Nadine Stephenson
Dr. sc. hum.

Development of a hierarchical stochastic search algorithm for model selection in the analysis of genetic epidemiologic studies

Promotionsfach: Epidemiologie
Doktormutter: Prof. Dr. sc. hum. Jenny Chang-Claude

This dissertation investigates stochastic model search methods that allow the incorporation of external biological information about the studied genetic polymorphisms into the statistical analysis of genetic epidemiologic data. We initially started with the application of an existing Bayesian Model Averaging algorithm and identified several disadvantages of the method that were addressed by the subsequently developed Hierarchical Stochastic Search algorithm. The method allows the inclusion of information like the presence of markers in metabolic pathways, location and functionality, and intermediate metabolic measurements via a hierarchical model to guide the stochastic search for subsets of variables that best explain the data.
The specific aims of the dissertation were three-fold.

1. Application of Bayes Model Averaging.
An existing Bayes Model Averaging approach was applied to data from a population-based case-control study investigating the influence of smoking and several genetic polymorphisms in tobacco-smoke-related metabolic pathways on breast cancer risk in Germany. The method included biological information by specifying prior distribution for model parameters and posterior estimates were obtained using a fully Bayesian approach averaging over multiple submodels. Findings were compared to results from logistic regression and regression with backward selection and confirmed a significant effect of the *NAT1**10 allele on breast cancer risk.
Methodological disadvantages of the method, such as long computation times due to the fully Bayesian approach and the inability to use prior information in a more flexible way via a hierarchical model, were discussed and led to the subsequent development of the Hierarchical Stochastic Search method.

2. Development, testing and application of Hierarchical Stochastic Search.
The method combines features from Bayes Model Averaging, hierarchical modeling and Shotgun Stochastic Search, and shares similarities with the established MCMC method Markov Chain Monte Carlo Model Composition. An empirical Bayes approach is used for parameter estimation to reduce computation time. A detailed description of the algorithm and information on its implementation are provided.
A hierarchical modeling approach is used to weight variables and thus to guide the stochastic model search. This led to shrinkage of the weights if variables were considered to be similar with respect to the biological prior information. The amount of shrinkage was investigated for different effect sizes and prior scenarios. Subsequently, the method was tested using two sets of simulated data, consisting of uncorrelated and correlated SNP variables. Details on prior information in the form of prior covariates was provided. The method was investigated with respect to model space exploration in terms of variable and model proposal frequencies and time to first proposal, as well as with respect to the ability to identify causal variables. Results

were compared to analysis using Markov Chain Monte Carlo Model Composition because of the structural similarity of the two methods. Hierarchical Stochastic Search was found to outperform Markov Chain Monte Carlo Model Composition with respect to model space exploration and performed comparably or better with respect to identification of causal variables.
Hierarchical Stochastic Search was applied to real data considering genetic effects on smoking cessation and confirmed the most significant SNP and multiple other significant findings from the published initial.
Overall, Hierarchical Stochastic Search worked well in the investigated scenarios. Aspects like performance in more complex scenarios and inclusion of interaction require further investigation. Promising areas for future research include parallelization of the method, application in a haplotype setting, and the analysis of rare variants.

3. Bayes Model Averaging using haplotypes.
We briefly explored a Bayes Model Averaging approach using haplotype information. A sparsity-inducing prior on model size was used and no model search was conducted; instead estimates were obtained by averaging over all models with a limited number of terms. The method was tested using simulated data and had better power to identify the causal marker, or variants in strong linkage disequilibrium with the causal marker, than a pointwise test using adjusted p-values. High posterior probability was assigned to sparse models containing these markers. Comparison with a corresponding SNP-based approach are necessary to draw further conclusions.
An extension of the method to include a model search in the spirit of Hierarchical Stochastic Search, and the subsequent application to define genetic regions for deep sequencing are potential future research topics.