A Bayesian View of Challenges in Feature Selection: Multilevel Analysis, Feature Aggregation, Multiple Targets, Redundancy and Interaction

Péter Antal ANTAL@MIT.BME.HU
András Millinghoffer MILLI@MIT.BME.HU
Gábor Hullám HULLAM@MIT.BME.HU
Dept. of Measurement and Information Systems
Budapest University of Technology and Economics
Csaba Szalai SZALAICS@GMAIL.COM
Inflammation Biology and Immunogenomics Research Group
Hungarian Academy of Sciences
András Falus FALAND@DGCI.SOTE.HU
Dept. of Genetics, Cell- and Immunobiology
Semmelweis University, Hungary

Editor: Saeys et al.

Abstract

In the paper we discuss applications of the Bayesian approach to new challenges in relevance analysis. Earlier, we formulated a Bayesian approach to Feature Subset Selection using Bayesian networks to jointly estimate the posteriors of Markov Blanket Memberships (MBMs), Markov Blanket Sets (MBSs), and Markov Blanket Graphs (MBGs) for a given target variable. These results of the Bayesian Multilevel Analysis of relevance (BMLA) correspond respectively to a model-based pairwise relevance, relevance of sets, and to the interaction models of relevant variables. Now we formulate refined levels in BMLA by introducing the concepts of k-MBSs and k-MBGs, which are intermediate, scalable model properties expressing relevance. We consider the extension of BMLA to multiple targets. We introduce and investigate a score for feature redundancy and interaction based on the decomposability of the structure posterior. Finally, we overview the problems of conditional and contextual relevance. We demonstrate the use of concepts and methods in the field of genomics of asthma.

1. Introduction

Earlier, we formulated generalizations of the feature subset selection problem in the Bayesian framework, based on structural properties of Bayesian networks (4). We presented the methodology of the Bayesian Multilevel Analysis (BMLA) of the relevance of input variables, which allows for the analysis of relevance at different abstraction levels (i.e., at the levels of Markov Blanket Memberships, Markov Blanket sets, and Markov Blanket graphs), and to express the sufficiency of the data and the uncertainty at the proposed multilevel representations.

However, there are many open issues in BMLA such as handling (1) more refined levels, (2) multiple target variables, (3) redundancy and interaction of features, (4) feature aggregation, (5) contextual relevance, and (6) predictive value of features. In this paper we discuss these extensions and experimentally investigate the first four issues. In Section 2 and Section 3 we overview the Bayesian approach to the Feature Subset Selection problem (FSS), including the concept of Markov Blanket Graph and Bayesian Multilevel Analysis. In Section 4 we formulate the concepts of k-MBS and k-MBG with scalable cardinality between pairwise relevance and complete subsets of relevant features, and between edges and subgraphs. In Section 5, we illustrate the concept of input feature aggregation. In Section 6, we discuss the concept of redundancy and interaction based on the decomposability of the structure posterior. In Section 7 we discuss the application of BMLA for multiple targets. In Section 8 we overview the concept of contextual relevance and its relation to relevance, conditional relevance, and interactions. In Section 10 we demonstrate the general concepts in a discrete, real-world application domain of the genomics of asthma using single nucleotide polymorphisms (SNPs), which are binary and ternary genomic variables responsible for most of the genetic variability (8). SNPs and genes are anonymized, because the biomedical publications of these results are still in progress.

2. Background

In the predictive approach, the concept of relevance can be defined specific to the applied model class used as a predictor, the optimization algorithm, the data set, and the loss function, whose generalization leads to the wrapper approach (13). In the filter approach, typically non-predictive methods rely on the following model-based, probabilistic definition of relevance for a set of variables $\mathbf{X}'$ (18).

Definition 1 (Markov boundary) *A set of variables $\mathbf{X}' \subseteq \boldsymbol{V}$ is called a* Markov blanket *set of X_i w.r.t. the distribution $p(\boldsymbol{V})$, if $(X_i \perp\!\!\!\perp V \setminus \mathbf{X}' | \mathbf{X}')_p$, where $\perp\!\!\!\perp$ denotes conditional independence. A minimal Markov blanket is called* Markov boundary*. Its indicator function is denoted by* $\mathrm{MBS}_p(X_i, \mathbf{X}')$.

