Bayesian data analytic knowledge bases for genetic association studies

P. Sarkozy[1], P. Marx[1], A. Millinghoffer[1], G. Varga[2], A. Szekely[2], Zs. Nemoda[3], Zs. Demetrovics[2], M. Sasvari-Szekely[3], P. Antal[1]

(1)Department of Measurement and Information Systems, Budapest University of Technology and Economics
(2)Institute of Psychology, Eötvös Loránd University, Budapest, Hungary
(3)Institute of Medical Chemistry, Molecular Biology and Pathobiochemistry, Semmelweis University, Budapest, Hungary

Abstract. Bayesian methods and Bayesian networks are increasingly popular in genetic association studies. We discuss the application of Bayesian networks to give a detailed characterization of relevance relations and their application in case of multiple outcome variables. These global properties of the relevance relations are investigated in the Bayesian statistical framework using a joint model, thus we can generate a coherent uncertainty measure for the results without post hoc corrections. We show the usefulness of the syntactic aggregation of the a posteriori distributions over the relevant variable sets, which allows the examination of the most relevant variables, variable pairs, and larger subsets. We present these methods as precursors for a unified framework of Bayesian data analytic knowledge bases describing the results of multiple Bayesian analysis of relevance. Concepts are demonstrated in the genetics of trait impulsivity.

1 Introduction

Genetic association studies face many challenges such as the poor description of phenotypes, presence of population confounding, effects of life style and environment, the seemingly non-functional nature of the factors found, the weak effect strength of the factors ("missing heritability"), but the most profound is the rapid increase of the number of potential predictors, which manifests itself as"the multiple hypothesis testing problem" in the frequentist framework. In response to this limit, more intensive usage of computational resources and background knowledge became central issues in biomedicine. In genetic association studies such approaches have emerged in various contexts to cope with the relative scarcity of the data such as the pooling of datasets in meta-analysis, pooling of the results in ad hoc repositories and knowledge bases, and the use of computation-intensive statistical approaches such as permutation testing, bootstrap, and Bayesian statistics.

In the paper we present elements of a Bayesian, global relevance analysis and show their application in the probabilistic knowledge fusion research direction in the following aspects:

1. *Partial (strong) relevance* We can infer the a posteriori probability the k variables are jointly strongly relevant for a given outcome potentially with further unspecificied variables.
2. *Type of relevance* We can infer posteriors for various types of relevance, e.g. strong relevance vs. association.
3. *Multi-target relevance* We can infer posteriors for strong relevance w.r.t. multiple outcome variables.

The advantages of Bayesian networks (BN) for representing global dependency maps and relevance relations are well-known, but their application was hindered in high-dimensional tasks by their high computational and sample (statistical) complexity. Motivated by this problem we proposed a Bayesian approach to the feature subset selection (FSS) problem and proposed the use of partial relevance and multi-target relevance [2]. In this paper we extend this approach by inferring and comparing posteriors for various subtypes of pairwise dependencies, such as association and strong relevance.

First in Section 2 we overview Bayesian network based concepts of relevance and earlier applications. In Section 3 we discuss the Bayesian approach to FSS, particularly the main assumption of its popular conditional version, which makes it different from the general, domain model based approaches, and summarize the Stochastic Search Variable Selection (SSVS), which is one of our evaluation methods. Then in Section 4 we overview earlier Bayesian network based methods in the Bayesian framework to analyze relevance and summarize our approach. Section 5 and Section 6 contains the results in impulsivity research and its discussion.

2 Bayesian network representation of relevance

There are many association analysis methods with different biases, advantages and disadvantages w.r.t the number of variables, sample size, quality and completeness of the data, loss function, time, and available computational resources. Thus an important point of reference is an asymptotic, loss-free, algorithm-free probabilistic concept of relevance, the Markov Blanket Set (MBS) [29]. It was connected to the Bayesian networks (BN), which became a central tool for the graphical representation of dependencies and optionally causation [24]. In the feature (attribute) learning context related univariate concepts of relevance, strong and weak relevance

was introduced [20]. To bridge the gap between the linear cardinality of the Markov blanket membership (MBM)s and exponential cardinality MBSs, we introduced the concept of partial (multivariate, strong) relevance (k-MBS) with scalable, intermediate polynomial cardinalities [2].

