}_\alpha$ probability one and $\mathbb{E}_{\mathcal{D}_\alpha}\left[\int f dP\right] = \int f d\mathbb{E}_{\mathcal{D}_\alpha}[P] = \int f d\alpha$. This fact in combination with an assumption that the state space is countable makes it easy to show that posterior consistency of the random measure P induces posterior consistency of the random mean of P. However, here neither the state space is assumed to be countable nor the random measure is assumed to possess a Dirichlet prior. Thus the posterior consistency of the random mean of G which is drawn according to a beta-Stacy process is studied in more depth and an affirmative result is given.

Lemma 3.3.9. *Suppose $G(\cdot)$ is drawn according to a beta-Stacy process with parameters (α, β), where α is a measure on $\mathbb{R}_+$ and $\beta(s) \ge 1$ such that*

(1.) $\int_{\mathbb{R}_+} [\beta(s)]^{-1} \alpha(ds) = \infty$

(2.) $\int_{\mathbb{R}_+} \exp\left[-\int_0^t [\beta(s)]^{-1} \alpha(ds)\right] dt < \infty.$

Then the posterior expectation of the random mean of G converges G_0^∞-a.s. to the true one. That is for G_0^∞-almost all sequences of data X_1^∞ it holds

$$\mathbb{E}_{BS;n}\left[\int_{\mathbb{R}_+} t dG(t)\right] := \mathbb{E}_{BS}\left[\int_{\mathbb{R}_+} t dG(t)\middle| X_1^n\right] \overset{n\to\infty}{\longrightarrow} \int_{\mathbb{R}_+} t dG_0(t).$$

Proof. First of all note that, due to Epifani et al. (2003), the first moment of the random mean under the prior does exist and is given as

$$\mathbb{E}_{BS;n}\left[\int_{\mathbb{R}_+} t dG(t)\right] = \int_0^\infty \exp\left[-\int_0^t \frac{\alpha(ds)}{\beta(s)+M_n(s)}\right] \times \left[\prod_{\{i:X_i\le t\}} \frac{\beta(X_i)+M_n(X_i)-1}{\beta(X_i)+M_n(X_i)}\right] dt,$$

where $M_n(s) = \sum_{j=1}^n \delta_{X_j}[s,\infty)$. Thus, letting $N_n(s) = \sum_{i=1}^n \delta_{X_i}[0,s)$, one obtains

$$\begin{aligned}&\mathbb{E}_{BS;n}\left[\int_{\mathbb{R}_+} t dG(t)\right]\\ &= \int_0^\infty \exp\left[-\int_0^t \frac{\alpha(ds)}{\beta(s)+M_n(s)}\right] \exp\left[-\int_0^t \log\left(\frac{\beta(s)+M_n(s)}{\beta(s)+M_n(s)-1}\right) dN_n(s)\right] dt.\end{aligned}$$

Since $\beta(s) \ge 1$, by elementary properties of the logarithm, it follows

$$\frac{1}{\beta(s)+M_n(s)} \le \log\left(\frac{\beta(s)+M_n(s)}{\beta(s)+M_n(s)-1}\right) \le \frac{1}{\beta(s)+M_n(s)-1},$$

which in turn implies

$$\begin{aligned}&\int_0^\infty \exp\left[-\int_0^t \frac{\alpha(ds)}{\beta(s)+M_n(s)}\right] \times \exp\left[-\int_0^t \frac{N_n(ds)}{\beta(s)+M_n(s)-1}\right] dt\\ &\le \mathbb{E}_{BS;n}\left[\int_{\mathbb{R}_+} t dG(t)\right]\\ &\le \int_0^\infty \exp\left[-\int_0^t \frac{\alpha(ds)+N_n(ds)}{\beta(s)+M_n(s)}\right] dt.\end{aligned}$$

such that it remains to show that the bounding terms converge to the mean of the true c.d.f.. From a straight-forward application of the monotone convergence theorem and by continuity of the exponential function, it follows

$$\begin{aligned}&\lim_{n\to\infty} \mathbb{E}_{BS;n}\left[\int_{\mathbb{R}_+} t dG(t)\right] = \int_0^\infty \exp\left[-\int_0^t \lim_{n\to\infty} \frac{1/nN_n(ds)}{1/n\beta(s)+1/nM_n(s)}\right] dt\\ &= \int_0^\infty \exp\left[-\int_0^t \frac{G_0(ds)}{1-G_0(s)}\right] dt.\end{aligned}$$

Now, note that $\frac{G_0(ds)}{1-G_0(s)} = \lambda(s)ds = \mathbb{P}(s < S \le s+\Delta s | S > s)$, where $\lambda(s) = \lim_{\Delta s\to 0} \frac{\mathbb{P}(s<S\le s+\Delta s)}{\Delta s[1-G_0(s)]}$ usually denotes the hazard function. Hence $\Lambda(t) = \int_0^t \lambda(s)ds = \int_0^t \frac{G_0(ds)}{1-G_0(s)}$ is the cumulative

hazard until $t \geq 0$. On the other hand $\Lambda(t) = -\log(1 - G_0(t))$. Thus,

$$\exp\left[-\int_0^t \frac{G_0(ds)}{1 - G_0(s)}\right] = \exp\left[-\int_0^t \lambda(s)ds\right] = \exp\left[-\Lambda(t)\right] = \exp\left[\log(1 - G_0(t))\right] = 1 - G_0(t),$$

which in turn completes the proof. □

The next lemma establishes the contraction of the mass of the posterior law of the random mean.

Lemma 3.3.10. *If, in addition to the assumptions of the previous lemma, the parameter* (α, β) *meets the condition*

$$\int_{\mathbb{R}_+} \exp\left[-\int_0^{\sqrt{t}} [\beta(s)]^{-1} \alpha(ds)\right] dt < \infty$$

then the posterior variance of the random mean $\int tG(dt)$ *vanishes* G_0^∞*-a.s. as the size of the data increases. That is, for* G_0^∞*-almost all sequences of data* X_1^∞ *it holds*

$$\mathbb{V}_{BS;n}\left[\int_{\mathbb{R}_+} tG(dt)\right] := \mathbb{V}_{BS}\left[\int_{\mathbb{R}_+} tG(dt)\middle| X_1^n\right] \xrightarrow{n\to\infty} 0.$$

Proof. Due to Lemma 3.3.9 and the continuous mapping theorem the claim follows if the second moment of the random mean under the posterior converges a.s. to the square of the mean of the true distribution function $G_0(\cdot)$. From Epifani et al. (2003) the second moment under the prior law exists and the second moment under the posterior law of the random mean is given by

$$\begin{aligned}
&\mathbb{E}_{BS;n}\left[\left(\int_0^\infty t dG(t)\right)^2\right] \\
&= 2\int_0^\infty \int_r^\infty \exp\left[-\int_0^r \frac{\alpha(dx)}{\beta(x) + M_n(x) + 1}\right] \exp\left[-\int_0^s \frac{\alpha(dx)}{\beta(x) + M_n(x)}\right] \\
&\qquad \times \left(\prod_{\{i:X_i \leq r\}} \frac{\beta(X_i) + M_n(X_i)}{\beta(X_i) + M_n(X_i) + 1}\right)\left(\prod_{\{j:X_j \leq s\}} \frac{\beta(X_j) + M_n(X_j) - 1}{\beta(X_j) + M_n(X_j)}\right) dsdr \\
&= 2\int_0^\infty \exp\left[-\int_0^r \frac{\alpha(dx)}{\beta(x) + M_n(x) + 1}\right] \exp\left[-\int_0^r \log\left(\frac{\beta(x) + M_n(x)}{\beta(x) + M_n(x) + 1}\right) dN_n(x)\right] \\
&\qquad \int_r^\infty \exp\left[-\int_0^s \frac{\alpha(dy)}{\beta(y) + M_n(y)}\right] \exp\left[-\int_0^s \log\left(\frac{\beta(y) + M_n(y)}{\beta(y) + M_n(y) - 1}\right) dN_n(y)\right] dsdr
\end{aligned}$$

Hence, by similar arguments as in the proof of the previous lemma and using Fubini's theorem, it follows that for G_0^∞−almost all sequences of data

$$\mathbb{E}_{BS;n}\left[\left(\int_0^\infty t dG(t)\right)^2\right] \overset{n\to\infty}{\longrightarrow} 2\int_0^\infty \exp\left[-\int_0^r \frac{dG_0(x)}{1-G_0(x)}\right]\int_r^\infty \exp\left[-\int_0^s \frac{dG_0(y)}{1-G_0(y)}\right] dsdr$$
$$= 2\int_0^\infty \int_r^\infty \exp\left[-\int_0^r \frac{dG_0(x)}{1-G_0(x)}\right]\exp\left[-\int_0^s \frac{dG_0(y)}{1-G_0(y)}\right] dsdr. \tag{\#}$$

Furthermore, a straight-forward application of Fubini's theorem yields

$$\int_0^\infty \int_0^s g(r,s)drds = \int_0^\infty \int_r^\infty g(r,s)dsdr,$$

where g is an arbitrary integrable function $g : [0,\infty)^2 \to \mathbb{R}_+$. Now, let $f(r,s)$ denote the integrand of the last integral of equation (#). Since f is symmetric, i.e. $f(r,s) = f(s,r)$, it follows from above equality and further application of Fubini's theorem

$$2\int_0^\infty \int_r^\infty f(r,s)dsdr = \int_0^\infty \int_0^s f(r,s)drds + \int_0^\infty \int_r^\infty f(r,s)dsdr$$
$$= \int_0^\infty \int_0^r f(s,r)dsdr + \int_0^\infty \int_r^\infty f(r,s)dsdr = \int_0^\infty \int_0^\infty f(r,s)drds.$$

Therefore,

$$\mathbb{E}_{BS;n}\left[\left(\int_0^\infty t dG(t)\right)^2\right] \overset{n\to\infty}{\longrightarrow} \left(\int_0^\infty t dG_0(t)\right)^2,$$

which completes the proof of the assertion. □

Theorem 3.3.11. *Under the assumptions of Lemma* 3.3.9 *and Lemma* 3.3.10 *the mean of the random c.d.f. G possesses the property of posterior consistency. That is for all $\epsilon > 0$ and G_0^∞-almost all sequences of data it holds*

$$\Pi_{BS}\left(\left|\int_{\mathbb{R}_+} tG(dt) - \int_{\mathbb{R}_+} tG_0(dt)\right| > \epsilon \middle| X_1^n\right) \overset{n\to\infty}{\longrightarrow} 0.$$

Proof. This is a direct consequence of the two previous lemmas and the Markov inequality. □

Remark 3.3.12. *A brief discussion of the assumptions of above lemmas and theorem, respectively, is given. Assumption* (1.) *of Lemma 3.3.9 is an artifact of survival analysis and ensures that the prior (α,β), that is to be chosen, actually leads to a cumulative*

hazard function. Assumption (2.) of the first lemma and the additional assumption of the second lemma, respectively, ensures the existence of the first, resp. second, moment of the posterior distribution of the random mean. These conditions are given in Epifani et al. (2003) Proposition 4.

Moreover, this work extends results which concern the existence of certain random functionals w.r.t. to a random c.d.f. drawn according to a NTR prior. Those results can be understood as a generalization of a work by Feigin and Tweedie (1989) which gives conditions under which certain functionals of a random measure drawn according to a Dirichlet prior exist in terms of the base measure (the prior parameter). At this place it should be emphasized that they investigated this problem by creating a new approach to the Dirichlet process which is often not mentioned in the literature. To be more precise, they show that the Dirichlet process can be extracted as the invariant distribution of a measure-valued Markov chain and exploit this theory to show sufficiency of the conditions of their existence theorem. Since Dirichlet priors are as well included in the family of NTR priors the extension by Epifani et al. seems natural.

The last assumption, i.e. that $\beta(s) \geq 1$ ensures posterior consistency of the random mean of the c.d.f. governed by the beta-Stacy prior and is used in the proofs of both lemmas. Hence, a prior for the random c.d.f. G is to be chosen, in terms of its parameter, such that the sequence of posteriors of the G-mean is consistent as long as this property is desired. It is interesting to see how advanced information on the posterior consistency of the random mean influences the prior knowledge of the random c.d.f. itself. To put it another way, information concerning the posterior consistency of the random mean is indeed prior information on G. However, the other way around is not true in general, i.e. posterior consistency of the random c.d.f. does not generally imply posterior consistency of the random mean. That is because this functional is not continuous. However, posterior consistency of G implies that of its truncated mean. That means integration is restricted to a compact set and would lead to a weaker form of consistency in the sense of compact convergence.

As an immediate consequence of the previous results one obtains the consistency of the traffic intensity of the $M/G/1$ queue.

Corollary 3.3.13. *Let $\mu := \int sG(ds)$, $\mu_0 := \int sG_0(ds)$ be the random and true mean of the service-time distribution and $\rho_0 = \lambda_0\mu_0$ the true traffic intensity. Further, let $\Pi_{BS\otimes\Gamma} = \Pi_{BS} \bigotimes \Pi_\Gamma$ be the prior on $(\mathcal{P}(\mathbb{R}_+), \mathbb{R}_+)$ that is formed by taking the product-measure of Π_{BS} and Π_Γ, $\Pi_{BS\otimes\Gamma;n}$ the posterior, respectively, and define $\hat{\lambda}_n := \mathbb{E}_{\lambda;n}[\lambda] := \mathbb{E}_\Gamma[\lambda|A_1^n]$, $\hat{\mu}_n := \mathbb{E}_{BS;n}\left[\int_0^\infty tdG(t)\right] := \mathbb{E}_{BS}\left[\int_0^\infty tdG(t)|S_1^n\right]$ and $\hat{\rho}_n := \mathbb{E}_{BS\otimes\Gamma;n}[\rho] := \mathbb{E}_{BS\otimes\Gamma}[\rho|A_1^n, S_1^n]$. Then, under the assumptions of Proposition 3.3.2 and Theorem 3.3.11, one has*

(i) for all $\epsilon > 0$

$$\Pi_{BS\otimes\Gamma}(|\rho - \rho_0| \geq \epsilon|S_1^n, A_1^n) \overset{n\to\infty}{\longrightarrow} 0,$$

for $P_{\lambda_0}^\infty \otimes G_0^\infty$-almost all data sequences (A_1^∞, S_1^∞),

(ii)

$$|\hat{\rho}_n - \rho_0| \overset{n\to\infty}{\longrightarrow} 0,$$

for $P_{\lambda_0}^{\infty} \otimes G_0^{\infty}$-almost all data sequences $(A_1^{\infty}, S_1^{\infty})$.

Proof. (i) Recalling the posterior consistency of the arrival rate λ in Proposition 3.3.2 and the service rate μ in Theorem 3.3.11 one has

$$\begin{aligned}
&\Pi_{BS\otimes\Gamma;n}(|\rho-\rho_0|\geq\epsilon) = \Pi_{BS\otimes\Gamma;n}(|\mu\lambda-\mu\lambda_0+\mu\lambda_0-\lambda_0\mu_0|\geq\epsilon)\\
&\leq \Pi_{BS\otimes\Gamma;n}(|\mu\lambda-\mu\lambda_0|+|\mu\lambda_0-\lambda_0\mu_0|\geq\epsilon)\\
&\leq \Pi_{BS\otimes\Gamma;n}\left(|\mu\lambda-\mu\lambda_0|\geq\epsilon/2\right)+\Pi_{BS\otimes\Gamma;n}\left(|\mu-\mu_0|\geq\frac{\epsilon}{2\lambda_0}\right)\\
&\leq \Pi_{BS\otimes\Gamma;n}\left((\mu_0+\delta)|\lambda-\lambda_0|\geq\epsilon/2,|\mu-\mu_0|<\delta\right)\\
&\qquad+\Pi_{BS;n}\left(|\mu-\mu_0|\geq\delta\right)+\Pi_{BS;n}\left(|\mu-\mu_0|\geq\frac{\epsilon}{2\lambda_0}\right)\\
&\leq \Pi_{\Gamma;n}\left(|\lambda-\lambda_0|\geq\frac{\epsilon}{2(\mu_0+\delta)}\right)+\Pi_{BS;n}\left(|\mu-\mu_0|\geq\delta\right)+\Pi_{BS;n}\left(|\mu-\mu_0|\geq\frac{\epsilon}{2\lambda_0}\right),
\end{aligned}$$

from which the assertion of (i) follows.

(ii) Straightforward one has a.s.

$$\begin{aligned}
|\hat{\rho}_n-\rho_0| &\leq |\hat{\mu}_n\lambda_0-\hat{\mu}_n\hat{\lambda}_n|+\lambda_0|\hat{\mu}_n-\mu_0|\\
&\leq (\mu_0+\gamma)|\lambda_0-\hat{\lambda}_n|+\lambda_0|\hat{\mu}_n-\mu_0|
\end{aligned}$$

by selecting $\gamma>0$ such that a.s. $|\hat{\mu}_n-\mu_0|<\gamma$, for all n sufficiently large. The assertion of (ii) then follows from the proof of Proposition 3.3.2 and Lemma 3.3.9 completing the proof. □

We are now in a position to state the main theorem of this section, i.e. the consistency of the estimators for the waiting time LST, the queue length p.g.f. and the system size p.g.f., defined in chapter 2 .

Theorem 3.3.14. *Under the assumptions of Proposition* 3.3.2, *Proposition* 3.3.7 *and Theorem* 3.3.11 *one has*

(i) for all $\epsilon>0$

$$\Pi_{BS\otimes\Gamma}\left(\sup_{0\leq z<\infty}|f(z)-f_0(z)|\geq\epsilon|S_1^n,A_1^n\right)\overset{n\to\infty}{\longrightarrow}0.$$

(ii)

$$\sup_{0\leq z<\infty}|f_n^*(z)-f_0(z)|\overset{n\to\infty}{\longrightarrow}0,$$

for $P_{\lambda_0}^{\infty}\otimes G_0^{\infty}$-almost all data sequences $(A_1^{\infty},S_1^{\infty})$, where $f(\cdot)\in\{n(\cdot),q(\cdot),w(\cdot)\}$.

Proof. We show the result for $f=w$, that is in the case of the p.g.f. of the waiting-time distribution. The other cases are treated similarly and we omit the details. We intent to

apply the continuous mapping theorem. Therefor note that the mapping $(g(\cdot), \lambda, \mu) \mapsto w(\cdot)$ is continuous with respect to the suitable topologies induced by the sup-norm. Hence, the assertion of the theorem is implied by the results of the present section. □

We shall stress that above theorem is important, especially for applications, since it states that one can reduce the difficult task of making Bayesian inference for queueing characteristics to that of making inference for the observables separately in order to compose it subsequently to that of the objects of interest. Qualitatively this will lead to reasonable asymptotic inference. This is appreciable because it is a non-feasible task to place a prior on the distributions of aforementioned characteristics in a way that it can be updated with data given by observations that we have access to.

3.4 Posterior Normality Results

Asymptotic normality results, so called Bernstein-von Mises theorems, for the posterior law serve as an additional validation of Bayesian procedures. Especially in situations where the exact posterior law is not available or hard to compute they are useful from a rather applied viewpoint to get an approximation of the posterior law. While section 3.3 has shown that the posterior concentrates around the true value when data increases, normality results give an idea how, asymptotically speaking, it does concentrate and how fluctuations around this centering appear. Often it is true that the posterior law, if centered and rescaled appropriately, resembles a centered Gaussian distribution with a certain covariance structure. For applications, special interest lies in this limiting covariance structure.

The earliest result dates back to de Laplace (1774) who approximates the posterior of a beta distribution by a normal integral. Sergei Bernstein and Richard von Mises gave a rather modern version of this approach by applying a general result about infinite products of functions to the sequence of posterior densities, see e.g. Johnson (1967) for a review and references. These approaches of expanding posterior densities coined the name. Nowadays, the parametric case is well understood and results including centering with the MLE or the Bayes estimate as well can be found in Schervish (1995) or Ghosh and Ramamoorthi (2003). However, the results needed for an asymptotic behavior of the posterior of infinite-dimensional parameters, i.e. in the nonparametric case, go deeper and deserve separate investigation. Posterior normality results involving Dirichlet process priors or NTR processes can be found in Conti (1999) who employs a result by Freedman (1963) on asymptotic normality in the finite-dimensional case.

We use a general result by Kim and Lee (2004) on the asymptotic normality of NTR processes to obtain the asymptotic behavior of the posterior law of the waiting time LST. This result is stated next. For the sake of clarity, it is stated for the non-censored case. However, it can be enlarged to the situation where data is right-censored. For a positive real number τ let $(D[0,\tau], \|\cdot\|_\tau)$ denote the space of all cadlag functions on $[0,\tau]$ equipped with the sup-norm and $\mathcal{L}[X]$ the law of a random object X.

Theorem 3.4.1 (Kim and Lee (2004)). *Suppose that $G(\cdot) = 1 - \exp[A(\cdot)]$ is a random c.d.f. drawn according to a NTR prior Π with corresponding increasing additive process $A(\cdot)$ whose Lévy measure L is given by*

$$L([0,t],B) = \int_0^t \int_B \frac{g_s(x)}{x} dx\lambda(s)ds,$$

where $\int_0^1 g_t(x)dx = 1$ for all $t \in \mathbb{R}_+$. Let the true c.d.f. G_0 be continuous and such that $G_0(t) < 1$ for all $t \in \mathbb{R}_+$. Moreover let the following conditions be fulfilled for a $\tau > 0$,

- $\sup_{t\in[0,\tau],x\in[0,1]} (1-x)g_t(x) < \infty$,
- *there is a function $q(t) : \mathbb{R}_+ \to \mathbb{R}$ such that $0 < \inf_{t\in[0,\tau]} q(t) < \sup_{t\in[0,\tau]} q(t) < \infty$ and, for some $\alpha > 1/2$ and $\epsilon > 0$,*

$$\sup_{t\in[0,\tau],x\in[0,\epsilon]} \left| \frac{g_t(x) - q(t)}{x^\alpha} \right| < \infty,$$

- *$\lambda(t)$ is bounded and positive on $(0,\infty)$.*

Then it holds that for G_0^∞-almost all sequences of data S_1^∞

$$\lim_{n\to\infty} \Pi\left(\sqrt{n}[G(\cdot) - \mathbb{E}_{\Pi;n}[G(\cdot)]] \middle| S_1^n\right) = \mathcal{L}[\mathcal{A}(\cdot)]$$

weakly on $(D[0,\tau], \|\cdot\|_\tau)$, where $\mathcal{A}(\cdot)$ denotes a centered Gaussian process given by $\mathcal{A}(t) = (G_0(t)-1)W(A_0(t))$ with a Brownian motion W. The covariance structure then is $h(u,v) := \mathsf{Cov}[\mathcal{A}(u),\mathcal{A}(v)] = (1-G_0(u))(1-G_0(v))min(A_0(u),A_0(v))$.

Proof. See Kim and Lee (2004) for the proof as well as for a discussion of the constraints of the theorem. □

Above theorem can well be stated as follows. The posterior distribution of the scaled and centered random process looks more and more like a Gaussian process as the sample size increases. The limiting process is centered, i.e. the expectation function is the constant function taking only the value zero. Next result ensures that above assertions hold for beta-Stacy processes.

Corollary 3.4.2 (Kim and Lee (2004)). *The assertion of Theorem 3.4.1 for the random c.d.f. G holds, with G being governed by a beta-Stacy process prior with parameters $(c(\cdot), H(\cdot))$, where H is a continuous distribution function with continuous density a with respect to the Lebesgue measure such that $0 < \inf_t c(t)[1-H(t)] < \sup_t c(t)[1-H(t)] < \infty$, i.e. for G_0^∞-almost all sequences of data S_1^∞ it holds*

$$\lim_{n\to\infty} \Pi_{BS;n}(\sqrt{n}[G(\cdot) - \hat{G}_n(\cdot)]) = \mathcal{L}[\mathcal{A}(\cdot)]$$

weakly on $(D^{(1)}[0,\infty), \|\cdot\|_\infty)$ with $\mathcal{A}(\cdot)$ as in Theorem 3.4.1, where $D^{(1)}[0,\infty)$ denotes the space of cadlag functions bounded by one on $[0,\infty)$.

Proof. By Theorem 5 of Dey et al. (2003), a beta-Stacy process is a transformed beta process. More precisely, a process Λ is a beta-Stacy process with parameters $(c(\cdot), H(\cdot))$ if and only if it is a beta process with parameters $\left(c(\cdot)[1-H(\cdot)], \int_0^{\cdot} \frac{dH(s)}{1-H(s)}\right)$. Further on, Kim and Lee (2004) showed that a beta process fulfills the constraints of the theorem as long as $\frac{a(s)}{1-H(s)}$ is positive and continuous on $\mathbb{R}_+$ and $c(t)[1-H(t)]$ is as required. □

Next, we turn to the asymptotic behavior of the LST g of the service time distribution G and its random mean $\mu := \int x dG(x)$. Since a plug-in estimator for g is employed, we will use the functional delta method to provide the asymptotic normality result. Recall that in section 3.3 a posterior consistency result was given for the random mean μ, i.e. it was shown that the posterior law of μ centers a.s. around its true value $\mu_0 := \int x G_0(x)$. This suggests the conjecture that a normality result might be present in this case as well. However, even if Regazzini et al. (2003) provide results to approximate the density of the random mean of random measures, we were not able to show the density in our case to be approximated by a normal density if the sample size increases. The main reason for this is that it does not seem to be straight-forward to obtain a suitable expansion of this approximated density.

Hence, in the following we make the rather technical assumption that there is prior knowledge of the kind that service times can not exceed a certain sufficiently large threshold $M \in \mathbb{R}_+$. From a practical point of view this is a rather gentle constraint. Let $\mathcal{F}_M := \mathcal{F}_M(\mathbb{R}_+) := \{F \in \mathcal{F}(\mathbb{R}_+) : F(t) = 1, \forall t \geq M\}$ be the space of all c.d.f.'s whose corresponding probability measure has support $[0, M]$. Since it is well known that the prior guess on the c.d.f. G under a beta-Stacy prior $BS(c, H)$ is given as $\mathbb{E}_{BS;n}[G] = H$, in the following we take $H \in \mathcal{F}_M$ such that H is continuous on $[0, M]$. Recall that $G(\cdot) = 1 - \exp[-A(\cdot)]$, where A is a non-negative increasing additive process with Lévy measure

$$dN_t(x) = \frac{dx}{1-e^{-x}} \int_0^t e^{-xc(s)[1-H(s)]} c(s) H(ds).$$

From the theory of increasing additive processes it is well known [see e.g. Sato (1999)] that for all $t > 0$ $A(t) = A_t$ is a random variable governed by a infinitely divisible distribution ϕ_t. Let $\hat{\phi}_t(\xi)$ denote the characteristic function of ϕ, i.e.

$$\hat{\phi}_t(\xi) = \int \exp[i\xi s]\phi(ds) = \exp\left[-\int_0^\infty (1-e^{i\xi x}) dN_t(x)\right].$$

Furthermore, since (A_t) has independent increments one has $\hat{\phi}_t(\xi) = \hat{\phi}_s(\xi)\hat{\phi}_{s,t}(\xi)$ for all $s < t$, where $\hat{\phi}_{s,t}(\xi)$ denotes the characteristic function of the difference $A_t - A_s$. Now, since $H \in \mathcal{F}_M$ it follows for $t > M$,

$$\hat{\phi}_t(\xi) = \exp\left[-\int_0^\infty \left(1-e^{i\xi x}\right) dN_t(x)\right] = \exp\left[-\int_0^\infty \frac{1-e^{i\xi x}}{1-e^{-x}} \int_0^t e^{-xc(s)[1-H(s)]} c(s)H(ds)dx\right]$$
$$= \exp\left[-\int_0^\infty \frac{1-e^{i\xi x}}{1-e^{-x}} \int_0^M e^{-xc(s)[1-H(s)]} c(s)H(ds)dx\right]$$
$$\cdot \exp\left[-\int_0^\infty \frac{1-e^{i\xi x}}{1-e^{-x}} \int_M^t e^{-xc(s)[1-H(s)]} c(s)H(ds)dx\right]$$
$$= \hat{\phi}_M(\xi),$$

the process $(A_t)_{t\geq M}$ is a.s. constant. That implies that the corresponding c.d.f. G is constant from M onwards. In order to ensure that it is indeed a distribution function, we set $G(t) := 1$, for all $t \geq M$. The support of the truncated prior $\Pi_{BS}^{(M)}$ law will still be all of $\mathcal{F}_M$.

In order to achieve a posterior normality result for the random LST g, we show that the mapping $\Phi : D^{(1)}[0,M] \to C^{(1)}[0,\infty); G \mapsto \int_0^\infty e^{-sz} dG(s)$ is Hadamard differentiable, where $D^{(1)}[0,M]$ and $C^{(1)}[0,\infty)$ denote the space of cadlag and continuous functions, bounded by one, respectively. Moreover, since we are solely interested in distribution functions on $\mathbb{R}_+$, w.l.o.g. it is assumed that $D^{(1)}[0,M]$ consists only of functions starting at zero. For a good reference on Hadamard differentiability and applications including the functional delta method see Kosorok (2008).

Lemma 3.4.3. *The mapping*

$$\Phi : (D^{(1)}[0,M], \|.\|_M) \to (C^{(1)}[0,\infty), \|.\|_\infty)$$
$$G \mapsto \Phi[G](\bullet) := \int_0^M e^{-\bullet s} G(ds)$$

is Hadamard differentiable.

Proof. Let $t \searrow 0$ and $h_t \in D^{(1)}[0,M]$, such that $h_t \xrightarrow{t\searrow 0} h \in D^{(1)}[0,M]$ w.r.t. the sup-norm. Then, by properties of the Riemann-Stieltjes integral, one has

$$\left\| \frac{\Phi[G+th_t](z) - \Phi[G](z)}{t} - \int_0^M e^{-zs} dh(s) \right\|_\infty$$
$$= \sup_{0\leq z<\infty} \left| \int_0^M [h_t(s) - h(s)](-z)e^{-zs} ds \right|$$
$$\leq \sup_{0\leq z<\infty} \left| \int_0^M \left(\sup_{0\leq x<M} |h_t(x) - h(x)| \right) (-z)e^{-zs} ds \right|$$
$$\leq \|h_t - h\|_M \sup_{0\leq z<\infty} \int_0^M z e^{-zs} ds$$
$$= \|h_t - h\|_M \xrightarrow{t\searrow 0} 0.$$

Define

$$\Phi'_G : D^{(1)}[0,M] \to C^{(1)}[0,\infty)$$
$$h \mapsto \Phi'_G[h](z) = \int_0^M e^{-zs} dh(s).$$

Since the Riemann-Stieltjes integral is linear in the integrator, the mapping Φ'_G is linear. Moreover, if $\sup_{0\le s<M} |F(s) - G(s)| < \delta$, it follows

$$\left| \int_0^M e^{-zs} dF(s) - \int_0^M e^{-zs} dG(s) \right| \le \|F - G\|_M \sup_{0\le z<\infty} \int_0^M z e^{-zs} ds < \delta,$$

thus the continuity of Φ'_G. Hence the mapping Φ is Hadamard differentiable with derivative Φ'_G. □

Now, the lemma will be applied in combination with the functional delta method to obtain the posterior normality of the service time LST centered suitably at its respective Bayes estimator. Write $\Pi^{(M)}_{BS;n}$ for the posterior law induced by the M-truncated beta-Stacy process.

Corollary 3.4.4. *Let $g^{*M}_n(z) = \int_0^M e^{-zs} d\mathbb{E}_{BS;n}[G](s)$. Under the assumptions of Theorem* 3.4.1 *and Corollary* 3.4.2*, it holds for G_0^∞-almost all sequences S_1^∞*

$$\lim_{n\to\infty} \Pi^{(M)}_{BS;n} \left(\sqrt{n} \left[g(\cdot) - g^{*M}_n(\cdot) \right] \right) = \mathcal{L}[\mathcal{G}(\cdot)]$$

on $C^{(1)}([0,\infty), \|.\|_\infty)$, where $\mathcal{G}(z)$ is a centered Gaussian process with covariance structure $\gamma(\cdot,\cdot)$ given by

$$\gamma(u,v) = \mathsf{Cov}[\mathcal{G}(u), \mathcal{G}(v)] = uv \int_0^M \int_0^M e^{-(us+vt)} h(u,v) du dv,$$

where $h(\cdot,\cdot)$ is defined in Theorem 3.4.1.

Proof. By the functional delta method applied to the mapping in the previous lemma one has

$$\mathcal{G}(z) = \Phi'_G \left[(G_0(\cdot) - 1) W(A_0(\cdot)) \right](z) = \int_0^M e^{-zs} d\left[(G_0(s) - 1) W(A_0(s)) \right].$$

Using a Riemann sum approximation for above integral, one concludes that the process $\mathcal{G}(\cdot)$ is a Gaussian process. Further, by well-known properties of the Riemann-Stieltjes integral, one gets

$$\mathcal{G}(z) = z \int_0^M (1 - G_0(s)) W(A_0(s)) \, e^{-sz} ds.$$

Using Fubini's theorem it is immediately seen that $\mathbb{E}[\mathcal{G}(z)] = 0$ for any $z \in \mathbb{R}_+$. Again

using Fubini's theorem, the covariance structure of $\mathcal{G}(\cdot)$ is obtained as

$$\begin{aligned}
&\mathsf{Cov}\left[\mathcal{G}(u), \mathcal{G}(v)\right] = E\left[\mathcal{G}(u)\mathcal{G}(v)\right] \\
&= uv\mathbb{E}\left[\int_0^M \int_0^M (1 - G_0(s))W(A_0(s))e^{-us}(1 - G_0(t))W(A_0(t))e^{-vt}dsdt\right] \\
&= uv \int_0^M \int_0^M e^{-(us+vt)}(1 - G_0(s))(1 - G_0(t))\mathbb{E}\left[W(A_0(s))W(A_0(t))\right]dsdt \\
&= uv \int_0^M \int_0^M e^{-(us+vt)}h(s,t)dsdt.
\end{aligned}$$

□

Next, we investigate the posterior normality of the mean of the random c.d.f. G. Since no exact results seem obtainable, we use the plug-in estimator $\mu_n^* := \int_0^M [1 - \mathbb{E}_{BS;n}[G(t)]]dt$ which, in general, does not equal $\mathbb{E}_{BS;n}\left[\int_0^M [1 - G(t)]dt\right]$. This estimator in combination with the M-truncated c.d.f.'s enables us to use the functional delta method for obtaining normality results.

Lemma 3.4.5. *Let M be an arbitrary positive real number. Then, the mapping*

$$\begin{aligned}
\Psi : (\mathcal{F}_M, \|\cdot\|_M) &\to ([0, M], |\cdot|) \\
G &\mapsto \Psi[G] := \int_0^M [1 - G(s)]ds
\end{aligned}$$

is Hadamard-differentiable with derivative $\Psi'[h] = -\int_0^M h(s)ds$.

Proof. Take $t \searrow 0$ and $h_t \in \mathcal{F}_M$, such that $h_t \xrightarrow{t\searrow 0} h \in \mathcal{F}_M$. Then

$$\left|\frac{\Psi[G + th_t] - \Psi[G]}{t} + \int_0^M h(s)ds\right| = \left|\int_0^M h(s) - h_t(s)ds\right| \leq M\left\|h - h_t\right\|_M \xrightarrow{t\searrow 0} 0.$$

Obviously, the derivative of Ψ is linear and continuous w.r.t. to the considered topologies. □

Corollary 3.4.6. *Under the assumptions of Theorem* 3.4.1 *and Corollary* 3.4.2 *it holds for G_0^∞ almost all data S_1^∞ that*

$$\lim_{n\to\infty} \Pi_{BS;n}^{(M)}\left(\sqrt{n}[\mu - \mu_n^*]\right) = \mathcal{L}[\mathcal{H}],$$

where $\mathcal{H}$ is a centered Gaussian random variable with variance $\eta := \mathsf{Var}[\mathcal{H}] = \int_0^M \int_0^M h(s,t)dsdt$.

