ADAPTIVE BAYESIAN SEQUENTIAL STUDY DESIGN

Gergely HAJÓS
Advisors: Péter ANTAL, Tadeusz DOBROWIECKI

I. Introduction

The relative scarcity of the results of genetic association studies (GAS) prompted many research directions and hypotheses. To address the multiple testing problem computer intensive statistical methods became widespread and as an ultimate solution genome-wide association studies (GWAS) had appeared.

In the case of partial genome association study (PGAS) contrary to GWAS only partial information is available about the genome of the participants. In PGAS, we attempt to discover from the subsequent measurements of well-selected blocks of variables the relevant genetic factors for a given target set with interim analyzis and meta-analyzis of the available aggregated data sets in order to interpret and guide further measurements (see Fig. 1). The phases are shown in Fig. 1, starting with the GWAS layer and the application of gene prioritization systems for the subjective, knowledge-rich initiation of our pruning process.

One of the main bottlenecks in genetic association studies today are the required large sample size and complex models (with larger computational resources). In the paper we jointly address both issues within context of sequential PGAS: we apply a Bayesian, model-based meta-analysis in using an adaptive study design for pruning the variables. This can ensure large sample size within a fixed budget. We evaluate typical sequential pruning policies in association studies based on interim Bayesian meta-analysis. The approximation of one-step look ahead of expected value of experiments was also investigated with a full Bayesian approach. Our application domain is the investigation of genetic background of asthma using PGASs.

II. Background

The objective of the *sequential study design* (SSD) is to retrieve statistical information from data collected sequentially, given a utility and cost with a budget constraint for the data collection. A possible application is to check the effectiveness of drug treatment or to find association between genetic factors and diseases, etc. In the field of sequential study design generally the standard hypotheses testing approach is used [1]. The aim is to collect the minimum amount of data necessary to make a decision between null and alternative hypotheses, since one wants to minimize the costs of measurements performed in the study design. In every step of the sequential study design the tests of hypotheseses are performed on a set of samples, if one of the hypotheseses is accepted the study design is stopped and a decision is made. If a decision cannot be made due to lack of enough information the study design continues and more data is collected [2, 3].

Instead of applying the standard statistical approach we present in this paper a multivariate extension based on Bayesian networks. In every step of the study design we first approximate the utility of the computed results based on the available data, second we predict the future data based on the available data using Bayesian model averaging over Bayesian networks. With the help of future data we predict the utility of the continuation of the study design. If the predicted utility of continuation is higher than the utility based on the available data then the sequential study design is continued, otherwise stopped and the last computed results are reported.

Besides selecting the sample size to minimize the cost, in this paper we present an *active learning*

approach to reduce the number of variables in every step of the sequential study design [4]. Since the algorithm in every step narrows down the set of the investigated variables, in the subsequent step just the last (i.e. the narrowest) set is used for further analysis. In this way in every subsequent step less and less measurements are performed.

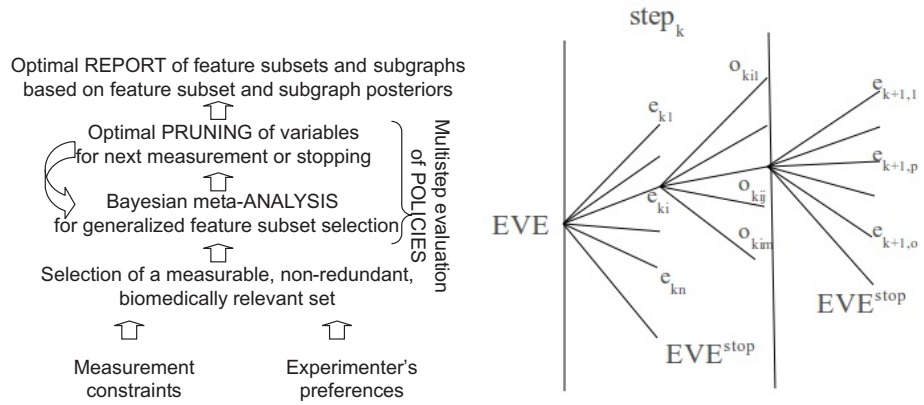

Figure 1: Left: The phases of sequential study design. Right: Expected value of experiment.