Because in our application domain the outcome variables are semantically related, we use the following acasual subtypes of relevance, which are derived from the combinations of {causal,confounded,conditional}, and {direct,indirect} relations and their aggregates, see Table 1 (for a causal interpretation under the Causal Markov Assumption, see e.g. [24]).

Table 1. Graphical model based definition of types of relevances and associations.

Relation	Abbreviation	Graphical
Direct causal relevance	DCR(X,Y)	There is an edge between X and Y
Transitive causal relevance	TCR(X,Y)	There is directed path between X and Y
Confounded relevance	ConfR(X,Y)	X and Y have Common ancestor
(Pairwise) Association	A	DCR or TCR or ConfR
Pure interactionist relevance	PIR(X,Y)	X and Y have common child
Strong relevance	SR(X,Y)	PIR or DCR

The ordering of relations in Table 1 indicates certain ontological, and practical aspects, but a hierarchy or ranking is problematic, because for example the standard concept of pairwise association (A) is narrower than strong relevance (it does not include Pure Interactionist Relevance). Further extension of relevance is possible, if there are multiple possible target variables $\boldsymbol{Y}$ which have to be examined together, thus we proposed the the concept of multi-target relevance [2].

The Markov Blanket Set and the Bayesian network representation induced many research direction in feature learning, in the feature subset selection problem, and in genetic association studies [11, 21, 34, 16, 1, 18, 36]. Because of the high computational complexity and particularly because of the high sample (statistical) complexity of learning complete Bayesian network models w.r.t number of variables these "local" approaches limit their scope, and focus on the identification of strongly relevant variables, and possibly their interaction and causal structure. Thus global and detailed characterization of relevance relations is not available. However as we will show the Bayesian statistical framework provides a normative solution for the high sample complexity and for medium sized problems with hundreds of variables the computational complexity is managable using high-throughput and high-performance computing resources.

3 The conditional Bayesian approaches and the SSVS

Bayesian methods are more and more popular in genetic association studies, and one of their advantage is their principled approach to model complexity and number of variables in case of relatively small sample size [6]. The infamous correction for multiple hypothesis testing with frequently ad hoc management - causing loss of significance and power - manifests itself in the Bayesian framework as a normative and inherent property, resulting in a more flat posterior for more complex models.

In the feature learning context a popular choice is the conditional Bayesian approach, which assumes independent beliefs corresponding to the modeling of the dependence of the output variable Y on X (i.e., without modeling the overall domain) [14]. Practically the conditional approach models the conditional distribution of Y given $\boldsymbol{X}$ using a parametric model class $S, \boldsymbol{\theta}$ as $p(Y = 1|X = x, S, \boldsymbol{\theta})$, for example using linear regression, logistic regression or multilayer perceptrons. The domain model based approach models the joint distribution of $Y, \boldsymbol{X}$ using a parametric model class $S, \boldsymbol{\theta}$ as $p(Y, \boldsymbol{X}|S, \boldsymbol{\theta})$, for example using Bayesian networks. In both cases using the posterior over model structure $p(S, \boldsymbol{\theta}|D_N)$ given a data set D_N we can induce a posterior for the relevance of a feature X_i and for the subset of features $\boldsymbol{X'}$. A fundamental difference between the conditional and domain model based approach is that in the conditional approach the presence of a variable in the model can not be interpreted as strong relevance (e.g. a highly predictive, but only weakly relevant factor can be present in the conditional model, if it is strongly associated through multiple paths, as we do not model the dependencies between the factors).

An early Bayesian conditional approach, the Stochastic Search Variable Selection puts the regression problem in a Bayesian statistical framework. This approach considers submodels with subsets of the predictor variables and estimates the a posteriori probability of the inclusion of a predictor and its corresponding strength parameters [15]. Bayesian variable selection method is based on assuming a normal prior distribution on the regression parameters. The variance of the distribution usually is a constant, but we can extend the model by estimating the variance as in case of SSVS. If we estimate the variance of normal prior, it helps tuning the parameters, because in the regression model the coefficient depends on the variance. In a heterogeneous problem, the variance can be set differently for all regression variables.