Proof. By the previous lemma and the functional delta method the limiting variable is given by $\Psi'[(G_0(\cdot) - 1)W(A_0(\cdot))] = \int_0^M (1 - G_0(s))W(A_0(s))ds$ which is seen to be centered Gaussian by a Riemann sum approximation in combination with Fubini's theorem. Moreover, again by Fubini' theorem

$$\begin{aligned}
&\mathsf{Var}[\mathcal{H}] = \mathbb{E}\left[\mathcal{H}^2\right] = \mathbb{E}\left[\int_0^M \int_0^M (1 - G_0(s))W(A_0(s))(1 - G_0(t))W(A_0(t))dsdt\right] \\
&= \int_0^M \int_0^M h(s,t)dsdt.
\end{aligned}$$

□

Next we consider the asymptotic normality of the arrival-rate when centering with its Bayes estimate.

Proposition 3.4.7. *Let $\hat{\lambda}_n := \mathbb{E}_{\Gamma;n}[\lambda]$. Then, for $P_{\lambda_0}^{\infty}$-almost all sequences of data A_1^{∞} one has*

$$\lim_{n\to\infty} \Pi_\Gamma\left(\sqrt{n}\left[\lambda - \hat{\lambda}_n\right] | A_1^n\right) = \mathcal{N},$$

where $\mathcal{N}$ is a centered Gaussian random variable with precision λ_0^2.

Proof. By Theorem 1.4.3. in Ghosh and Ramamoorthi (2003) the convergence of the posterior distribution of $\sqrt{n}\left[\lambda - \hat{\lambda}_n\right]$ to a centered normal distribution directly follows. Moreover, the variance of this limiting Gaussian variable is given by the inverse Fisher information at the true arrival rate. Checking the necessary conditions for interchanging integral and derivative is left to the interested reader. The Fisher information is obtained as

$$\mathcal{I}(\lambda_0) = \mathbb{E}_{\lambda_0}\left[\left(\frac{\partial}{\partial\lambda}\log\left(\lambda e^{-\lambda A}\right)\right)^2\right] = -\mathbb{E}_{\lambda_0}\left[\frac{\partial^2}{\partial\lambda^2}[\log(\lambda) - \lambda A]\right]_{|_{\lambda=\lambda_0}} = \lambda_0^{-2}.$$

□

We are now in a position to formulate the posterior normality of the waiting-time LST. However, the same techniques can be applied to show posterior normality of several other queueing characteristics like e.g. for the queue length p.g.f. or the sojourn time LST. The waiting-time distribution is of special interest since it gives a qualitative idea about the loss of information that can occur in a $M/G/1$ system and thus helps to ensure a well-working system. However, since the exact posterior law of the waiting time distribution is not obtainable in closed form, asymptotic approximations are given. These results extend results of section 3.3 where it was shown that the plug-in estimator is reasonable to make inference. Roughly, we prove that the posterior law of the LST follows, asymptotically speaking, a Gaussian quantity that is centered around the plug-in estimator. Let $\Pi_{BS\otimes\Gamma} = \Pi_{BS}^{(M)} \otimes \Pi_\Gamma$ denote the prior on the parameter space $\mathbb{R}_+ \times \mathcal{F}_M$.

Theorem 3.4.8. *Let $w(z) = \frac{z(1-\rho)}{z-\lambda(1-g(z))}$ be the LST of the waiting time distribution as given in chapter 2 and $w_n^{*M}(z) = \frac{z(1-\hat{\lambda}_n\mu_n^*)}{z-\hat{\lambda}_n(1-g_n^{*M}(z))}$ be its plug-in estimator. Then, under the assumptions of Theorem* 3.4.1 *and Corollary* 3.4.2*, for $G_0^{\infty}\otimes P_{\lambda_0}^{\infty}$-almost all sequences of data $(S,A)_1^{\infty}$ it holds that*

$$\lim_{n\to\infty} \Pi_{BS\otimes\Gamma}\left(\sqrt{n}\left[w(\cdot) - w_n^{*M}(\cdot)\right] | A_1^n, S_1^n\right) = \mathcal{L}[\mathcal{Z}(\cdot)]$$

weakly on $C([0,\infty), \|.\|_\infty)$, where $\mathcal{Z}(\cdot)$ is a centered Gaussian process with covariance

structure $\zeta(u,v) = \mathsf{Cov}[\mathcal{Z}(u), \mathcal{Z}(v)]$ *given by*

$$\begin{aligned}\zeta(u,v) = &\frac{w_0(u)w_0(v)}{(1-\rho_0)^2}\left[\lambda_0^2\eta + \frac{\mu_0^2}{\lambda_0^2}\right] + \lambda_0^{-2}\frac{w_0(u)(1-g_0(u))}{\lambda_0(1-g_0(u))-u} \times \frac{w_0(v)(1-g_0(v))}{\lambda_0(1-g_0(v))-v}\\ &+ \frac{w_0(u)w_0(v)\lambda_0^2}{[\lambda_0(1-g_0(u))-u][\lambda_0(1-g_0(v))-v]}\gamma(u,v)\\ &- \frac{\mu_0 w_0(u)w_0(v)}{[\lambda_0(1-\rho_0)]^2}\left[\frac{w_0(u)(1-g_0(u))}{u} + \frac{w_0(v)(1-g_0(v))}{v}\right]\\ &- \frac{\lambda_0 w_0(u)w_0(v)}{(1-\rho_0)^2}\left[w_0(u)\int\limits_{[0,M]^2} e^{-us}h(s,t)d(s,t) + w_0(v)\int\limits_{[0,M]^2} e^{-tv}h(s,t)d(s,t)\right],\end{aligned}$$

where $\gamma(\cdot,\cdot)$ *is given in Corollary 3.4.4.*

Proof. Taking a similar route of proving as in the proof of Theorem 3 in Conti (1999), we begin with a decomposition of $\sqrt{n}\left[w(z) - w_n^{*M}(z)\right]$. The decomposition yields

$$\begin{aligned}\sqrt{n}\left[w(z) - w_n^{*M}(z)\right] =& \frac{-w(z)}{1-\rho}\sqrt{n}\left[\lambda\mu - \hat{\lambda}_n\mu_n^*\right]\\ &+ z(1-\hat{\lambda}_n\mu_n^*)\sqrt{n}\left[\frac{1}{z-\lambda(1-g(z))} - \frac{1}{z-\hat{\lambda}_n(1-g(z))}\right]\\ &+ z(1-\hat{\lambda}_n\mu_n^*)\sqrt{n}\left(\frac{1}{z-\hat{\lambda}_n(1-g(z))} - \frac{1}{z-\hat{\lambda}_n(1-g_n^{*M}(z))}\right)\\ &=: Z_{1;n}(z) + Z_{2;n}(z) + Z_{3;n}(z)\end{aligned}$$

Now, the three terms of the sum are investigated separately and it will be shown that they possess the same asymptotic distribution as objects whose asymptotic is easier to obtain. These objects will be tagged by an additional * superscript. First write

$$Z_{1;n}(z) = \frac{-w(z)}{1-\rho}\sqrt{n}\left[\mu_n^*(\lambda - \hat{\lambda}_n) + (\mu - \mu_n^*)(\lambda - \lambda_0) + \lambda_0(\mu - \mu_n^*)\right]$$

and note that by the uniform posterior consistency results of section 3.3 and the continuity of the mapping $G \mapsto \int_0^M (1-G(s))ds$ one has

$$\begin{aligned}&\lim_{n\to\infty}\Pi_{BS\otimes\Gamma}\left[Z_{1;n}(z)\middle|A_1^n, S_1^n\right] = \lim_{n\to\infty}\Pi_{BS\otimes\Gamma}\left[\frac{-w_0(z)}{1-\rho_0}\sqrt{n}(\mu_0(\lambda-\hat{\lambda}_n) + \lambda_0(\mu-\mu_n^*))\middle|A_1^n, S_1^n\right]\\ &=: \lim_{n\to\infty}\Pi_{BS\otimes\Gamma}\left[Z_{1;n}^*(z)\middle|A_1^n, S_1^n\right] =: \mathcal{L}[\mathcal{Z}_1(z)].\end{aligned}$$

For $Z_{2;n}(z)$, note that the mapping $\lambda \mapsto [z-\lambda(1-g(z))]^{-1}$ is analytic in a suitably chosen neighborhood of λ_0. Its derivative is given by

$$\lambda \mapsto \frac{1-g(z)}{[z-\lambda(1-g(z))]^2}.$$

Thus, a Taylor expansion of that mapping yields

$$Z_{2;n}(z) = (1 - \hat{\lambda}_n \mu_n^*) z \frac{1 - g(z)}{\left[z - \bar{\lambda}(1 - g(z))\right]^2} \sqrt{n} \left(\lambda - \hat{\lambda}_n\right),$$

for a suitably chosen $\bar{\lambda} \in [\lambda_0, \hat{\lambda}_n]$. Therefore, using consistency results and continuous mapping, one gets

$$\begin{aligned}
&\lim_{n\to\infty} \Pi_{BS\otimes\Gamma}\left[Z_{2;n}(z) \middle| S_1^n, A_1^n\right] = \lim_{n\to\infty} \Pi_\Gamma \left[(1-\rho_0) z \frac{1 - g_0(z)}{\left[z - \lambda_0(1 - g_0(z))\right]^2} \sqrt{n} \left(\lambda - \hat{\lambda}_n\right) \middle| A_1^n \right] \\
&=: \lim_{n\to\infty} \Pi_\Gamma \left[Z_{2;n}^* \middle| A_1^n \right] =: \mathcal{L}[\mathcal{Z}_2(z)].
\end{aligned}$$

Another Taylor expansion for the mapping $x \mapsto \left[z - \hat{\lambda}_n(1-x)\right]^{-1}$ and analogous reasoning as before yields

$$\begin{aligned}
&\lim_{n\to\infty} \Pi_{BS\otimes\Gamma}\left[Z_{3;n}(z) \middle| S_1^n, A_1^n\right] = \lim_{n\to\infty} \Pi_{BS}^{(M)} \left[-\frac{w_0(z)\lambda_0}{\lambda_0(1 - g_0(z)) - z} \sqrt{n} \left[g(z) - g_n^{*M}(z)\right] \middle| S_1^n \right] \\
&=: \lim_{n\to\infty} \Pi_{BS}^{(M)} \left[Z_{3;n}^* \middle| S_1^n \right] =: \mathcal{L}[\mathcal{Z}_3(z)].
\end{aligned}$$

Now, the convergence of the posterior law of the waiting time LST follows from the previous results of the present section. What remains is the calculation of the covariance structure. This, in turn, is easily obtained by above decomposition and the assumed independence of the arrivals and services or their prior laws, respectively.

$$\begin{aligned}
\mathsf{Cov}\left[\sum_{i=1}^3 \mathcal{Z}_i(u), \sum_{i=1}^3 \mathcal{Z}_i(v)\right] &= \sum_{i=1}^3 \mathsf{Cov}\left[\mathcal{Z}_i(u), \mathcal{Z}_i(v)\right] + \sum_{i\neq j} \mathsf{Cov}\left[\mathcal{Z}_i(u), \mathcal{Z}_j(v)\right] \\
&= \sum_{i=1}^3 \mathsf{Cov}\left[\mathcal{Z}_i(u), \mathcal{Z}_i(v)\right] \\
&\quad + \mathsf{Cov}\left[\mathcal{Z}_1(u), \mathcal{Z}_2(v)\right] + \mathsf{Cov}\left[\mathcal{Z}_2(u), \mathcal{Z}_1(v)\right] \\
&\quad + \mathsf{Cov}\left[\mathcal{Z}_1(u), \mathcal{Z}_3(v)\right] + \mathsf{Cov}\left[\mathcal{Z}_3(u), \mathcal{Z}_1(v)\right]
\end{aligned}$$

By the previous results of this section it follows

$$\begin{aligned}
\sum_{i=1}^3 \mathsf{Cov}\left[\mathcal{Z}_i(u), \mathcal{Z}_i(v)\right] &= \frac{w_0(u)w_0(v)}{(1-\rho_0)^2} \left[\lambda_0^2 \eta + \frac{\mu_0^2}{\lambda_0^2}\right] \\
&\quad + \lambda_0^{-2} \frac{w_0(u)(1 - g_0(u))}{\lambda_0(1 - g_0(u)) - u} \times \frac{w_0(v)(1 - g_0(v))}{\lambda_0(1 - g_0(v)) - v} \\
&\quad + \frac{w_0(u)w_0(v)\lambda_0^2}{\left[\lambda_0(1 - g_0(u)) - u\right]\left[\lambda_0(1 - g_0(v)) - v\right]} \gamma(u, v).
\end{aligned}$$

Furthermore, by the independence assumption of the prior laws of the inter-arrival rate

and the service time distribution, one has

$$\mathsf{Cov}\left[\mathcal{Z}_1(u), \mathcal{Z}_2(v)\right] = \frac{\mu_0}{\lambda_0^2(1-\rho_0)} \times \frac{w_0(u)w_0(v)(g_0(v)-1)}{v-\lambda_0(1-g_0(v))}.$$

Furthermore, using the previous results of the present section and Fubini's theorem, one has

$$\begin{aligned}
\mathsf{Cov}\left[\mathcal{Z}_1(u), \mathcal{Z}_3(v)\right] &= \frac{w_0(u)w_0(v)\lambda_0^2}{(1-\rho_0)(\lambda_0(1-g_0(v))-v)}\mathbb{E}[\mathcal{H}\times\mathcal{G}(v)] \\
&= \frac{w_0(u)w_0(v)\lambda_0^2}{(1-\rho_0)(\lambda_0(1-g_0(v))-v)} \\
&\quad \times \mathbb{E}\left[\int_0^M (1-G_0(s))W(A_0(s))ds \times v\int_0^M (G_0(t)-1)W(A_0(t))e^{-tv}dt\right] \\
&= \frac{vw_0(u)w_0(v)\lambda_0^2}{(1-\rho_0)(\lambda_0(1-g_0(v))-v)}\int_0^M\int_0^M e^{-tv}h(s,t)dsdt \\
&= -\frac{w_0(u)w_0(v)}{(1-\rho_0)^2}\lambda_0^2 w_0(v)\int_{[0,M]^2} e^{-tv}h(s,t)d(s,t).
\end{aligned}$$

Finally, compounding above covariance structures yields $\zeta(\cdot,\cdot)$. □

The covariance structure $\zeta(\cdot,\cdot)$ depends on the unknown objects. However, one can use the suggested estimators and plug them in place of the true ones. This might be helpful to implement the problem and the provided consistency results ensure accuracy as long as the sample size is large enough.

4 Inference for the Service Distribution in $M/G/1$ based on Observations of the Departure Process

4.1 Introduction

Chapter 3 dealt with the issue of making Bayesian inference for characteristics of the $M/G/1$ system as for instance the traffic intensity. Therefor it was assumed that the inner of the system can be observed, i.e. the service time data of customers was available. However, in some situations this is a rather unrealistic assumption. For instance think of a machine in an automatized production process executing different tasks which randomly depend on the item in progress. If the items cannot be distinguished in advance, the assumption of a general service time distribution may be justified. Moreover, if one has no access to the execution process of the machine the system might be assumed to be a black box. In such situations one can merely observe the departure process. However, it will turn out that this is not enough to make inference for the service time distribution such that an additional observation is required. This additional observation amounts to a kind of counter that yields the number of items waiting for service at the instance a certain item leaves the system. Hence, a marked departure process is used in order to make inference for the random service time distribution.

Based on these ideas the present chapter deals with nonparametric Bayesian statistical inference for the service time distribution. The chapter is organized as follows. In section 4.2 the the necessary preliminaries and assumptions are given that extend the survey given in the first chapter. Section 4.3 is devoted to finding a suitable prior distribution for the problem. This will come along with the study of the statistical structure of the space of probability measures that govern the data of the marks of the departure process. This structure is shown to lie in between of usual exchangeability and partial exchangeability. Additionally there are some examples presented that shall clarify the ideas. In section 4.4 an explicit estimator is given as well as its justification in form of posterior consistency results. The last point of section 4.4 is devoted to a Bernstein-von Mises type result further describing the centering process of the posterior law similarly as in section 3.4.

4.2 Preliminaries

The model under consideration is again the single server $M/G/1$ queueing model under a first-in-first-out policy. It will be assumed throughout, that the queueing system has reached its stable equilibrated behavior, thus the process of successive waiting times of customers forms a stationary process. Recall from chapter 2 the necessary and sufficient conditions for the system to reach steady state. Moreover, recall the features of $M/G/1$ which will be used in this chapter to infer the service time distribution. These are PASTA, LAA, LC and time-reversibility of the underlying birth-death process. Also bear in mind that the process of the system's occupation is a semi-Markov process. This semi-Markov process possesses an embedded Markov chain consisting of the magnitude of the system at departures, i.e. at time points customers have just received their complete service. The embedded Markov chain is governed by a Markov law which, in turn, is parametrized by an infinite stochastic matrix of particular shape. The matrix belongs to the family of stochastic delta matrices, as defined in Abolnikov and Dukhovny (1991), and is of the form

$$\begin{pmatrix} a_0 & a_1 & a_2 & a_3 & a_4 & a_5 & \cdots \\ a_0 & a_1 & a_2 & a_3 & a_4 & a_5 & \cdots \\ 0 & a_0 & a_1 & a_2 & a_3 & a_4 & \cdots \\ 0 & 0 & a_0 & a_1 & a_2 & a_3 & \cdots \\ 0 & 0 & 0 & a_0 & a_1 & a_2 & \cdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix},$$

where a_j denotes the probability that $j \in \mathbb{N}_0$ customers enter the system during a service time.

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be an arbitrary probability space which represent the domain of all random quantities involved. Furthermore, let for $n \in \mathbb{Z}$ denote S_n the random service time of the n-th customer, T_n, the instance of time at which customer n departs from the system and $D_n := T_{n+1} - T_n$ be the inter-departure time between the n-th and the $(n+1)$-th customer. Furthermore, let A_n denote the inter-arrival time between the n-th and $(n+1)$-th customer and A_{S_n} the number of customers entering the system during the service of customer n. Since all of the above random variables do not depend on a specific customer, it is common to write them without index. Note, that by properties of $M/G/1$ mentioned above, one has $\mathcal{L}[A] = \mathcal{L}[D] = \mathcal{E}[\lambda]$, where $\mathcal{L}$ stands for the the law of a random quantity and $\mathcal{E}(\lambda)$ for the exponential distribution with rate $\lambda > 0$. Concerning the marks of the departure process, let $\{N_m\}_{m\in\mathbb{Z}}$ denote the embedded Markov chain. Therefore, for any $m \in \mathbb{Z}$ there is an integer n such that $N_m = N_{T_n}$. Besides of the observation processes, let $\mathcal{P}(\mathcal{S})$ denote the space of all probability measures on some Polish space $\mathcal{S}$ and $G \in \mathcal{P}(\mathbb{R}_+)$ be the general distribution of the service times.

All of the aforementioned facts lead us to the following assumption on the accessible observations.

Data is collected from observations of the stochastic process $\{N(T_n), T_n\}_{n\in\mathbb{N}}$. (O)

Assumption (O) means that we observe the instants of time at which the consecutively served customers depart from the system as well as the number of customers they leave behind in the system. Thus, one can imagine the $M/G/1$ system as a perfect black box. The only thing to which we have access is the departure process being a marked point process. This point process is a marked Poisson process with rate $\lambda > 0$ and with marks consisting of the system size at departure time points. Notice that the marks do not directly depict the service times since they are possibly corrupted by idle times under the promises that $\rho < 1$.

We are now interested in making Bayesian statistical inference for several characteristics of the system on grounds of these observations. These characteristics are e.g. the unknown service time distribution, waiting time distribution and the distribution of the busy and idle times, c.f. chapter 2.

4.3 Prior Assignments and the Statistical Structure of the Law of the Embedded Markov Chain

In this section we assign prior distributions to the laws of several random quantities needed for statistical inference in the Bayesian paradigm. Assumption (O) on the observations already indicates that the issue is twofold. On the one hand we can extract the inter-departure times $D_n := T_n - T_{n-1}$ from the observations. As argued in the previous section, these can be viewed as independent and exponentially distributed random quantities with mean λ^{-1}, leading to a parametric inference problem. On the other hand, the marks $\{N(T_n)\}_n$ were argued to form a Markov chain with stochastic matrix $\Delta_{1,1}$-matrix, leading to a non-parametric inference problem. We now regard the both issues separately.

4.3.1 Prior for Inter-Departure Time Distribution

By the properties of $M/G/1$ reviewed in chapter 2, the process of the departure instances is distributed the same way as the arrival process, i.e. both are homogenous Poisson processes with intensity $\lambda > 0$. Therefore, the same theoretical considerations as in section 3.2 can be employed leading to the following statistical setup

$$\begin{aligned} D_1^\infty|\lambda &\overset{i.i.d.}{\sim} \bigotimes_{\mathbb{N}} \mathcal{E}(\lambda) \\ \lambda &\sim \Gamma(a,b) \end{aligned}$$

Thereby, $\Gamma(a,b)$ denotes the gamma distribution with parameters $a, b > 0$. Since the gamma distribution is well known to be a conjugate prior for the rate of exponential

distributions, it easily follows

$$\begin{aligned}
\lambda|(a,b), D_1^n &\sim \Gamma\left(a+n, b+\sum_{i=1}^n D_i\right) \\
\mathbb{E}_\Gamma[\lambda|D_1^n,(a,b)] &= \frac{a+n}{b+\sum_{i=1}^n D_i} \\
f(a_{n+1}|D_1^n,(a,b)) &= \frac{(a+n)\left(b+\sum_{i=1}^n D_i\right)^{a+n}}{\left(b+\sum_{i=1}^n D_i+a_{n+1}\right)^{a+n+1}} \\
\mathbb{E}[D_{n+1}|(a,b),D_1^n] &= \frac{1}{a+n-1}\sum_{i=1}^n D_i + \frac{b}{a+n-1}.
\end{aligned}$$

Recall that the latter equation again reflects the learning process which is not directly available in the frequentistic approach in and that for $n=1$ it is given by $\mathbb{E}[D_2|D_1,(a,b)] = 1/aD_1 + b/a$ as required in above assumption on the shape of the prior.

4.3.2 On the statistical structure of the law of the embedded Markov chain

Eliciting a prior distribution for the law of the marks $\{N(T_n)\}_n$ of the marked departure process described in section 4.2 is a more involved task since the data can not longer be considered as conditional i.i.d. or equivalently as exchangeable. However, Diaconis and Freedman (1980) have shown hat an analogue of the de Finetti theorem remains to hold for laws fulfilling an invariance principle [c.f. Kallenberg (2006)] that is more general than exchangeability and appropriate for the case of Markov chains. Since this theory will be used throughout the present chapter, it is briefly reviewed.

Let $Y_1^\infty : \Omega \to \mathbb{N}_0^{\mathbb{N}}$ be a discrete-time stochastic process with state space $\mathbb{N}_0$. Y_1^∞ is called partially exchangeable if for all $n \in \mathbb{N}$

$$\mathcal{L}(Y_1^n | t_n(Y_1^n) = r) = \mathcal{U}_{t_n^{-1}(r)}. \tag{PE}$$

Thereby, $Y_1^n = (Y_1,\dots,Y_n)$, $\mathcal{U}_B$ denotes the uniform law on a discrete set B and $t := \{t_n\}_{n\in\mathbb{N}}$ is a certain statistic, i.e. a family of measurable mappings, defined by

$$t_n : (i_1,i_2,\dots,i_n) \mapsto t_n((i_1,i_2,\dots,i_n)) := (i_1,T)),$$

where $T = (t_{rs})_{r,s\in\mathbb{N}_0}$ is the transition count matrix defined by

$$t_{rs} := \#\{j : (i_j,i_{j+1}) = (r,s), j = 1,\dots,n-1\}.$$

Condition (PE) is a another way to say that the law of the process Y_1^∞ is summarized by the statistic $\{t_n\}_n$, c.f. Freedman (1962). If in addition to (PE) recurrence holds for the process Y_1^∞, i.e. $\mathbb{P}\left(\limsup_{n\to\infty}\{Y_n = Y_1\}\right) = 1$, then Diaconis and Freedman (1980, Theorem

(7)) shows that there is a unique measure $\mu \in \mathcal{P}(\mathcal{P}(\mathbb{N}_0) \times \mathfrak{S})$ such that for all $n \in \mathbb{N}$

$$\mathbb{P}(Y_i = y_i; i = 1, \dots, n) = \int\limits_{\mathcal{P}(\mathbb{N}_0)\times\mathfrak{S}} p_{y_1} \prod_{i=1}^{n-1} m_{y_i, y_{i+1}} \mu(dp, dM).$$

Thus, any law of a recurrent process also fulfilling (PE) is a mixture of Markov measures. Note that recurrence is necessary to exclude pathologies from the mixture, see e.g. Diaconis and Freedman (1980, Example (19)). However, an earlier result appeared in Freedman (1962) for measures being shift-invariant, a property which clearly implies recurrence of all states whenever all of the state space is supported by the push-forward of the particular shift-invariant measure under coordinate-projections. (Note that more recently Fortini et al. (2002) introduced a stronger version of recurrence and showed that laws of processes whose successor matrix is row-wise exchangeable always fulfill this strong recurrence.) Furthermore, Freedman (1962) showed an even more general result for stationary probabilities being summarized by a S-structure statistic and noted that the family $t = \{t_n\}_n$ has S-structure. Since the $M/G/1$ system is assumed to be in equilibrium, we feel that this approach is appropriate for the embedded Markov chain of the underlying queueing system. Roughly, this implies for above de Finetti-style theorem for Markov chains that it suffices to regard the mixing measure μ having support merely consisting of $\mathfrak{S}$ since the corresponding invariant distribution p is uniquely determined by the stochastic matrix M [see e.g. Chung (1967) or Freedman (1983)]. In the case of the embedded Markov chain of $M/G/1$ the M-invariant distribution is expressible explicitly in terms of M, see chapter 2.

In analogy to the previous subsection, we examine under what constraints the support of $\mu \in \mathcal{P}(\mathfrak{S})$ can be shrunk to the set of Δ matrices governing the embedded Markov chains of $M/G/1$ systems. Of course, $\mathcal{L}(N)$ is summarized by the family $\{t_n\}_n$ since it is Markovian. Yet, $\{t\}_n$ is not minimal sufficient, a fact that yields a different grouping of data strings of equal length in equivalence classes, see examples below. Clearly, there are two properties of M being not fulfilled for arbitrary infinite stochastic matrices. Namely Δ-shape and homogeneity. They should be reflected in the appropriate statistic summarizing a mixture of laws of embedded Markov chains of $M/G/1$.
For any positive integer n, let $\mathcal{D}_n := \{(a_i) \in \mathbb{N}_0^n : a_{i+1} - a_i < 2, \forall 1 \le 1 \le n-1\}$ be the n-dimensional down-skip-free subspace of $\mathbb{N}_0^n$, $\mathcal{D}_\infty \subset \mathbb{N}_0^{\mathbb{N}}$ the down-skip-free sequence space and $\mathcal{D} := \bigcup_{n\in\mathbb{N}} \mathcal{D}_n$ the collection of down-skip-free strings of any length. Moreover, let $\tau := (\tau_n)_{n\in\mathbb{N}}$ be a family of measurable mappings, each τ_n operating on $\mathcal{D}_n$ through

$$\begin{aligned} \tau_n : \mathcal{D}_n &\to \mathbb{N}_0 \times \{0, 1, \dots, n\} \times \mathbb{N}_0^{\mathbb{N}} \\ a_1^n &\mapsto \left(a_1, \sum_{k=1}^{n} \delta_{a_k 0}, I(a_1^n) \right), \end{aligned}$$

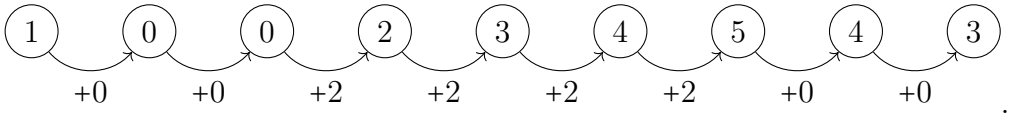
where $I(a_1^n) := (\iota_r(a_1^n))_{r\in\mathbb{N}_0} := \left(\#\left\{j : a_{j+1} - a_j + (1 - \delta_{a_j 0}) = r, j = 1, \dots, n\right\}\right)_{r\in\mathbb{N}_0}$ are the ("zero-adjusted") increments of the data string a_1^n. Thus, τ_n records the length, the initial state, the number of zeros and the increments of the down-skip-free data string a_1^n. Call two strings $a_1^n, b_1^n \in \mathcal{D}_n$ τ-equivalent and write $a_1^n \sim_\tau b_1^n$ if and only if $\tau_n[a_1^n] = \tau_n[b_1^n]$.

Following examples are given to clarify the meaning of the statistic τ.

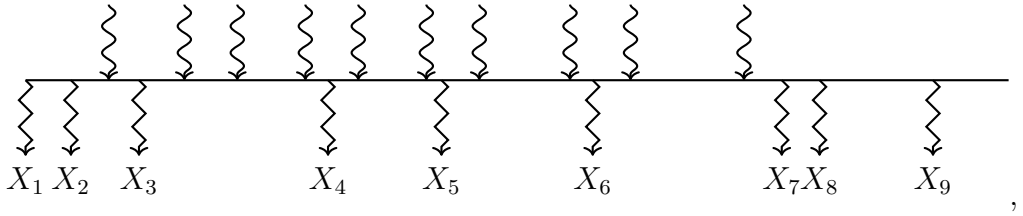
Example 4.3.1. *Define the following elements of* $\mathcal{D}_9$ *by*

$$\begin{aligned} a_1^9 &= 121211100, & b_1^9 &= 102232101, & c_1^9 &= 123332100 \\ d_1^9 &= 100234543, & e_1^9 &= 121002343, & f_1^9 &= 102102323 \\ g_1^9 &= 100234543, & h_1^9 &= 100002345, & i_1^9 &= 210101100 \end{aligned}$$

For instance, g_1^9 *can be viewed as*

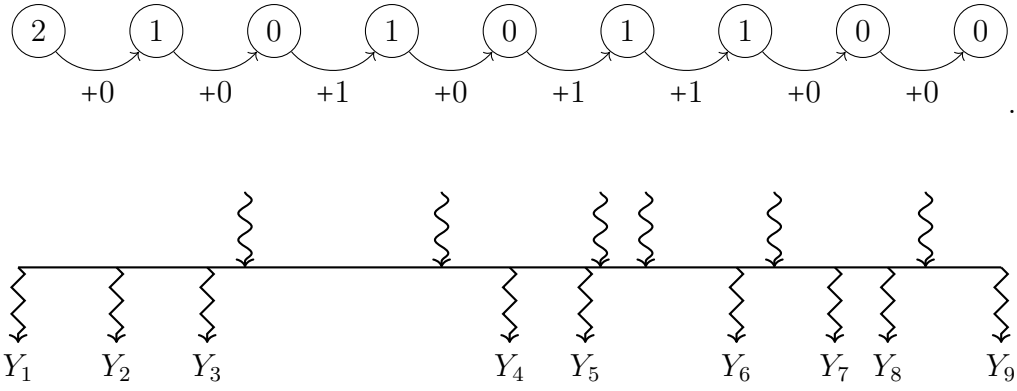


This corresponds to the possible path (depending on the times T_1^n*) given trough*



where a snake arrow depicts an arrival, a zigzag arrow a departure and $X_i = (T_i, N(T_i))$*. Note that the system is idle after the second and the third departure, respectively, and occupied apart from that.*

We contrast above by depicting i_1^9 *the same way.*



We contrast g_1^9 *and* i_1^9 *due to the following reasons. Having a look at the path of* g_1^9 *and* i_1^9*, respectively, one observes that the i-path has more idle times as the g-path. Roughly speaking, that is because the g-path is governed by more increments of higher magnitude. Thus, the probability that the system is unoccupied is lower. This responds to the fact that the arrival and the departure stream of g do not look like having the same intensity rate, while the streams of i rather do. However, this is a basic assumption of ours, since the system is assumed to has run an infinitely long time. So, if* g_1^9 *would be a "typical" path, the embedded Markov chain is rather likely to be transient. Further, keep in mind that only the stream below the horizontal line is communicated to us as data and notice that the only equivalences among above data examples are*

$$a_1^9 \sim_\tau b_1^9 \sim_\tau c_1^9 \text{ and } d_1^9 \sim_\tau e_1^9 \sim_\tau f_1^9 \sim_\tau g_1^9.$$

The string i_1^9 cannot be a member of any equivalence class of the other strings appearing in the example since it has initial state departing from all the other strings.

Remark 4.3.2. *(i) In a sense, the statistic τ can be seen to lie in between of t and o, o denoting order statistic, but closer to t in the following sense. t-equivalence clearly implies τ-equivalence, since one can recover the number of appearance of states among data, as well as the increments ι. The converse is not true nor does o-equivalence imply τ-invariance or the other way around, respectively. Of course, t-equivalence implies o-equivalence, see e.g. Diaconis and Freedman (1980, Proposition (27)).*

(ii) For a data string of finite length, it is obvious that the number of elements of the corresponding equivalence class is bounded. However, eliciting the exact number of elements included in this class seems to be an interesting but hard task, as it is in the case of t. We leave this as an open combinatorial problem, see also section 5.4 of the next chapter.

We continue with mentioning an observation concerning the number of arrivals and departures that occur within a time horizon of observation $T_n - T_1$. By definition, it is clear that one observes n departures. However, even if the statistic τ keeps track of the number of increments of the process, it may happen that the number of arrivals differ in τ-equivalent strings. For instance note that in above example the number of arrivals in b_1^9 exceeds that of a_1^9 by one despite $a_1^9 \sim_\tau b_1^9$. We pin that fact as a proposition.

Proposition 4.3.3. *The number of departing customers during $T_n - T_1$ is an invariant of τ, but the number of arriving customers is not.*

Now, we go on to shed more light on the aforementioned fact that the numbers of arrivals in equivalent strings can differ. This fact corresponds in a sense to the one that equivalent stings may not end with the same symbol. However, in the case of mixtures of general Markov chains this is not true, i.e. the terminal state x_n of a data string x_1^n is completely determined by x_1 and the transition-counts. For a proof see e.g. Martin (1967, Lemma 6.1.1.) and notice that it continues to hold true for countable infinite state spaces.

Lemma 4.3.4. *Let for $a_1^n, b_1^n \in \mathcal{D}_n$ hold $a_1^n \sim_\tau b_1^n$. Then*

(i) $a_n \in \{0,1\} \Leftrightarrow b_n \in \{0,1\}$,

(ii) $a_n = r > 1 \Leftrightarrow b_n = r$.

Proof. (i) By contradiction.
Suppose $a_n \in \{0,1\}$ and $b_n \geq 2$. Since $a_1^n \sim_\tau b_1^n$, one necessarily has

$$\sum_{r\in\mathbb{N}_0} \iota_r(a_1^n) = \sum_{r\in\mathbb{N}_0} \iota_r(b_1^n).$$

Thus,

$$\sum_{k=1}^{n-1} \left[a_{k+1} - a_k + (1 - \delta_{a_k 0})\right] = \sum_{k=1}^{n-1} \left[b_{k+1} - b_k + (1 - \delta_{b_k 0})\right], \qquad (*)$$

which in turn yields

$$a_n + \sum_{k=1}^{n-1} \delta_{a_k 0} = y_n + \sum_{k=1}^{n} \delta_{b_k 0}, \qquad (**)$$

since $a_1 = b_1$ and $b_n \neq 0$ by assumption.
Now, treat the two possible cases $a_n \in \{0, 1\}$ separately.