Bayesian networks (BNs) and their properties offer a wide range of options for representing relevance (18). The following theorem gives a sufficient condition for the unambiguous BN representation of the relevant features.

Theorem 2 *For a distribution p defined by Bayesian network (G, θ) the variables* $\mathrm{bd}(Y, G)$ *form a Markov blanket of Y, where* $\mathrm{bd}(Y, G)$ *denotes the set of parents, children and the children's other parents for Y (18). If the distribution p is stable w.r.t. the DAG G, then* $\mathrm{bd}(Y, G)$ *forms a unique and minimal Markov blanket of Y,* $\mathrm{MBS}_p(Y)$ *and $X_i \in \mathrm{MBS}_p(Y)$ iff X_i is strongly relevant (22).*

We also refer to $\mathrm{bd}(Y, G)$ as the Markov blanket set for Y in G using the notation $\mathrm{MBS}(Y, G)$ by the implicit assumption that p is Markov compatible with G[1]. The induced

1. Note that in typical Bayesian scenarios (e.g., in case of Dirichlet distributions applied in the paper to specify $p(\theta|G)$), the graph-theoretic neighborhood $\mathrm{bd}(Y, G)$ is the unique Markov Boundary with probability 1, i.e. the parameterizations encoding independencies have measure 0 (17).

(symmetric) pairwise relation $\mathrm{MBM}(Y, X_j, G)$ w.r.t. G between Y and X_j is called *Markov blanket membership.*

$$\mathrm{MBM}(Y, X_j, G) \Leftrightarrow X_j \in \mathrm{bd}(Y, G) \tag{1}$$

3. Bayesian Multilevel Analysis of Relevance

Earlier works on using Bayesian network properties in relevance analysis include the Markov Blanket Approximating Algorithm (15), its recent extensions (24), the IAMB algorithm and its variants (2; 22; 23). Beside these deterministic, maximum likelihood, or maximum a posteriori (MAP) identification methods, stochastic and Bayesian approaches were proposed as well (for an ad hoc randomized approach, see (20)). In the computationally more demanding, Bayesian approach we are interested in the posteriors for various model properties expressing relevance for a given target variable Y. In earlier works the goal was the overall characterization of the domain using edge and MBM posteriors (11; 14; 16).

To extend the scope of the FSS problem we proposed the use of Markov Blanket Graph (MBG) feature (property), a.k.a. classification subgraph (1; 4).

Definition 3 (Markov Blanket Graph) *A subgraph of Bayesian network structure G is called the Markov Blanket Graph or Mechanism Boundary Graph* $\mathrm{MBG}(Y, G)$ *of variable Y if it includes the nodes in the Markov blanket defined by* $\mathrm{bd}(Y, G)$ *and the incoming edges into Y and into its children.*

For a probabilistic and causal interpretation of MBGs, a representation of observation equivalent MBGs, bounds for their cardinality and use in prediction, see (1; 4). An important property of the MBG is that it is sufficient for relevance analysis in case of complete data (which is the direct consequence of Th. 2). Unfortunately, the MBG posterior is not tractable computationally, but it is easy to show that the ordering-conditional posterior can be computed in polynomial time, which can be exploited in ordering-MCMC methods (4).

Note that the MBM and the MBS or MBG concepts reflect two different approaches to Bayesian network properties. The first approach provides an overall characterization as a fragmentary representation, and the number of features and feature values are tractable (e.g. linear or quadratic in the number of variables). Such features are pairwise edges, compelled edges, and Markov blanket relations. At the other extreme of feature learning we find the identification of arbitrary subgraphs with statistical significance (19). This is close to our approach to Bayesian network features investigated in the paper, but we restrict the subgraphs to Markov Blanket Graphs to have a focused representation from a single, but complex point of view (i.e., from the point of view of the FSS problem) and we use the Bayesian framework instead of the frequentist framework.