III. Methods

A. Calculation of expected value of experiment with multilevel analysis

In each step of a SSD researchers set an experiment. We can say researchers in the step k choose experiment e_{1i} from set of possible experiments $(e_{k1}, e_{k2}, \ldots, e_{kn})$. Each experiment has an outcome o_{kij} with probability $p(o_{kij})$. The *expected value of experiment* (EVE) in step k is defined by $EVE(e_{ki}) = \max(EVE^{stop}(exp), \sum_i \sum_j p(o_{kij}) EVE(e_{k+1,j}))$. Where EVE^{stop} is the utility of stopping the experiment with the actual results. The expected value of an experiment in step k depends on the expected value in step $k+1$. In this paper we want to maximize the EVE by selecting the optimal set of variables for experiment (see Fig. 1).

We assume that there is a special set of target variables, and the goal is the identification of an optimal set of relevant variables, and their interactions (for an overview of feature subset selection (FSS) problem, see e.g. [5]). The goal of the analogous Bayesian FSS can be defined as the computation of the posteriors for the pair wise relations, relevant sets, and interactions, which can be formalized as the posteriors for Markov Blanket Membership (MBM), Markov Blanket set (MBS), Markov Blanket subgraph (MBG) [6].

Respectively, we assume that utility function for the model can be additively decomposed into three parts, specifically for the MBM, MBS, and MBG levels. Note that given the utility function and the posterior over the model space in step i, the expected utility of reporting a structural model $\hat{M}$ is computable. The model with maximal utility can be determined as:

$$M^* = \arg\max_{\hat{M}} E_{p(M|D_{<i})}[U(\hat{M}|M)].$$

where M is a model with probability $p(M|D_{<i})$, and $U(\hat{M}|M)$ is the utility of $\hat{M}$ in case of M.

We evaluate typical policies in association studies based on interim Bayesian meta-analysis, and also the performance of a one-step look ahead approximation of the expected value of the experiments in the full Bayesian approach. Our application domain is the investigation of the genetic background of asthma using PGAS, where the costs of sample collection and genotyping are considerable, and a multivariate approach is essential due to complex, weak interactions behind multifactorial diseases.

B. *Bayesian sequential study design and variable pruning*

Since beside the selection of the number of samples, in case of variable pruning we also narrow down the number of the variables, we define the following: a prior $p(M)$ for the generative models; variable set S_i, where S_i contains only variables present in step i due to the reduction of variables; a corresponding likelihood $p(D_i|M)$ for the ith step and data set D where D_i represents the data set narrowed down to variable set S_i with the samples N_i collected in step i; a set of actions, continuing sequential study design consisting of both the selection of S_{i+1}, N_{i+1} or reporting actions (stop experiment and report the last computed model); a context-free, e.g. timeless, cost $C_{N_i}^{S_i}$ of measuring (observing) D_i. $D_{<i}$ represents the data set narrowed down to variable set S_i with all the samples collected in steps $< i$ ($N_{<i} = \sum_{j=0}^{i-1} N_j$).

In the optimal Bayesian approach, at step $1 < i$ one possibility is to stop and to report the optimal maximal utility model M^* with utility

$$U_i^{\text{report}} = E_{p(M|D_{<i})}[U(M^*|M)] - \sum_{j=1}^{i-1} C_{N_j}^{S_j}. \tag{1}$$

The other option is to continue by selecting the next, optimal experiment defined by the selection of N_i, with utility

$$U_i^{\text{cont}} = U(N_i) - \sum_{j=1}^{i-1} C_{N_j}^{S_j}, \tag{2}$$

where $U(N_i)$ denotes EVE. In case of a decision problem with finite horizon, backward induction can be applied to calculate Eq. 2. The exponential number of the potential future subsequent data makes however the estimation of these expectations computationally prohibitive (for evaluation of the value-of-information, see [7, 8]). The one-step approximation of $U(N_i)$ is as follows

$$U(N_i) \approx E_{p(D_i|D_{<i})}[E_{p(M|D_{\leq i})}[U(M^*|M)]], \tag{3}$$

which means that after the first step, the optimal Bayesian decision, (reporting M^* or continuing with measuring N_i) can be determined by comparing U_i^{report} to U_i^{cont} (see [9]). Note that the framework of Markov decision processes is not directly applicable to this context, because of the dynamic state space.

IV. Results

During the evaluation of the pruning algorithm (Section B.), results showed that the main bottleneck was the underlying Bayesian FSS algorithm (Section A.). We found that the Markov chain Monte Carlo (MCMC) based multilevel analysis (see [10]) does not converge to the real posterior distribution in the planned settings. The applied MCMC uses multiple chains with different heat temperature [11]. Our aim was to validate the MCMC algorithm and find proper temperature parametrization for stable simulation and fast convergence.