Case 1:($a_n = 0$) By $(**)$, one has

$$\begin{aligned} 0 - b_n &= \sum_{k=1}^{n} \delta_{b_k 0} - \sum_{k=1}^{n} \delta_{a_k 0} + \underbrace{\delta_{a_n 0}}_{=1} \\ \Rightarrow b_n &= -1. \end{aligned}$$

Case 2:($a_n = 1$) Again by $(**)$, one has

$$\begin{aligned} 1 - b_n &= \sum_{k=1}^{n} \delta_{b_k 0} - \sum_{k=1}^{n} \delta_{a_k 0} + \underbrace{\delta_{a_n 0}}_{=0} \\ \Rightarrow b_n &= 1. \end{aligned}$$

This shows necessity. By interchanging the roles of a_1^n and b_1^n, sufficiency is proven the same way. Thus, (i) follows.

(ii) To prove (ii), note that

$$\sum_{k=1}^{n-1} \delta_{a_k 0} = \sum_{k=1}^{n} \delta_{a_k 0} = \sum_{k=1}^{n} \delta_{b_k 0} = \sum_{k=1}^{n-1} \delta_{b_k 0},$$

exploiting (i) and $a_1^n \sim_\tau b_1^n$. Thus, $(*)$ yields

$$a_n - a_1 = b_n - b_1.$$

□

The lemma states that equality of the terminal state must only hold if the terminal state exceeds 1. This reflects the fact that the laws over which one mixes are governed by $\Delta_{1,1}$-matrices. As a consequence of the Lemma 4.3.4 we have that the statistic τ possesses a certain kind of algebraic structure which reflects stability of the equivalence classes induced by the statistic with respect to extension of the data. This structure was discovered in Freedman (1962) who called it S-structure. Here we consider its analog on the space $\mathcal{D}$. That is a statistic $\sigma = \{\sigma_n\}_{n\in\mathbb{N}}$ is said to have S-structure on $\mathcal{D}$ if for $a_1^n, b_1^n \in \mathcal{D}_n$ and $x_1^m, y_1^m \in \mathcal{D}_m$ such that $a_1^n \sim_\tau b_1^n$, $x_1^m \sim_\tau y_1^m$ and $a_1^n x_1^m, b_1^n y_1^m \in \mathcal{D}_{n+m}$ it holds true that $a_1^n x_1^m \sim_\tau b_1^n y_1^m$, $n, m \in \mathbb{N}$. S-structure will then enable us to identify stationary measures summarized by τ in a unique way as mixtures of laws of embedded Markov chains of $M/G/1$.

Proposition 4.3.5. *τ has S-structure on $\mathcal{D}$.*

Proof. By above Lemma 4.3.4, nothing has to be shown if $a_n > 1$. However, if $a_n \in \{0,1\}$ and $a_n \neq b_n$ then, again by the lemma, one has $b_n = 1 - a_n$. Since the stacked data strings of length $n+m$ are down-skip-free by assumption, one has $a_1^n x_1^m \sim_\tau b_1^n y_1^m$. Indeed, seeing an increment from 0 to $r \in \mathbb{N}_0$ provides the same τ-information as from 1 to r by definition of τ. □

Notice that without recording the number of zeros among the data, one would not have S-structure for τ, i.e. $\tilde{\tau}[a_1^n] := (a_1, I(a_1^n))$ does not possess S-structure. For a counter example take g_1^9 and h_1^9 from previous example and note that both have 1 as initial state and four increments of magnitude 0 and 2, respectively. Hence, $g_1^9 \sim_{\tilde{\tau}} h_1^9$ but $g_1^9 44$ is not $\tilde{\tau}$-equivalent to $h_1^9 44$. Roughly speaking, seeing how often the system is idle is informative with respect to the $M/G/1$ system.

The following lemma states an invariance property of laws summarized by τ with respect to scaling the magnitude of the coordinate process. It is necessary for the proof of the subsequent theorem since it will establish the homogeneity of the stochastic matrix.

Lemma 4.3.6. *Let $P \in \mathcal{P}(\mathcal{D})$ be summarized by τ. Furthermore, let $x_1^n \in \mathcal{D}_n$ with $x_i > 0$, $\forall i = 1, \dots, n$ and for $r \geq 1$ let $y_1^n = x_1^n + r_1^n$, where $r_1^n = (r, r, \dots, r)$. Then it holds true that $P(x_1^n | x_1) = P(y_1^n | y_1)$.*

Proof. First of all note that $y_1^n \in \mathcal{D}_n$ and that $\sum_{i=1}^n \delta_{x_i 0} = 0 \Leftrightarrow \sum_{i=1}^n \delta_{y_i 0} = 0$. Moreover,

$$
\begin{aligned}
&P\left(x_1^n | x_1\right) = P\left(x_1^n | x_1, \textstyle\sum \delta_{x_i 0} = 0\right) = P\left(x_1^n | \tau_n(x_1^n)\right) P\left(I(x_1^n) | x_1, \textstyle\sum \delta_{x_i 0} = 0\right) \\
&= P\left(x_1^n | x_1, \textstyle\sum \delta_{x_i 0} = 0, \left(\#\{j : x_{j+1} - x_j = s, j = 1, \dots, n\}\right)_{s \in \mathbb{N}_0}\right) P\left(I(x_1^n) | x_1, \textstyle\sum \delta_{y_i 0} = 0\right) \\
&= P\left(y_1^n | x_1 + r, \textstyle\sum \delta_{y_i 0} = 0, \left(\#\{j : y_{j+1} - y_j = s, j = 1, \dots, n\}\right)_{s \in \mathbb{N}_0}\right) P\left(I(y_1^n) | x_1 + r, \textstyle\sum \delta_{y_i 0} = 0\right) \\
&= P\left(y_1^n | \tau_n(y_1^n)\right) P\left(I(y_1^n) | x_1, \textstyle\sum \delta_{y_i 0} = 0\right) = P\left(y_1^n | y_1\right).
\end{aligned}
$$

□

We are now in position to state the mixing theorem for $M/G/1$, which will give rise to a prior distribution that is concentrated on the subspace of Markov measures (MM) that are governed by stochastic matrices which are of the shape discussed in the present section. Denoting this space as $MM[\Delta_{1,1}^{(h)}]$ and equipping it with the sigma field induced by weak convergence turns it into a measurable space.

Theorem 4.3.7. *Let $P \in \mathcal{P}(\mathcal{D})$ be a shift-invariant probability summarized by τ. Then, P is uniquely representable as a convex mixture of Markov measures governed by homogenous $\Delta_{1,1}$ stochastic matrices. That is, there is an unique measure $\mu \in \mathcal{P}(\Delta_{1,1}^{(h)})$ such that*

$$
P(\cdot) = \int_{MM[\Delta_{1,1}^{(h)}]} Q(\cdot) \mu(dQ).
$$

Proof. By David Freedman's S-structure Theorem [Freedman (1962, Theorem 1)] it holds that P is a mixture of shift-ergodic laws being themselves summarized by τ. Since τ is

a function of t, P is a mixture of MM [Freedman (1962, Theorem 2)]. Moreover, the space of MM's supporting the mixing measure μ consists of laws governed by stochastic matrices possessing Δ-shape since $supp(P) = \mathcal{D}$. By above lemma, homogeneity of these matrices follows for all row indexes $i \geq 1$. To see, that the zeroth row has to equal the first, just note that $0101 \sim_\tau 0110 \Rightarrow m_{01} = m_{11}$. Using this and $01010 \sim_\tau 01100$ it follows $m_{00} = m_{10}$. Now, since τ has S-structure, $0101r \sim_\tau 0110r$ for $r > 1$. But then, $Q \in MM$ being summarized by τ yields $Q(0101r) = Q(0110r)$ and this, in turn, $m_{0r} = m_{1r}$. □

To state Theorem 4.3.7 in the language of Choquet theory, see chapter 5, the space of stationary measures that are summarized by the statistic τ is a simplex with boundary consisting of all Markov measures governed by homogenous $\Delta_{1,1}$-matrices. Any nontrivial convex mixture thus gives a barycenter in the interior of this simplex. Using an obvious parametrization, one can state the result rather statistically.

Corollary 4.3.8. *Let X_1^∞ be a sequence of stationary data with state space $\mathbb{N}_0$ inducing a joint distribution which is summarized by τ. Then for $n \in \mathbb{N}$ and all strings of data x_1^n one has*

$$\mathbb{P}(X_j = x_j; j = 1, \ldots, n) = \int\limits_{\Delta_{1,1}^{(h)}} p_{x_1} \prod_{i=1}^{n-1} m_{x_i, x_{i+1}} \tilde{\mu}(dp, dM).$$

The corollary states that the problem of finding a prior distribution, modeling the mixing measure μ, can be reduced to that of finding a random object which takes a.s. values in the space $\mathcal{P}(\mathbb{N}_0) \times \Delta_{1,1}^{(h)}$ and whose distribution is analytically tractable. Obviously the random objects ν and M are dependent which complicates the model in general. However, since ν is the unique invariant distribution with respect to M, it is fully determined by M and thus can be viewed as an injective function of M. This simplifies the mixture in Corollary 4.3.8 in the way that one merely has to take into account the distribution of the random stochastic $\Delta_{1,1}^{(h)}$ matrix. That is

$$\mathbb{P}(X_j = x_j; j = 1, \ldots, n) = \int\limits_{\Delta_{1,1}^{(h)}} \nu_{x_1}(M) \prod_{i=1}^{n-1} m_{x_i, x_{i+1}} \hat{\mu}(dM),$$

for a $\hat{\mu}$ suitably related to μ. We stress again, that the particular shape of the Δ-matrices allow for an explicit form of $p(M)$ in terms of probability generating functions, see Harris (1967).

It is known that some measures summarized by a certain statistic can be described in terms of dynamical systems whose dynamics are induced by an associated transformation of the underlying space. The bridge from statistics to dynamical systems is built by ergodic theory, see e.g. Maitra (1977). For instance, exchangeable probability measures, i.e. measures summarized by the order statistic, are invariant with respect to transformations induced by finite permutations. Recall that exchangeable probability measures are automatically stationary. Furthermore, stationary measures summarized by transition counts can be argued to be invariant with respect to transformations induced by switching certain blocks, see below. In both of these cases, there is a fact that simplifies the investigation of classes of equivalent strings, namely the multi-set of symbols through which a string

passes is the same for all equivalent strings. However, in the here considered situation this is not necessarily true as above example shows. Hence, a description in terms of certain permutation-like transformations of the corresponding sequence of increments seems more useful since their multi-sets are invariants for τ. What can be certainly stated is that under τ-equivalence invariance of the probability measure with respect to "more" than only block-switch transformations holds. It seems natural to investigate transformations induced by changes in the increments. In order to formalize this approach, denote for a sting x_1^n the string of ordered increments occurring in x_1^n as $I[x_1^n] := i_1^{n-1}(x_1^n)$.

Proposition 4.3.9. *If P is a stationary probability measure summarized by τ, then P is invariant with respect to the transformations induced by the following operations*

(i) switching two blocks whenever these have the same initial state and (a) end with the same symbol or (b) one ends with a 0 and the other with a 1,

(ii) for a permutation σ of $(k+1)$ elements, permuting a block of positive increments i_m^{m+k} into $i_{\sigma(m)}^{\sigma(m+k)}$.

Proof. One has to argue that value of the statistic τ remains the same under above transformations. Since an increment from 0 has the same observable character than an increment from 1, the assertion of (i) follows from Lemma 4.3.4. For (ii) note that permuting the order of increments within a block of positive increments does not change the accumulated increment over this block. Thus, the number of 0's of the string remains the same. □

The transformations induced by (i) and (ii) of Proposition 4.3.9 give necessary conditions. However, these are not sufficient, i.e. for two stings that are equivalent with respect to τ it is in general not possible to turn one into the other by only applying transformations of the mentioned types. As an example regard the strings $2324321 \sim_\tau 2354321$. Note that it is not possible to generate state 5 in the first string by solely applying transformations of type (i) and (ii). The reason is that also block-switch transformations of the increments are allowed that keep the number of zeros among the sting constant. However, one can hardly formalize those transformations in a neat way and should stick to the description using the statistic τ.

A further description of measures being summarized by the statistic τ can be given using a common ergodic theoretical categorization. This again reflects the fact that τ lies, in a sense, in-between the order statistic and the transition counts. Therefor, keep in mind that all probabilities summarized by the order statistic are invariant under finite permutations, and note from Diaconis and Freedman (1980) that stationary probabilities summarized by transition counts are invariant with respect to a certain subgroup of permutations. This subgroup consists of all permutations that can be described as transformations of blocks that begin with the same symbol and end with the same symbol and hence do not affect the transition counts.

4.3.3 A prior for the law of the embedded Markov chain

Having clarified the statistical structure of the data we are observing, we continue by finding a suitable prior, i.e. by modeling the mixing measure μ in above theorem. We motivate this modeling by an urn process. For urn processes yielding suitable prior distributions for Bayesian statistics see e.g. Blackwell and MacQueen (1973), Hoppe (1984) , Fortini and Petrone (2012a) and references therein. However, our prior can not be chosen in a way such that the rows of the stochastic matrix are seen to be independently sampled. This is due to the following fact [see Fortini and Petrone (2014, Corollary 1)]. The rows of M are stochastically independent with respect to μ if and only if transition counts together with recording the first state are predictive sufficient, i.e. if and only if the probability of observing the next datum only depends on the observation of the last state and the observed number of transitions out of this last state. This clearly fails in our context since transitions of the same magnitude are informative no matter what the starting state of these transitions was.

Consider the following situation. Suppose there is a countable infinite set C called the color space. Without loss of generality, take $C = \mathbb{N}_0$. Furthermore, suppose there is an urn U_i associated to each color $i \in C$, i.e. think of U_i being colored by color i. Let U_i contain initially α_i black balls and start drawing a black ball from urn U_{x_1}, where x_1 is chosen according to a (stationary) start distribution $p_0 \in \mathcal{P}(\mathbb{N}_0)$. Then, having drawn the black ball, replace it together with a ball of color x_2 sampled by a color distribution $c_{x_1} \in \mathcal{P}(\mathbb{N}_0)$ and move to urn U_{x_2}. Once a colored ball is sampled from an urn, replace it together with another ball of this color and move to the urn of this color. Otherwise, continue as before. This is the general definition of a reinforced Hoppe urn process. However, in the here considered situation slight modifications are needed.
Proceed as described but with the following **constraints**. **Firstly**, the initial number of black balls is the same for all urns, i.e. $\alpha_i = \alpha$, for all $i \in \mathbb{N}_0$. **Secondly**, if one draws a ball from urn $i \in \mathbb{N}_0$, then the support of the color sampling distribution c_i is shrunk to $supp(c_i) = \mathbb{N}_0 \setminus \{0, 1, \dots, i-2\}$ for $i > 1$ and $supp(c_i) = \mathbb{N}_0$ for $i = 0, 1$. Moreover, the color sampling distributions fulfill the shift condition $c_i(\{j\}) = c_0(\{j - i + 1\})$. **Thirdly**, not only the present urn is reinforced but all urns are reinforced the following way. If one draws from U_i a ball of color j then replace it together with an additional ball of the same color and add to U_k, $k \neq i$, an additional ball of color $j - i + 1 - \delta_{k0} - \delta_{i0}$ and move to U_j. If a black ball is drawn from U_i, sample a color l and replace the black ball together with the ball of that color. Additionally, add a ball of color $l - i + 1 - \delta_{k0} - \delta_{i0}$ to U_k, $k \neq i$.
Now, let X_1^∞ denote the process of the colors successively sampled according to above urn process. That yields the predictive scheme

$$X_{n+1} = \bullet | X_1^n \sim \frac{\alpha}{\alpha + n - 1} c_0(\{\bullet - X_n + 1\}) + \frac{1}{\alpha + n - 1} \sum_{i=1}^{n-1} \delta_{X_{i+1} - X_i + (1 - \delta_{X_i 0})}(\{\bullet\}). \tag{P}$$

Now, it is shown in Fortini and Petrone (2012a) that a reinforced Hoppe urn process as introduced above is partially exchangeable. Further, it is well known from Blackwell and MacQueen (1973) that the right hand side of (P) converges for $n \to \infty$ to a Dirichlet process with base measure $\alpha c_0(\cdot)$. However, this is essentially the same whatever value X_n takes except of the different shifting of c_0 in (P). This motivates the following choice of a model for the prior μ.

Let $\mu \in \mathcal{P}(\Delta_{1,1}^{(h)})$ be the distribution on the space of homogenous $\Delta_{1,1}$ stochastic matrices such that the 0^{th} row of $M \in \Delta_{1,1}^{(h)}$ is sampled according to a Dirichlet process with base measure $\alpha c_0(\cdot)$ and the i^{th} row, $i \geq 1$, is a copy of the 0^{th} row but shifted $i-1$ times to the right and the resulting "empty" entries of the row filled with zeros. Furthermore, (P) tells one how to update that prior distribution by seeing the data X_1^n.

To summarize the present section, assume that the data X_1^∞ forms an infinitely extendable stationary process with law summarized by τ. Thus, X_1^∞ is a mixture of stationary Markov chains governed by homogenous $\Delta_{1,1}$ stochastic matrices. The distribution of the random stochastic matrix, i.e. the prior, is such that it makes rows dependently sampled from a Dirichlet process with parameters $\alpha > 0$ and $c_0(\cdot) \in \mathcal{P}(\mathbb{N}_0)$. Symbolically we write

$$\begin{aligned} M &\sim Dir^{(\Delta)}(\alpha c_0) \\ X_1^\infty | M &\overset{\mathsf{MM}}{\sim} M. \end{aligned}$$

The posterior of M after having seen data X_1^n is given by

$$M | X_1^n \sim Dir^{(\Delta)}(c_n),$$

where c_n is the discrete measure given through

$$c_n(\{k\}) = \alpha c_0(\{k\}) + \sum_{i=1}^{n-1} \delta_{X_{i+1}-X_i+(1-\delta_{X_i 0})}(\{k\}).$$

Thus, the posterior guess on the stochastic matrix M is given by

$$\mathbb{E}[M|X_1^n] = \begin{pmatrix} \bar{c}_n(\{0\}) & \bar{c}_n(\{1\}) & \bar{c}_n(\{2\}) & \bar{c}_n(\{3\}) & \bar{c}_n(\{4\}) & \dots \\ \bar{c}_n(\{0\}) & \bar{c}_n(\{1\}) & \bar{c}_n(\{2\}) & \bar{c}_n(\{3\}) & \bar{c}_n(\{4\}) & \dots \\ 0 & \bar{c}_n(\{0\}) & \bar{c}_n(\{1\}) & \bar{c}_n(\{2\}) & \bar{c}_n(\{3\}) & \ddots \\ 0 & 0 & \bar{c}_n(\{0\}) & \bar{c}_n(\{1\}) & \bar{c}_n(\{2\}) & \ddots \\ 0 & 0 & 0 & \bar{c}_n(\{0\}) & \bar{c}_n(\{1\}) & \ddots \\ 0 & 0 & 0 & 0 & \bar{c}_n(\{0\}) & \ddots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots \end{pmatrix},$$

where $\bar{c}_n(\{\cdot\}) = \frac{c_n(\{\cdot\})}{\alpha+n-1}$. The prior guess on M is given in a similar way.

4.4 Inference for the Hidden Service Time Distribution

4.4.1 Estimators for queueing characteristics

In the previous section a posterior law of parametric form for the inter-departure time distribution was obtained. Moreover, a non-parametric posterior law for the stochastic matrix M governing the embedded Markov chain of the $M/G/1$-system was obtained based on a Dirichlet process sampling the 0^{th} row of M, which describes the probability for the number of customers who enter the system during a service time. Now, we will use both to obtain an estimator for the service time distribution. Obtaining a direct and tractable closed-form prior to posterior analysis for the service time distribution based on observations given through the marked departure process as described before seems hardly possible. However, a natural approach can be given by the connection provided by the functional relation ship

$$g(z) = a\left(1 - \frac{z}{\lambda}\right).$$

Exploiting this, we define a plug-in estimator for the service-time LST $g(\cdot)$ after having seen data $(T, N(T))_1^n$ by

$$\hat{g}_n(z) := \gamma_n\left(1 - \frac{z}{\bar{\lambda}_n}\right),$$

where $\bar{\lambda}_n = \mathbb{E}_\Gamma[\lambda|(T_{i+1} - T_i)_1^n]$ denotes the posterior expected value of the variable λ under the prior specified in section 4.3 and

$$\gamma_n(z) = \sum_{k=0}^{\infty} z^k \bar{c}_n(\{k\}) = \mathbb{E}_{\mathcal{D}(\Delta)}\left[\sum_{k=0}^{\infty} z^k A(\{k\})|X_1^n\right]$$

denotes the posterior expected value of $a(\cdot)$, the p.g.f. of the discrete distribution of A_S, which is denoted as $A(\{\cdot\})$ and itself is regarded as being random. Notice that the interchange of the sum and the limit is justified since $\mathbb{E}_{\mathcal{D}}[a(z)] = \mathbb{E}_{\mathcal{D}}\left[\int_{\mathbb{N}_0} z^k P^{A_S}(dk)\right]$ and for any $z \in [0,1]$ the mapping $k \mapsto z^k$ is a real valued measurable function and $\int_{\mathbb{N}_0} z^k \alpha(dk) < \infty$ by assumption. Thus, one has $\Pi_{\mathcal{D}}\left(\int_{\mathbb{N}_0} z^k P(dk) < \infty\right) = 1$ and $\mathbb{E}_{\mathcal{D}}\left[\int_{\mathbb{N}_0} z^k P(dk)\right] = \int_{\mathbb{N}_0} z^k \mathbb{E}_{\mathcal{D}}[P](dk)$, see e.g. Feigin and Tweedie (1989) or Phadia (2015).

Based on $\hat{g}_n(\cdot)$, one is able to give estimators for other values of interest. One of those is the traffic intensity $\rho := \mathbb{E}[S]/\mathbb{E}[T_{i+1} - T_i] = \lambda\sigma$, where σ denotes the mean service time and λ^{-1} is the mean inter-departure time. The traffic intensity appears in further characteristics as the LST of the waiting-time distribution or the p.g.f. of the queue-length distribution and hence is of particular interest. An immediate approach is given by defining a plug-in estimator for ρ through $\hat{\rho}_n := \bar{\lambda}_n \hat{\sigma}_n$, where $\hat{\sigma}_n$ is given by $\hat{\sigma}_n = -\left[\frac{\partial}{\partial z}\hat{g}_n(z)\right]_{|z=0}$, i.e. the negative of the derivative of $\hat{g}_n(\cdot)$ elaborated at $z = 0$. Such estimators can be problematic with respect to translating (uniform) large sample results for $\gamma_n(\cdot)$ to that for $\hat{\sigma}_n$. However, in the here considered situation things become easier since ρ has a direct relation to the

random variable A_S which we have access to through the observations. Indeed, one has

$$\hat{\rho} = \hat{\lambda}_n \hat{\sigma}_n = \hat{\lambda}_n \left[\sum_{k=0}^{\infty} k \left(1 - \frac{z}{\hat{\lambda}_n}\right)^{k-1} \frac{1}{\hat{\lambda}_n} \bar{c}_n(\{k\}) \right]_{|_{z=0}} = \sum_{k=1}^{\infty} k \bar{c}_n(\{k\}) = \mathbb{E}_{\mathbb{E}_{\mathcal{D};n}[P^{A_S}]}[A_S].$$

Based on above estimators, we can define estimators for further queueing characteristics as e.g. the waiting-time distribution, the busy-time distribution and duration-time distribution exploiting similar functional relationships. For details on these relationships see e.g. Nelson (2013, chapter 7). Define estimators for the following queueing characteristics

- p.g.f. of number of customers in queue: $\hat{q}_n(z) = \frac{(1-\hat{\rho}_n)(1-z)}{\hat{g}_n(\bar{\lambda}_n(1-z))-z}$,
- p.g.f. of number of customers in the system: $\hat{m}_n(z) = \hat{g}_n(\bar{\lambda}_n(1-z))\hat{q}_n(z)$,
- LST of waiting time of a customer in queue: $\hat{w}_n(s) = \frac{s(1-\hat{\rho}_n)}{s-\bar{\lambda}_n+\bar{\lambda}_n\hat{g}_n(s)}$.

For the number of customers served in a busy period as well as the length of the busy period itself only estimates for the associated functional equation can be given, i.e.

- LST of busy period: $b(s) = \hat{g}_n(s + \bar{\lambda}_n[1 - b(s)])$,
- p.g.f. of number of customers served in a busy period: $m_b(z) = z\hat{g}_n(z)(\bar{\lambda}_n[1 - m_b(z)])$.

Solutions to these equations may be understood as estimators for the busy time LST and the p.g.f. of the number of customers served in a busy period. However, the goodness of those estimators w.r.t. large samples is in question not only from a applied point of view, i.e. due to deviations appearing from numerical approximations, but also from a theoretical viewpoint since it is not known if minor changes in λ and γ do lead to minor changes in the solution to the equations. Put another way, it is not clear whether the mapping that maps λ and $g(\cdot)$ onto $b(\cdot)$ and $m_b(\cdot)$, respectively, is continuous.

We continue by emphasizing the role of the special form of the stochastic matrix M with respect to the M-invariant distribution p of the Markov chain X_1^{∞}. We point out that the specific appearance of M allows to write down explicitly the invariant distribution as a function of M in form of their transforms. That is, the diagram

$$\begin{array}{ccc} M & \xrightarrow{\phi} & p \\ \psi \uparrow\downarrow & & \uparrow\downarrow \psi \\ a(\cdot) & \xrightarrow[\xi]{} & \pi(\cdot) \end{array}$$

commutes. Therein, for the sake of brevity, ψ on the left-hand side denotes the composition of the mapping that extends the distribution of A_S appearing in the 0^{th} row of M to

the whole of M and the mapping that maps the distribution of A_S onto its p.g.f., while on the right-hand side it just describes the mapping that maps the distribution p onto its p.g.f. . Recall from chapter 2 that the mapping ξ is given through

$$\xi : a(z) \mapsto a(z)\frac{(1-z)(1-a'(1))}{a(z)-z} =: \pi(z).$$

Certainly, such a description is not possible in general. It even fails for the case of a non-homogenous $\Delta_{1,1}$ stochastic matrix which governs the embedded Markov chain of $M/G/1$ with state-dependent service, see Harris (1967, equation (4)). This special feature of standard $M/G/1$ enables us to give a direct estimator for the p.g.f. of the distribution of the system size at instants of departing customers which is, by the PASTA property, the same for any arbitrary instant of time. This estimator is given by

$$\hat{\pi}_n(z) = \gamma_n(z)\frac{(1-z)(1-\gamma_n'(1))}{\gamma_n(z)-z}.$$

4.4.2 Posterior consistency

The estimators just defined are obvious ones, yet deserve some further theoretical justification. This will be given by posterior consistency which, roughly speaking, states that the mass of the posterior law will center around the true data-generating measure. To be more precise, let for a random probability measure $P \in \mathcal{P}^\Omega$ a prior $\Pi \in \mathcal{P}(\mathcal{P})$ be given. Further, let data Y_1^∞ be given such that $Y|P \overset{iid}{\sim} P$. Then, denote by $(\Pi_n)_{n\in\mathbb{N}_0}$ the sequence of posterior laws of P given observed data Y_1^n, i.e. $\Pi_n(C) = \Pi(P \in C|Y_1^n)$ and for the sake of completeness $\Pi_0 := \Pi$, for all sets C in the sigma field induced by weak convergence of measures. The sequence $(\Pi_n)_{n\in\mathbb{N}_0}$ is called consistent at the true data-generating distribution P_0 if for P_0-almost all data sequences it holds that $\Pi_n \overset{w,n\to\infty}{\longrightarrow} \delta_{P_0}$. First of all we state posterior consistency of the the parametric sequence of posteriors for the inter-arrival rate λ.

Lemma 4.4.1. *For almost all sequences T_1^∞ and for any $\epsilon > 0$ it holds that*

$$\Pi_{\Gamma;n}([\lambda_0-\epsilon, \lambda_0+\epsilon]) \overset{n\to\infty}{\longrightarrow} 1,$$

where Π_Γ denotes the prior distribution for λ as specified in section 4.3 and λ_0 the true inter-arrival rate.

Proof. Let $\{Di\}_i$, $D_i := T_{i+1} - T_i$ be the exponentially distributed inter-departure time data. By conjugacy of the gamma-distribution with respect to exponential likelihoods and well known properties of the gamma distribution, the posterior expected value of the arrival rate is given by

$$\mathbb{E}_\Gamma[\lambda|D_1^n] = \frac{a+n}{b+\sum_{i=1}^n [D_i]},$$

where $(a,b) \in \mathbb{R}_+^2$ are the prior parameters. Moreover, the posterior variance is given by

$$\mathbb{V}_\Gamma[\lambda|D_1^n] = \frac{a+n}{\left(b+\sum_{i=1}^n [D_i]\right)^2}.$$

Thus, by means of the SLLN and the continuous mapping theorem, $\mathbb{E}_\Gamma[\lambda|D_1^n] \overset{n\to\infty}{\longrightarrow} 1/\lambda_0$ and $\mathbb{V}_\Gamma[\lambda|D_1^n] \overset{n\to\infty}{\longrightarrow} 0$ such that the assertion of the lemma follows from a straight forward application of the Markov inequality. □

Remark 4.4.2. *The proof of the previous lemma shows that the Bayes estimator is a consistent (in the usual sense) estimator. However, for a Bayesian this is not enough. It is required to know that fluctuations of the random intensity around its estimate become smaller when sample size increases.*

Next, we study the posterior consistency of the random stochastic matrix $M \in \left[\Delta_{1,1}^{(h)}\right]^\Omega \subset \mathfrak{S}^\Omega$. Virtually, this task requires an extended definition of posterior consistency. So, let M be a random matrix and let Y_1^∞ be a Markov chain with countable state space that, given M, is governed by M, i.e. $Y_1^\infty \mid M \overset{MC}{\sim} M$. Let a prior Π be given for M and, as before, denote by $(\Pi(\cdot|Y_1^n))_{n\in\mathbb{N}_0} = ((\Pi_n(\cdot))_{n\in\mathbb{N}_0}$ the sequence of posterior laws of M. Call $(\Pi_n)_{n\in\mathbb{N}_0}$ consistent if for M_0-almost all sequences of data Y_1^∞ it holds that $\Pi_n(C_0) \overset{n\to\infty}{\longrightarrow} 1$, for all sets C_0 in the sigma field on $\mathfrak{S}$ induced by coordinate-wise convergence containing M_0, the true stochastic matrix governing the data. Here, M_0-almost all sequences of data means the smallest set of data strings which has full mass under the stationary Markov probability measure that is induced by M_0.

Next, we show the posterior consistency of the random matrix M.

Lemma 4.4.3. *For almost all data sequences $N(T)_1^\infty$ and all measurable neighborhoods C_0 of the true stochastic matrix M_0 governing the embedded Markov chain of $M/G/1$ it holds for the prior $\Pi_{\mathcal{D}(\Delta)}$ specified in section 4.3 that*

$$\Pi_{\mathcal{D}^\Delta;n}(C_0) \overset{n\to\infty}{\longrightarrow} 1.$$

Proof. Since $M_0 \in \Delta_{1,1}^{(h)}$, it suffices to regard all neighborhoods of M_0 contained in the trace sigma filed induced by $\Delta_{1,1}^{(h)}$. But then, using mapping ψ in above diagram, it is enough to show consistency for the 0^{th} row of M, which is nothing but the distribution of the variable A_S. Since the posterior, emerging form the Dirichlet process prior updated in a manner as described before by data X_1^n, is as well a Dirichlet process with updated base measure

$$c_n(\{k\}) = \alpha c_0(\{k\}) + \sum_{i=1}^{n-1} \delta_{X_{i+1}-X_i+(1-\delta_{X_i 0})}(\{k\}),$$

posterior consistency of M follows from convergence properties of that prior process, see e.g. Ghosh and Ramamoorthi (2003, chapter 3). □

The posterior consistency of the random matrix immediately yields the consistency of the Bayes estimator for the stochastic matrix and for the p.g.f. of A_S, respectively.

Corollary 4.4.4. *For M_0-almost all sequences of data $X_1^n := N(T)_1^n$, it holds that*

(i) $\mathbb{E}_{\mathcal{D}^\Delta}[M \mid X_1^n] \overset{n\to\infty}{\longrightarrow} M_0$,

(ii) $\mathsf{Var}_{\mathcal{D}^\Delta}[M \mid X_1^n] \xrightarrow{n\to\infty} 0$,

(iii) for any $R>0$, $\sup_{z\in[0,R]} |\gamma_n(z) - a_0(z)| \xrightarrow{n\to\infty} 0$.

Proof. The assertions of (i) and (ii) just follows as necessary consequences of Lemma 4.4.3. For (iii) note that one has

$$\begin{aligned}
&\lim_{n\to\infty} \sup_{z\in[0,R]} |\gamma_n(z) - a_0(z)| \le \lim_{n\to\infty} \sup_{z\in[0,R]} \sum_{k=0}^{\infty} z^k \left|\bar{c}_n(\{k\}) - P_0^{A_S}(\{k\})\right| \\
&\le \lim_{n\to\infty} \sum_{k=0}^{\infty} R^k \left|\bar{c}_n(\{k\}) - P_0^{A_S}(\{k\})\right| \le \lim_{n\to\infty} \sum_{k=0}^{\infty} R^k \mathbb{E}_{\mathcal{D}^\Delta}\left[\left|P^{A_S}(\{k\}) - P_0^{A_S}(\{k\})\right| \mid X_1^n\right],
\end{aligned}$$

such that the assertion of (iii) follows from (i) and the monotone convergence theorem. □

So far we established posterior consistency with respect to the direct observables. Now we show that the indirect estimator defined above possesses certain consistency properties as well. Since the LST of the service time distribution is expressed in terms of the p.g.f. of the distribution A_S, which in turn is a power series, it seems natural to undertake the investigation of posterior consistency within a framework that reflects this analytic approach. Hence, posterior consistency of the LST of the service time distribution will be stated as a kind of a.s. compact convergence inside the posterior law. Call a series of functions compact convergent to a limit function if its restriction to compact sets converges uniformly. Therefor, let $g_0(\cdot)$ denote the the LST of the true service time distribution $G_0(\cdot)$, i.e. $g_0(z) = \int_0^\infty e^{-zs} dG_0(s)$.