The Bayesian Multilevel Analysis of relevance goes one step further and it yields a comprehensive view of multiple levels. It allows for the calculation and crosslinking of the posteriors corresponding to features X_i, sets of features, and (sub)graph models of features and a target variable Y. Following our assumption about the underlying BN representation, this implies the calculation of the posteriors for the Markov Blanket Memberships, Markov Blanket sets, and Markov Blanket graphs. Further levels would be also possible either using domain specific knowledge for defining groups of variables w.r.t. their types, or collapsing

the MBG space to the space of class-focused restricted partially directed acyclic graphs (C-RPDAGs) (1). Note that the MBM, MBS, and MBG features form a hierarchy of increasing complexity ($|MBM| \ll |\,\mathrm{MBS}\,| < |\,\mathrm{MBG}\,| < |BN|$).

4. Multivariate Scalability: k-MBS and k-MBG Features

The multiple levels in BMLA offer a wide range of analysis at multiple abstraction levels (i.e., with varying complexity). However, the MBG and MBS features are much more expressive than the edge and MBM features, e.g. their cardinalities are superexponential, exponential, and linear for a given target respectively. Consequently, the MBG and MBS posteriors are often too "flat" (i.e. there are hundreds of MBS or MBG features with moderately high posteriors), even when the MBM posteriors are peaked (for further details see (4)). Typically, —even in the "flat" posterior case— the most probable MBS and MBG feature values often share a significant common part. To handle this we define concepts between MBMs and MBSs, and edges and MBGs, which are focused on target variables and they have intermediate, scalable complexities.

Definition 4 (k-MBS) *For a distribution $p(\boldsymbol{V})$ ($|\boldsymbol{V}| = n$), if all the variables $X_i \in \boldsymbol{s}$, where $\boldsymbol{s} \subseteq \boldsymbol{V}$, are members of a Markov Boundary set* mbs *and $|\boldsymbol{s}| = k$, then $\boldsymbol{s}$ is called a k-ary Markov Boundary subset*[2] $\text{k-MBS}_p(\boldsymbol{s}, Y) \Leftrightarrow (\exists\,\text{mbs} : \text{MBS}_p(\text{mbs}, Y), \boldsymbol{s} \subseteq \text{mbs}$.

Definition 5 (k-MBG) *A subgraph g of Bayesian network structure G is called the k-ary Markov Blanket Graph* $\text{k-MBG}(g, Y, G)$ *of variable Y if it includes k edges of the* $\text{MBG}(Y, G)$[3].

The graph-theoretic characterization of the k-MBS_p concept is as follows.

Proposition 6 *For a stable distribution p defined by Bayesian network (G, θ) s is k-ary Markov Boundary subset* $\text{k-MBS}_p(s, Y)$*, iff $s \subseteq \text{bd}(Y, G)$ and $|s| = k$ (otherwise* $\text{bd}(Y, G)$ *may not be minimal)*[2].

The k-MBS and k-MBG offer scalable features for the analysis of relevance, as their cardinalities are polynomial $\mathcal{O}(n^k)$. In practice this means, that we can analyze the most probable $\text{k-MBS}(Y)$ and $\text{k-MBG}(Y)$ feature values in a relatively large range of k values. The posteriors for k-MBS and k-MBG can be derived off-line from the estimates for the MBS and MBG posteriors. The maximum value of k, at which model properties (feature values) with high probability are present is problem dependent. Reasonable limits can be found either by a bottom-up or a top-down approach starting from $k = 1$ or $k = |\boldsymbol{V}|$ respectively (note that for intermediate values of k the number of feature values is computationally not tractable, e.g. $\binom{n}{k}$ for k-MBS).

2. Because p is stable with probability 1 in case of Dirichlet distributions applied in the paper to specify $p(\theta|G)$ (17), we also use the indicator function $\text{k-MBS}(s, Y, G)$ assuming that p is compatible with G. However, in regard to the possible non-stable cases with potential non-minimality of s, we call these sets in general k-ary Markov Blanket subsets.
3. The posterior for the presence of a given edge e in the complete domain model G is different from the posterior for the presence of e in $\text{MBG}(Y, G)$, because the presence of an edge in $\text{MBG}(Y, G)$ may depend on the presence of other edges.