Other Bayesian conditional methods e.g. using logistic regression or multilayer perceptrons, are widely used in biomedicine and in GASs (e.g., see [3, 27, 6, 31, 28, 35, 32, 12]). Although the conditional approach is capable for multivariate analysis including interactions, the domain model based approach allows better characterization of both local and global dependencies.

4 Bayesian analysis of relevance using Bayesian networks

The local "causal" discovery methods limit their scope to the strongly relevant variables to reduce the high computational complexity and particularly the high sample (statistical) complexity of learning complete Bayesian network models, i.e. to avoid the learning of a global and detailed characterization of relevance relations [1]. However the Bayesian statistical framework provides a normative solution for the high sample complexity and for medium sized problems with hundreds of variables the computational complexity is managable using high-throughput and high-performance computing resources. Thus the BN based Bayesian approach can ensure global, potentially causal characterization of the dependencies and normative characterization of weakly significant results.

The Bayesian inference over structural properties of Bayesian networks was proposed in [7, 9]. In [22], Madigan et al. proposed a Markov Chain Monte Carlo (MCMC) scheme to approximate such Bayesian inference. In [13], Friedman et al. reported an MCMC scheme over the space of orderings. In [19], Koivisto et al. reported a method to perform exact full Bayesian inference over modular features. An ad hoc randomized approach were reported in [30]. For the application of Bayesian networks in the Bayesian framework we reported specialized ordering MCMC methods to efficiently estimate posteriors over structural model properties, particularly over Markov Blanket Graphs (MBG) (the ordering-conditional posterior of an MBG can be computed in polynomial time, which can be exploited in ordering-MCMC methods [4]). Based on these concepts we proposed a Bayesian network based Bayesian multilevel analysis of relevance (BN-BMLA), which estimates posteriors of hierarchic, interrelated hypotheses, e.g. for partial strong relevance for all subsets. Partial strong relevance is particularly useful, because it defines an embedded hypotheses space with varying complexity, i.e. sets of k predictors that are strongly relevant [2].

The posteriors for the hierarchic, interrelated hypotheses of the BN-BMLA methodology are estimated in a two-step process to support post-

hoc analysis. First we estimate posteriors over the MBS, MBG, and for the pairwise relations in Table 1 for the target variables. In the second phase we use these posteriors as a probabilistic knowledge-base to estimate various posteriors and discover interesting and significantly confirmed hypotheses. In the first phase we applied MCMCM method over the Bayesian network structures (i.e. over directed acyclic graphs, DAGs) without limiting the maximum number of parents. We used both the Cooper-Herskovits (CH) and the observationally equivalent BDeu parameter priors with various virtual sample sizes (VSS=1,10,50,100), but from the point of view of biomedical relevance we found that the theoretically preferable BDeu prior is more sensitive to "small-sample" anomalies, thus we report results for the CH and VSS=1 setting. The structure prior was uniform. The length of the burn-in and MC simulation is 10^6 and 5×10^6, the probability of the DAG operators is uniform [8]. In the second step we computed offline the k-MBS posterior values from the MBS posterior, the posteriors over the types of the dependency relations in Table 1, and the posteriors for multi-target relevance.

5 Results

Impulsivity or impulsiveness is a personality trait defined as a predisposition toward rapid, unplanned reactions. We investigated a combined set of serotonergic (HTR1A–1019 C/G, HTR1B 1997 C/G, 5-HTTLPR in SLC6A4 gene) and dopaminergic (COMT Val158Met, DRD4 48bp VNTR, DRD2/ANKK1 Taq A1) polymorphisms. The sample size was 561, which included only complete records from a preliminary dataset of a larger study. The impulsivity phenotype was measured by the Hungarian version of the Barratt Impulsivity Scale (BIS-11) originally published by Patton and colleagues [23]. The instrument consists of 30 items, scored on a four point scale. The three main impulsivity factors are: Cognitive, Motoric, and Nonplanning impulsiveness. The total score is the sum of all items.