Theorem 4.4.5. *For almost all data sequences* $(T, N(T))_1^\infty$, *all* $R>0$ *and all* $\epsilon>0$ *it holds true that*

$$\mathbb{P}\left(\sup_{z\in[0,R]} |\hat{g}_n(z) - g_0(z)| \ge \epsilon \mid (T,N(T))_1^n\right) \xrightarrow{n\to\infty} 0.$$

Proof. Let $R>0$ and $\epsilon>0$ be arbitrarily chosen real numbers. Define

$$\begin{aligned}
X &:= \sup_{z\in[0,R]} \left|\sum_{k=0}^{\infty}\left(1-\frac{z}{\bar{\lambda}_n}\right)^k \bar{c}_n(\{k\}) - \sum_{k=0}^{\infty}\left(1-\frac{z}{\bar{\lambda}_n}\right)^k A_0(\{k\})\right|, \\
Y &:= \sup_{z\in[0,R]} \left|\sum_{k=0}^{\infty}\left(1-\frac{z}{\bar{\lambda}_n}\right)^k A_0(\{k\}) - \sum_{k=0}^{\infty}\left(1-\frac{z}{\lambda_0}\right)^k A_0(\{k\})\right|
\end{aligned}$$

Then one has

$$\begin{aligned}
&\mathbb{P}\left(\sup_{z\in[0,R]} |\hat{g}_n(z) - g_0(z)| \ge \epsilon \mid (T,N(T))_1^n\right) \\
&= \mathbb{P}\left(\sup_{z\in[0,R]} \left|\gamma\left(1-\frac{z}{\bar{\lambda}_n}\right) - a_0\left(1-\frac{z}{\lambda_0}\right)\right| \ge \epsilon \mid (N,T(N))_1^n\right) \\
&\le \mathbb{P}\left(X+Y\ge\epsilon, Y\ge\epsilon/2 | (N,T(N))_1^n\right) + \mathbb{P}\left(X+Y\ge\epsilon, Y<\epsilon/2 \mid (N,T(N))_1^n\right) \\
&\le \mathbb{P}\left(Y\ge\epsilon/2 | (N,T(N))_1^n\right) + \mathbb{P}\left(X\ge\epsilon/2 \mid (N,T(N))_1^n\right).
\end{aligned}$$

Exploiting the independence assumption between λ and M, for the first addend it follows

$$\begin{aligned}&\mathbb{P}\left(Y \geq \epsilon/2 \,|\, (N, T(N))_1^n\right)\\ &\leq \Pi_\Gamma\left(\sum_{k=0}^{\infty} A_0(\{k\}) \sum_{i=0}^{k} R^{k-i} \left|\bar{\lambda}_n^{-(k-i)} - \lambda_0^{-(k-i)}\right| \geq \epsilon/2 \,|\, T_1^n\right),\end{aligned}$$

while for the second one has

$$\begin{aligned}&\mathbb{P}\left(X \geq \epsilon/2 \,|\, (N, T(N))_1^n\right)\\ &\leq \mathbb{P}\left(\sum_{k=0}^{\infty} \sup_{z\in[0,R]} \left|1 - \frac{z}{\bar{\lambda}_n}\right|^k |\bar{c}_n(\{k\}) - A_0(\{k\})| \geq \epsilon/2 \,|\, (T, N(T))_1^n\right)\\ &\leq \Pi_{\mathcal{D}(\Delta)}\left(\sum_{k=0}^{\infty}\left(1 + \frac{R}{\lambda_0 - O(n^{-\kappa})}\right)^k |\bar{c}_n(\{k\}) - A_0(\{k\})| \geq \epsilon/2 \,|\, N(T)_1^n\right),\end{aligned}$$

for some suitably chosen $\kappa > 0$. Hence the assertion of the theorem follows from Lemma 4.4.1 and Corollary 4.4.4. □

As an immediate consequence, one has the a.s. uniform convergence of the estimator $\hat{g}_n(z)$ on sets of the form $\{z \in \mathbb{R}_+ : z \leq R\}$ for some positive real number R.

Theorem 4.4.6. *For almost all data sequences $(T, N(T))_1^\infty$ and all $R > 0$ it holds true that*

$$\sup_{z\in[0,R]} |\hat{g}_n(z) - g_0(z)| \overset{n\to\infty}{\longrightarrow} 0.$$

Proof. Let $R > 0$ be an arbitrarily fixed positive real number. Then one has

$$\begin{aligned}&\sup_{z\in[0,R]} |\hat{g}_n(z) - g_0(z)| = \sup_{z\in[0,R]} \left|\gamma_n\left(1 - \frac{z}{\bar{\lambda}_n}\right) - a_0\left(1 - \frac{z}{\lambda_0}\right)\right|\\ &= \sup_{z\in[0,R]} \left|\sum_{k=0}^{\infty}\left(1 - \frac{z}{\bar{\lambda}_n}\right)^k \bar{c}_n(\{k\}) - \sum_{k=0}^{\infty}\left(1 - \frac{z}{\lambda_0}\right)^k A_0(\{k\})\right|\\ &\leq \sup_{z\in[0,R]} \sum_{k=0}^{\infty}\left|\left(1 - \frac{z}{\bar{\lambda}_n}\right)^k \bar{c}_n(\{k\}) - \left(1 - \frac{z}{\lambda_0}\right)^k A_0(\{k\})\right|\\ &\leq \sum_{k=0}^{\infty} \sup_{z\in[0,R]} \left|1 - \frac{z}{\bar{\lambda}_n}\right|^k |\bar{c}_n(\{k\}) - A_0(\{k\})| + \sum_{k=0}^{\infty} A_0(\{k\}) \sum_{i=0}^{k} R^{k-i} \left|\bar{\lambda}_n^{-(k-i)} - \lambda_0^{-(k-i)}\right|.\end{aligned}$$

Hence, the assertion of the theorem follows using above lemmas in combination with monotone convergence and continuous mapping theorems. □

Next, we point out that similar consistency properties hold for several derivatives of $\hat{g}_n(\cdot)$ as well as for other queueing characteristics mentioned earlier.

Theorem 4.4.7. *Let $f \in \{w, q, m\}$, $\hat{f}_n(z)$ be one of the estimators defined at the beginning of this section and $f_0(z)$ the true transform. Then, for any $R > 0$ and ϵ, one has*

$$\mathbb{P}\left(\sup_{0\leq z\leq R} \left|\hat{f}_n(z) - f_0(z)\right| > \epsilon \,|\, (T, N(T))_1^n\right) \overset{n\to\infty}{\longrightarrow} 0.$$

Moreover it holds

$$\sup_{0\le z\le R}\left|\hat{f}_n(z)-f_0(z)\right|\overset{n\to\infty}{\longrightarrow}0.$$

Proof. The assertion of the theorem follows directly by applying above results together with the continuous mapping theorem. The technical detail of the proof are similar to that of the proofs in section 3.3 and therefore omitted. □

4.4.3 Posterior normality

Another frequentist large sample property, which can be used as another justification of a certain estimator, is given by posterior consistency. Roughly speaking, this means that the posterior law of the object of interest, centered at its estimate and rescaled suitably, looks more and more like a Gaussian distribution. Result of that kind are useful for simulations and give first insight into convergence rates of the posterior. The first result in this direction was obtained in Conti (1999) where the author has proven that the posterior law of the p.g.f. of a random law drawn according to a Dirichlet process and centered at its Bayesian estimate converges towards a centered Gaussian process possessing a certain covariance structure. To be more precise, in the notation of the present work, it was obtained that under suitable constraints it holds that

$$\mathcal{L}\left(\sqrt{n}\left[a(z)-\gamma_n(z)\right]|X_1^n\right)\overset{n\to\infty}{\longrightarrow}\mathcal{L}\left(X(z)\right),$$

where $X(\cdot)$ is a centered Gaussian process with covariance structure $H(u,v)=a_0(uv)-a_0(u)a_0(v)$. Weak convergence, thereby, is considered on the space of continuous functions equipped with the sup-norm and the theorem holds for almost all data sequences X_1^∞, see also chapter 3. Moreover, it is easy to show in the parametric situation of the departure rate that for almost all data sequences T_1^∞ it holds that

$$\mathcal{L}\left(\sqrt{n}\left[\lambda-\bar{\lambda}_n\right]|T_1^n\right)\overset{n\to\infty}{\longrightarrow}\mathcal{N}(0,\lambda_0^{-2}).$$

Thus, combining these two results, one has

Theorem 4.4.8. *For almost all data sequences $(T,N(T))_1^\infty$ it holds that*

$$\mathcal{L}\left(\sqrt{n}\left[g(z)-\hat{g}_n(z)\right]|(T,N(T))_1^n\right)\overset{n\to\infty}{\longrightarrow}\mathcal{L}\left(G(z)\right),$$

on the space of continuous functions equipped with the sup-norm. Here, $G(z)$ is a centered Gaussian process with covariance structure

$$K(u,v)=H\left(1-\frac{u}{\lambda_0},1-\frac{v}{\lambda_0}\right)+uv\lambda_0^{-6}a_0'\left(1-\frac{u}{\lambda_0}\right)a_0'\left(1-\frac{v}{\lambda_0}\right).$$

Proof. We omit technicalities by mentioning that the proof works along the lines of proof of Conti (1999, Theorem 3). Thus, the assertion of the theorem follows from results obtained earlier. □

Applying this result, one is able to give similar results for the centering of the estimators $\hat{f}_n$ appearing in Theorem 4.4.7. In order to do so, the main work to do is to apply previous results and subsequently calculate the particular covariance structure analogously as in the proof of Theorem 3.4.8 However, since these are rather non-telling calculations which are similar to those already presented in section 3.4, we omit the details and leave them to the interested reader.

5 Inference for Stationary Data motivated by the $M/G/\infty$ Queue

5.1 Introduction

The first two parts of this thesis were devoted to the continuous-time queueing model with homogenous Poisson input and a server executing tasks of generally distributed random time intervals. However, there are situations where one server (or more generally finitely many servers) are not appropriate to think of. A classical example is that of cars entering and leaving a sparsely crowded motorway suggesting that the system is modeled without interactions of the customers. Furthermore, the emerging $M/G/\infty$ system might be taken in situations where one is certain in advance that the number of servers is so large that the system can provide immediate service to any customer. Since customers entering the system are served immediately, no queue builds up at all. Therefore the interest is not in waiting times of the customers but rather in the occupation of the system, i.e. how many cars are on the road, and the customers duration in the system, i.e. how long does a particular car stay on the motorway. Of course the duration then just amounts to the service time time which assumed to be distributed according to the general distribution G. Inference for G on basis of direct observations is a rather easy task which can be dealt with as in chapter 3 of the thesis. However, direct observations are often not appropriate to think of. This is mainly due to the facts that either the number of customers in the system is so large that it becomes impossible to track them all or just due to technical limitations. Going back to the example of cars on a motorway it might be both. So, one is interested in a way of making inference for G on basis of observations which can easily be accessed. However, data one has direct access to or which can be collected efficiently most often amounts to raw data which have to be worked up. Such data for instance might consist of records of the instants of arrivals and departures of customers, i.e. of instants when cars enter and leave the motorway. This setup was dealt with by Brown (1970) from a purely frequentist viewpoint. As a result that improves the possibility of making inference for G in an indirect way by great extent, Brown showed that G is representable as a function of the c.d.f. of a certain functional of the raw data consisting of arrival and departure instants. This functional is called the sequence of differences, which consists of the differences of the departure instances and the arrivals instances which occur directly before a particular departure. Plainly, those differences do not match the customers service times in general. Consciously, the term *functional* is used since one can think of the data to emerge from a function that maps two divergent sequences of real numbers onto another sequence of reals. The sequence of differences then was shown to be a stationary and ergodic sequence of random variables. Exploiting the obtained relationship of G and the c.d.f. of the sequence of differences, Brown defined estimators for G. However, these estimators are based on an estimator for the c.d.f. of the sequence of differences which,

roughly speaking, is some version of the empirical measure. Put differently, Brown pretended as the data was i.i.d. which it is certainly not. He remarked that fact at the end of his work by stating that *"it is clearly not the best estimator in any sense because we do not use all the information"*. Obviously the information which wasn't used consists of the interplay of the projections of the data. Simply put, the way he used the data only asked how often does a particular length of data appear rather then when (in the sense of the structure of the data) do a certain length appear. From that perspective the estimator provided by Brown is the most workable one but at the same time it is the worst one can think of.

The problem of embedding these ideas into a Bayesian framework is manifold. The major point is that a Bayesian's considerations have to start where Brown's came to an end. This is quite natural and due to the fact that a Bayesian has to clarify the mixing measure before being able to model it in some workable way. Anyway, from a Bayesian perspective, the data in form of the sequence of differences is not ergodic at all. As observed data one has to express her uncertainty in the measure that might have generated the sequence of differences in form of an integral mixture. A theoretically deep problem emerges from the fact that the space of possible measures, i.e. the space of shift-ergodic measures, is enormous. Another problem is that the integration with respect to some appropriate mixing measure has to be clarified. This is usually done by a suitable parametrization. However, those problems become even harder if the space the data takes values in becomes more general. Hence, in order to not lose track of the problem, the state space will be assumed to be a finite set. For the original problem that means that the length of the differences has to be finitely categorized. A parametrization is then given by using the algebraic idea of the inverse limit, which yields an appropriate parameter space. Subsequently, a reasonable prior is provided under several assumptions on the sampling scheme of the data and its update to the posterior is defined.

The chapter is organized in several sections. Section 5.2 recalls the paper of Brown (1970) in full extent. In section 5.3 the mathematical preliminaries are presented in order to clarify the problem and to describe the difficulty of making Bayesian inference for stationary data. Section 5.4 is devoted to the issue of describing the magnitude of equivalence classes of binary that emerge from a statistical judgment which departs from independence. This generalizes the well known fact that the equivalence classes of exchangeable binary data have $\binom{n}{k}$ elements, where $k \leq n$ is the number of "successes" among a data string of length n. The subsequent sections deal with the issue of finding a suitable parameter space of the shift-ergodic measures under consideration (section 5.5) as well as an explicit model for a prior distribution (section 5.6). In section 5.6 it is also demonstrated how the prior may be updated through observed stationary data.

5.2 Background from Brown's work on $M/G/\infty$ queues

The statistical analysis of stochastic systems often becomes a sophisticated task due to relatively general structures appearing for which one would have to provide tools for statistical inference. The more is known about the structure in advance, i.e. "smaller" the

model is in a particular sense of dimension which is to specify, the more manageable the statistical evaluations become. That is due to the fact that often "finite-dimensional" models do possess a neat parametrization of the stochastic mechanism generating the data. However, if the data being observed can not be assumed to be generated by a "finite generator" it is in question what meaning a parametrization or a parameter, respectively, does have.

The theory of queueing systems will serve as a motivating field for the rather theoretical considerations in this text. In queueing theory general stationary processes do appear frequently and from a statistical viewpoint one is often intended to make inference for them. However, directly eliciting a suitable prior distribution for laws governing these processes is apparently a difficult task as well as the formalization of the learning process induced by observation of data. The main reason for this is that one has to express one's uncertainty in the structure among the data itself if it is not known in advance.
For instance, if one is to make inference for the waiting time distribution or the queue length distribution in a $M/G/1/\infty$ system by only observing the service times and the inter-arrival times of customers, one can employ indirect methods which involve particular transforms of those distributions. This approach, which consists of a functional relationship, allows to bridge inference methods for the distributions of the observables to the distributions of interest by continuity and differentiability properties of the functional. The point is that the relatively simple stochastic structures of the input variables are exploited to provide inference procedures for the structurally complicated ones. However, neither gives this a possibility for making inference for the structure of the target variables, nor is such an approach always available. An example where it is not is provided by the following (frequentistic) statistical setup of an $M/G/\infty$ system which is due to Brown (1970) and which is briefly reviewed in the following.

Suppose an $M/G/\infty$ system is given. Since the number of servers is arbitrarily large, each customer can be viewed to have her own server. Hence, no queue builds up and no customers waiting times arise. Instead, the interest may lie in the number of customers being present in the system or the duration of a customer, respectively. The original motivation in the paper of Brown was to model cars on a highway. The only observation one has access to are the times when a car enters and leaves the highway, respectively, and the main interest is in the service time distribution, i.e. in the distribution that governs the time a car spends on the highway. For the sake of simplicity interactions of cars as in the form of traffic jams are omitted. On a first glance the problem hardly seems to be solvable. Yet, Brown found an ingenious way to handle it. The basic idea is to schedule the instants of arrivals to the system and departures from the system the following way. Let $(\Omega, \mathcal{A}, \mathbb{P})$ denote an abstract probability space. Furthermore, let $(A_i : \Omega \to \mathbb{R})_{i\in\mathbb{Z}}$ and $(D_i : \Omega \to \mathbb{R})_{i\in\mathbb{Z}}$ denote the sequence of random time points of customer's arrivals and departures, respectively, occurring to a $M/G/\infty$ system. Thus If a departure D is observed, it is matched to the nearest arrival being less than D which will be denoted as A_D. Thus, $A_D = \sup\{A_i : A_i < D\}$. Define the sequence of differences $(Z_i)_{i\in\mathbb{Z}}$ by $Z_i := D_i - A_{D_i}$. The assumption of infinitely extendability of the system indeed ensures that the sequence (Z_i) is actually infinitely long. Plainly, (Z_i) does not necessarily reproduce the duration of a customer in the system since the customer departed at D does not have to have arrived to the system at time point A_D. This is an issue being dealt with later on. First note a statement about the sequence (Z_i).

Lemma 5.2.1 (Brown). *The sequence of differences $(Z_i)_{i\in\mathbb{Z}}$ is stationary and ergodic.*

This is a remarkable result even if not that surprising on a second glance. Roughly, since the input stream is a Poisson process with rate $\lambda > 0$, i.e. the inter-arrival times of the customers are assumed to be independent and identically distributed according to an exponential distribution with rate λ, and the service times are assumed to form an i.i.d sequence following a general distribution the functional (Z_i) has no chance to depart from stationarity. Keep in mind that any i.i.d. sequence of random variables is stationary (to be defined rigorously below), which is, roughly speaking, due to the factorization of their joint law into a product measure. For a random variable $X : \Omega \to \mathbb{R}$ denote by $\mathcal{L}[X]$ the law of X. Put another way, $\mathcal{L}[X]$ is the push-forward of $\mathbb{P}$ under the mapping X. By stationarity, it holds that $\mathcal{L}[Z_i]$ is the same for all $i \in \mathbb{Z}$. Hence, the c.d.f. of Z_i is the same for all $i \in \mathbb{Z}$ and will be denoted by $H : \mathbb{R}_+ \to [0, 1]$.

Brown's paper moves on by proving a functional relation ship between the H and the distribution G of the time spent by the customers in the system.

Lemma 5.2.2 (Brown). $G(x) = 1 - (1 - H(x))e^{\lambda x}$.

This functional relationship enables one to transfer statistical inference for H to that of G which was of original interest. For the statistical evaluation of H, he chose a purely frequentist approach. The method is a nonparametric one that uses $\hat{H}_n(x) := 1/n \sum_{k=0}^{n-1} 1_{[0,Z_i]}(x)$ as an estimator for $H(x)$. Further, $\hat{V}_n(x) := 1 - (1 - H_n(x))e^{\hat{\lambda}_n x}$ and $\hat{G}_n(x) := \sup_{0\le y\le x} \hat{V}_n(y)$ serves as a plug-in estimate for G, where $\hat{\lambda}_n$ is the sample intensity. The estimator is then justified by a consistency result.

Theorem 5.2.3 (Brown). *(i) $\hat{H}_n \xrightarrow{n\to\infty} H$ a.s. uniformly,*

(ii) $\hat{V}_n \xrightarrow{n\to\infty} G$ a.s. uniformly on compact intervals,

(iii) $\hat{G}_n \xrightarrow{n\to\infty} G$ a.s. uniformly.

Thus, the defined estimators are able to recover G arbitrarily precise in a certain metric as the amount of data increases. The proof of (i) of that theorem mainly relies on exploiting of the ergodic theorem [see e.g. Petersen (1989)] applied to the function $1_{[0,Z_i]}(x)$. Notice that it is possible to apply the ergodic theorem since (Z_i) was shown to be stationary and ergodic. Assertion (ii) is an easy consequence of (i) and (iii) is a dodge which allows to define an estimator that converges "globally uniformly" instead of only "locally uniformly". See the paper of Brown for further details.

In a brief discussion at the end of his paper Brown himself pointed out that *it is clearly not the best estimator in any sense because we do not use all the information.* Even if Brown did not further specify what information he exactly meant to be dropped, it is almost obvious what it is. The estimator $\hat{H}_n$ only processes "absolute" information about the length of an observed difference Z but not the information "when" a certain length occurs with respect to the other observations. More precisely, the estimator pretends as

the observed data $Z_1^n := (Z_1, \ldots, Z_n)$ were i.i.d. and leaves out the dependence structure among the projections Z_i. Thus, if one is to improve the estimator one would have to embed the problem into a larger framework in a way that includes the dependency of the data.

5.3 Theoretical Preliminaries

If one would like to undertake inferential analysis from the Bayesian viewpoint for the statistical setup just presented, the problem indicated by Brown appears at the very first rather than at the end of some frequentistic procedure. The reason therefor is the subjectivistic approach of probability which builds the theoretical fundament of Bayesian statistics. It forces one to approach the problem from a less direct and more theoretical direction because a Bayesian has to elicit a prior distribution on the space of measures that possibly generated the data. Since the data forms a (general) stationary sequence of random variables, this space cannot be shrunk to that of all i.i.d measures in general as was done tacitly by Brown's frequentist approach. This rather heuristical explanation will be described more precisely subsequently in the sequel.

Suppose a sequence of random quantities $(X_i : \Omega \to \mathcal{X})_{i\in\mathbb{N}}$ with state space $\mathcal{X}$ is given. In order to ensure $\mathcal{X}$ to possess necessary measure-theoretical structure, throughout it will be assumed that $\mathcal{X}$ is a Polish space. That is a completely metrizable separate topological space, i.e. its topology is induced by a certain metric, any Cauchy sequence w.r.t. this metric has a limit point in $\mathcal{X}$ and the topology has a countable basis. Endow $\mathcal{X}$ with the Borel σ-field $\mathfrak{B}_\mathcal{X}$ in order to make it a measurable space. Call the sequence (X_i) stationary if $\mathcal{L}[X_1, \ldots, X_n] = \mathcal{L}[X_k, \ldots, X_{k+n}]$ for all $n \in \mathbb{N}$ and $k \in \mathbb{N}_0$. That is (X_i) is stationary if the joint law of any of its subfamily is the same as the joint law of the subfamily after having shifted the index set arbitrarily. This perception allows a more general description in terms of ergodic theory which is preferable. Let $\mathcal{X}^\mathbb{N}$ denote the sequence space with entries in $\mathcal{X}$. As a product of Polish spaces, $\mathcal{X}^\mathbb{N}$ is Polish itself and one can take the Borel σ-field $\mathfrak{B}_{\mathcal{X}^\mathbb{N}}$ as measurable structure on the sequence space. Recall $\mathfrak{B}_{\mathcal{X}^\mathbb{N}} = \bigotimes_{k\in\mathbb{N}} \mathfrak{B}_\mathcal{X}$ and that a compatible metric is given by the Fréchet-metric

$$d(x, y) = \sum_{k=1}^{\infty} \frac{d_\mathcal{X}(x_k, y_k)}{2^{k+1}(1 + d_\mathcal{X}(x_k, y_k))}.$$

Let $\mathcal{P}(\mathcal{X}^\mathbb{N})$ be the space of all probability measures on $\mathcal{X}^\mathbb{N}$ and endow this space with the topology induced by weak convergence. That is take as neighborhood basis of $\mu \in \mathcal{P}(\mathcal{X}^\mathbb{N})$ sets of the form $\{\nu \in \mathcal{P}(\mathcal{X}^\mathbb{N}) : |\int f_i d\nu - \int f_i d\mu| < \epsilon; i = 1, \ldots, m; f_i \in \mathcal{C}_b; \}$. The topology defined that way in turn induces the Borel σ-field $\mathfrak{B}_\mathcal{P}$ which is known to be the smallest σ-field which makes the mappings $\mu \mapsto \mu(A)$ measurable for all $\mu \in \mathcal{P}(\mathcal{X}^\mathbb{N})$ and all $A \in \mathfrak{B}_{\mathcal{X}^\mathbb{N}}$. Moreover, it is well known that $\mathcal{P}(\mathcal{X}^\mathbb{N})$ is a Polish space if $\mathcal{X}$ is, see e.g. Kechris (1995). A compatible metric is given through the celebrated Prohorov-metric, c.f. Billingsley

(1999), or the metric defined by

$$d_{\mathcal{P}}(\mu,\nu) = \sum_{k=1}^{\infty} \frac{|\int f_k d\mu - \int f_k d\nu|}{\|f\|_\infty},$$

where $\{f_k\} \subset \mathcal{U}_d \subset \mathcal{C}_b$ is a suitably chosen subset of uniformly continuous (w.r.t. d) and bounded functions, c.f. Kechris (1995), and $\|f\|_\infty$ denotes the sup-norm of f.

Next, define the *shift* T on $\mathcal{X}^{\mathbb{N}}$ to be the mapping

$$\begin{aligned} T : \mathcal{X}^{\mathbb{N}} &\to \mathcal{X}^{\mathbb{N}} \\ x = (x_1, x_2, \dots) &\mapsto T(x) := (x_2, x_3, \dots), \end{aligned}$$

i.e. T affects the indices of the sequence x such that $k \mapsto k+1$ which results in a shifting of x to the left. T is a continuous onto map as is easily seen using the distance d. Therefore T is measurable with respect to $\mathfrak{B}_{\mathcal{X}^{\mathbb{N}}}$.

The mapping T naturally induces an operator on $\mathcal{P}(\mathcal{X}^{\mathbb{N}})$ by forming push-forward measures under T. This approach gives the so called *shift-operator*. More precisely, define the operator

$$\begin{aligned} T : \mathcal{P}(\mathcal{X}^{\mathbb{N}}) &\to \mathcal{P}(\mathcal{X}^{\mathbb{N}}) \\ P &\mapsto T[P] := TP, \end{aligned}$$

where the measure TP is defined as $TP(B) := P(T^{-1}B) := P(\{x \in \mathcal{X}^{\mathbb{N}} : T(x) \in B\})$ for all $B \in \mathfrak{B}_{\mathcal{X}^{\mathbb{N}}}$. Since there is hardly a danger of confusion we will use T for both the mapping on $\mathcal{X}^{\mathbb{N}}$ and $\mathcal{P}(\mathcal{X}^{\mathbb{N}})$. It is easily seen that the operator T is continuous and affine, i.e. it holds that $T[\sum_{i\in I} a_i\mu_i] = \sum_{i\in I} a_i T\mu_i$ for any $a_i \in [0,1]$, $\sum_{i\in I} a_i = 1$ and $\mu_i \in \mathcal{P}(\mathcal{X}^{\mathbb{N}})$, where I is an arbitrary index set. A physical interpretation is of this is that T preserves barycenters if the involved measures are stationary.

Call a $\mu \in \mathcal{P}(\mathcal{X}^{\mathbb{N}})$ *stationary* if $T\mu = \mu$, in words if the measure is invariant with respect to the shift operator. Let $\mathcal{P}(\mathcal{X}^{\mathbb{N}}, T) \subset \mathcal{P}(\mathcal{X}^{\mathbb{N}})$ denote the set of stationary measures. Note that there is a geometrical description of the appearance and a topological description of the size of $\mathcal{P}(\mathcal{X}^{\mathbb{N}}, T)$ with respect to $\mathcal{P}(\mathcal{X}^{\mathbb{N}})$. That is the set of stationary measures is a compact convex and nowhere-dense subset of $\mathcal{P}(\mathcal{X}^{\mathbb{N}})$. Nowhere-density of a set $K \subset S$ relative to S means that the interior of the closure of K is just the empty set, $\bar{K}^o = \varnothing$. Thus a non-stationary probability measure cannot be approximated arbitrarily close by stationary measures.

As an introductory example of a stationary measure take $\delta_{\bar{x}}(\cdot)$, where $\bar{x} := (x, x, x, \dots)$ for a $x \in \mathcal{X}$ and $\delta_{\bar{x}}$ denotes the Dirac measure with atom 1 at $\bar{x}$. Thus for any $B \in \mathfrak{B}_{\mathcal{X}^{\mathbb{N}}}$ it is $\delta_{\bar{x}}(B) = 1$ if and only if (iff) $\bar{x} \in B$ and $T\delta_{\bar{x}}(B) = \delta_{\bar{x}}(\{y \in \mathcal{X}^{\mathbb{N}} : y \in T^{-1}B\}) = 1$ iff $\bar{x} \in T^{-1}B$. Since $\bar{x} \in B \Leftrightarrow \bar{x} \in T^{-1}B$ stationarity of $\delta_{\bar{x}}$ is readily obtained. Stationary measures of that certain form are of special interest because they form the set of extreme measures of $\mathcal{P}(\mathcal{X}^{\mathbb{N}})$ which are stationary at the same time. Call a measure $\mu \in \mathcal{Q} \subset \mathcal{P}(\mathcal{X}^{\mathbb{N}})$

extreme (in $\mathcal{Q}$) if all convex mixtures of measures in $\mathcal{Q}$ yielding μ are trivial ones, more formally if $\mu = \sum_{i\in I} a_i\mu_i \Rightarrow \mu_i = \mu, \forall i \in I$, where $\{a_i\}$ is as above and $\mu_i \in \mathcal{Q}$ for all $i \in I$. The set of all extremes in $\mathcal{P}(\mathcal{X}^{\mathbb{N}})$ is given by $ex[\mathcal{P}(\mathcal{X}^{\mathbb{N}})] = \{\delta_x : x \in \mathcal{X}^{\mathbb{N}}\}$ while $ex[\mathcal{P}(\mathcal{X}^{\mathbb{N}})] \cap \mathcal{P}(\mathcal{X}^{\mathbb{N}}, T) = \{\delta_{\bar{x}} : x \in \mathcal{X}\}$. Plainly, one also has $\{\delta_{\bar{x}} : x \in \mathcal{X}\} \subset ex[\mathcal{P}(\mathcal{X}^{\mathbb{N}}, T)]$.

5.3.1 Choquet Theory

Extreme points of a certain set are of special interest because they generate the convex hull of this set by mixing up extreme elements properly and, applied to sets of measures, have a particular statistical meaning. The first-mentioned is made more precise by the theory of Choquet simplices. As a result one has the celebrated representation theorem named after Choquet. For a proof and more details see e.g. Phelps (2001).

Theorem 5.3.1 (Choquet, 1956)**.** *Let $K \subset X$ be a metrizable compact and convex subset of a separable locally convex space X. Then, for every point $x \in X$ there is a unique probability measure $m \in \mathcal{P}(K)$ with $supp[\rho] = ex[K]$ such that ρ represents x, that is*

$$x = \int_{ex[K]} ym(dy).$$

If applied to $X = \mathcal{P}(\mathcal{X}^{\mathbb{N}})$, which is seen as embedded in the locally convex vector space of all signed measures, and $K = \mathcal{P}(\mathcal{X}^{\mathbb{N}}, T)$, Theorem 5.3.1 yields a mixing measure that is supported by the extremes of $\mathcal{P}(\mathcal{X}^{\mathbb{N}}, T)$. However, if one is to make Bayesian statistical inference for stationary data, there is need for a more statistical interpretation of the theory just presented. A first step is to identify $ex[\mathcal{P}(\mathcal{X}^{\mathbb{N}}, T)]$ with the set of so called ergodic measures. In order to give a definition call a set $B \in \mathfrak{B}_{\mathcal{X}^{\mathbb{N}}}$ invariant with respect to the mapping T (or in short T-invariant or just invariant) if $T^{-1}B = B$ and notice that the collection of such sets forms a σ-field denoted by $\mathcal{I} = \mathcal{I}_T$ usually called the invariant σ-field. A stationary measure P is clearly invariant on $\mathcal{I}$ since for $I \in \mathcal{I}$ it holds $TP(I) = P(T^{-1}I) = P(I)$. However, this not true for all $B \in \mathfrak{B}_{\mathcal{X}^{\mathbb{N}}}$ in general. Now, call a stationary measure P *ergodic* with respect to T (in short T-ergodic or just ergodic) if for all $I \in \mathcal{I}$ it holds $P(I)P(I^c) = 0$. Thus for an ergodic measure either an invariant event or its complement, respectively, is certain. Write $\mathcal{P}^e(\mathcal{X}, T)$ for the set of all T-ergodic measures. Regard above exemplary measure $\delta_{\bar{x}}$ and let I be an invariant set. By definition $\delta_{\bar{x}}(I) \in \{0, 1\}$ such that $\delta_{\bar{x}}$ is ergodic. This rises the question if it is true for all extreme points of $\mathcal{P}(\mathcal{X}^{\mathbb{N}}, T)$ to be ergodic. It is well known that this question has an affirmative answer. Due to the author's lack of an appropriate source for citation the proof is given. However, it can be found in the literature.

Lemma 5.3.2. $ex[\mathcal{P}(\mathcal{X}, T)] = \mathcal{P}^e(\mathcal{X}, T)$.

Proof. Let $\mu \in ex[\mathcal{P}(\mathcal{X}^{\mathbb{N}}, T)]$ and $I \in \mathcal{I}$ with $\mu(I) > 0$. Define for $B \in \mathfrak{B}_{\mathcal{X}^{\mathbb{N}}}$ the trace measure of I which is given by $\nu(B) := \frac{\mu(B\cap I)}{\mu(I)}$. Then $\nu \in \mathcal{P}(\mathcal{X}^{\mathbb{N}}, T)$. Suppose $\mu(I) < 1$, then $\mu(\mathcal{X}^{\mathbb{N}} \setminus I) > 0$ and such that $\rho(B) := \frac{\mu(B\setminus I)}{\mu(\mathcal{X}^{\mathbb{N}}\setminus I)} \in \mathcal{P}(\mathcal{X}^{\mathbb{N}}, T)$ is properly defined. Hence, $\mu(\cdot) = \mu(I)\nu(\cdot) + (1 - \mu(I))\rho(\cdot)$ which is a contradiction. For the converse, suppose $\mu \in \mathcal{P}^e(\mathcal{X}, T)$ and regard the non-trivial (i.e.$0 < a < 1$) representation $\mu = a\nu + (1 - a)\rho$.

Then $\nu, \rho << \mu$ and the Radon-Nikodym derivatives $f_\nu := \frac{d\nu}{d\mu}$ and $f_\rho := \frac{d\rho}{d\mu}$ exist. Since $I_+ := \{f_\nu - f_\rho > 0\}, I_- := \{f_\nu - f_\rho < 0\} \in \mathcal{I}$, one has $\mu(I_+), \mu(I_-) \in \{0,1\}$. Assume w.l.o.g. that $\mu(I_+) = 1 (\Rightarrow \mu(I_-) = 0)$, then

$$1 = \nu(\mathcal{X}^{\mathbb{N}}) = \int_{\mathcal{X}^{\mathbb{N}}} f_\nu d\mu = \int_{I_+} f_\nu d\mu > \int_{I_+} f_\rho d\mu = \int_{\mathcal{X}^{\mathbb{N}}} f_\rho d\mu = \rho(\mathcal{X}^{\mathbb{N}}) = 1.$$

Thus $f_\nu = f_\rho$ μ-a.s. which in turn implies $\nu = \rho$. □

5.3.2 Ergodic Decomposition

By Theorem 5.3.1 in combination with Lemma 5.3.2 one has that any $\mu \in \mathcal{P}(\mathcal{X}^{\mathbb{N}}, T)$ is representable as integral mixture of T-ergodic measures. This is also known as ergodic decomposition and serves a a certain kind of integral-limit interchanging result. As an example, suppose that for $0 < a < 1$ a stationary measure is given by $\mu = a\nu + (1-a)\rho$, where and $\nu, \rho \in \mathcal{P}^e(\mathcal{X}^{\mathbb{N}}, T)$. Thus, $\mu(\cdot) = \int_{\mathcal{P}^e(\mathcal{X}^{\mathbb{N}}, T)} \sigma(\cdot) m(d\sigma)$, where the mixing measure is given as $m = a\delta_\nu + (1-a)\delta_\rho$. Then, for an μ integrable function f it holds that

$$\int_{\mathcal{X}^{\mathbb{N}}} f(y)\mu(dy) = \int_{\mathcal{X}^{\mathbb{N}}} f(y)[a\nu + (1-a)\rho](dy) = a\int_{\mathcal{X}^{\mathbb{N}}} f(y)\nu(dy) + (1-a)\int_{\mathcal{X}^{\mathbb{N}}} f(y)\rho(dy)$$
$$= \int_{\mathcal{P}^e(\mathcal{X}^{\mathbb{N}}, T)} \left[\int_{\mathcal{X}^{\mathbb{N}}} f(y)\sigma(dy)\right] m(d\sigma)$$

and the questions arises if a similar result remains to be true for more general mixtures of ergodic measures. The ergodic decomposition theorem gives an affirmative answer.