For testing convergence of the MCMC algorithm the multiple chain based Gelman-Rubin R score (see [12]) was used. For multiple parametrization five independent parallel simulations were performed.

The single chain Geweke's Z score (see [13]) was used to test convergence of confidence of the MCMC simulations. Z score was calculated for MBM and MBS features, our results show that over 5×10^6 step of burn-in, the calculated Geweke score of most of the features fall into the acceptable region (for illustration, see Fig. 2).

Regarding the confidence scores we face the multiple testing problem (MTP): the confidence of the simulation is estimated based on the confidence score of numerous (dependent or independent)

features, thus we have to apply a correction, e.g. the Bonferroni correction. In the case of MBGs an extreme case of MTP occurs, because the typical number of MBG features is extremely high.

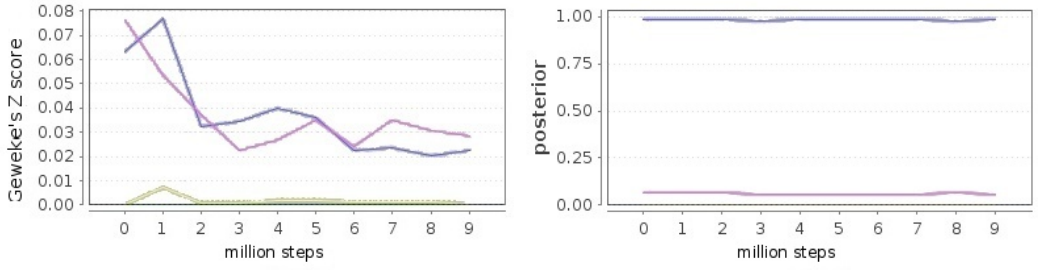

Figure 2: The horizontal axis: the simulation steps after burn-in. Left: Geweke's score of three features with the same parametrization. Right: The posteriors of the same features (one is constant zero).

V. CONCLUSION

In the paper we investigated the decision support for sequential partial genetic association studies, which are essential methods to ensure high sample size for more targeted statistical analysis after GWAS-based explorations. During the convergence of confidence investigations we determined a default parametrization for temperature settings and burn-in length. The MCMC algorithm was improved, the simulation is stable with the proper settings. Further research needed to find the utility function for one-step look ahead and calculate non-myopic look ahead.

References

[1] I. R. Konig and A. Ziegler, "Group sequential study designs in genetic-epidemiological case-control studies," *Hum. Hered.*, 56:63–72, 2003.

[2] D. J. Spiegelhalter, K. R. Abrams, and J. P. Myles, *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*, John Wiley ans Sons, 2003.

[3] S. M. Berry, B. P. Carlin, J. J. Lee, and P. Muller, *Bayesian Adaptive Methods for Clinical Trials*, Chapman and Hall, 2010.

[4] J. Li, "Prioritize and select snps for association studies with multi-stage designs," *Journal of Computational Biology*, 15(3):241–257, 2008.

[5] Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, 23(19):2507–2517, 2007.

[6] P. Antal, A. Millinghoffer, G. Hullám, C. Szalai, and A. Falus, "A bayesian view of challenges in feature selection: Feature aggregation, multiple targets, redundancy and interaction," *JMLR Proceeding*, 4:74–89, 2008.

[7] H. D., E. Horvitz, and B. Middleton, "An approximate non-myopic computation for value of information," in *Proc. of the 7th Conf. on Uncertainty in Artificial Intelligence (UAI'91)*, pp. 101–107. Morgan Kaufmann, 1991.

[8] W. Liao and Q. Ji, "Efficient non-myopic value-of-information computation for influence diagrams," *International journal of approximate reasoning*, 49:436–450, 2008.

[9] J. M. Bernardo, *Bayesian Theory*, Wiley & Sons, Chichester, 1995.

[10] P. Giudici and R. Castelo, "Improving Markov Chain Monte Carlo model search for data mining," *Machine Learning*, 50:127–158, 2003.

[11] G. Altekar, S. Dwarkadas, J. P. Huelsenbeck, and F.Ronquist, "Parallel metropolis coupled markov chain monte carlo for bayesian phylogenetic inference," *Bioinformatics*, 20(3):407–415, 2004.

[12] A. Gelman and D. B. Rubin, "Inference from iterative simulation using multiple sequences," *Statistical Science*, 7(4):457–511, 1992.

[13] J. Geweke, "Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments," in *Bayesian Statistics 4*, J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Eds., pp. –. Clarendon Press, Oxford, UK, 1992.