To cope with multiple predictors with potentially weak effects we applied the stochastic search variable selection method (SSVS) to the genotypes, Sex and Age data, while normalizing the scale targets. We used the SSVS for quantile regression implementation in the MCMCpack package [25] in R [26]. We ran the algorithm for the different scale targets with the same parameter setting to get comparable results. We set the shape parameters of the beta distribution to 10 and 2. We ran 100000 iterations and 10000 for the burn-in period. Two of the predictor variables was

found significant in case of all target variables (Fig. 1). These two variables (HTR1B and DRD4) have significantly higher marginal inclusion probability. The regression coefficients with the highest absolute value belongs also to these two predictors HTR1B and DRD4 (Fig. 5).

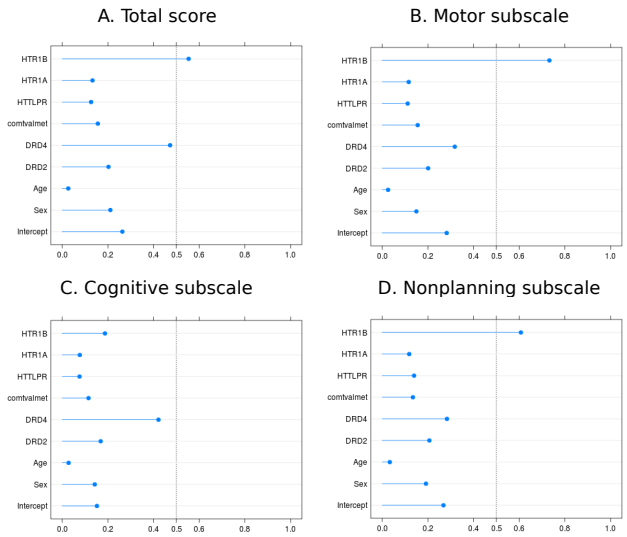

Fig. 1. Marginal posterior probability for all predictors. The posterior probabilities (X axis) for the total score (A), for the motor subscale (B), for the cognitive subscale (C) and for the nonplanning subscale (D).

To investigate the interactions; types of the relevance of the predictors; and their relevance in a joint analysis of multiple target variables we applied the BN-BMLA method. It was applied for the 5-HTTLPR genotypes and five other grouped genotype categories, as well as Sex, and Age. Outcome (target) variable was the BIS Total score or the scale variables separately and jointly. The scale variables and Age were discretized into three bins with uniform univariate frequencies.

The identification of an overall domain model was not possible, because considerable uncertainty remained at the level of full multivariate analysis (see Fig. 5). Therefore we computed the aggregate posterior probabilities for variables, pairs of variables, and triplets of variables. Fig. 5 reports a comparative overview of peakness of the posteriors for uni-, bi-, and trivariate partial strong multi-target relevance. It shows that DRD4 is strongly relevant (with 0.575 posterior probability), the DRD4 and HTR1B pair is among the strongly relevant variables (with 0.251 pos-

Table 2. The regression coefficients for the predictors in case of the three subscales and the total score

Predictors	Total score	Motor subscale	Cognitive subscale	Nonplanning subscale
Intercept	-5.821e-03	-2.730e-02	0.01	-0.008
Sex	2.264e-02	4.126e-03	0.022	0.0151
Age	-4.644e-05	6.403e-05	-0.0006	0.0003
DRD2	-6.593e-03	-1.093e-02	-0.016	-0.0069
DRD4	-1.074e-01	-4.822e-02	-0.09	-0.04
COMT	-1.122e-02	-8.690e-03	-0.012	-0.0038
5-HTTLPR	4.721e-03	1.588e-03	-0.0045	0.006
HTR1A	4.670e-03	-5.480e-04	-0.004	0.0002
HTR1B	-1.391e-01	-2.003e-01	-0.023	-0.16

terior probability). Posterior probabilities of three member variable sets showed no marked features. The standard errors of the estimated posteriors are below 0.001. Fig. 5 shows the most probable MBGs with posterior larger than 0.001 (the width of the edges indicate their aggregate posteriors).