Theorem 5.3.3 (Ergodic Decomposition)**.** *The state space $\mathcal{X}^{\mathbb{N}}$ admits an essentially unique decomposition map, i.e. a measurable mapping $g : \mathcal{X}^{\mathbb{N}} \to \mathcal{P}^e(\mathcal{X}^{\mathbb{N}}, T)$ for which it holds*

(i) g is T-invariant, that is $g = g \circ T^{-1}$,

(ii) for any $m \in \mathcal{P}^e(\mathcal{X}^{\mathbb{N}}, T)$ the pre-image $g^{-1}(\mu)$ is a measurable set with full m-mass,

(iii) for all $\mu \in \mathcal{P}(\mathcal{X}^{\mathbb{N}}, T)$ and all $B \in \mathfrak{B}_{\mathcal{X}^{\mathbb{N}}}$ it holds

$$\mu(B) = \int_{\mathcal{X}^{\mathbb{N}}} g_x(B)\mu(dx).$$

Moreover, for any μ-integrable function $f : \mathcal{X} \to \mathbb{R}$ one has

$$\int_{\mathcal{X}^{\mathbb{N}}} f(y)\mu(dy) = \int_{\mathcal{X}^{\mathbb{N}}} \left[\int_{\mathcal{X}^{\mathbb{N}}} f(y) g_x(dy)\right] \mu(dx).$$

A proof of Theorem 5.3.3 can be found in Varadarajan (1963) and Maitra (1977) as well as further generalizations to more general spaces and groups of transformations. In the

proof it becomes clear how one can think about the decomposition map and what exactly is meant by "essentially unique". More precisely, $g_x(\cdot)$ can be identified with the regular conditional probability $\mu(\cdot \mid \mathcal{I})(x)$ which is μ-a.s. unique. Thus, the extreme points of $\mathcal{P}(\mathcal{X}^{\mathbb{N}}, T)$ can be identified with "measures having seen an infinite amount of information", i.e. the joint distribution $\mathcal{L}[X]$ of a sequence $X = X_1^\infty$ is extreme iff $\mathcal{L}[X]$ is a.s. non-random (as a variant of Kallenberg (2006, Proposition 1.4)).

5.3.3 Statistical Interpretations and Definitions

Here the main similarity and the main difference of the frequentistic and subjectivistic interpretation of statistics appears. Both factions assume that the data being observed is generated from an ergodic measure. But while a frequentist says that this generating measure is "fixed but unknown", this is not satisfactory to a Bayesian since she wants to express her uncertainty in the generating measure. This uncertainty is expressed by the mixing measure $m \in \mathcal{P}(\mathcal{P}^e(\mathcal{X}^{\mathbb{N}}, T))$ which appears in Theorem 5.3.1 and which is just a Dirac measure with atom at the "unknown" probability measure in the frequentist framework. As a result, a Bayesian feels like observing (non-ergodic) stationary data while the same data is interpreted as T-ergodic from a frequentist viewpoint. The task of a Bayian statistician now is to "model" m suitably such that one is able to update it by observed data. This is a hard task in general since $\mathcal{P}^e(\mathcal{X}^{\mathbb{N}}, T)$ is a very large space and it is not known how to calculate the mixing integral in general. Even if additional structural assumptions on the data can shrink the set of all ergodic measures to one that is well manageable, such an assumption is not appropriate in general. However, studying such situations can give an idea how one would have to proceed in the general case.

The easiest example arises when the additional structure is given in form of an invariance assumption with respect to the group of all index permutations Π. Roughly speaking, while T-invariance means that time does not matter, i.e. "it does not matter when data is collected", invariance with respect to Π means that in addition "it does not matter in which order data is observed". This additional judgement lets the support of the mixing measure m shrink to the set of ergodic measures which support this judgment. By the Hewitt-Savage 0/1-law [c.f. Hewitt and Savage (1955)] these are just the i.i.d. measures, i.e. the measures that make the projections of elements of $\mathcal{X}^{\mathbb{N}}$ i.i.d. variables. The property of independence lets an i.i.d. measure μ factorize while the assumption of identically distributed projections ensures that the components of this factorization are all the same. So that $\mu = \bigotimes_{i \in \mathbb{N}} \tilde{\mu}$ for a particular $\tilde{\mu} \in \mathcal{P}(\mathcal{X})$ and it suffices to make inference for $\tilde{\mu}$ by considering an updating mixture measure $\tilde{m} \in \mathcal{P}(\mathcal{P}(\mathcal{X}))$. By suitable parametrization of $\mathcal{P}(\mathcal{X})$, for instance by the dense subspace of all discrete measures, one can obtain statistically workable mixing distributions (priors), see e.g. Ferguson (1973) and Sethuraman (1994). It is worth mentioning that the i.i.d. measures are the ergodic measures $g_x(\cdot) = \mu(\cdot \mid \mathcal{E})(x)$ in the decomposition theorem, where $\mathcal{I} \subset \mathcal{E}$ is the exchangeable σ-field, c.f. Kallenberg (2006).

This example shows that a suitable parametrization can simplify the task of making Bayesian inference considerably. However, further examples, which generalize the invari-

ance property of the data, are rather complicated to study for general state spaces $\mathcal{X}$. Thus, in the subsequent it will be assumed that $\mathcal{X}$ is a finite set, often $\mathcal{X} = \{0,1\}$ which leads to the Cantor space $\mathcal{C} = \{0,1\}^{\mathbb{N}}$ and accounts to binary response of an experiment. Admittedly, this is a slight loss in generality and lets one depart from the original task (of making inference for the distribution of the sequence of differences). However there will be a big gain in the presentation and clarifying of a parametrization of the set of ergodic measures in general. The above considerations for Polish spaces clearly remain true if $|\mathcal{X}| < \infty$.

As already mentioned above, the key for Bayesian statistics is to give a suitable parametrization of the set of ergodic measures one wishes to define a prior distribution on. Thus, the task is to head for a parametrization of $\mathcal{P}^e(\mathcal{X}^{\mathbb{N}}, T)$. If $\mathcal{X} = \{0,1\}$ the i.i.d. measure coincide with the so called Bernoulli-measures which possess a simple parametrization. To be more precise, let for $c_1, \dots, c_k \in \{0,1\}$ the cylinder set generated by the c_i's be given through $[c_1, \dots, c_k] = \{x \in \mathcal{C} : x_i = c_i, \forall i = 1, \dots, k\}$. It is well known that the collection of all possible cylinder sets $\mathcal{Z}$ is a semi-algebra which generates $\mathfrak{B}_{\mathcal{C}}$. Then, define

$$\mu_k([c_1, \dots, c_k]) := \prod_{i=1}^{k} p_{c_i},$$

where $p_1 = 1 - p_0 = p$ for a number $p \in [0,1]$. The family $\{\mu_k\}_{k \geq 1}$ can be uniquely extended to a probability $\mu \in \mathcal{P}^e(\mathcal{C}, T)$ That means $\mathcal{P}^e(\mathcal{C}, \Pi)$ is parametrized by $[0,1]$. In combination with Theorem 5.3.1 the parametrization yields for $P \in \mathcal{P}(\mathcal{C}, \Pi)$ the unique existence of $m \in \mathcal{P}(\mathcal{P}^e(\mathcal{C}, \Pi))$ with

$$\begin{aligned} P([c_1, \dots, c_k]) &= \int_{\mathcal{P}^e(\mathcal{C},\Pi)} Q([c_1, \dots, c_k]) m(dQ) = \int_{[0,1]} \prod_{i=1}^{k} p_{c_i} \tilde{m}(dp) \\ &= \int_{[0,1]} p^{\sum_{i=1}^{k} c_i} (1-p)^{k - \sum_{i=1}^{k} c_i} \tilde{m}(dp), \end{aligned}$$

which amounts to the classical de Finetti theorem. From a statistical viewpoint any distribution on $[0,1]$ can be used as a prior to express one's state of information which is then updated by observed data through Bayes theorem. The Beta distribution plays a special role since it is conjugate with respect to Bernoulli-data and it supports all of the unit interval. Before continuing with structurally more complicated considerations, an exact definition of the terms *parametrization* and *parameter space* shall be given.

Definition 5.3.4. *Let $\mathcal{P}^e(\mathcal{C}, G)$ be the set of measures which are ergodic with respect to a suitable family of transformations G. Further, let $(S, \mathcal{S})$ be a measurable space. A mapping $\phi : \mathcal{P}^e(\mathcal{C}, G) \to S$ is called a parametrization if it is bimeasurable (one-to-one and both ϕ and ϕ^{-1} are measurable). In that case $(S, \mathcal{S})$ is called a parameter-space of $\mathcal{P}^e(\mathcal{C}, G)$.*

So, in above example of mixtures of Bernoulli measures, $[0,1]$ is a parameter space with corresponding parametrization ϕ given by

$$\begin{aligned} \phi^{-1} : [0,1] &\to \mathcal{P}^e(\mathcal{C}, \Pi) \\ p &\mapsto \phi^{-1}(p) := \bigotimes_{i \in \mathbb{N}} [p\delta_1(\cdot) + (1-p)\delta_0(\cdot)] \end{aligned}$$

and $\tilde{m} = \phi m$. Moreover, notice that from a statistical point of view the points $0, 1$ might be excluded from $supp[\tilde{m}]$ since they correspond to statistically uninteresting measures, i.e. to reducible Bernoulli-measures. Those are Bernoulli-measures for which the state space can be reduced without any effect, e.g. notice that $supp[\phi^{-1}(0)] = \bar{0}$. Furthermore, suppose a prior distribution with an atom on 0 is given. Then updating to the posterior by observing a 1 at any place would let vanish this atom.

5.3.4 Turning to Dependent Data

Next, further families of measures shall be described which are invariant with respect to T but possess "less" additional invariance than exchangeables. Think of these measures to be of the form $\mu(\cdot \mid \mathcal{I} \bigvee \mathcal{G})$ for a particular σ-field $\mathcal{G}$ such that $\mathcal{I} \subset \mathcal{I} \bigvee \mathcal{G} \subset \mathcal{I} \bigvee \mathcal{E} = \mathcal{E}$. Put another way, the invariance property is relaxed to a subfamily $\beta \subset \Pi$, where Π denotes the family of all coordinate permutations. More precisely, this subfamily β consists of all finite permutations that swap two blocks of symbols which begin and end with the same symbol. Since those transformations leave the first symbol as well as the number of transitions from a particular state into another one fixed, one is led to think about mixtures of Markov laws. The work of Diaconis and Freedman (1980) does confirm this idea. However, this paper deals with laws which are recurrent but not necessarily stationary from a rather stochastic viewpoint. There was an earlier approach appearing in Freedman (1962) which deals solely with stationary measures rather from a statistical perspective. Moreover, it contains a result that clarifies mixtures possessing a certain structure, the so called S-structure. The definition is as follows. Let t be a statistic, i.e. a family of measurable mappings $t = \{t_n : \mathcal{X}^n \to E_n\}_{n \in \mathbb{N}}$ for some suitable subspace E_n of an Euclidean space which is equipped with a measurable structure $\mathcal{E}_n$. Note that for many statistical problems it is true that $t_{n+1}(x_1^{n+1}) = t_n(x_1^n) + f_{n+1}(x_1^n, x_{n+1})$ for a suitable family of functions $f = \{f_n\}$. The family f then indicates the structure among the data, for instance the length of memory the data possesses. A more precise way to express this is to say that there is a family of transition kernels $r = \{r_n : \mathfrak{B}_{\mathcal{X}} \times E_{n-1} \to [0,1]\}_{n \in \mathbb{N}}$ such that for any $n > 1$ and $e^{(n-1)} \in \mathcal{E}_{n-1}$ it holds that $r_n(t_n^{-1}(e^{(n-1)}), e^{(n-1)}) = 1$ and if $r_{n+1}(\cdot, t)$ is a distribution of the data X_{n+1} for some $t \in \mathcal{E}_n$ then $r_n(\cdot, e^{(n-1)})$ is a regular conditional distribution of X_n given $t_{n-1}(X_1^{n-1}) = e^{(n-1)}$ for all $e^{(n-1)} = t_n(pr_{1:n}(t_{n+1}^{-1}(\{t\})))$, where $pr_{1:n}$ denotes the projection onto the first n symbols. For more details see e.g. Schervish (1995). The statistic t induces an equivalence relation $\sim_t$ on the data through $a_1^n \sim_t b_1^n \Leftrightarrow t_n(a_1^n) = t_n(a_1^n)$. Now, t is said to have *S-structure* if for all $n, m \in \mathbb{N}$ and $a_1^n \sim_t b_1^n$, $x_1^m \sim_t y_1^m$ it follows that $a_1^n x_1^m \sim_t b_1^n y_1^m$. The relation $\sim_t$ can now be used to classify stationary measures. Say that $\mu \in \mathcal{P}(\mathcal{X}^{\mathbb{N}}, T)$ is summarized by t if $\mu(a_1^n) = \mu(b_1^n)$ for all $a_1^n \sim_t b_1^n$. For finite state spaces $\mathcal{X}$ this is tantamount to saying that $[pr_n\mu](\cdot \mid t_n(y_1^n)) = \mathcal{U}_{\{x_1^n \in \mathcal{X}^n : x_1^n \sim_t y_1^n\}}(\cdot)$ for all $n \in \mathbb{N}$ and for all cylinders y_1^n, where $\mathcal{U}_A$ denotes the uniform distribution on a finite set A. Then, Freedman proved the following mixing result.

Theorem 5.3.5 (Freedman's S-structure theorem)**.** *Let t be a statistic which has S-structure. Any $\mu \in \mathcal{P}(\mathcal{X}^{\mathbb{N}}, T)$ that is summarized by t is a mixture of T-ergodic measures summarized by t.*

The converse is trivially fulfilled. By taking t to be the transition counts, Freedman readily showed that so called *partially exchangeable* stationary measures are given as mixtures of

stationary Markov measures. More precisely, let $k := \#\mathcal{X}$ and define for a stochastic vector $p = (p_i)_{i\in\{1,\dots,k\}}$ and a stochastic matrix $M = \left(p^i_j\right)_{i,j\in\{1,\dots,k\}}$ the (p, M)-Markov measure as

$$\mu_n([c_1, \dots, c_n]) := p_{c_1} \prod_{m=1}^{n-1} p^{c_m}_{c_{m+1}},$$

for all cylinder sets $[c_1, \dots, c_n]$ and all $n \in \mathbb{N}$. By means of the Kolmogoroff extension theorem the family $\{\mu_n\}_{n\geq 1}$ is uniquely extendable to a probability $\mu \in \mathcal{P}(\mathcal{X}^{\mathbb{N}})$. In that case call μ the (p, M)-Markov measure. Thus, a Markov measure is eventually seen to be parametrized by $(k+1)$ units of unit-simplices of dimension $(k-1)$ denoted by Δ^{k-1}. In contrast to the case of Bernoulli measures not every Markov measure is stationary, i.e. not all combinations of stochastic vectors and stochastic matrices lead to a stationary Markov measure. The space of stationary Markov measures is obtained as a suitable subspace of $\Delta^{k-1} \times \mathfrak{S}(\mathcal{X})$, where $\mathfrak{S}(\mathcal{X}) = \mathfrak{S}$ denotes the space of all stochastic matrices on the state space. This subspace can be obtained as an algebraic constraint involving the roots of certain polynomials in the components of p. Hence, one can think of this subspace as an algebraic variety, yet under some additional (probabilistic) constraints. From the theory of Markov chains it is well known that these polynomials are given by the equation $pM = p$, where p is interpreted as a row vector. If one has binary response, i.e. if the state space is $\mathcal{X} = \{0, 1\}$, this equation gives rise to the polynomials

(I) $p_0 p^0_0 + p_1 p^1_0 - p_0$,

(II) $p_0 p^0_1 + p_1 p^1_1 - p_1$.

Thus, a given stochastic matrix

$$M = \begin{pmatrix} p^0_0 & p^0_1 \\ p^1_0 & p^1_1 \end{pmatrix}$$

uniquely determines a stochastic vector through the roots of above polynomials under the constraint $p_0 + p_1 = 1$ which then is given by $p_0 = \frac{p^1_0}{p^0_1 + p^1_0}$ as long as a certain requirement for M hold. This is essentially given through irreducibility. Recall the following notions from the theory of Markov chains, c.f. Freedman (1983) and Seneta (1981). For a finite state space $\mathcal{X}$ call a stochastic matrix $M = (m_{ij})_{i,j\in\mathcal{X}}$ *irreducible* if for any pair $i, j \in \mathcal{X}$ there is a positive integer l such that $(M^l)_{ij} > 0$, i.e. any two states communicate. Let $h_i = h_i(M)$ denote the *period* of state $i \in \mathcal{X}$ with respect to M, that is the greatest common divisor of positive integers l such that $(M^l)_{ii} > 0$. In case that the state space is finite, irreducibility implies *positive recurrence*, i.e. $\sum_{l\in\mathbb{N}} (M^l)_{ii} = \infty$ and the mean recurrence time is finite for all $i \in \mathcal{X}$, and that the period h is the same for all states. Call M *primitive* (eventually positive) if there is a positive integer l such that $M^l > 0$.

Theorem 5.3.6 (Perron-Frobenius)**.** *Let $M \in \mathfrak{S}(\mathcal{X})$ be irreducible. Then, for the spectral radius it holds $\rho = \rho(M) = 1$ and there are h complex eigenvalues e_j, $j = 1\dots, h$ with $|e_j| = 1$. If in addition M is primitive, then 1 is the only eigenvalue on the spectral radius and the associated left eigenvector p (resp. right eigenvector q) is strictly positive and unique up to positive multiplicity. If p (resp. q) is properly normalized, i.e. $p \in \mathcal{P}(\mathcal{X})$, then it is called the M-invariant distribution. The matrix $qp \in \mathfrak{S}(\mathcal{X})$ is the projection onto the ρ-eigenspace and is called the Perron-projection. It is given by $qp = \lim_{l\to\infty} M^l = \lim_{l\to\infty} \frac{1}{l} \sum_{k=1}^{l} M^k$. Moreover, each row of qp equals p.*

For a proof of the Perron-Frobenius theorem as well as for a good source for the the theory of stochastic matrices see Seneta (1981).

Note that the Césaro-limit always exists, even for periodic matrices. However, in general periodicity affects the convergence towards the M-invariant distribution p. That is the convergence in form of $\lim_{l\to\infty} M^l$ does only hold if the matrix is irreducible, positive recurrent and aperiodic ($h = 1$, e.g. implied by primitivity). These attributes are together sometimes, in abuse of notation, called "ergodic" which has nothing to do with the T-ergodicity of the Markov law implied by M as following proposition shows.

Proposition 5.3.7. *Let $\mathcal{X}$ be finite. A stationary Markov measure with a primitive stochastic matrix is T-ergodic.*

Proof. Define for $n_A, n_B \in \mathbb{N}$ the cylinder sets $A := [a_1, \dots, a_{n_A}]$ and $B := [b_1, \dots, b_{n_B}]$. Then for $s > n_A$ one has

$$\mu(A \cap T^{-s}B) = p_{a_1}\left(\prod_{i=1}^{n_A-1} m_{a_i a_{i+1}}\right)(M^s)_{a_{n_A} b_1}\left(\prod_{i=1}^{n_B-1} m_{b_i b_{i+1}}\right).$$

By the Perron-Frobenius theorem, this yields $\lim_{s\to\infty} \frac{1}{s}\sum_{i=1}^{s} \mu(A \cap T^{-i}B) = \mu(A)\mu(B)$ which in turn implies ergodicity. □

For $\mathcal{X} = \{0, 1\}$, similarly to the case of Bernoulli measures, the interest from a statistical point of view ought to be in those stationary Markov measures that are parametrized by a positive stochastic matrix. At the same time such an assumption would exclude (a.s.) the reducible matrices $\begin{pmatrix} 1 & 0 \\ p_1^0 & p_1^1 \end{pmatrix}, \begin{pmatrix} p_0^0 & p_1^0 \\ 0 & 1 \end{pmatrix}$ and the only periodic matrix $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ from the inference procedure, where p_j^i is the conditional probability of jumping in state j from state i. Above fact can be depicted as follows, making obvious that the set of binary stationary Markov measures is nowhere-dense in the set of all binary Markov measures.

The entire unit cube in three dimensions depicts the parameter space of all binary Markov measures. Solving for the system of equations which emerges from the constraint of stationarity yields (whenever well defined) $p_0 = \frac{p_0^1}{p_0^1+(1-p_0^0)}$. Thus the invariant probability vector is obtained as function of the stochastic matrix.

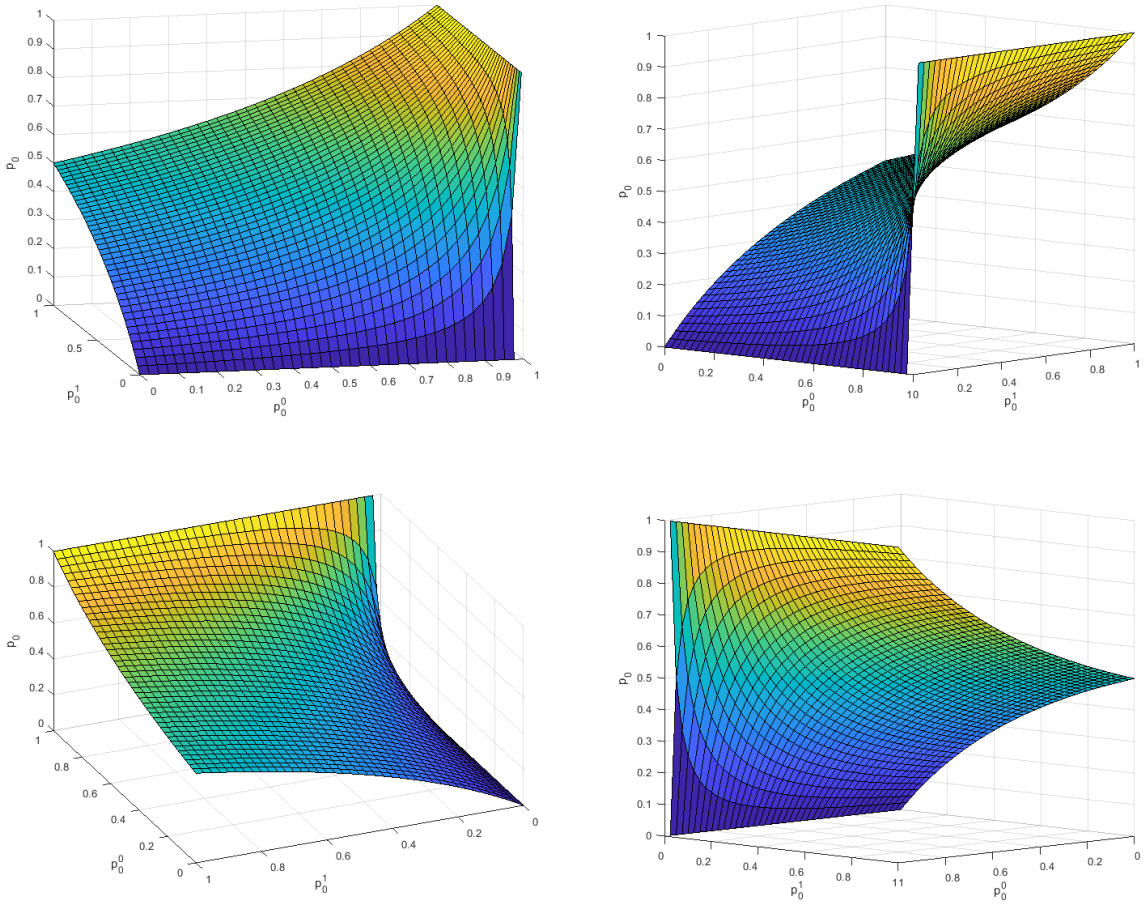


Figure 5.1: Embedding of stationary Markov measures in all Markov measures

This function can be visualized as a surface embedded into the unit cube making obvious that the stationary Markov measures are nowhere-dense in all Markov measures.Moreover, one can discover the "singularity" of the surface which corresponds to the identity matrix. Moreover, one can discover the "singularity" of the surface which corresponds to the identity matrix.

In addition to above visualization, it is possible to depict the Bernoulli measures embedded into the surface representing the stationary Markov measures. Clearly, the Bernoulli measures are the Markov measures for which it holds $p_0^0 = p_0^1$, which also yields $p_0 = p_0^0$. This means that one can discover the Bernoulli measure as diagonal of the unit cube. Since this diagonal is embedded into the surface, this visualization makes obvious the fact that Bernoulli measures are automatically stationary.

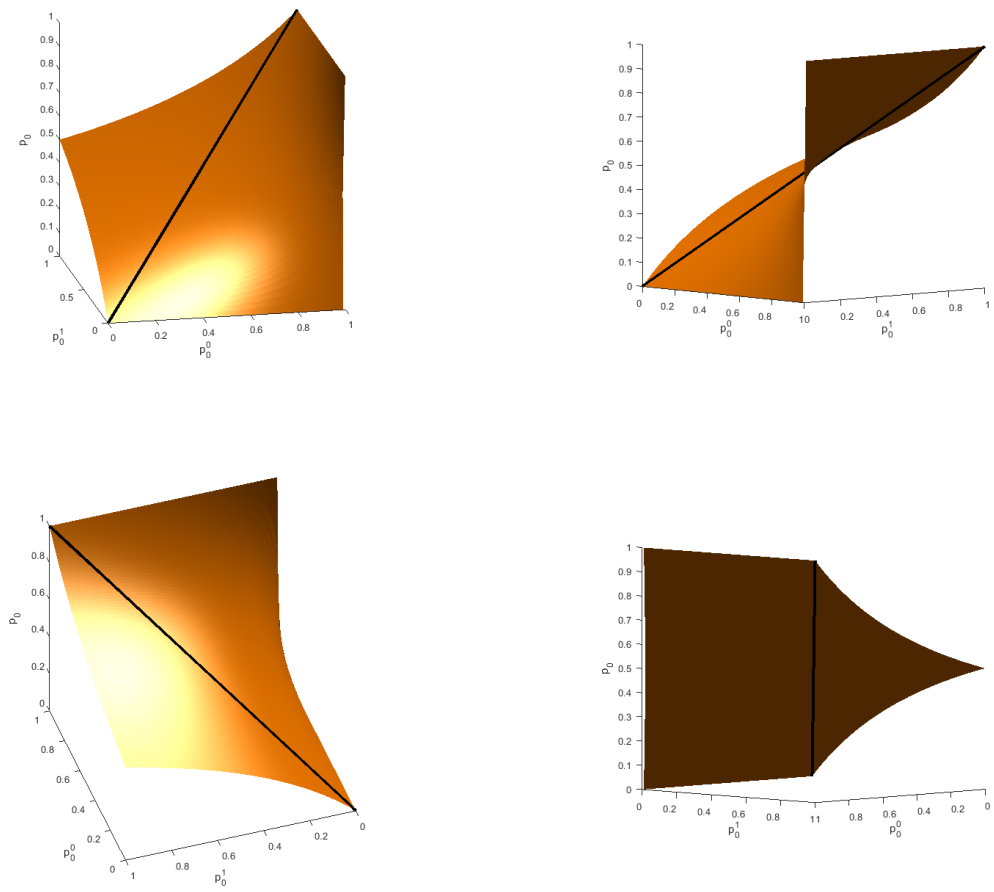


Figure 5.2: Embedding of Bernoulli measures in the stationary Markov measures

If the state space has more than two elements, above visualizations become difficult as well as in the case of dependencies with a wider range, c.f. section 5.5 . However, one may bear in mind above pictures when dealing with those generalizations in order to have an intuitive comprehension when dealing with them in higher dimensional simplices.

Now, let t be the family of measurable functions given through

$$t_n : \mathcal{X}^n \to \mathcal{X} \times \mathbb{N}_0^{k \times k}$$
$$[c_1, \ldots, c_n] \mapsto t_n([c_1, \ldots, c_n]) := \left(c_1, \left[\#\left\{m = 1, \ldots, n-1 : (c_m, c_{m+1}) = (i,j)\right\}\right]_{i,j \in \mathcal{X}}\right),$$

which defines the *transition counts*. This particular statistic is easily seen to have S-structure such that Theorem 5.3.5 holds. The ergodic measures being summarized by transition counts are of certain form.

Theorem 5.3.8 (Freedman (1962))**.** *If a measure $\mu \in \mathcal{P}^e(\mathcal{X}^{\mathbb{N}}, T)$ is summarized by transition counts, then it is a Markov measure.*

That means stationary measures which are summarized by transition counts, sometimes called *partially exchangeable* or nowadays often *Markov exchangeable*, see e.g. Fortini et al. (2002), are mixtures of stationary Markov measures which in turn are parametrized by a stochastic matrix. Thus, in order to make Bayesian inference for partially exchangeable measures it suffices to model a prior distribution on the space of stochastic matrices wich is equipped with the Borel σ-field $\mathfrak{B}_{\mathfrak{S}}$ induced by the topology of coordinate-wise convergence. This is tantamount to defining a random stochastic matrix which is a.s. irreducible with respect to an appropriate (prior) distribution. Moreover, the prior should be chosen in a way that allows to update this prior appropriately by (Markovian) data. As mentioned earlier, it is reasonable to let the set of matrices which have zero entries be a null set with respect to the prior as long as no additional information on the observation process (cyclic behavior etc.) is available in advance. However, by definition of the support of a probability measure, such a prior will have support of the entire space of stochastic matrices as long as all positive matrices are supported. The update procedure of such a prior strongly depends on how the prior does sample the rows of the random stochastic matrix. This in turn is known to be influenced by the predictive sufficient statistic, c.f. Fortini et al. (2000); Fortini and Petrone (2012b, 2014). Roughly speaking, call a statistic $t = \{t_n\}_{n \in \mathbb{N}}$ *predictively sufficient* if prediction of future observations on basis of t are as good as on basis of the original data. That is the future observation X_{n+1} is conditionally independent of the past X_1^n given $t_n(X_1^n)$, or in symbols $X_{n+1} \perp\!\!\!\perp X_1^n \mid t_n(X_1^n)$. See e.g. Dawid (1979) for an exhaustive treatment of conditional independency in statistical theory. The following well known result on mixtures of Markov measures, when t is taken to be transition counts, clarifies the independence among the rows of a sampled stochastic matrix.

Proposition 5.3.9. *The rows of the random stochastic matrix $M \in \mathfrak{S}^{\Omega}$ are sampled independently iff transition counts are predictively sufficient.*

A rigorous proof can be found in Fortini and Petrone (2014). Note that for a string x_1^n the terminal state x_n is fully determined by $t_n(x_1^n)$. Hence, if on has at hand $t_n(X_1^n)$, then only the terminal state and the number of transitions from X_n will influence the prediction of X_{n+1}, if transition counts are judged to be predictively sufficient. Note that this is not always the case for Bayesian inference problems for Markov processes. A counter example is given by a random stochastic Δ-matrix [Abolnikov and Dukhovny (1991)], c.f. chapter 4 of the thesis. In that case, since certain rows are essentially obtained as repetitions of others, classical transition counts are not predictively sufficient, nor is the support of a suitable prior full. However, in such a case a lot of prior information is given in the form that one knows that data e.g. stems from a queueing system. Usually

such information is not available such that predictive sufficiency of transition counts is a reasonable assumption that simplifies Bayesian inference for Markov measures considerably. This simplification is die to the fact that rows are updated separately by data only affecting a particular row. Suppose $\mathcal{X} = \{0,1\}$, since the Bernoulli measures can be rediscovered in the space of stationary stochastic matrices (i.e. those for which $p_0^0 = p_1^0$) one can think of two extremes with respect to the dependency of the the rows. On the one hand one has totally independent rows while on the other, one has totally dependent (i.e. identical) rows.

5.4 On the Size of Equivalence Classes of Partially Exchangeable Measures in the Binary Case

Sufficiency and predictive sufficiency in the case of a finite state space can also be described by how fine equivalence classes becomes or in how many equivalence classes the data splits. Since the general case is difficult to describe, merely the case of binary data is considered. Suppose that $\mathcal{X} = \{0,1\}$ and that the order statistic is judged to be sufficient which is well known to be equivalent to $\Sigma(X_1^n) := \Sigma_n(X_1^n) := \sum_{i=1}^n X_i$, that is the number of 1's appearing in a data string, being sufficient in the case of binary response. Now, given $\Sigma(X_1^n) = m$, by easy combinatoric arguments one has that the class of strings which possess the certain value of Σ, i.e. $\Sigma^{-1}(m)$, has $\binom{n}{m}$ elements. Thus, given $\Sigma(X_1^n) = m$ one has a discrete uniform distribution on some space with $\binom{n}{m}$ elements. Moreover, $(n+1)$ of such uniform distributions exist. If the statistic equals transition counts t rather than Σ, an analog result can be given which is done here. For a given $n \in \mathbb{N}$ let $t_j^i := t_j^i(x_1^n) := \#\{l = 1,\dots,n-1 : (x_l, x_{l+1}) = (i,j)\}$ denote the number of jumps the process makes from state i to j among the string x_1^n.

Proposition 5.4.1. *Let $\mathcal{X} = \{0,1\}$ and $t = \{t_n\}_{n\in\mathbb{N}}$ be the statistic of transition counts. Then*

(a) The total number of equivalence classes given $\Sigma(x_1^n) = m \in \{0,\dots,n\}$ evolves as follows. Let $n = 2\eta$ for $\eta \in \mathbb{N}$ then

$$\#\left\{t_n\left(\Sigma_n^{-1}(\eta)\right)\right\} = 2(2\eta - 1).$$

Moreover, for all $l \geq 2\eta + 1$ it holds

$$\#\left\{t_l\left(\Sigma_l^{-1}(\eta)\right)\right\} = \#\left\{t_l\left(\Sigma_l^{-1}(l-\eta)\right)\right\} = 4\eta - 1.$$

(b) The total number of equivalence classes is given by

$$\#\{t_n(x_1^n) : x_1^n \in \mathcal{X}^n\} = n(n-1) + 2.$$

(c) the number of elements in a particular equivalence class is given by

$$\#t_n^{-1}(h) = \begin{cases} \binom{t_0^0+t_1^0-\delta_{x_n}(\{1\})}{t_1^0-\delta_{x_n}(\{1\})}\binom{t_1^1+t_0^1-\delta_{x_n}(\{0\})}{t_0^1-\delta_{x_n}(\{0\})} & , \text{ if } h = \left(x_1\middle|\left[t_0^0,t_1^0,t_0^1,t_1^1\right]\middle|x_n\right) \in t_n(\mathcal{X}^n) \\ 0 & , \text{ if } h \notin t_n(\mathcal{X}^n), \end{cases}$$

where $\binom{-1}{-1} := 1$.