5. A Knowledge-rich Aggregation of Input Features

An attractive property of the Bayesian approach to relevance is that the model posterior can be transformed and interpreted without theoretical restrictions. In our case, using the space of Bayesian network structures, it means that the posterior $p(G|D_N)$ can be aggregated by any partitioning over model structures G, where each partitioning offers a potentially different interpretation. However, only few partitions have a general or domain-specific meaning.

Beside noninformative model aggregation, the prior domain knowledge can be used as well to define interesting partitions. As with the noninformative aggregation, such an aggregation can (1) provide a more general description of relevance relations in the domain, and (2) yield more confident numerical results. E.g., a straightforward way to augment the SNP space is to introduce the level of genes, because many SNPs are related to a given gene. On the level of genes, we have calculated the aggregated versions of the Markov blanket membership and Markov blanket set relations. The corresponding equations are derived from their counterparts belonging to the more specific SNP level, e.g. (where Y, g, s respectively denote the target, gene, and SNP variables)

$$p(MBM(Y, g|D)) = \sum_{G:\exists s:onGene(g,s))\wedge MBM(Y,s,G)} p(G|D). \tag{2}$$

6. Interaction, Redundancy based on Posterior Decomposition

In relevance analysis we typically focus on high-scoring subfeatures, although low probabilities may also indicate important relations, because composite measures representing high-level semantic properties can be constructed. Such a score for the discovery of interaction and redundancy can be constructed using the exact k-MBS posterior and its approximations as the product of the Markov Blanket Membership probabilities of each member variable X_i in the given k-MBS, as if their occurrences were independent

$$p(\text{k-MBS}(\mathbf{X}', Y, G)|D_n) \approx \prod_{X_i \in \mathbf{X}'} p(\text{MBM}(Y, X_i, G)|D_n), \tag{3}$$

These approximations related to the decomposability of the structure posterior enable a direct Bayesian approach to the concept of redundancy and interaction. If the higher-order k-MBS posterior is larger than the approximation based on lower-order k-MBS posteriors, it may indicate that the subset has interacting features. In the opposite case, it may indicate the redundancy of features. This is formalized in the following definition, which can be generalized to multiple variables and orders higher than 1 (i.e., not only for MBMs, which are $1 - MBS$s).

Definition 7 (Interaction and redundancy) *The features $\mathbf{X}' = \{X_{i_1}, \ldots, X_{i_k}\}$ are 1,k-product interacting (redundant), if the posterior $p(\text{k-MBS}(\mathbf{X}', Y, G)|D_N)$ is larger (less) than $\prod_j p(\text{MBM}(X_{i_j}, Y, G)|D_N)$.*

The task of finding redundant subfeatures can be regarded as the complement of finding stable subfeatures, e.g. in the first case we are looking for those elements which often supplement the stable parts of features.

7. Relevance for Multiple Targets

If there are multiple possible target variables $\boldsymbol{Y}$ which have to be examined together and the relations among them are irrelevant one may ask for the variables relevant to the target set. Note, that this is similar to the aggregation of input features in Section 5, but in this case the target variables are "aggregated". Fortunately, the basic concepts of relevance discussed earlier can easily be extended to use target sets instead of a single target node.

Definition 8 (Multi-target relevance) *A feature (stochastic variable) X_i is strongly (weakly) relevant to $\boldsymbol{Y}$, if it is strongly (weakly) relevant to any $Y_i \in \boldsymbol{Y}$*

It is easy to see that the union of the MBSs of the targets, except the elements of the target set itself, is a Markov Blanket set for the target set.