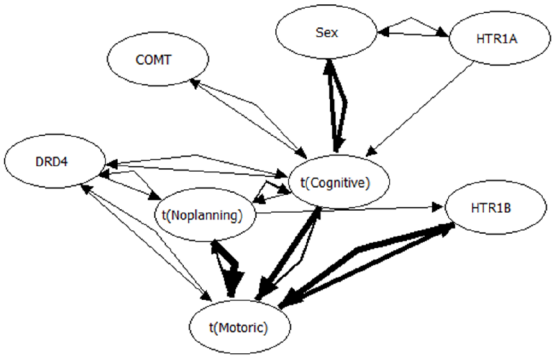

Fig. 2. The most probable Markov Blanket Graphs with posterior larger than 0.001 (the edge posteriors are indicated by their width).

We also computed the posteriors for the types of the relevance of the predictors in Table 1. As expected it provides a useful, detailed characterization, e.g. in case of DRD2 the a posteriori probability of association between DRD2 and bisTotal is 0.4926, but its strong relevance is only 0.0219. Note that such interpretation and decomposition is not available with SSVS.

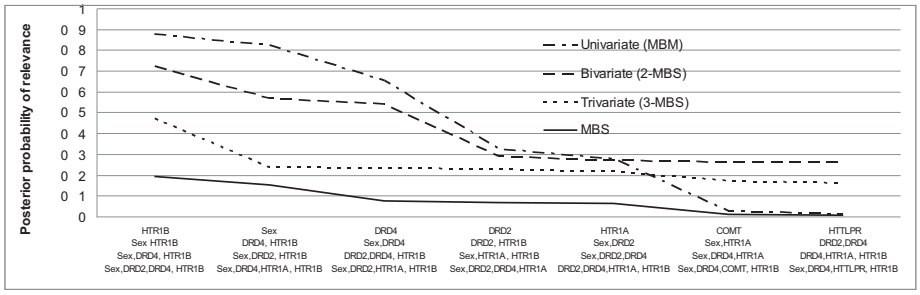

Fig. 3. The most probable uni-, bi-, and tri-variate subsets with decreasing multi-target relevance.

Finally we performed a refined analysis for multiple outcomes and applied the BN-BMLA method for the three subscales separately, also for the BIS Total score, and also jointly to compute the multi-target relevance. This confirmed that HTR1B is strongly relevant for the motoric and noplanning subscales, DRD4 has somewhat weaker, but similar multiple effect, but interestingly Comt and HTR1A is strongly relevant only for the cognitive subscale.

6 Discussion

The applied methods (SSVS, BN-BMLA) gave similar results. Both confirmed that the serotonergic and dopaminergic polymorphisms affect the trait impulsivity scores, specifically DRD4 and HTR1B. Furthermore BN-BMLA provided a coherent characterization of the system of dependencies and a detailed picture of the genetic background of the subscales which makes it a promising option in genetic studies. The Bayesian network based analysis confirmed association of DRD4 with BIS Total, moreover it was also weakly linked to all three subscales. With respect to this variable set the effect of DRD4 was direct, i.e. it was strongly relevant. HTR1B showed marked effects only towards the BIS Total score and towards the Motor subscale. The analysis showed that there was no statistical interaction between these two variables, which was confirmed by posterior decomposition analysis [2].

We analyzed partial multivariate strong relevances, because the Bayesian statistical framework allows the calculation of posteriors for the strong relevance of variables, pairs of variables, triplets of variables, etc. This is more flexible than the complete relevance patterns of all the variables, because it allows the selection of appropriate level of complexity of hypothe-

ses. As shown in Fig 5 the relevance of such subsets of variables exhibit differently peaked distributions, which are in close correspondence with feature complexity. These results indicated weak associations for HTR1A and Sex.