Proof. (a) Let $n = 2\eta$ and $m = \eta$ and indicate by an overlined sequence $c_1, \ldots, c_l$ of symbols with an index I the string which appears by repeating the sequence i times, i.e.

$$\overline{c_1,\ldots,c_l}_I = \underbrace{[c_1,\ldots,c_l]\cdots[c_1,\ldots,c_l]}_{I-times}.$$

Due to symmetry it suffices to regard strings with 1 as initial symbol. Then, start with the equivalence class given by the only element $\overline{1}_{(\eta)}\overline{0}_{(\eta)}$ and observe that the only transitions are given by $\eta-1$ transitions $1\to1$, $\eta-1$ transitions $0\to0$ and one transition $1\to0$. To create new equivalence proceed inductively by transposing "forbidden" blocks as follows. First transpose the first block $[10]$ in $\overline{1}_{(\eta)}\overline{0}_{(\eta)}$ with the second $[0]$ appearing which yields $\overline{1}_{(\eta-1)}[01]\overline{0}_{(\eta-1)}$. Then, transpose the first block $[10]$ and the third 0 appearing in the new sting to obtain $\overline{1}_{(\eta-2)}[0101]\overline{0}_{(\eta-2)}$. Proceed that way until the string $\overline{10}_{(\eta)}$ is reached. Plainly, this yields η equivalence classes since every single string possesses e.g. a different number of transitions $1\to1$. Now, create the remaining $\eta-1$ equivalence classes beginning and ending with 1 by the following scheme. First, transpose the last 1 and the last 0 in $\overline{10}_{(\eta)}$ to get $\overline{10}_{(\eta-1)}[01] = \overline{10}_{(\eta-2)}[1001]$. Then, transpose the first 1 and the block $[00]$ of the block $[1001]$ for having $\overline{10}_{(\eta-3)}[100011]$. Proceed that way until the sting $[1]\overline{0}_{(\eta)}\overline{1}_{(\eta-1)}$ is produced. Note that this algorithm is fully exhaustive since it produces all admissible combinations of transitions $0\to0$ and $1\to1$. Hence, by symmetry (regard now all stings of length 2η having η symbols 1 and initial state 0) there are $2[\eta+(\eta-1)]$ equivalence classes.
Now, consider the case $l = 2\eta+1$ and regard the set of equivalence classes of length l having η symbols 1. By S-structure of the transition counts, essentially the same number of equivalence classes evolves up to the exception $[0]\overline{01}_\eta$ which is only one additional equivalence class. Therefore, the total number of equivalence classes is given by $4\eta-1$. Note that the same reasoning holds for all $l \geq 2\eta+1$ and, by symmetry, as well for $m \geq \lfloor l/2\rfloor + 1$.
(b) Let $n = 2\eta$ be an even number. Then, by (a) one has

$$\begin{aligned} \#\{t_n(x_1^n) : x_1^n \in \mathcal{X}^n\} &= 2\left[1+\sum_{m=1}^{\eta-1}(4m-1)\right] + 2(2\eta-1) \\ &= 2\left[1+2\eta(\eta-1)-(\eta-1)\right] + 4\eta - 2 \\ &= (2\eta-2)(2\eta-1) + 4\eta \\ &= 4\eta^2 - 2\eta + 2 \\ &= n(n-1)+2. \end{aligned}$$

On the other hand, for $n = 2\eta + 1$, one has

$$\begin{aligned}\#\{t_n(x_1^n) : x_1^n \in \mathcal{X}^n\} &= 2\left[1 + \sum_{m=1}^{\eta}(4m-1)\right] \\ &= 2 + 4\eta(\eta+1) - 2\eta \\ &= 4\eta^2 + 2\eta + 2 \\ &= n(n-1) + 2.\end{aligned}$$

(c) By well known properties of partially exchangeable sequences, among others c.f. Diaconis and Freedman (1980) and Zabell (1995), one can think of a partially exchangeable sequence as several exchangeable sequences "properly interwoven". In case of binary data, there are two of those, the so called sequences of 0-successors and 1-successors. Hence, the number of elements in a certain equivalence class is determined by the different ways the 0-successors and 1-successors can be arranged. The total number of 0-successors and 1-successors is given by $t_0^0+t_1^0$ and $t_0^1+t_1^1$, respectively. Therefore the possibilities to arrange succeeding 1's among the 0-successors and 0's among the 1-successors are essentially given by $\binom{t_0^0+t_1^0}{t_1^0}$ and has $\binom{t_0^1+t_1^1}{t_0^1}$, respectively. Suppose some string ends with 1 and $t_0^0 + t_1^0 = 0$, then the string has to equal $\overline{1}_n$ and the assertion is true by definition. (The same holds for the string $\overline{0}_n$.) If a sting ends with 1 and has $t_0^0 + t_1^0 > 0$, then $t_1^0 \geq 1$. But then the last transition $0 \to 1$ has to appear right in front of the last block of 1's such that there are just $\binom{t_0^0+t_1^0-\delta_{x_n}(\{0\})}{t_1^0-\delta_{x_n}(\{0\})}$ possibilities left to schedule the 0-succeeding 0's and 1's. Since the same is true for strings having 0 as terminal state, the assertion follows. □

Proposition 5.4.1 gives a full description of the equivalence classes which emerge from the assumption that transition counts are sufficient. Using it one can get an idea how the refinement of the clusters that build up from order statistics to transition counts evolves.

In order to depict the theory just developed, all the equivalence classes are drawn for binary data strings up to length five. Notice that all the equivalence classes associated to the order statistic can be recovered by summarizing the respective columns. The image is believed to give an idea of the ongoing splitting process of the equivalence classes when moving to higher dependencies.

$\underline{\{0,1\}^1}$:

(0|0,0,0,0|0) 0	**(1|0,0,0,0|1)** 1

$\underline{\{0,1\}^2}$:

(0|1,0,0,0|0) 00	**(1|0,0,1,0|0)** 10	**(1|0,0,0,1|1)** 11
	(0|0,1,0,0|1) 01	

$\underline{\{0,1\}^3}$:

(0|2,0,0,0|0) 000	**(0|1,1,0,0|1)** 001	**(0|0,1,0,1|1)** 011	**(1|0,0,0,2|1)** 111
	(1|1,0,1,0|0) 100	**(1|0,0,1,1|0)** 110	
	(0|0,1,1,0|1) 010	**(1|0,1,1,0|1)** 101	

$\underline{\{0,1\}^4}$:

(0|3,0,0,0|0) 0000	**(1|2,0,1,0|0)** 1000	**(0|1,1,0,1|1)** 0011	**(0|0,1,0,2|1)** 0111	**(0|3,0,0,0|0)** 0000
	(0|2,1,0,0|1) 0001	**(1|1,0,1,1|0)** 1100	**(1|0,0,1,2|0)** 1110	
	(0|1,1,1,0|0) 0100 0010	**(1|1,1,1,0|1)** 1001	**(1|0,1,1,1|1)** 1011 1101	
		(0|0,1,1,1|0) 0110		
		(0|0,2,1,0|1) 0101		
		(1|0,1,2,0|0) 1010		

$\underline{\{0,1\}^5}$:

(0|4,0,0,0|0) 00000	**(1|3,0,1,0|0)** 10000	**(1|2,0,1,1|0)** 11000	**(0|1,1,0,2|1)** 00111	**(0|0,1,0,3|1)** 01111	**(1|0,0,0,4|1)** 11111
	(0|3,1,0,0|1) 00001	**(0|2,1,0,1|1)** 00011	**(1|1,0,1,2|0)** 11100	**(1|0,0,1,3|0)** 11110	
	(0|2,1,1,0|0) 01000 00100 00010	**(1|2,1,1,0|1)** 10001	**(0|0,1,1,2|0)** 01110	**(1|0,1,1,2|1)** 10111 11011 11101	
		(0|0,2,2,0|0) 01010	**(1|0,2,2,0|1)** 10101		
		(1|1,1,1,1|0) 10100 10010	**(0|0,2,1,1|1)** 01011 01101		
		(0|1,2,1,0|1) 01001 00101	**(1|0,1,2,1|0)** 10110 11010		
		(1|1,1,1,1|0) 01100 00110	**(1|1,1,1,1|1)** 10011 11001		

In addition, the equivalence classes can be further summarized by stationarity, i.e. by the additional invariance $T\mu = \mu$. More precisely, stationarity yields a coarser splitting of the spaces of finite strings in equivalence classes. From a more theoretical perspective it is explainable that this must happen since $\mathcal{B} \subset \mathcal{B} \bigvee \mathcal{I}$, where $\mathcal{B}$ denotes the σ-field of sets being

invariant with respect to the family of admissible block transformations β c.f. Diaconis and Freedman (1980, Proposition (27)). Roughly, this means that the additional (prior) information encoded in $\mathcal{I}$ must lead to "less" measures fulfilling the associated invariance condition. Indeed, for the associated extreme points of the family of measures fulfilling above invariances, one has $\{\mu(\cdot|\mathcal{B}\bigvee\mathcal{I})\} \subset \{\mu(\cdot|\mathcal{B})\}$. This behavior can neatly be depicted in the case of binary data by a proper extension of above picture. By stationarity the parameters binary Markov measure μ has to fulfill the additional constraint (see also the subsequent section 5.5, especially Proposition 5.5.3, for further explanations)

$$p_0 p_0^0 + p_1 p_0^1 = p_0,$$

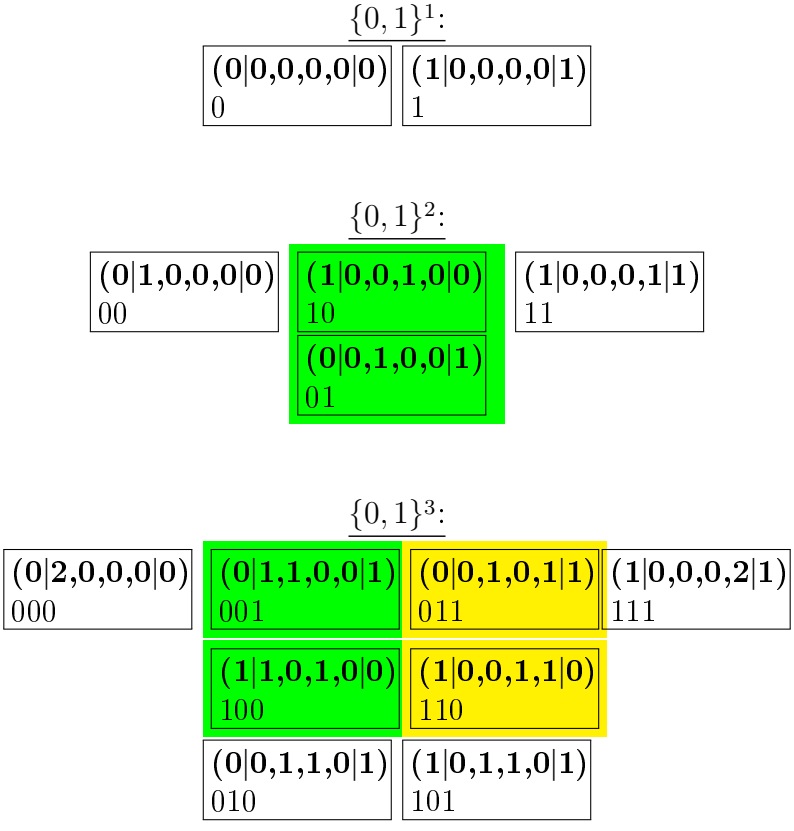
where $p = (p_0, p_1)$ is the stochastic vector and $\begin{pmatrix} p_0^0 & p_1^0 \\ p_0^1 & p_1^1 \end{pmatrix}$ is the stochastic matrix which together parametrize the Markov measure μ. By straight-forward manipulations, above constraint becomes

$$p_0 p_1^0 = p_1 p_0^1.$$

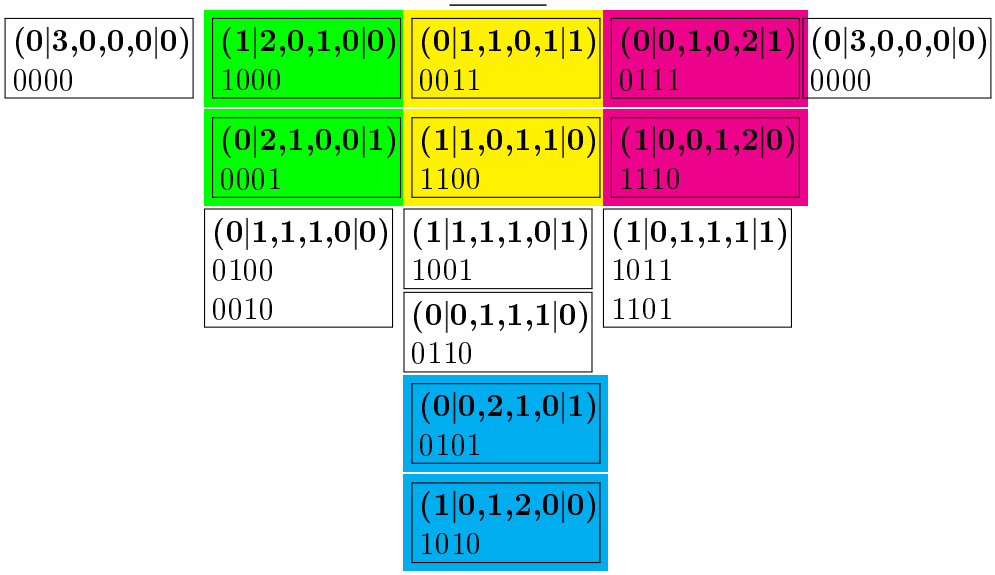
This in turn yields e.g.

$$\mu([0,1]) = \mu([1,0]), \quad \mu([0,0,1]) = \mu([1,0,0]), \quad \mu([0,1,1]) = \mu([1,1,0]),$$

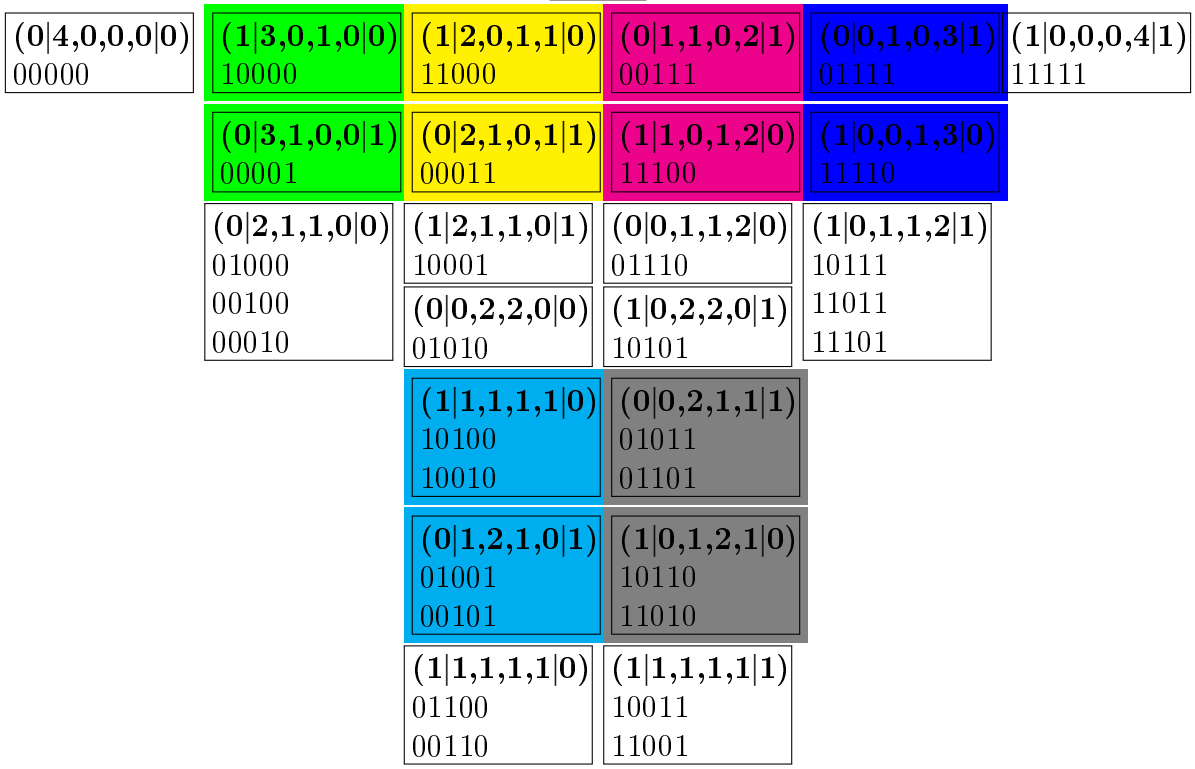
etc., which amounts to a kind of "mirroring-invariance". Above figure depicts the melting of different equivalence classes under β-invariance to joint equivalence classes under $\{\beta, T\}$ by assigning a joint color to the respective classes.



$\{0,1\}^4$:



$\{0,1\}^5$:



However, it is not known yet how this coarsening of equivalence classes induced by stationarity affects the frmulas given in Proposition 5.4.1. An appropriate result would be interesting and finding it is left as future work at this point.

5.5 Higher-Order Dependencies

In order to deal with higher-order dependencies, most often the theory of Markov chains is employed by modeling processes with long memories as Markov chains in larger state spaces. For example, if $\mathcal{X} = \{0,1\}$ and the process to be modeled has length of memory 2, then one might take the process to the "higher-dimensional" state space $\mathcal{X}^{[2]} := \{(00),(01),(10),(11)\}$ and let it evolve by a 4×4 stochastic matrix with entries of 0 whenever $i_2 \neq j_1$ for some $i, j \in \mathcal{X}^{(2)}$. The theory behind this idea is given by symbolic topological dynamics. For a comprehensive introduction see e.g. Lind and Marcus (1995). Particularly, it uses the embedding of a higher-order Markov chain into a larger space by exploiting a so called higher block code. Hence, a Markov chain of order N on $\mathcal{X}$ evolves as a usual Markov chain in the N^{th} higher block shift $\mathcal{X}^{[N]}$. Some drawback appears since a stationary probability measure for a Markov chain on $\mathcal{X}^{[N]}$ which is governed by a $\#\mathcal{X}^N \times \#\mathcal{X}^N$ stochastic matrix will be a probability vector on $\mathcal{X}^{[N]}$ rather than on $\mathcal{X}$. Furthermore, there are probability measures whose associated process does not evolve as a usual Markov chain on any higher block shift. An early and prominent example dates back to Blackwell (1957) which uses the idea of collapsing states, which briefly recalled.

Example 5.5.1 (Blackwell (1957))**.** *Let S be the shift map on $\{a,b,c\}$ and $\mu \in \mathcal{P}(\{a,b,c\},S)$ be a Markov measure governed by the stochastic matrix*

$$M = \begin{pmatrix} 0 & 2/3 & 1/3 \\ 1/3 & 2/3 & 0 \\ 2/3 & 0 & 1/3 \end{pmatrix}$$

and M-invariant probability vector p which is given by $p = (p_a, p_b, p_c) = (2/7, 4/7, 1/7)$. Now, let the two states b and c collapse into the same state. More precisely, apply the factor map (a continuous and onto map) f, which is given by

$$\begin{aligned} f : &\{a,b,c\} \to \{0,1\} \\ &a \mapsto 0; \quad b,c \mapsto 1. \end{aligned}$$

It can be argued that the push-forward of μ under f is stationary as well, i.e. $\nu := f\mu \in \mathcal{P}(\{0,1\},T)$. However, it can be shown that ν is not Markovian of any finite order.

Hence, the question arises what can be inferred from binary data with respect to the measure ν. The "goodness" of an estimate for ν will clearly depend on the "size" of the model space one does work in. For instance, if one has prior information that the data can only stem from Markov measures on $\{a,b,c\}$ hidden by a suitable factor map, the model space is rather manageable and it might be possible to have access to some "finite inference procedure". See e.g. Marcus et al. (2011) for more information on hidden Markov structures as well as for conditions ensuring their "finite origin". However, if such prior information is not available, one would have to embed the inference procedure into a larger framework which allows to enlarge the model space step-wise if data increases. Since Bayesian statistics require a reasonable parametrization of the measure one wishes to infer on, it is now headed for such a parametrization and parameter space, respectively.

First of all a well known assertion about creating probability measures is recalled.

Theorem 5.5.2 (Ionescu-Tulcea). *Let $(\mathcal{X}_n, \mathcal{A}_n)_{n\in\mathbb{N}}$ be a sequence of measurable spaces,*

$$\left(\kappa_n : \prod_{i=1}^{n-1} \mathcal{X}_i \times \mathcal{A}_n \to [0,1]\right)_{n\geq 2}$$

a sequence of stochastic kernels and $\kappa_1 \in \mathcal{P}(\mathcal{X}_1)$. Then, there is a unique probability measure $\mu \in \mathcal{P}\left(\prod_{n\in\mathbb{N}} \mathcal{X}_n\right)$ such that $\mu = \bigotimes_{n\in\mathbb{N}} \kappa_n$, i.e. for all $n \in \mathbb{N}$ and for all $A_i \in \mathcal{A}_i; i = 1, \ldots, n$ it holds that

$$\mu\left(A_1 \times \cdots \times A_n \times \prod_{i=n+1}^{\infty} \mathcal{X}_i\right) = \int_{A_1} \ldots \int_{A_{n-1}} \kappa_n\left[(x_1, \ldots, x_{n-1}), A_n\right] \kappa_{n-1}[(x_1, \ldots, x_{n-2}), dx_{n-1}] \ldots \kappa_1[dx_1].$$

Hence, one might say that a general probability $\mu \in \mathcal{P}\left(\mathcal{X}^{\mathbb{N}}\right)$, i.e. $\mathcal{X}_n = \mathcal{X}$, is parametrized by a sequence of stochastic kernels. Now, let $\mathcal{X}$ be finite. If μ is a stationary Markov measure then for all $n > 2$ the kernels are determined by κ_2 through an appropriate repetitive embedding of κ_2. Moreover, under sufficient conditions, the probability vector κ_1 is uniquely determined by κ_2 as well by the system of algebraic constraints given by $\kappa_1 * \kappa_2 = \kappa_1$, where κ_2 is considered as a matrix. As mentioned earlier, then it would be sufficient to make statistical inference for the stochastic matrix κ_2 and subsequently to transfer it to κ_1. The question is if this holds true for higher dependencies as well. To put it another way, under what constraints is it true that κ_N determines κ_n for all $n \neq N$? Plainly, one such constraint ought to be stationarity. Thus a result is recorded which gives a characterization of stationarity in the case that $\mathcal{X}$ is finite (or at least countable) for measures carrying longer memories. Since no appropriate source for a citation is known to the author, a proof is given.

Proposition 5.5.3. *For a countable state space $\mathcal{X}$ and $\mu \in \mathcal{P}(\mathcal{X}^{\mathbb{N}})$ the following is equivalent*

(i) $\mu \in \mathcal{P}(\mathcal{X}^{\mathbb{N}}, T)$

(ii) for all cylinder sets $[c_1, \ldots, c_m]$ it holds

$$\mu([c_1, \ldots, c_m]) = \sum_{c\in\mathcal{X}} \mu([c, c_1, \ldots, c_m]).$$

Proof. Let $T\mu = \mu$ and $[c_1, \ldots, c_m]$ be a cylinder set. Then

$$\mu([c_1, \ldots, c_m]) = T\mu([c_1, \ldots, c_m]) = \mu\left(T^{-1}[c_1, \ldots, c_m]\right) = \mu\left(\bigcup_{c\in\mathcal{X}} [c, c_1, \ldots, c_m]\right) = \sum_{c\in\mathcal{X}} \mu([c, c_1, \ldots, c_m]).$$

On the other hand, if above is true for some set function and for all cylinder sets, by Caratheodory's extension theorem, the invariance property extends to all Borel sets since the set of cylinders is a semi-ring generating the Borel σ-field. □

Regarding a (p, M)-Markov measure μ, it is possible to interpret the stochastic matrix M as a device that maps a probability measure on the finite state space $\mathcal{X}$ onto another one. If the probability measure in form of a vector is the right eigenvector of M, the image is just the same and by Proposition 5.5.3 μ is readily seen to be stationary. A similar description is possible when dealing with higher dependencies although the notion of an eigenvector is more delicate. More precisely, the higher kernels can be regarded as functions whose arguments are lower kernels. As an example, let $\mathcal{X} = \{0, 1\}$ and let $\mu \in \mathcal{P}(\mathcal{C})$ have order of dependency be $N = 2$. Then, a probability κ_0, and the first two kernels κ_1, κ_2 appearing in Theorem 5.5.2 which parametrize μ are given as a vector $p = (p_0, p_1)$, a matrix $M = \begin{pmatrix} p_0^0 & p_0^1 \\ p_1^0 & p_1^1 \end{pmatrix}$ and a third device that maps M onto another matrix M', i.e.

$$M \overset{\kappa_2}{\mapsto} M' := \begin{pmatrix} p_0^0 p_{00}^{00} + p_1^0 p_{10}^{01} & p_0^0 p_{01}^{00} + p_1^0 p_{11}^{01} \\ p_0^1 p_{00}^{10} + p_1^1 p_{10}^{11} & p_0^1 p_{01}^{10} + p_1^1 p_{11}^{11} \end{pmatrix},$$

where $p_{jk}^{ij} := \mathbb{P}(X_n = k \mid X_{n-1} = j, X_{n-2} = i)$ is given through the kernel κ_2. The image matrix M' encodes the two-step jump probabilities that for a usual Markov chain are just given through M^2. This motivates the definition of a stochastic tensor following the notion of a stochastic matrix. Note that a matrix can be regarded as a tensor, too. To be more precise, let V be a finite-dimensional linear space over a field $\mathbb{K}$ which is taken to be the reals. Further, let V^* denote the dual space of V, that is the space of all linear mappings $l : V \to V$. Let $r, s \in \mathbb{N}_0$.

Definition 5.5.4 (Tensor). *A (r, s)-tensor τ is a $(r + s)$ multi-linear map*

$$\tau : \underbrace{V^* \times \cdots \times V^*}_{r\text{-}times} \times \underbrace{V \times \cdots \times V}_{s\text{-}times} \to \mathbb{K}.$$

Further, let $\mathcal{T}_s^r := \mathcal{T}_s^r(V)$ denote the space of all (r, s) tensors, which is itself a linear space with $\dim[\mathcal{T}_s^r] = \dim[V]^{r+s}$. This linear space possesses a basis which is of the form

$$B_s^r = \left\{ e_{i_1} \otimes \cdots \otimes e_{i_r} \otimes e^{j_1} \otimes \cdots \otimes e^{j_s} : i's, j's = 1, \ldots, \dim[V] \right\},$$

where a basis element is defined by

$$e_{i_1} \otimes \cdots \otimes e_{i_r} \otimes e^{j_1} \otimes \cdots \otimes e^{j_s} : \underbrace{V^* \times \cdots \times V^*}_{r\text{-}times} \times \underbrace{V \times \cdots \times V}_{s\text{-}times} \to \mathbb{K}$$

$$(v^1, \ldots, v^r, w_1, \ldots, w_s) \mapsto \prod_{k=1}^{r} < v^i, e_{i_k} > \prod_{l=1}^{s} < v_j, e^{e^{j_l}} >,$$

where $< -, \cdot >: V^* \times V \to \mathbb{K}$ denotes a standard inner product.

Using the multi-linearity, a (r, s)-tensor τ can be decomposed as

$$\tau = \sum_{i's, j's = 0, \ldots, \dim[V]} a_{j_1 \cdots j_s}^{i_1 \cdots i_r} e_{i_1} \otimes \cdots \otimes e_{i_r} \otimes e^{j_1} \otimes \cdots \otimes e^{j_s},$$

which can be abbreviated by Einstein's summation convention as

$$\tau = a^{i_1\cdots i_r}_{j_1\cdots j_s} e_{i_1} \otimes \cdots \otimes e_{i_r} \otimes e^{j_1} \otimes \cdots \otimes e^{j_s},$$

which means summation over the "cross-indices". Call $(i_1\cdots i_r)$ the contra-variant and $(j_1\cdots j_s)$ the co-variant index of the coefficients $a^{i_1\cdots i_r}_{j_1\cdots j_s}$ of τ. Notice that for instance $V \cong \mathcal{T}^1_0$, $V^* \cong \mathcal{T}^0_1$, $M(2\times 2,\mathbb{K}) \cong \mathbb{L}(V,V) \cong \mathbb{L}(V^*,V^*) \cong \mathcal{T}^1_1$, $\mathbb{L}(\mathbb{L}(V,V),\mathbb{L}(V,V)) \cong \mathbb{L}(\mathbb{L}(V^*,v^*),\mathbb{L}(V^*,V^*)) \cong \mathcal{T}^2_2$ etc., where $\mathbb{L}(V,W)$ denotes the space of linear mappings from some linear space V to another linear space W. See e.g. Bowen and Wang (2008) for more details on multi-linear algebra and tensor algebra, respectively.

Apparently, a tensor is a multi-dimensional array of numbers which is able to carry a lot of information. This is the reason for its frequently usage e.g. in theoretical physics, particularly in relativity. The idea is that for the statistician it might be useful because a higher-dimensional tensor can carry information about the unfolding of of a stationary stochastic process. The next definition which is in the spirit of a stochastic matrix that is used to encode one-step jump probabilities of a stochastic process clarifies this idea.

Definition 5.5.5 (Stochastic Tensor)**.** *For a positive integer N, call a (N,N)-tensor which is represented by $\tau = p^{i_1\cdots i_N}_{j_1\cdots j_N} \bigotimes_{k=1}^M e_{i_k} \otimes \bigotimes_{l=1}^M e^{j_l}$ stochastic if the following conditions are fulfilled.*

1. $p^{i_1\cdots i_N}_{j_1\cdots j_N} \in [0,1]$,
2. $\sum_{j\in\mathcal{X}} p^{i_1\cdots i_N}_{j_1\cdots j_{N-1}j} = 1$,
3. $p^{i_1\cdots i_N}_{j_1\cdots j_N} = 0$, *whenever* $(i_2,\ldots,i_N) \neq (j_1,\ldots,j_{N-1})$.

Moreover, call a stochastic (N,N)-tensor τ_N positive (and write $\tau_N > 0$) if $p^{i_1\cdots i_N}_{j_1\cdots j_N} > 0$ whenever $(i_2,\ldots,i_N) = (j_1,\ldots,j_{N-1})$. Denote by $\mathfrak{S}^N_N$ the space of stochastic (N,N)-tensors and by $\mathfrak{S}^N_N(+)$ the space of positive stochastic tensors.

A tensor can be represented by a so called *flattening*, c.f. Landsberg (2012). Such a flattening maps a tensor, which is thought of as a multi-dimensional array of numbers, onto a matrix. This is a frequently used way to depict a tensor as a matrix, yet it is certainly not unique. However, it is reasonable to take that squared flattening that allows for the usual interpretation of multi-dependence jump probabilities in terms of powers of stochastic matrices For instance, if $\mathcal{X} = \{0,1\}$ a stochastic $(2,2)$-tensor τ_2 can be represented by the matrix

$$\tilde{\tau} = \begin{pmatrix} p^{00}_{00} & p^{00}_{01} & 0 & 0 \\ 0 & 0 & p^{01}_{10} & p^{01}_{11} \\ p^{10}_{00} & p^{10}_{01} & 0 & 0 \\ 0 & 0 & p^{11}_{10} & p^{11}_{11} \end{pmatrix} \Rightarrow \tilde{\tau}^2 = \begin{pmatrix} p^{00}_{00}p^{00}_{00} & p^{00}_{00}p^{00}_{01} & p^{00}_{01}p^{01}_{10} & p^{00}_{01}p^{01}_{11} \\ p^{01}_{10}p^{10}_{00} & p^{01}_{10}p^{10}_{01} & p^{01}_{11}p^{11}_{10} & p^{01}_{11}p^{11}_{11} \\ p^{10}_{00}p^{00}_{00} & p^{10}_{00}p^{00}_{01} & p^{10}_{01}p^{01}_{10} & p^{10}_{01}p^{01}_{11} \\ p^{11}_{10}p^{10}_{00} & p^{11}_{10}p^{10}_{01} & p^{11}_{10}p^{10}_{00} & p^{11}_{11}p^{11}_{11} \end{pmatrix}.$$

The flattenings of stochastic tensors make obvious the embedding of measures having lower-order dependencies in spaces of measures having higer-order dependencies, e.g. the

Bernoulli measures have stochastic tensors

$$\tilde{\tau}_0 = \begin{pmatrix} p_0 & p_1 \end{pmatrix}, \tilde{\tau}_1 = \begin{pmatrix} p_0 & p_1 \\ p_0 & p_1 \end{pmatrix}, \tilde{\tau}_2 = \begin{pmatrix} p_0 & p_1 & 0 & 0 \\ 0 & 0 & p_0 & p_1 \\ p_0 & p_1 & 0 & 0 \\ 0 & 0 & p_0 & p_1 \end{pmatrix}, \tilde{\tau}_3 = \begin{pmatrix} p_0 & p_1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & p_0 & p_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & p_0 & p_1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & p_0 & p_1 \\ p_0 & p_1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & p_0 & p_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & p_0 & p_1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & p_0 & p_1 \end{pmatrix},$$

etc.

For similar reasons as mentioned earlier for Markov measures, from a statistical viewpoint, it is reasonable to consider merely positive stochastic tensors. In addition, $\mathfrak{S}_N^N(+) \subset \mathfrak{S}_N^N$ is obviously dense with respect to the topology of coordinate-wise convergence. Notice that a flattening of above stochastic type for a positive stochastic (N, N)-tensor τ_N leads to a primitive stochastic $(\#\mathcal{X})^N \times (\#\mathcal{X})^N$ matrix $\tilde{\tau}_N$. Indeed, it is readily seen that $\tau_N > 0 \Rightarrow (\tilde{\tau}_N)^N > 0$.

If $\mathcal{X}$ is finite and one concentrates on positive stochastic matrices, i.e. positive stochastic tensors of degree $1 + 1$, the associated stochastic vector is uniquely defined by such a matrix. In the binary case, the appropriate $M = \begin{pmatrix} p_0^0 & p_1^0 \\ p_1^0 & p_1^1 \end{pmatrix}$-invariant probability vector is given as $(p_0, p_1) = \left(\frac{p_0^1}{p_1^0 + p_0^1}, \frac{p_1^0}{p_1^0 + p_0^1} \right)$. It is obvious that the mapping which maps the positive stochastic matrix onto its invariant probability is injective and continuous with respect to the topologies of coordinate-wise convergence. Moreover, from the theory of Markov chains it is clear that the corresponding stochastic vector is strictly positive. This motivates the question if that is also true for measures describing higher-order dependencies. Plainly speaking, is it true that stationarity of a probability measure possessing order N dependencies gives rise to the following properties affecting its parametrization. (i) Does a positive stochastic (N, N)-tensor uniquely determine the corresponding lower-dimensional tensors? (ii) Are all lower-dimensional tensors strictly positive? (iii) If so, is the mapping which gives the lower dimensional tensors continuous?

Before answering these questions, an exemplary observation is given for the case of binary data, which is stated as a proposition.

Proposition 5.5.6. *Let $\mu \in \mathcal{P}(\mathcal{C})$ be a probability measure which has the order of dependency be $N = 3$. Then, under the hypotheses that μ is stationary and its highest stochastic tensor τ_3 is positive, the associated lower-dimensional stochastic tensors are uniquely determined by τ_3, they are positive and the mappings which maps τ_3 onto the lower-dimensional tensors is continuous.*

Proof. Using Proposition 5.5.3, stationarity of μ implies the following system of 14 polynomial equations to hold, where the coefficients of the tensors τ_0, τ_1, τ_2 are regarded as variables and those of the tensor τ_3 as coefficients of the polynoms.