Proposition 9 *If* $\mathrm{MBS}(\mathbf{Y}) = (\bigcup_{Y_i \in \mathbf{Y}} \mathrm{MBS}_p(Y_i)) \setminus \mathbf{Y}$, *then* $\mathrm{MBS}(\mathbf{Y})$ *is a Markov blanket for $\boldsymbol{Y}$ w.r.t. distribution p.*

An equivalent proposition can be stated for Markov boundaries, although the effects of logical dependencies should be handled appropriately. Note, that the posterior for a given target set $\boldsymbol{Y}$ cannot be calculated from the posteriors corresponding to the members of any partitioning of $\boldsymbol{Y} = \bigcup_i \boldsymbol{Y}_i$, because of the dependencies. However posteriors corresponding to subsets of the target set can be used for an approximation. In case of MBMs and singular variables Y_i, e.g.

$$p(MBM(X_j, \boldsymbol{Y}, g)|D_N) \approx 1 - \prod_i (1 - p(MBM(X_j, Y_i, g)|D_N)) \tag{4}$$

Still, in case of MBMs, if the posteriors are available for all of the subsets $\boldsymbol{Y'} \subseteq \boldsymbol{Y}$, then for any $\boldsymbol{Y''} \subseteq \boldsymbol{Y}$, using inductively $p(A \cap B) = p(A) + p(B) - p(A \cup B)$, we can compute the posterior probability that X_j is a Markov blanket member for each $Y_i \in \boldsymbol{Y}$.

The extension of the MBG concept for multiple targets is similarly straightforward, which again defines the necessary and sufficient dependency structure and parameters for predicting the targets under general conditions.

8. Conditional and Contextual Relevance

The fundamental definitions of relevance in Def. 1 are based on the general concept of conditional independence. However, as conditional independence can be made more specific by introducing *contextual independence*, we can introduce the concept of *contextual relevance* to support more refined analysis. Recall that contextual independence is a specialized form of conditional independence, i.e when conditional independence is valid only for a certain value $\boldsymbol{c}$ of another disjoint set $\boldsymbol{C}$ (for its use in the context of Bayesian networks, see e.g. (6)). Let us denote the *contextual independence* of $\boldsymbol{X}$ and $\boldsymbol{Y}$ given $\boldsymbol{Z}$ and context c with $(\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} | \boldsymbol{Z}, \boldsymbol{c})$, that is

$$(\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} | \boldsymbol{Z}, \boldsymbol{c}) \textit{ iff } (\forall \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}\ p(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{c}, \boldsymbol{x}) = p(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{c}) \textit{ whenever } p(\boldsymbol{z}, \boldsymbol{c}, \boldsymbol{x}) > 0). \tag{5}$$

An analogous extension for relevance is as follows.

Definition 10 (Contextual Irrelevance) *Assume that* $\boldsymbol{X} = \boldsymbol{X'} \cup \boldsymbol{C''}$ *is relevant for* $\boldsymbol{Y}$*, that is* $(\boldsymbol{Y} \not\perp\!\!\!\perp (\boldsymbol{X'} \cup \boldsymbol{C''})$*, and* $(\boldsymbol{X'} \cap \boldsymbol{C''}) = \emptyset)$*. We say that* $\boldsymbol{X'}$ *is contextually irrelevant if there exists some* $\boldsymbol{c''}$ *for which* $(\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{X'} | \boldsymbol{c''})$*.*

For completeness, recall the definition of conditional relevance

Definition 11 (Conditional Relevance) *Assume that* $\boldsymbol{X} = \boldsymbol{X'} \cup \boldsymbol{C'}$ *is relevant for* $\boldsymbol{Y}$*, that is* $(\boldsymbol{Y} \not\perp\!\!\!\perp (\boldsymbol{X'} \cup \boldsymbol{C'}))$*, and* $(\boldsymbol{X'} \cap \boldsymbol{C'} = \emptyset)$*. We say that* $\boldsymbol{X'}$ *is conditionally relevant if* $(\boldsymbol{X'} \not\perp\!\!\!\perp \boldsymbol{Y})$*, but* $(\boldsymbol{X'} \not\perp\!\!\!\perp \boldsymbol{Y} | \boldsymbol{C'})$*.*

This definition applies to both weak and strong relevance. Note that conditional relevance and contextual irrelevance are independent, although typically somewhat opposite concepts. In case of conditional relevance, we have to know a value of a relevant feature C' to ensure the relevance of an otherwise irrelevant feature X'. Whereas in case of contextual irrelevance there should be a value c'' whose knowledge makes an otherwise relevant feature irrelevant.