These preliminary results and other applications indicate that Bayesian networks offers a rich language for the detailed representation of types of relevance, including causal, acausal, and multi-target aspects. Additionally Bayesian statistics offers an automated and normative solution for the multiple hypothesis testing problem, thus using high-throughput and high-performance computing resources posteriors for global(!), detailed characterization of relevance relations can be estimated in medium sized problems (i.e. for hundreds of variables). This Bayesian statistical, global relevance analysis extends the scope of local "causal" discovery methods and because of the direct interpretation of Bayesian posteriors contrary to p-values from the frequentist approach, it is an ideal candidate for creating probabilistic knowledge bases to support off-line meta-analysis and fusion of background knowledge.

The coherent characterization of the uncertainties over the detailed types of relevances offers the opportunity to interpret the results of a Bayesian GAS analysis as a "Bayesian data analytic knowledge base". Currently we are working on techniques to allow the fusion of multiple Bayesian data analytic knowledge bases in related domains and support offline meta-analysis.

7 Acknowledgements

P.M. performed the SSVS based computations, P.S. performed the BN based computations and postprocessing. A.M. implemented the DAG-MCMC methods and algorithms for partial relevance, multiple relevance, and relevance types. A.Sz.,G.B. collected the data. M.S. performed the genotyping. P.A. designed the Bayesian analysis of partial relevance, multiple relevance, and relevance types. All the authors agree on contents of this paper. This work was supported by the NIH R03 TW007656 Fogarty International Research grant to Maria Sasvari-Szekely and the following Hungarian Scientific Research Funds: OTKA K81466 to Maria Sasvari-Szekely; OTKA-PD-76348 and NKTH TECH 08-A1/2-2008-0120 (Genagrid) to P. Antal. Anna Szekely and Peter Antal acknowledge the financial support of the János Bolyai Research Fellowship by the Hungarian Academy of Sciences. The authors thank Gabriella Kolmann for her technical assistance.

References

1. C. F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani, and X.D. Koutsoukos. Local causal and markov blanket induction for causal discovery and feature selection for classification. *Journal of Machine Learning Research*, 11:171–284, 2010.
2. ANTAL, P., MILLINGHOFFER, A., HULLAM, G., SZALAI, C. and FALUS, A. 2008. A Bayesian View of Challenges in Feature Selection: Feature Aggregation, Multiple Targets, Redundancy and Interaction. JMLR Workshop and Conference Proceedings.
3. P. Antal, G. Fannes, D. Timmerman, Y. Moreau, and B. De Moor. Bayesian applications of belief networks and multilayer perceptrons for ovarian tumor classification with rejection. *Artificial Intelligence in Medicine*, 29:39–60, 2003.
4. P. Antal, G. Hullám, A. Gézsi, and A. Millinghoffer. Learning complex bayesian network features for classification. In *Proc. of third European Workshop on Probabilistic Graphical Models*, pages 9–16, 2006.
5. D. J. Balding. A tutorial on statistical methods for population association studies. *Nature*, 7:781–91, 2006.
6. Stephens M. and Balding D.J. Bayesian statistical methods for genetic association studies. *Nature Review Genetics*, 10(10):681–690, 2009.
7. W. L. Buntine. Theory refinement of Bayesian networks. In *Proc. of the 7th Conf. on Uncertainty in Artificial Intelligence (UAI-1991)*, pages 52–60. Morgan Kaufmann, 1991.
8. P. Giudici and R. Castelo. Improving Markov Chain Monte Carlo model search for data mining. *Machine Learning*, 50:127–158, 2003.
9. G. F. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347, 1992.
10. COOPER, G. F. & HERSKOVITS, E. 1992. A Bayesian method for the induction of probabilistic networks from data. Machine Learning, 9, 309-347.
11. G. F. Cooper, C.F. Aliferis, R. Ambrosino, J. Aronis, B. G. Buchanan, R. Caruana, M. J. Fine, C. Glymour, G. Gordon, B. H. Hanusa, J. E. Janosky, C. Meek, T. Mitchell, T. Richardson, and P. Spirtes. An evaluation of machine-learning methods for predicting pneumonia mortality. *Artificial Intelligence in Medicine*, 9:107–138, 1997.
12. B. I. Fridley. Bayesian variable and model selection methods for genetic association studies. *Genetic Epidemiology*, 33:27–37, 2009.
13. N. Friedman and D. Koller. Being Bayesian about network structure. *Machine Learning*, 50:95–125, 2003.
14. A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall, London, 1995.
15. E. I. GEORGE, R. E. MCCULLOCH, 1993. Variable Selection Via Gibbs Sampling,Journal of the American Statistical Association, 88(423):881-889.
16. B. Han, M. Park, and X. Chen. A markov blanket-based method for detecting causal snps in gwas. *BMC Bioinformatics*, 11(3):5, 2010.
17. HULLAM, G., ANTAL, P., SZALAI, C. & FALUS, A. 2010. Evaluation of a Bayesian model-based approach in GA studies. JMLR Workshop and Conference Proceedings.
18. X. Jiang, M. M. Barmada, and S. Visweswaran. Identifying genetic interaction in genome-wide data using bayesian networks. *Genetic Epidemiology*, 34:575–581, 2010.
19. M. Koivisto and K. Sood. Exact bayesian structure discovery in bayesian networks. *Journal of Machine Learning Research*, 5:549–573, 2004.