(I) $p_{c_1} - \sum_{j\in\{0,1\}} p_j p_{c_1}^j = 0;\ \forall c_1 \in \{0,1\}$

(II) $p_{c_1} p_{c_2}^{c_1} - \sum_{j\in\{0,1\}} p_j p_{c_1}^j p_{c_1 c_2}^{j c_1} = 0;\ \forall c_1, c_2 \in \{0,1\}$

(III) $p_{c_1} p_{c_2}^{c_1} p_{c_2 c_3}^{c_1 c_2} - \sum_{j\in\{0,1\}} p_c p_{c_1}^j p_{c_1 c_2 c_3}^{j c_1 c_2} = 0;\ \forall c_1, c_2, c_3 \in \{0,1\}.$

As one readily checks, the unique solution to above system of equations is given in the following recursive form. Clearly, stochasticity and positivity of the tensor τ_3 is exploited in order to reach this unique solution.

$$\tilde{\tau_2}(\tau_3) = \begin{pmatrix} p_{00}^{00} & p_{01}^{00} & 0 & 0 \\ 0 & 0 & p_{10}^{01} & p_{11}^{01} \\ p_{00}^{10} & p_{01}^{10} & 0 & 0 \\ 0 & 0 & p_{10}^{11} & p_{11}^{11} \end{pmatrix}$$

$$:= \begin{pmatrix} \frac{p_{000}^{100}}{p_{000}^{100}+p_{001}^{000}} & \frac{p_{001}^{000}}{p_{000}^{100}+p_{001}^{000}} & 0 & 0 \\ 0 & 0 & \frac{p_{010}^{101}+p_{100}^{110}(p_{010}^{001}-p_{010}^{101})}{1+(p_{010}^{101}-p_{010}^{001})(p_{100}^{010}-p_{100}^{110})} & \frac{p_{011}^{101}-p_{100}^{010}(p_{010}^{001}-p_{010}^{101})}{1+(p_{010}^{101}-p_{010}^{001})(p_{100}^{010}-p_{100}^{110})} \\ \frac{p_{100}^{110}+p_{010}^{101}(p_{100}^{010}-p_{100}^{110})}{1+(p_{010}^{101}-p_{010}^{001})(p_{100}^{010}-p_{100}^{110})} & \frac{p_{101}^{110}-p_{010}^{001}(p_{100}^{010}-p_{100}^{110})}{1+(p_{010}^{101}-p_{010}^{001})(p_{100}^{010}-p_{100}^{110})} & 0 & 0 \\ 0 & 0 & \frac{p_{110}^{111}}{p_{110}^{111}+p_{111}^{011}} & \frac{p_{111}^{011}}{p_{110}^{111}+p_{111}^{011}} \end{pmatrix},$$

$$\tilde{\tau_1}(\tau_2) = \begin{pmatrix} p_0^0 & p_1^0 \\ p_0^1 & p_1^1 \end{pmatrix} := \begin{pmatrix} \frac{p_{00}^{10}}{p_{01}^{00}+p_{00}^{10}} & \frac{p_{01}^{00}}{p_{01}^{00}+p_{00}^{10}} \\ \frac{p_{10}^{11}}{p_{10}^{11}+p_{11}^{01}} & \frac{p_{11}^{01}}{p_{10}^{11}+p_{01}^{11}} \end{pmatrix},$$

$$\tilde{\tau_0}(\tau_1) = (p_0, p_1) := \left(\frac{p_0^1}{p_0^1+p_1^0}, \frac{p_1^0}{p_0^1+p_1^0} \right).$$

The only cases where stochasticity and positivity is not seen at a first glance are the entries of the second and third rows of the tensor $\tilde{\tau_2}(\tau_3)$. For stochasticity, regard the sum of the entries of the second row, which is given by

$$\frac{p_{010}^{101} + p_{100}^{110}(p_{010}^{001} - p_{010}^{101})}{1 + (p_{010}^{101} - p_{010}^{001})(p_{100}^{010} - p_{100}^{110})} + \frac{p_{011}^{101} - p_{100}^{010}(p_{010}^{001} - p_{010}^{101})}{1 + (p_{010}^{101} - p_{010}^{001})(p_{100}^{010} - p_{100}^{110})} = \frac{1 + (p_{010}^{001} - p_{010}^{101})(p_{100}^{110} - p_{100}^{010})}{1 + (p_{010}^{101} - p_{010}^{001})(p_{100}^{010} - p_{100}^{110})} = 1.$$

Furthermore, suppose that

$$\frac{p_{010}^{101} + p_{100}^{110}(p_{010}^{001} - p_{010}^{101})}{1 + (p_{010}^{101} - p_{010}^{001})(p_{100}^{010} - p_{100}^{110})} \leq 0.$$

By positivity of τ_3, the denominator is clearly positive. Hence, it follows

$$p_{010}^{101} + p_{100}^{110}(p_{010}^{001} - p_{010}^{101}) \leq 0 \Leftrightarrow p_{010}^{101}(1 - p_{100}^{110}) + p_{100}^{110} p_{010}^{001} \leq 0 \Leftrightarrow p_{010}^{101} p_{101}^{110} + p_{100}^{110} p_{010}^{001} \leq 0,$$

which is a contradiction. The same reasoning holds for the third row of $\tilde{\tau}_2$. Moreover, the mappings $\tau_i \mapsto \tilde{\tau}_{i-1}$ are obviously continuous for $i = 1, 2, 3$. □

Note that the system of polynomial equations appearing in the proof of Proposition 5.5.6 can be reduced to 7 equations by omitting redundancies. These arise since the equations for $[0, \dots, c_i] = [1, \dots, c_i]$, $i = 2, 3$, possesses the same amount of information. Proposition 5.5.6 is of particular statistical interest since, once it is clarified that the order of dependency of the data is $N = 3$ and the rows of $\tilde{\tau}_3$ are assumed independent under the prior, it suffices to make inference for the entries of $\tilde{\tau}_3$ separately. More precisely, an order-3 Markov measure μ, which is parametrized by positive stochastic tensors, is solely parametrized by the highest positive stochastic tenosr τ_3. Hence, it suffices to put a prior distribution on $\mathfrak{S}_3^3(+)$ which is then updated appropriately. If required, this inference procedure can be transferred to lower tensors using Proposition 5.5.6. Hence, the learning process for a stationary measure μ which is parametrized by positive stochastic tensors is determined by the learning process for τ_3 alone. For the next higher order of dependencies $N = 4$ a similar description is expected to hold. Actually, it is easy to obtain $p_{000}^{000} = \frac{p_{0000}^{1000}}{p_{0000}^{1000}+p_{0001}^{0000}}$ and $p_{111}^{111} = \frac{p_{1111}^{0111}}{p_{1111}^{0111}+p_{1110}^{1111}}$. However, obtaining further solutions is costly due to the overwhelming complexity of the system of 30 (respectively 15) equations. Hence, the search in on for a rather theoretical approach to solutions of above kind.

As a first step, a result of Freedman (1962) is generalized. Therefore, transitions counts of length N are introduced. For $N \in \mathbb{N}$ call the statistic $t^{(N)} = (t_n^{(N)})_{n\in\mathbb{N}}$, which for $n > N$ is given through

$$t_n^{(N)} : \mathcal{X}^n \to \mathcal{X}^N \times \mathcal{T}_N^N$$
$$(c_1, \dots, c_n) \mapsto \left((c_1, \dots, c_N), [\#\{m = 1, \dots, n-N : (c_m, \dots, c_{m+N}) = (i_1, \dots, i_N)\}]_{i_1,\dots,i_N\in\mathcal{X}}\right)$$

and through $(c_1, \dots, c_n)$ if $n \le N$. Analogously to the case of usual transition counts, for $n > N$ the "terminal state" $(c_{n-N+1}, \dots, c_n)$ can be seen to be uniquely determined by $t_n^{(N)}(c_1, \dots, c_n)$. The proof works along the lines of Martin (1967, Lemma 6.11). Furthermore the statistic can easily be seen to possess S-structure. Then, one has the following result.

Lemma 5.5.7. *Let A probability measure $\mu \in \mathcal{P}^e(\mathcal{X}^{\mathbb{N}}, T)$ which is summarized by $t^{(N)}$ is an order-N Markov measure.*

Proof. The proof is in the fashion of the proof of Freedman (1962, Theorem 2). Define

$$A := [a_1, \dots, a_{n-N}, c_1 \dots, c_N]$$
$$B := [c_1, \cdots, c_N, b_1, \cdots, b_{m-N}].$$

Then, since μ is summarized by $t^{(N)}$, one has for $i, j \ge N$

$$\mu\left([c_1, \cdots, c_N] \bigcap T^{-j}\left[A \cap T^{-(j+n)}B\right]\right)$$
$$= \mu\left(\left([c_1, \cdots, c_N] \cap \left[\bigcup_{x_1,\dots,x_{i+N+1}} [x_1, \dots, x_{i+N+1}, c_1, \dots, c_N]\right]\right) \bigcap T^{-(j+i+N)}\left[A \cap T^{-(n-N)}B\right]\right).$$

Now, ergodicity of μ implies

$$\begin{aligned}&\mu([c_1,\cdots,c_N])\,\mu\left(A\cap T^{-(j+n)}B\right)\\&=\mu\left([c_1,\cdots,c_N]\cap\left[\bigcup_{x_1,\ldots,x_{i+N+1}}[x_1,\ldots,x_{i+N+1},c_1,\ldots,c_N]\right]\right)\mu\left(A\cap T^{-(n-N)}B\right).\end{aligned}\tag{*}$$

Now, suppose first that $\mu([c_1,\cdots,c_N])=0$ which clearly implies $\mu(A)=0$ by stationarity of μ. Hence

$$0=\mu\left(A\cap T^{-n}[x]\right)\mu(A)=\mu([c_1,\ldots,c_N])\mu\left(A\cap T^{-n}[x]\right)=0.$$

Otherwise, i.e. if $\mu([c_1,\cdots,c_N])>0$, it follows by (*) for $l\in\mathbb{N}$ large enough that

$$\mu([c_1,\cdots,c_N])\,\mu\left(A\bigcap T^{-l}[c_1,\ldots,c_N,x]\right)=\mu\left(A\cap T^{-n}[x]\right)\mu\left([c_1,\ldots,c_N]\cap T^{-l}[c_1,\ldots,c_N]\right).$$

Letting $l\to\infty$ and again exploiting ergodicity of μ, this becomes

$$\mu(A)\,\mu([c_1,\ldots,c_N,x])=\mu\left(A\cap T^{-n}[x]\right)\mu([c_1,\ldots,c_N]),$$

which yields the N-order Markov property

$$\frac{\mu\left(A\cap T^{-n}[x]\right)}{\mu(A)}=\frac{\mu([c_1,\ldots,c_N,x])}{\mu(([c_1,\ldots,c_N])}.$$

□

Hence, by Theorem 5.3.5 it is true that all stationary measures which are summarized by the statistic $t^{(N)}$ are representable as mixtures of order-N Markov measures. Next, an analog result to Proposition 5.5.6 shall be given in the case of an arbitrary finite state space and an arbitrary order of dependence.

Theorem 5.5.8. *Let $\mathcal{X}$ be a finite state space and N be a positive integer. Furthermore, let $\mu\in\mathcal{P}^e(\mathcal{X}^{\mathbb{N}},T)$ be Markovian of order N with highest positive stochastic tensor $\tau_N\in\mathfrak{S}_N^N(+)$, $\tilde{\tau}_N$ denote the stochastically flattened version of τ_N and $p^{(N)}:=\left(p_{(c_1,\ldots,c_N)}\right)_{c's\in\mathcal{X}}=(\mu([c_1,\ldots,c_N]))_{c's\in\mathcal{X}}\in\Delta^{N-1}$ denote the stochastic vector which is invariant with respect to $\tilde{\tau}_N$. Then, a solution of τ_{N-1} is given by*

$$p_{c_2\ldots c_N}^{c_1\ldots c_{N-1}}=\frac{p_{(c_1,\ldots,c_N)}}{\sum\limits_{j\in\mathcal{X}}p_{(j,c_1,\ldots,c_{N-1})}}.$$

Moreover, the solution is unique and positive. The solutions of the lower-dimensional tensors which fully parametrize μ are given recursively the same way.

Proof. Recall that the flattened form of a positive stochastic tensor τ_N gives rise to a primitive $N\times N$ stochastic matrix. Hence, by the theory of Markov chains, the invariant distribution $p^{(N)}$ is uniquely determined by τ_N and all the entries of $p^{(N)}$ are positive. The same reasoning holds as well for the lower-dimensional tensors. It remains to show that the sequence of tensors generated this way is actually a solution. However, this is

clear since

$$p_{c_2\dots c_N}^{c_1\dots c_{N-1}} = \frac{p_{(c_1,\dots,c_N)}}{\sum\limits_{j\in\mathcal{X}} p_{(j,c_1,\dots,c_{N-1})}} \Leftrightarrow p_{c_2\dots c_N}^{c_1\dots c_{N-1}} \sum_{j\in\mathcal{X}} p_{(j,c_1,\dots,c_{N-1})} = p_{(c_1,\dots,c_N)},$$

which, by Proposition 5.5.3, is nothing but

$$p_{c_2\dots c_N}^{c_1\dots c_{N-1}} \sum_{j\in\mathcal{X}} \mu([j,c_1,\dots,c_{N-1}]) = \mu([c_1,\dots,c_{N-1}])p_{c_2\dots c_N}^{c_1\dots c_{N-1}} = \mu([c_1,\dots,c_N]).$$

By an recursion argument the assertion follows also for all lower-dimensional stochastic tensors. □

The statement of Theorem 5.5.8 lets one think about a "larger parameter space" in which all the positive parametrization of the finite order Markov measures can be embedded in a natural way in order to accomplish a parametrization of more general stationary (and ergodic) measures. Since Theorem 5.5.8 induces an inverse scheme of the kind

$$\tau_0 \xleftarrow{\phi_1} \tau_1 \xleftarrow{\phi_2} \cdots \xleftarrow{\phi_{N-1}} \tau_{N-1} \xleftarrow{\phi_N} \tau_N \tag{D}$$

one might think of an inverse limit as a suitable parameter. However it is necessary to have a topological or measurable structure on such a space since Bayesian statistics force one to define a prior distribution on the parameter space. This will be achieved by a continuity result for the mappings ϕ in diagram (D), which was already indicated in Proposition 5.5.6.

Theorem 5.5.9. *The mappings $\phi_l : \mathfrak{S}_l^l(+) \to \mathfrak{S}_{l-1}^{l-1}(+)$, $i = 1,\dots,N$ are continuous with respect to the topologies of coordinate-wise convergence.*

Proof. It is known from Schweitzer (1968) that the mapping which maps a stochastic matrix onto its invariant probability vector is continuous as long as the mapping is considered on the space of stochastic matrices that possess a single irreducible set of states. Indeed, it is shown that if $pM_1 = p$, $qM_2 = q$ and $p1 = 1 = q1$, then $p = qH(M_1, M_2)$ for some suitably chosen matrix H which depends on the two irreducible stochastic matrices M_1 and M_2. Plainly the conditions of Schweitzer (1968) are fulfilled for the flattenings of positive stochastic tensor, the mapping that maps a flattened tensor onto its invariant stochastic vector is continuous. Thus, the mapping from Theorem 5.5.8,

$$\tau_N \mapsto \tilde{\tau}_N \mapsto \tilde{\tau}_{N-1} := \left(p_{c_1\dots c_{N-1}}^{c_2\dots c_N}\right)_{c_i\in\mathcal{X}} := \left(\frac{p_{(c_1,\dots,c_N}}{\sum_{j\in\mathcal{X}} p_{(j,c_1,\dots,c_{N-1})}}\right)_{c_i\in\mathcal{X}}$$

is continuous, which proves the assertion of the theorem. □

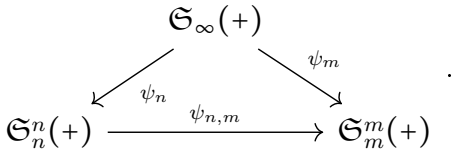
Using Theorem 5.5.8 and Theorem 5.5.9 the following setup for a parametrization of positive T-ergodic measures, i.e. measures which give positive mass to any cylinder set, can be given. An algebraic approach will be employed.

Therefor let for $N \in \mathbb{N}$ the mapping $\phi_N : \mathfrak{S}_N^N(+) \to \mathfrak{S}_{N-1}^{N-1}(+)$ be given as above and define

for $n, m \in \mathbb{N}_0$, $m < n$ the mappings

$$\psi_{n,m} : \mathfrak{S}_n^n(+) \to \mathfrak{S}_m^m(+)$$
$$\psi_{n,m} := \phi_{m+1} \circ \phi_{m+2} \circ \cdots \circ \phi_{n-1} \circ \phi_n.$$

Let $\prod_{l\in\mathbb{N}_0} \mathfrak{S}_l^l(+)$ be the space of sequences of positive stochastic tensors increasing by dimensionality. Furthermore, let for $n \in \mathbb{N}$ the projection onto the n^{th} coordinate be denoted by $pr_n : \prod_{l\in\mathbb{N}_0} \mathfrak{S}_l^l(+) \to \mathfrak{S}_n^n(+)$. Define $\mathfrak{S}_\infty(+) \subset \prod_{l\in\mathbb{N}_0} \mathfrak{S}_l^l(+)$ to be the subset of sequences $\tau = (\tau_0, \tau_1, \dots)$ for which it holds true that $pr_m(\tau) = \psi_{n,m} \circ pr_n(\tau)$ for all $m \leq n$. Now, $\mathfrak{S}_\infty(+) = \varprojlim(\mathfrak{S}_l^l(+), \psi_{n,m})_{m,n,l}$ is the inverse limit of the family of sets $\left\{\mathfrak{S}_l^l(+)\right\}_{l\in\mathbb{N}_0}$ with respect to the family of mappings $\{\psi_{n,m}\}_{n,m\in\mathbb{N}_0, m\leq n}$ and it will be called *positive stochastic unfolding*. For any $n \in \mathbb{N}$ call the restriction of the projections to $\mathfrak{S}_\infty(+)$ the canonical mappings and denote them by $\psi_n := (pr_n)_{|\mathfrak{S}_\infty(+)}$. Hence, for all $m < n$, the following diagram commutes



Clearly, the inverse limit is non-empty since all trivial extensions of i.i.d. measures are contained. Furthermore, it is well known that an inverse limit is essentially unique (up to homeomorphism), so is the unfolding. For further details see e.g. Bourbaki (1968). Moreover, by Theorem 5.5.9, all the topologies on the $\mathfrak{S}_l^l(+)$ with respect to coordinatewise convergence induce a topology on $\mathfrak{S}_\infty(+)$, the inverse topology. That is the topology induced by the product topology on $\prod_{l\in\mathbb{N}_0} \mathfrak{S}_l^l(+)$ and is the coarsest topology that makes the mappings $\{\psi_l\}$ continuous. Note that a basis of the inverse topology is given by the family of sets

$$\mathcal{U} = \left\{\psi_n^{-1}(B_l) : n \in \mathbb{N}, B_l \in U_l\right\},$$

where U_l is a basis of the topology on $\mathfrak{S}_l^l(+)$, c.f. Bourbaki (1969). Hence, by taking the Borel σ-field induced by the inverse topology, denoted by $\mathfrak{B}_\infty$, one eventually obtains a measurable structure on $\mathfrak{S}_\infty(+)$. That makes $(\mathfrak{S}_\infty(+), \mathfrak{B}_\infty)$ a measurable space which serves as the parameter space.

5.6 Bayesian Statistical Inference for Stationary Data

As mentioned earlier, in order to make Bayesian inference for a certain class of measures, it is reasonable to have access to a parametrization of that class. This is due to the fact that in most cases it enables one to execute the integration appearing in the de Finetti theorem or the ergodic decomposition, respectively, on the parameter space rather than on an abstract space of measures. However, while the integration space can be clarified by a suitable parametrization, it remains the question about the specific measure with

respect to which this integration is done. This is the actual statistical model in form of a prior measure and forms the core of Bayesian statistics. Since a prior measure is required to have nice properties with respect to the update mechanism, it operates on an appropriate subspace of the parameter space. For instance, the celebrated Dirichlet process concentrates on the space of discrete measures since a general Polish space gives rise to a measure space that is "too rich". In a way, this is the price one has to pay for a analytically feasible update procedure. The same issue also appears for Bayesian inference for stationary data but in an even more delicate manner. The previously given parametrization of positive stationary measures does provide an appropriate parameter space. Any measure in $\mathcal{P}(\mathfrak{S}_\infty)$ could be used as a prior. However, as long as one wishes to make statistical inference in a closed form with a support as large as possible, some restrictions in form of a model appear to be accepted. The model itself should be chosen such that it supports most of reasonable prior knowledge. One such assumption is that the data which is observed is generated by an positive ergodic measure that possesses a finite order of dependency, i.e. from an order-N Markov measure which is parametrized by positive stochastic tensors. A second assumption is that, given the true order of dependence N_0, the prior samples the rows of $\tilde{\tau}_N$ independently. Put another way, given N_0 the N_0-step transition counts are assumed to be predictively sufficient. A great advantage of this second assumption is that it gives a clear-cut rule how one can learn from observed data, i.e. a clear-cut procedure for updating the prior. This is mainly due to the fact that updating the prior can be done by updating the rows of $\tilde{\tau}_{N_0}$ by the number of the successors, which are associated to the particular rows, separately. Then, by results from the previous section 5.5, the inference procedure is uniquely given by merely updating the (N_0, N_0)-tensor and transferring inference to lower-dimensional tensors through the mappings ϕ. Notice that row-independence is clearly a model assumption that excludes the possibility of learning from transitions corresponding to different rows. This is not an appropriate assumption in some situations in which one must be able to learn for rows from different transitions. For instance, Bayesian inference for stochastic Δ-matrices, which occur in chapter 4 requires such "cross-learning procedures". However, these special learning processes occur in situations which can be judged to correspond to additional prior information, e.g. such as the data is generated from a queueing system. Such additional prior information is not assumed to be available here such that the setup just presented is felt to be the best balance of a large support of the prior and suitable updating on the one hand side and being most "uninformative" on the other.

An actual way to provide inference is to split it in two stages. In the first stage one chooses the order of dependence that, in a certain way, is most likely to govern the data observed so far. Subsequently, the second stage provides inference for the parameters of the measures having a particular dependence structure. While the second stage is fairly clear under the model assumptions, the second deserves further considerations. A Bayesian way to make inference for the order of dependence could consist in placing a prior on the non-negative integers and updating the weights by the data according to some rule. This updating rule should incorporate some properties which come from the fact that solely finite data strings are observed. More precisely, it seems pointless to make "direct" inference for dependency structures of order greater than $n \in \mathbb{N}$ when a data string X_1^n of length n is observed. Furthermore, the dependency structure of the true data generating measure which possesses an order that is relatively small (in comparison to n) should be better exposed from X_1^n than one that is large. This is in full correspondence to information criterions as for instance BIC, i.e. over-fitting of the model should be punished.

Roughly speaking, this means that very high orders of dependency ought to loose mass by observing relatively short data strings but must be able to regain mass when the length of the data string increases considerably.

Before giving a proposal for a direct approach to a prior to posterior update mechanism which allows to infer the order of dependence, two different approaches are briefly recalled. The first traces back to De Finetti (1938). Therein de Finetti intended to make structural inference for contingency tables for which he searched for a description of the data to be "close to exchangeability". As a classical example suppose the situation where one is given data of the form $(X_1, Y_1), \ldots, (X_n, Y_n)$. For the sake of clarity, suppose that the X's and Y's are of binary response. One might think of the X's being the positive or negative response of a medical treatment of women and the Y's of men, respectively. The question may come up if the X's and Y's behave similarly or if they depart from each other considerably. In other words, is the sequence of data $X_1, Y_1, X_2, Y_2, \ldots, X_n, Y_n$ to be treated as an exchangeable sequence, or is it more reasonable to think of the data possessing a rather partial exchangeability property. Latter would mean that merely the X's and Y's are exchangeable among each others separately. Under suitable conditions, both situation provide de Finetti type results which appear as

$$\mathbb{P}\left((X_i, Y_i) = (x_i, y_i); i = 1, \ldots, n\right) = \int\limits_{[0,1]} p^{\sum\limits_{k=1}^{n} x_i + \sum\limits_{k=1}^{n} y_i} (1-p)^{2n - \left[\sum\limits_{k=1}^{n} x_i + \sum\limits_{k=1}^{n} y_i\right]} \mu(dp),$$

for the first and

$$\mathbb{P}\left((X_i, Y_i) = (x_i, y_i); i = 1, \ldots, n\right) = \int\limits_{[0,1]^2} p^{\sum\limits_{k=1}^{n} x_i} (1-p)^{n - \sum\limits_{k=1}^{n} x_i} q^{\sum\limits_{k=1}^{n} y_i} (1-p)^{n - \sum\limits_{k=1}^{n} y_i} \nu(dp, dq),$$

for the second. In a sense these both integral representation depict two extremes. While the first one represents the conditional independence, the second represents the conditional Markov property. Recall that if in the second situation transition counts are judged to be predicively sufficient the mixing measure ν factorizes. Furthermore, if ν concentrates on the diagonal $p = q$ one is in the first situation. With respect to the relationship of the rows of the stochastic matrix these are the two extremes that can happen, i.e. total independence of the rows versus complete dependence in the sense of spitting image of each other. De Finetti suggested to use a prior which is properly centered around the diagonal $p = q$ and might be given as

$$\nu(dp, dq) = C^{-1} e^{-D(p-q)^2} dpdq,$$

where D serves as a scaling parameter and C is the appropriate normalization factor. The power appearing in the exponent is motivated by the central limit and provides an approximation to the posterior law. More precisely, for large samples X_1^n and under relatively mild condition it can be argued that the posterior law can be approximated by

$$\nu(dp, dq) = \tilde{C}^{-1} e^{-\frac{1}{2}\left[(p-q)^2 + nP(p,q)\right]},$$

where P is a polynomial in p and q which provides an update involving empirical versions of the "success"-probabilities under X and Y, respectively, as well as empirical versions

of the deviations from those. For further details and more examples see Bacallado et al. (2015) who recently took up that approach from a rather computational viewpoint using MCMC for calculations of the posterior. They also emphasize the relation to algebraic statistics since for the systems of occurring polynomials it might be useful to calculate Gröbner bases in order to have a conventionally unified approach. In the case considered in this thesis one has systems of polynomials, too. They form the constraints of the parametrization of the positive measures. Hence, one might think of an analogous approach. However, since such methods are based on MCMC procedures, the may become intensive not only from a view point of computational power but also from a viewpoint of implementation. Furthermore, they are based on the CLT which imposes additional constraints and moreover gives just an approximation to the problem.

A second method was indicated by Strelioff et al. (2007) who investigate the problem from a viewpoint which is apparently closer to physics. They also decide to make Bayesian inference N-order Markov chains in two steps. The first step consists in making inference for the transition probabilities. Their method is based on a Dirichlet distribution prior which is put on each row of the flattening of the corresponding stochastic tensor. Hence, even if they do not emphasize that fact, they tacitly assume N-step transition counts to be predictively sufficient. As mentioned earlier, this seems to be the most reasonable approach for at least two reasons. The first is that, from a rather philosophical point of view, knowledge about ability of learning across the rows would amount to additional prior knowledge. Thus, the independence assumption might be seen to be relatively "non-informative". Certainly, by analogy of taking a uniform prior as a non-informative one in a suitably parametric problem, this is not how the story is going to end. The second is a rather practical one. That is an appropriate prior distribution emerges as the product of, in their framework, Dirichlet distributions, as well as the posterior does. So, by taking independent Dirichlet distributions one can set up a model which can be finely tuned by known theory and which allows the usual prior to posterior update using Bayes' theorem. In a second step they infer the order of dependence. Since they intend to exploit Bayes' theorem for their method, this already indicates that some constraint must come into play. More precisely, since Bayes' theorem generally does not hold in non-parametric situations, its usage indicates a rather parametric approach. This fact manifests itself in form of an assumption on the finiteness of the order of dependence. Put another way, they take the maximal order of dependence as known which, in turn, yields a finitely supported prior on the order of dependence. This might be seen as a drawback since it amounts to additional prior knowledge one usually does not have access to, even if it simplifies prior to posterior computations. Two typical priors they present in this framework are a uniform and one that gives a punishment which increases exponentially in the number of parameters.

Here, I suggest a different kind of inference procedure, which will be based on the idea of the update mechanism of the Dirichlet process. The Dirichlet process, which is the most famous non-parametric prior for conditional i.i.d. data in Bayesian statistics, places prior weights on elements of the (countable) parameter space in form of a properly normalized finite measure. These weights are updated by means of the empirical measure. Roughly speaking, this means that weights corresponding to rare events lose mass in favor of events occurring relatively often which gain mass. Thereby, the numbers involving the update are easily obtained by the assumption of conditional independence, i.e. they are just the frequency counts. Notice that a classical form of the Bayes' theorem is not available in

this non-parametric approach in general, see e.g. Schervish (1995) for more details. The basic idea is to generate for each $0 < N < n$ a hypothetical empirical N-transition count matrix under the hypotheses that the order of dependence is $N-1$ and subsequently to measure some distance of the observed N-transition count matrix and the hypothetical one. This approach gives a number which can then be used to update the prior on the orders of dependence. Since the assumption of conditional i.i.d. data clearly fails in the case considered here, the numbers involved in the update will become more complicated and obtaining them is computationally intensive. However, nowadays there is usually a huge amount of computational capacity available which still keeps on growing. Once suitable numbers are available, the target is rather to find a proper update mechanism which is workable in the following sense. It should emerge from an algorithm which is relatively simple to implement and which can be taught to a machine. It should respect the point stated earlier. And it should enable one to update priors with infinite support.

In order to clarify the idea, first regard the following example. Suppose one is given the following data strings of length $n = 20$.

(a) $X_1^{20} = 11111111111111111111$

(b) $X_1^{20} = 11110011011010101111$

(c) $X_1^{20} = 11111110000111100111$

(d) $X_1^{20} = 01010101010101010101$

(e) $X_1^{20} = 11011011011011011011$.

The data string (a) is generated from a Bernoulli measure with success probability $p_1 = 1-10^{-10}$ and the string (b) from a Bernoulli measure with $p_1 = 3/4$. The strings (c) and (d) are generated from stationary Markov measures with parametrization $(1/4, 3/4), \begin{pmatrix} 7/10 & 3/10 \\ 1/10 & 9/10 \end{pmatrix}$ and $(1/2, 1/2), \begin{pmatrix} 1/100 & 99/100 \\ 99/100 & 1/100 \end{pmatrix}$, respectively. The data string (d) is generated from a stationary 2-Markov measure with parametrization $\begin{pmatrix} 51/151 & 100/151 \end{pmatrix}, \begin{pmatrix} 1/50 & 50/51 \\ 1/2 & 1/2 \end{pmatrix}, \begin{pmatrix} 1/2 & 1/2 & - & - \\ - & - & 1/100 & 99/100 \\ 1/100 & 1/100 & - & - \\ - & - & 99/100 & 1/100 \end{pmatrix}$. If one is intended to measure a distance of an empirical transition count matrix under the hypothesis that the data is generated from a Bernoulli measure and the actually observed one, the strings should give larger distances in increasing order up to (e). Indeed, already on a first glance the data (d) looks more likely to stem from a Markov measure as the data (a). Let for $j \in \{0,1\}$ the empirical measure be denoted as $e_n(j) = e_n^{(0)} = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}(\{j\})$. Define the empirical transition count matrix under the hypothesis that the data is generated from a

Bernoulli measure by

$$\hat{tc}_n = (n-1)\begin{pmatrix} e_n(0)e_n(0) & e_n(0)e_n(1) \\ e_n(1)e_n(0) & e_n(1)e_n(1) \end{pmatrix}.$$

For above data strings that becomes

(a) $\hat{tc}_n = 19\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 19 \end{pmatrix}$

(b) $\hat{tc}_n = 19\begin{pmatrix} 6/20 * 6/20 & 6/20 * 14/20 \\ 14/20 * 6/20 & 14/20 * 14/20 \end{pmatrix} = \begin{pmatrix} 1.71 & 3.99 \\ 3.99 & 9.31 \end{pmatrix}$

(c) $\hat{tc}_n = 19\begin{pmatrix} 6/20 * 6/20 & 6/20 * 14/20 \\ 14/20 * 6/20 & 14/20 * 14/20 \end{pmatrix} = \begin{pmatrix} 1.71 & 3.99 \\ 3.99 & 9.31 \end{pmatrix}$

(d) $\hat{tc}_n = 19\begin{pmatrix} 1/2 * 1/2 & 1/2 * 1/2 \\ 1/2 * 1/2 & 1/2 * 1/2 \end{pmatrix} = \begin{pmatrix} 4.75 & 4.75 \\ 4.75 & 4.75 \end{pmatrix}$

(e) $\hat{tc}_n = 19\begin{pmatrix} 6/20 * 6/20 & 6/20 * 14/20 \\ 14/20 * 6/20 & 14/20 * 14/20 \end{pmatrix} = \begin{pmatrix} 1.71 & 3.99 \\ 3.99 & 9.31 \end{pmatrix}$

while the actually observed transition count matrices are given by

(a) $tc_n = \begin{pmatrix} 0 & 0 \\ 0 & 19 \end{pmatrix}$

(b) $tc_n = \begin{pmatrix} 1 & 5 \\ 5 & 8 \end{pmatrix}$

(c) $tc_n = \begin{pmatrix} 4 & 2 \\ 2 & 13 \end{pmatrix}$

(d) $tc_n = \begin{pmatrix} 0 & 10 \\ 9 & 0 \end{pmatrix}$

(e) $tc_n = \begin{pmatrix} 0 & 6 \\ 6 & 7 \end{pmatrix}$.

Using entry-wise absolute distance $d_1^{(n)}$ or entry-wise quadratic distance $d_2^{(n)}$, respectively, one obtains

(a) $d_1^{(n)}(tc_n, \hat{tc}_n) = 0.00$; $d_2^{(n)}(tc_n, \hat{tc}_n) = 0.0$

(b) $d_1^{(n)}(tc_n, \hat{tc}_n) = 4.04$; $d_2^{(n)}(tc_n, \hat{tc}_n) \approx 2.1$

(c) $d_1^{(n)}(tc_n, \hat{tc}_n) = 9.96$; $d_2^{(n)}(tc_n, \hat{tc}_n) \approx 5.2$

(d) $d_1^{(n)}(tc_n, \hat{tc}_n) = 19.0$; $d_2^{(n)}(tc_n, \hat{tc}_n) \approx 9.5$

(e) $d_1^{(n)}(tc_n, \hat{tc}_n) = 8.04$; $d_2^{(n)}(tc_n, \hat{tc}_n) \approx 4.0$.

So, the distances behave as they were expected to, where it should be stressed that the string (e) gives less evidence to the 1-Markov measures as string (c) does. This is due to the fact that string (e) is generated from a 2-Markov measure. To continue the example, an analogous approach is given for $N = 2$. Therefor define the empirical transition probabilities as

$$e_n^{(1)}(j_1, j_2) := \begin{cases} \frac{\sum_{i=1}^{n-1} \delta_{(X_i, X_{i+1})}(\{(j_1, j_2)\})}{n e_n^{(0)}(j_1)}, & e_n^{(0)}(j_1) \neq 0 \\ 0, & e_n^{(0)}(j_1) = 0 \end{cases}$$

and the emprical 2-transition count matrix under the hypothesis that the data is 1-Markov as

$$\hat{tc}_n^{(2)} := (n-2) \times \begin{pmatrix} e_n^{(0)}(0)e_n^{(1)}(0,0)e_n^{(1)}(0,0) & e_n^{(0)}(0)e_n^{(1)}(0,0)e_n^{(1)}(0,1) & - \\ - & e_n^{(0)}(0)e_n^{(1)}(0,1)e_n^{(1)}(1,0) & e_n^{(0)}(0)e_n^{(1)}(0,1)e_n^{(1)}(1,1) \\ e_n^{(0)}(1)e_n^{(1)}(1,0)e_n^{(1)}(0,0) & e_n^{(0)}(1)e_n^{(1)}(1,0)e_n^{(1)}(0,1) & - \\ - & e_n^{(0)}(1)e_n^{(1)}(1,1)e_n^{(1)}(1,0) & e_n^{(0)}(1)e_n^{(1)}(1,1)e_n^{(1)}(1,1) \end{pmatrix}.$$

For above strings, the 2-transition count matrices and the empirical ones under the 1-Markov hypothesis are given by

(a) $tc_n^{(20)} = \begin{pmatrix} 0 & 0 & - & - \\ - & - & 0 & 0 \\ 0 & 0 & - & - \\ - & - & 0 & 18 \end{pmatrix}$; $\hat{tc}_n^{(2)} = \begin{pmatrix} 0 & 0 & - & - \\ - & - & 0 & 0 \\ 0 & 0 & - & - \\ - & - & 0 & 18 \end{pmatrix}$

$\Rightarrow d_1^{(n)}\left(\hat{tc}_n^{(2)}, tc_n^{(2)}\right) = 0 = d_2^{(n)}\left(\hat{tc}_n^{(2)}, tc_n^{(2)}\right)$

(b) $tc_n^{(20)} = \begin{pmatrix} 0 & 1 & - & - \\ - & - & 2 & 3 \\ 1 & 4 & - & - \\ - & - & 3 & 4 \end{pmatrix}$; $\hat{tc}_n^{(2)} \approx \begin{pmatrix} 0.15 & 0.75 & - & - \\ - & - & 1.73 & 2.77 \\ 0.81 & 4.04 & - & - \\ - & - & 3 & 4.7 \end{pmatrix}$

$\Rightarrow d_1^{(n)}\left(\hat{tc}_n^{(2)}, tc_n^{(2)}\right) = 1.83$; $d_2^{(n)}\left(\hat{tc}_n^{(2)}, tc_n^{(2)}\right) \approx 0.86$

(c) $tc_n^{(20)} = \begin{pmatrix} 2 & 2 & - & - \\ - & - & 0 & 2 \\ 2 & 0 & - & - \\ - & - & 2 & 8 \end{pmatrix}$; $\hat{tc}_n^{(2)} \approx \begin{pmatrix} 2.4 & 1.2 & - & - \\ - & - & 0.27 & 1.5 \\ 1.3 & 0.64 & - & - \\ - & - & 1.64 & 9 \end{pmatrix}$

$\Rightarrow d_1^{(n)}\left(\hat{tc}_n^{(2)}, tc_n^{(2)}\right) \approx 4.4$; $d_2^{(n)}\left(\hat{tc}_n^{(2)}, tc_n^{(2)}\right) \approx 1.66$

(d) $tc_n^{(20)} = \begin{pmatrix} 0 & 0 & - & - \\ - & - & 9 & 0 \\ 0 & 9 & - & - \\ - & - & 0 & 0 \end{pmatrix}$; $\hat{tc}_n^{(2)} = \begin{pmatrix} 0 & 0 & - & - \\ - & - & 9 & 0 \\ 0 & 9 & - & - \\ - & - & 0 & 0 \end{pmatrix}$

$\Rightarrow d_1^{(n)}\left(\hat{tc}_n^{(2)}, tc_n^{(2)}\right) = 0 = d_2^{(n)}\left(\hat{tc}_n^{(2)}, tc_n^{(2)}\right)$

(e) $tc_n^{(20)} = \begin{pmatrix} 0 & 0 & - & - \\ - & - & 0 & 6 \\ 0 & 6 & - & - \\ - & - & 6 & 0 \end{pmatrix}$; $\hat{tc}_n^{(2)} = \begin{pmatrix} 0 & 0 & - & - \\ - & - & 2.5 & 3 \\ 0 & 6 & - & - \\ - & - & 3 & 3.5 \end{pmatrix}$

$\Rightarrow d_1^{(n)}\left(\hat{tc}_n^{(2)}, tc_n^{(2)}\right) = 12$; $d_2^{(n)}\left(\hat{tc}_n^{(2)}, tc_n^{(2)}\right) \approx 6$.

It is interesting that string (d) which has the strongest evidence of being order 1-Markovian at the same time has no evidence of being 2-Markovian. The evidence of strings (b) and (c) being 2-Markovian is less than that of being 1-Markovian. However, it does not vanish since it might stem from a higher-order which did not reveal itself yet in the data of length n. That is the same for the string (a). If n increases observing a 1 in string (a) might change ones opinion in the order of dependence dramatically. For instance if (a) is extended to $\overline{(\overline{0}_{20}1)}_m$ for a suitable $m \in \mathbb{N}$ one might focus on an order of dependence of $N = 20$ even if for $n = 20$ it is most reasonable to give high posterior weight on $N = 0$.

For a general approach proceed as in the example. For $l > 0$ and $j's \in \mathcal{X}$ define by

$$e_n^{(l)}(j_1, \ldots, j_{l+1}) := \begin{cases} \frac{\sum_{i=1}^{n-l} \delta_{(X_i,\ldots,X_{i+l})}(\{(j_1,\ldots,j_{l+1})\})}{\sum_{i=1}^{n-(l-1)} \delta_{(X_i,\ldots,X_i)}(\{(j_1,\ldots,j_l)\})} & , \sum_{i=1}^{n-(l-1)} \delta_{(X_i,\ldots,X_i)}(\{(j_1,\ldots,j_l)\}) \neq 0 \\ 0 & , \sum_{i=1}^{n-(l-1)} \delta_{(X_i,\ldots,X_i)}(\{(j_1,\ldots,j_l)\}) = 0 \end{cases}$$

the empirical conditional probability of a transition $(j_1 \dots, j_l) \to (j_2, \dots, j_{l+1})$. Furthermore, define the l-transition count tensor under the hypothesis of a $(l-1)$-Markov measure as

$$\hat{tc}_n^{(l)} := \hat{tc}^{(l)}(X_1^n) := (n-l)\left(\prod_{i=1}^{l-1} e_n^{(i)}(j_1, \dots, j_{i+1}) \times e_n^{(l-1)}(j_2, \dots, j_l)\right)_{j's \in \mathcal{X}}$$

and the actually observed one as $tc_n^{(l)} = tc^{(l)}(X_1^n)$. Let $d^{(l)} : \mathcal{T}_l^l \times \mathcal{T}_l^l \to \mathbb{R}_+$ be an appropriate distance and define $d_l^{(n)} := \max_{X_1^n \in \mathcal{X}^n} d^{(l)}\left(tc^{(l)}(X_1^n), \hat{tc}^{(l)}(X_1^n)\right)$ as well as $w_l^{(n)} := d^{(l)}\left(tc_n^{(l)}, \hat{tc}_n^{(l)}\right)$. For a finite measure α on the non-negative integers let $a := \alpha(\mathbb{N}_0)$ such that $\pi(\bullet) := \frac{\alpha(\bullet)}{a} \in \mathcal{P}(\mathbb{N}_0)$ is a proper prior on the orders of dependence. Observing the data X_1^n define the update $\pi \overset{X_1^n}{\to} \pi^{(n)} = \pi(\bullet | X_1^n)$ for $1 < l < n-1$ as

$$\pi_l \overset{X_1^n}{\to} \frac{1}{a + (n-1)}\left[\alpha_l + \frac{w_l^{(n)}}{d_l^{(n)}} \sum_{i=l}^{n-1} \prod_{m=l+1}^{i} \frac{d_m^{(n)} - w_m^{(n)}}{d_m^{(n)}}\right],$$

where the empty product is taken to equal 1. Moreover, define $\pi_m^{(n)} := \frac{\alpha_m}{a+(n-1)}$ for all $m \geq n$ and $\pi_0^{(n)} := \frac{1}{a+(n-1)} \sum_{i=1}^{n-1} \prod_{m=1}^{i} \frac{d_m^{(n)} - w_m^{(m)}}{d_m^{(n)}}$ for $l = 0$. In full detail this amounts to the following scheme

$$\begin{aligned}
\pi_0^{(n)} &= \frac{1}{\alpha + (n-1)}\left[\alpha_0 + \frac{d_1^{(n)} - w_1^{(n)}}{d_1^{(n)}}\left(1 + \frac{d_2^{(n)} - w_2^{(n)}}{d_2^{(n)}} + \dots + \prod_{i=2}^{n-1} \frac{d_i^{(n)} - w_i^{(n)}}{d_i^{(n)}}\right)\right] \\
\pi_1^{(n)} &= \frac{1}{\alpha + (n-1)}\left[\alpha_1 + \frac{w_1^{(l)}}{d_1^{(n)}}\left(1 + \frac{d_2^{(n)} - w_2^{(n)}}{d_2^{(n)}} + \dots + \prod_{i=2}^{n-1} \frac{d_i^{(n)} - w_i^{(n)}}{d_i^{(n)}}\right)\right] \\
\pi_2^{(n)} &= \frac{1}{\alpha + (n-1)}\left[\alpha_2 + \frac{w_2^{(l)}}{d_2^{(n)}}\left(1 + \frac{d_3^{(n)} - w_3^{(n)}}{d_3^{(n)}} + \dots + \prod_{i=3}^{n-1} \frac{d_i^{(n)} - w_i^{(n)}}{d_i^{(n)}}\right)\right] \\
&\vdots \\
\pi_{n-2}^{(n)} &= \frac{1}{a + (n-1)}\left[\alpha_{n-2} + \frac{w_{n-2}^{(n)}}{d_{n-2}^{(n)}}\left(1 + \frac{d_{n-1}^{(n)} - w_{n-1}^{(n)}}{d_{n-1}^{(n)}}\right)\right] \\
\pi_{n-1}^{(n)} &= \frac{1}{a + (n-1)}\left[\alpha_{n-1} + \frac{w_{n-1}^{(n)}}{d_{n-1}^{(n)}}\right] \\
\pi_m^{(n)} &= \frac{\alpha_m}{a + (n-1)}, \forall m \geq n.
\end{aligned}$$

To briefly comment on the update scheme, the idea is the following. Analogously as for the Dirichlet process, observing a data string of length n brings in additional total weight of $(n-1)$. However, in the case considered here the single weights do not amount to rather direct observations which, for the Dirichlet process, account for the assumption of exchangeable data. Instead the weights stem from a measurement of the distance between several classes of measures carrying different orders of dependence. Once the weights are available, they are split according to their evidence data of length n is able to provide. More precisely, the weight $w_l^{(n)}$ is used to update $\pi_l^{(n)}$ rather direct way by the

fraction $\frac{w_l^{(n)}}{d_l^{(n)}}$, while the remaining mass is spread to the lower-dimensional classes according to their empirical evidence of being the true one. Thereby, as usual the prior guess is washed away by increasing information in form of data. Moreover, there seems to be hope for posterior consistency since one would expect the updating weights to converge to 1 or 0, respectively, depending on they correspond to the true data generating order or not.

To sum it up, this gives the following model for sampling stationary data

$$\begin{aligned} N &\sim \pi \in \mathcal{P}(\mathbb{N}_0) \\ \tau_N \mid N &\sim \otimes_{i=1}^{k^N} \mathsf{Dir}_i\left(\overline{h_i}^k\right) \\ X_1^n \mid N, \tau_N &\sim \mu_{\tau_N}^{(N)}, \end{aligned}$$

where $k = \#\mathcal{X}$, $\mu_{\tau_N}^{(N)}$ denotes the Markov measure of order N, which is parametrized by $(\tau_0(\tau_N), \tau_1(\tau_N), \ldots, \tau_{N-1}(\tau_N), \tau_N)$ and $\mathsf{Dir}_i\left(\overline{h_i}^k\right) \in \mathcal{P}(\Delta_{k-1})$ means a Dirichlet prior for the i^{th} row of the flattening of τ_N with hyper-parameters $\overline{h_i}^k := \left(h_i^{(1)}, \ldots, h_i^{(k)}\right) \in \Delta_{k-1}$ for all $i \in k^N$. Hence, after a suitable parametrization, a prior $\eta \in \mathcal{P}(\mathcal{P}_+^e(\mathcal{X}^{\mathbb{N}}, T))$ on the space of positive shift-ergodic measures is provided through

$$\tilde{\eta}(B) = \sum_{N \in \mathbb{N}_0} \pi_N \int_{\mathfrak{S}_N^N(+)} 1_{\bar{\mathfrak{S}}_N^N(+) \cap B}(\tau_N) \left[\otimes_{i=1}^{k^N} \mathsf{Dir}_i\left(\overline{h_i}^k\right)\right](d\tau_N),$$

where $B \subset \mathfrak{S}_\infty(+)$ and $\bar{\mathfrak{S}}_N^N(+)$ is the natural embedding of $\mathfrak{S}_N^N(+)$ into $\mathfrak{S}_\infty(+)$. Note the difference of $\bar{\mathfrak{S}}_N^N(+) \cap B$ and $\psi_N(B)$; for the latter the independence assumption of the Dirichlet distributions for dimensions exceeding N becomes vacuous in general. The corresponding posterior measure is provided by above update mechanism which, under the assumption of row-wise independence for the flattened stochastic tensors, yields $\eta(\bullet) \overset{X_1^n}{\to} \eta^{(n)}(\bullet) := \eta(\bullet \mid X_1^n)$ with

$$\tilde{\eta}^{(n)}(B) = \sum_{N \in \mathbb{N}_0} \pi_N^{(n)} \int_{\mathfrak{S}_N^N(+)} 1_{\bar{\mathfrak{S}}_N^N(+) \cap B}(\tau_N) \left[\otimes_{i=1}^{k^N} \mathsf{Dir}_i\left(\overline{h_i^{(n)}}^k\right)\right](d\tau_N),$$

where $\overline{h_i^{(n)}}^k$ denotes the hyper-parameters updated by N-transition counts in the common way.

6 Conclusions and Outlook

In this thesis, I dealt with the issue of making statistical inference for continuous-time queueing systems from the Bayesian perspective. From an applied point of view, queueing systems are vivid stochastic models which are often set up with a certain intention in mind or at least with a initial idea of its future behavior. This amounts to prior knowledge which can be represented within the framework of Bayesian statistics. Observing the system or parts of it, the scientist can improve the prediction about the system's future behavior by updating the prior knowledge through the data. Formalizing this update mechanism, which may be seen as a learning process projecting past observations onto future prediction, is the major assignment of Bayesian statistics. From a rather theoretical viewpoint, queueing systems can be seen as functionals of several stochastic input-processes which characterize the queueing system. One such input is the arrival stream of customers to the system. Throughout the thesis, this was taken to be a Poisson process for at least two reasons. One is that Poisson arrivals fit most situations fairly well. Another is that this assumption ensures the system to possess several nice properties which can be exploited to increase the range of statistical inference. As adequate this particular parametric assumption on the arrival stream is, as inadequate often is any parametric assumption on the serving process. From the subjectivist point of view, the main reason for this is that a parametric assumption would correspond to additional prior knowledge which is usually not available and thus would lead to poor inference procedures. Hence, one has to model the service time nonparametrically, leading to the $M/G/c$ model, where c is the number of serving stations. The output of the queueing functional consists of further stochastic processes as for instance the occupation of the system or the waiting time process of the customers. However, often the process one wishes to infer is not directly observable. For that reason one has to develop statistical methods which enable one to make inference for objects of interest using observations of different processes. These statistical methods base on the usage of functional relationships between the observables and the object of interest. While the first two parts of the thesis were devoted to the issue of indirect Bayesian nonparametric inference for the single server model $M/G/1$, in the last part I examined related questions for the $M/G/\infty$ model.

In chapter 3 of the thesis, I developed indirect inference methods for the continuous-time $M/G/1$ queueing model. The object of interest was the distribution of the size of the system and the waiting time distribution. Even if those processes were directly observable, it is not clear at all how to make Bayesian inference for them. The main reason for this is that, under some reasonable constraints, they form general stationary processes. However, the aim was to make indirect inference which, at the same time, provides a solution to aforementioned issue. The methods were based on the observations of the arrival stream and the service times, respectively. Since the $M/G/1$ queue evolves in continuous time, the law of the random service time c.d.f. is chosen to have a large support in order to express one's prior knowledge about G as well as possible. Therefore,

a certain neutral-to-right prior was assigned to the distribution of G, namely the beta-Stacy process. The family of beta-Stacy processes, which contains the Dirichlet process as a special case, is well understood since it bases on the theory of increasing additive processes. Moreover, explicit closed-form formulas can be given describing the update to the associated posterior law and posterior consistency results are known for that class of priors under relatively mild conditions. However, it appears that one is interested in statistics for the traffic coefficient of the queue, thus also in the expected value of G, which is random itself. Since G does not have a finite support in general, consistency results for G can not be easily transferred to its mean. Hence, I gave a direct proof of the fact that the distribution of the random mean associated to G does actually center around its true value. Using this result, I showed posterior consistency and posterior normality results for the estimators of the Laplace transform of the waiting-time distribution and the probability generating function of the system size, respectively.

Chapter 4 was also dedicated to the $M/G/1$ system but under a different observational setup. In contrast to chapter 3 where the input, in form of arrivals and services, was assumed to be observable, here the queueing system is assumed to be a complete black box. The only thing which is assumed to be observable is the departure stream of the customers. In that situation one is usually interested in making inference for the service time distribution G. By a known result from queueing theory, G relates to the parameter of the Poisson arrival process and the distribution of the size of the system at instances of departures, which can be seen as marks of the marked departure process. Since, in steady state, the departure process behaves as the arrival stream, inferring the intensity of the arrival stream is a minor problem. In contrast to that, the marks of the departure process deserve some deeper study. It is well known that these marks form a Markov chain on the non-negative integers, the so called embedded Markov chain of the $M/G/1$ system. Hence, from a subjectivist point of view, a random Markov measure needs to be given which almost surely reflects the $M/G/1$ origin of the data and which is possible to be updated by observed data consisting of the marks. In order to clarify the problem and to find a satisfying solution, I presented a deeper study concerning the statistical structure of the probability measures governing the embedded Markov chain. By a new result obtained in this thesis, the measures governing the embedded Markov chain are summarized by a certain statistic which in turn possesses S-structure. Together with Freedman's S-structure theorem, this result paved the way for the theoretical clarification of the existence of a prior on the space of Markov measures governing the embedded chain. As a particular model for that prior, a certain functional of the Dirichlet process was taken that lead to consistency and normality results for the emerging posterior of the random stochastic matrix. These results were finally used to obtain analog properties of the Bayesian estimator for G.

Chapter 5 departed from the previous two since the continuous-time queueing model $M/G/\infty$ with Poisson input and generally distributed service times, but with infinitely many servers was considered. The basic assumption was that only time points when customers enter and leave the system, respectively, can be recorded. By a known result, G relates to the the distribution of some functional of the observations. However, this functional amounts to a general stationary process and Bayesian statistics for stationary data taking values in arbitrary Polish spaces is a difficult task which has not been studied intensively, yet. Hence, in a first step, I simplified the problem by assuming the

data to take values in a state space with finitely many elements in order to clarify the problem. The theory then presented shed light on the difficulties of Bayesian inference for stationary data taking values in a finite state space by developing a parametrization of a statistically appropriate subset of the shift-ergodic measures. This parametrization relied on the algebraic idea of an inverse limit which is taken with respect to measurable mappings between suitable parameter spaces of measures carrying dependencies of fixed finite range. A Bayesian statistical model, in the first instance, requires to elicit a prior distribution that is able to express one's à priori knowledge. Simply put, the model reveals the mechanism which generates the data. After having clarified the associated parameter space for stationary/ergodic measures, the model was chosen in a way that allows for inference for stationary data sampled according to this mechanism. The data generating mechanism itself evolved in three consecutive stages. In the first one, the order of dependence of the measure that eventually generates the data was sampled from a distribution on the non-negative integers. Thus, one model assumption was that the true order of dependence is finite but by no means bounded. In the second stage, given the order of dependence just sampled, a stochastic tensor, which amounts to a natural generalization of stochastic matrices encoding the long range dependence transition probabilities, was sampled. Here a second crucial assumptions appeared. Namely, the independence assumption among sectors of the stochastic tensors that relate to different predecessor states of the transition probabilities. This assumption allowed to sample "row-wise" from Dirichlet distributions which induce a probability measure on the space of stochastic tensors. By the theoretical considerations of the given parametrization, a stochastic tensor of a certain dimension fully determines all the lower-dimensional ones and therefore the measure which in turn generates the data. Sampling the data from this measure was the last stage.

Certainly, this model is not the only possible. However, it has several advantages. Firstly it is workable in the sense that the mechanism the data is sampled from becomes lucid. Secondly, from a rather philosophical viewpoint, it relates to a model which is (in abuse of language) merely sparsely informative. This is mainly due to the row-wise independence assumption that does not allow for learning across the rows, since the opportunity of such procedures might be regarded as additional information. The third and probably the most striking advantage is that it allows for a clear-cut update by observed data. Therefor the stage-wise sampling scheme is exploited by updating the Dirichlet distributions separately by higher order transition counts in a first step to be followed by updating the prior on the order of dependence in a second. While the first one is well known, the second relies on a specific non-trivial topological measurement of the empirical strength of belief in the order of dependence.

As an outlook it is worth mentioning the examination of the stated update mechanism with respect to posterior consistency. More precisely to answer the question whether the posterior law does actually center around the parameter of the true data generating measure which is sampled according to the aforementioned scheme. Therefor, due to the topological considerations which the posterior builds on, this is supposed to come along with further sophisticated theory. But there seems to be hope for an affirmative answer since a Doob-type posterior consistency result was developed in Lijoi et al. (2007). However, since Doob-type consistency results classically rely on reverse martingale techniques they do not give much insight into what is happening in more detail. Hence, there is additional need for consistency results for explicit and workable models.

Furthermore, it would be interesting to extend the theory to more general situations. Thereby, the most interesting generalizations would of course be with respect to the state space on the one hand side and with respect to the model itself. Extending the presented theory to countably infinite or even uncountable state spaces seems fairly difficult since the mappings which are used to construct the parameter space of the positive shift-ergodic measures are not easily seen to be continuous or measurable, respectively. For the case of countably infinite state truncation techniques as in Gibson and Seneta (1987) are expected to be of help. The case of infinite state spaces is presumably even harder since finite approximations to the kernels determining the shift-ergodic measures would be necessary. The work of Fortini et al. (2002) may give an idea of such approximations. Moreover, perturbation techniques of general operators as in Kato (1995) might be helpful. Merely considering usual Markov measures (i.e. the dependency range of the induced process is unity) on infinite state spaces, a basic question will be how to model the learning procedure for the Markov kernel. More precisely, what one can learn about transitions of the process into some Borel set given the process was located to a certain region of the state space based on observations that belong to distinct regions. In order to visualize the issue, I personally find it helpful to imagine pebbles thrown onto the surface of some fluid, the pebbles standing for the observed data. How long can the repercussion of the emerging circles be modeled? How can one build further statistical models out of that?
Extending the model itself can be done in several ways and this depends on the certain situation one wishes to model. Basically, any probability measure one the developed parameter space of the shift-ergodic measures can be used as a prior. However, generally it is not easy to give a full description of its push-forwards under finite-dimensional projections nor of the appropriate posterior measure whenever the row-wise independence assumption is dropped. Perhaps further rigid models with known update mechanisms can be build in order to mix over them subsequently. This might create an additional level of modeling as well as expressing one's uncertainty. Then the question would naturally arise to the mathematician whether there is a reasonable family of explicitly describable models that is dense in all models sampling stationary data or at least in a statistically appropriate subset of them.

Bibliography

L. Abolnikov and A. Dukhovny. Markov chains with transition Δ-matrix: ergodicity conditions, invariant probability measures and applications. *International Journal of Stochastic Analysis*, 4(4):333–355, 1991.

C. Armero. Bayesian analysis of $M/M/1/\infty/FIFO$ queues. *Bayesian Statistics*, 2:613–618, 1985.

C. Armero and D. Conesa. Inference and prediction in bulk arrival queues and queues with service in stages. *Applied Stochastic Models and Data Analysis*, 14(1):35–46, 1998.

C. Armero and D. Conesa. Statistical performance of a multiclass bulk production queueing system. *European Journal of Operational Research*, 158(3):649–661, 2004.

C. Armero and D. Conesa. Bayesian hierarchical models in manufacturing bulk service queues. *Journal of Statistical Planning and Inference*, 136(2):335–354, 2006.

S. Asmussen. *Applied Probability and Queues*, volume 51. Springer Science & Business Media, 2008.

M. C. Ausín, R. E. Lillo, and M. P. Wiper. Bayesian estimation for the $M/G/1$ queue using a phase-type approximation. *Journal of Statistical Planning and Inference*, 118 (1):83–101, 2004.

M. C. Ausín, R. E. Lillo, and M. P. Wiper. Bayesian control of the number of servers in a $GI/M/c$ queueing system. *Journal of Statistical Planning and Inference*, 137(10): 3043–3057, 2007.

S. Bacallado, P. Diaconis, and S. Holmes. De Finetti priors using Markov chain Monte Carlo computations. *Statistics and Computing*, 25(4):797–808, 2015.

P. Billingsley. *Convergence of Probability Measures*. John Wiley & Sons, 1999.

D. Blackwell. The entropy of functions of finite-state Markov chains. In *Trans. First Prague Conf. Information Thoery, Statistical Decision Functions, Random Processes*, pages 13–20, 1957.

D. Blackwell and J. B. MacQueen. Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, pages 353–355, 1973.

N. Bourbaki. *Theory of Sets*. Springer Science & Business Media, 1968.

N. Bourbaki. *General Topology*. Springer Science & Business Media, 1969.

R. M. Bowen and C.-C. Wang. *Introduction to vectors and tensors*, volume 2. Courier Corporation, 2008.

P. H. Brill. *Level Crossing Methods in Stochastic Models*, volume 123. Springer Science & Business Media, 2008.

M. Brown. An $M/G/\infty$ estimation problem. *The Annals of Mathematical Statistics*, 41 (2):651–654, 1970.

K. L. Chung. *Markov Chains with Stationary Transition Probabilities*. Springer, 1967.

P. L. Conti. Large sample Bayesian analysis for $Geo/G/1$ discrete-time queueing models. *The Annals of Statistics*, pages 1785–1807, 1999.

A. P. Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–31, 1979.

B. De Finetti. Sur la condition d'equivalence partielle, Actualit & Scientifques et Industrielles, No. 739 (Paris: Hermann & Cie). Translated by P. Benacerraf and R. Jeffrey as "On the Condition of Partial Exchangeability". *Studies in Inductive Logic and Probability*, 2:193–205, 1938.

B. De Finetti. *Theory of Probability*, volume 1. Wiley New York, 1974.

P. de Laplace. Mémoire sur les suites récurro-récurrentes et sur leurs usages dans la théorie des hasards. *Mém. Acad. Roy. Sci. Paris*, 6:353–371, 1774.

J. Dey, R. Erickson, and R. Ramamoorthi. Some aspects of neutral to right priors. *International Statistical Review*, 71(2):383–401, 2003.

P. Diaconis and D. Freedman. De Finetti's theorem for Markov chains. *The Annals of Probability*, pages 115–130, 1980.

P. Diaconis and D. Freedman. On the consistency of Bayes estimates. *The Annals of Statistics*, pages 1–26, 1986.

P. Diaconis and D. Ylvisaker. Quantifying prior opinion. *Bayesian Statistics*, 2:133–156, 1985.

K. Doksum. Tailfree and neutral random probabilities and their posterior distributions. *The Annals of Probability*, pages 183–201, 1974.

J. L. Doob. Application of the theory of martingales. *Le calcul des probabilites et ses applications*, pages 23–27, 1949.

I. Epifani, S. Fortini, and L. Ladelli. A characterization for mixtures of semi-Markov processes. *Statistics & Probability Letters*, 60(4):445–457, 2002.

I. Epifani, A. Lijoi, and I. Prünster. Exponential functionals and means of neutral-to-the-right priors. *Biometrika*, 90(4):791–808, 2003.

P. D. Feigin and R. L. Tweedie. Linear functionals and Markov chains associated with Dirichlet processes. *Mathematical Proceedings of the Cambridge Philosophical Society*, 105(03):579–585, 1989.

T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, pages 209–230, 1973.

S. Fortini and S. Petrone. Hierarchical reinforced urn processes. *Statistics & Probability Letters*, 82(8):1521–1529, 2012a.

S. Fortini and S. Petrone. Predictive construction of priors in Bayesian nonparametrics. *Brazilian Journal of Probability and Statistics*, pages 423–449, 2012b.

S. Fortini and S. Petrone. Predictive characterization of mixtures of Markov chains. *arXiv preprint arXiv:1406.5421*, 2014.

S. Fortini, L. Ladelli, and E. Regazzini. Exchangeability, predictive distributions and parametric models. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 86–109, 2000.

S. Fortini, L. Ladelli, G. Petris, and E. Regazzini. On mixtures of distributions of Markov chains. *Stochastic Processes and their Applications*, 100(1):147–165, 2002.

D. Freedman. *Markov Chains*. Springer Science & Business Media, 1983.

D. A. Freedman. Invariants under mixing which generalize de Finetti's theorem. *The Annals of Mathematical Statistics*, 33(3):916–923, 1962.

D. A. Freedman. On the asymptotic behavior of Bayes' estimates in the discrete case. *The Annals of Mathematical Statistics*, pages 1386–1403, 1963.

J. K. Ghosh and R. Ramamoorthi. *Bayesian Nonparametrics*. Springer Series in Statistics, 2003.

D. Gibson and E. Seneta. Augmented truncations of infinite stochastic matrices. *Journal of Applied Probability*, 24(03):600–608, 1987.

R. D. Gill and S. Johansen. A survey of product-integration with a view toward application in survival analysis. *The Annals of Statistics*, pages 1501–1555, 1990.

D. Gross, S. John, J. Thomson, and C. Harris. *Fundamentals of Queueing Theory*. Wiley New York, 2008.

R. Grübel and S. M. Pitts. A functional approach to the stationary waiting time and idle period distributions of the $GI/G/1$ queue. *The Annals of Probability*, pages 1754–1778, 1992.

F. Guillemin and A. Simonian. Transient characteristics of an $M/M/\infty$ system. *Advances in Applied Probability*, 27(03):862–888, 1995.

F. M. Guillemin, R. R. Mazumdar, and A. D. Simonian. On heavy traffic approximations for transient characteristics of $M/M/\infty$ queues. *Journal of Applied Probability*, 33(02): 490–506, 1996.

J. Haigh. *Probability Models*. Springer Science & Business Media, 2004.

C. M. Harris. Queues with state-dependent stochastic service rates. *Operations Research*, 15(1):117–130, 1967.

E. Hewitt and L. J. Savage. Symmetric measures on Cartesian products. *Transactions of the American Mathematical Society*, 80(2):470–501, 1955.

N. L. Hjort. Nonparametric Bayes estimators based on beta processes in models for life history data. *The Annals of Statistics*, pages 1259–1294, 1990.

F. M. Hoppe. Pólya-like urns and the Ewens' sampling formula. *Journal of Mathematical Biology*, 20(1):91–94, 1984.

D. R. Insua, M. Wiper, and F. Ruggeri. Bayesian analysis of $M/Er/1$ and $M/H\_k/1$ queues. *Queueing Systems*, 30(3-4):289–308, 1998.

R. A. Johnson. An asymptotic expansion for posterior distributions. *The Annals of Mathematical Statistics*, pages 1899–1906, 1967.

O. Kallenberg. *Probabilistic Symmetries and Invariance Principles*. Springer Science & Business Media, 2006.

T. Kato. *Perturbation Theory for Linear Operators*. Springer Science & Business Media, 1995.

A. S. Kechris. *Classical Descriptive Set Theory*. Springer, 1995.

D. G. Kendall. Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain. *The Annals of Mathematical Statistics*, pages 338–354, 1953.

Y. Kim and J. Lee. On posterior consistency of survival models. *The Annals of Statistics*, pages 666–686, 2001.

Y. Kim and J. Lee. A Bernstein-von Mises theorem in the nonparametric right-censoring model. *The Annals of Statistics*, pages 1492–1512, 2004.

J. F. Kingman. Uses of exchangeability. *The Annals of Probability*, pages 183–197, 1978.

L. Kleinrock. *Queueing Systems*. Wiley New York, 1976.

C. Knessl and Y. P. Yang. Asymptotic expansions for the congestion period for the $M/M/\infty$ queue. *Queueing Systems*, 39(2):213–256, 2001.

M. Kosorok. *Introduction to Empirical Processes and Semiparametric Inference*. Springer, New York, 2008.

V. G. Kulkarni. *Modeling and analysis of stochastic systems*. CRC Press, 2009.

J. M. Landsberg. *Tensors: Geometry and Applications*, volume 128. American Mathematical Society Providence, RI, USA, 2012.

A. Lijoi, I. Prünster, and S. G. Walker. Bayesian consistency for stationary models. *Econometric Theory*, 23(04):749–759, 2007.

D. Lind and B. Marcus. *An Introduction to Symbolic Dynamics and Coding*. Cambridge university press, 1995.

A. Maitra. Integral representations of invariant measures. *Transactions of the American Mathematical Society*, 229:209–225, 1977.

B. Marcus, K. Petersen, and T. Weissman. *Entropy of Hidden Markov Processes and Connections to Dynamical Systems: Papers from the Banff International Research Station Workshop*, volume 385. Cambridge University Press, 2011.

J. J. Martin. *Bayesian decision problems and Markov chains*. Wiley, 1967.

M. F. McGrath and N. D. Singpurwalla. A subjective Bayesian approach to the theory of queues (II)—inference and information in $M/M/1$ queues. *Queueing Systems*, 1(4): 335–353, 1987.

M. F. Mcgrath, D. Gross, and N. D. Singpurwalla. A subjective Bayesian approach to the theory of queues (I)—modeling. *Queueing Systems*, 1(4):317–333, 1987.

J. Medhi. *Stochastic Models in Queueing Theory*. Academic Press, 2002.

J. A. Morrison, L. A. Shepp, and C. J. Van Wyk. A queueing analysis of hashing with lazy deletion. *SIAM Journal on Computing*, 16(6):1155–1164, 1987.

R. Nelson. *Probability, Stochastic Processes, and Queueing Theory: the mathematics of computer performance modeling*. Springer Science & Business Media, 2013.

G. Newell. The $M/G/\infty$ queue. *SIAM Journal on Applied Mathematics*, 14(1):86–88, 1966.

K. E. Petersen. *Ergodic Theory*, volume 2. Cambridge University Press, 1989.

E. G. Phadia. *Prior Processes and their Applications*. Springer, 2015.

R. R. Phelps. *Lectures on Choquet's theorem*. Springer Science & Business Media, 2001.

E. Regazzini, A. Lijoi, and I. Prünster. Distributional results for means of normalized random measures with independent increments. *The Annals of Statistics*, pages 560–585, 2003.

K. Sato. *Lévy Processes and Infinitely Divisible Distributions*. Cambridge University Press, 1999.

M. J. Schervish. *Theory of Statistics*. Springer Science & Business Media, 1995.

P. J. Schweitzer. Perturbation theory and finite Markov chains. *Journal of Applied Probability*, 5(02):401–413, 1968.

E. Seneta. *Non-negative matrices and Markov chains*. Springer Science & Business Media, 1981.

J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, pages 639–650, 1994.

C. C. Strelioff, J. P. Crutchfield, and A. W. Hübler. Inferring Markov chains: Bayesian estimation, model comparison, entropy rate, and out-of-class modeling. *Physical Review E*, 76(1):011106, 2007.

V. S. Varadarajan. Groups of automorphisms of borel spaces. *Transactions of the American Mathematical Society*, 109(2):191–220, 1963.

S. Walker and P. Muliere. Beta-stacy processes and a generalization of the Pólya-urn scheme. *The Annals of Statistics*, pages 1762–1780, 1997.

M. Wiper. Bayesian analysis of $Er/M/1$ and $Er/M/c$ queues. *Journal of Statistical Planning and Inference*, 69(1):65–79, 1998.

R. W. Wolff. Poisson arrivals see time averages. *Operations Research*, 30(2):223–231, 1982.

S. L. Zabell. Characterizing Markov exchangeable sequences. *Journal of Theoretical Probability*, 8(1):175–178, 1995.