20. R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97:273–324, 1997.
21. D. Koller and M. Sahami. Toward optimal feature selection. In *International Conference on Machine Learning*, pages 284–292, 1996.
22. D. Madigan, S. A. Andersson, M. Perlman, and C. T. Volinsky. Bayesian model averaging and model selection for Markov equivalence classes of acyclic digraphs. *Comm.Statist. Theory Methods*, 25:2493–2520, 1996.
23. PATTON, J. H., STANFORD, M. S. & BARRATT, E. S. 1995. Factor structure of the Barratt impulsiveness scale. J Clin Psychol, 51, 768-74.
24. PEARL, J. 2000. Causality: Models, Reasoning, and Inference, Cambridge University Press.
25. A. D. Martin, K. M. Quinn, J. H. Park, 2011. MCMCpack: Markov Chain Monte Carlo in R, Journal of Statistical Software, 42(9):1-21
26. R Development Core Team 2011, R: A Language and Environment for Statistical Computing,
27. C. Kooperberg and I. Ruczinski. Identifying interacting snps using monte carlo logic regression. *Genet Epidemiol*, 28(2):157–170, 2005.
28. M. Y. Park and T. Hastie. Penalized logistic regression for detecting gene interactions. *Biostatistics*, 9(1):30–50, 2007.
29. J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Francisco, CA, 1988.
30. J.M. Pena, R. Nilsson, J. Bjrkegren, and J. Tegnér. Towards scalable and data efficient learning of markov boundaries. *International Journal of Approximate Reasoning*, 45:211–232, 2007.
31. M. A. Province and I. B. Borecki. Gathering the gold dust: Methods for assessing the aggregate impact of small effect genes in genomic scans. In *Proc. of the Pacific Symposium on Biocomputing (PSB08)*, volume 13, pages 190–200, 2008.
32. B. Servin and M. Stephens. Imputation-based analysis of association studies: candidate genes and quantitative traits. *PLoS Genetics*, 3(7):e114, 2007.
33. I. Tsamardinos and C. Aliferis. Towards principled feature selection: Relevancy,filters, and wrappers. In *Proc. of the Artificial Intelligence and Statistics*, pages 334–342, 2003.
34. Lei Yu and Huan Liu. Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research*, 5:1205–1224, 2004.
35. Y. Zhang and J. S Liu. Bayesian inference of epistatic interactions in case-control studies. *Nature Genetics*, 39(9):1167–1173, 2007.
36. H. Xing, P. D. McDonagh, J. Bienkowska, T. Cashorali, K. Runge, R. E. Miller, D. DeCaprio adn B. Church, R. Roubenoff, I. Khalil, and J. Carulli. Causal modeling using network ensemble simulations of genetic and gene expression data predicts genes involved in rheumatoid arthritis. *PLoS Computational Biology*, 7(3):1001105, 2011.