The BMLA method based on standard BNs allows for a model-based Bayesian inference about conditional relevance (see Section 11). However, to handle contextual relevances, a Bayesian network representing contextual dependencies is necessary, e.g. using decision trees as local dependency models (6).

9. Algorithmic Aspects

The Bayesian inference over structural properties of Bayesian networks was proposed in (7; 9). In (16), Madigan et al. proposed a Markov Chain Monte Carlo (MCMC) scheme to approximate such Bayesian inference. The MCMC method over the DAG space was improved by Castelo et al. (12). In (11), Friedman et al. reported an MCMC scheme over the space of orderings. In (14), Koivisto et al. reported a method to perform exact full Bayesian inference over modular features.

To estimate the posteriors we applied both a DAG-based and an ordering-based Metropolis Coupled Markov Chain Monte Carlo (MC^3) method (4; 11; 12). Because of their correspondence, and the direct applicability of the DAG-MC^3 in the proposed extensions, we report results only from this method.

The settings are as follows. The number of chains is 10, the temperature parameter T is 1 (3). The length of the burn-in and MC simulation is 10^6 and $5 \cdot 10^6$, the probability of the DAG operators is uniform (12). The CH parameter prior and the uniform structure prior are used (9). The maximum number of parents is 4.

10. Results

We demonstrate the newly proposed general concepts in discrete domains using a small artificial model and a realistic reference model (see Fig. 6). The small model (G_0^-, θ_0^-) contains a binary target variable, its four ternary children, of which one has two other parents, and for the purpose of testing a completely independent uniform ternary variable was included. Its parameters are set so that dependence is weakening between C_1 and its parents T, P_1, P_2 in this order; and similarly between T and its children C_1, C_2, C_3,

C_4. The realistic reference model (G_0^+, θ_0^+) was learned from a real data set containing 1117 samples (5), which was slightly modified to test special cases of relevance. The model contains three clinical variables (*Asthma*, *Allergy*, *Rhinitis*) and forty-six SNPs selected from the asthma susceptibility region of chromosome 11q13 (21). We generated 10,000-10,000 complete random samples D^- and D^+ from the reference models. A data set D_N with size N is defined as the first N samples. We always use the corresponding set of variables for the data set (i.e., the domain is always identified implicitly by the data set in the paper).

Table 1 and Table 2 show a summary of the MAP outputs of Bayesian methods with MBM posteriors. For each set the corresponding members are marked with '1'. The MBG rows show a more refined picture. They not only identify the relevant variables, but also the type of interaction between them. Members of an MBG for a given target variable can be classified into the following types with decreasing precedence:

- sp: single parent of the target variable.
- mp: one of multiple parents of the target variable.
- sc: child of the target variable with no other parents.
- mc: child of the target variable with no parents not connected to the target.
- ic: child of the target with at least one parent not connected to the target.
- op: other parent (i.e. another parent of the target node's child(ren)) .

The distinction between single and multiple parental status is important, because of the v-structure formed by the multiple parents and the target. The last three cases are noteworthy, as they indicate various roles in interactions.

To evaluate the sufficiency of the data for the multivariate analysis of relevance, we report the ranked MBS posteriors in Fig. 1. To indicate the relations within the multilevel approach, Fig. 1 also shows the MBM-based approximations of the posteriors calculated as

$$p(\mathrm{MBS}(\boldsymbol{X}, Y, G|D_N) \approx \prod_{X_i \in \boldsymbol{X}} p(\mathrm{MBM}(Y, X_i)|D_N) \prod_{X_i \notin \boldsymbol{X}} (1 - p(\mathrm{MBM}(Y, X_i)|D_N)). \quad (6)$$

For more information about the power of the data in the multivariate case, we computed the posteriors of the Markov blanket sets of the target variable for increasing data size (see Fig. 2). Furthermore, to provide quantitative measures characterizing the uncertainty, Fig. 2 shows the entropy for the MBM, MBS, and MBG posteriors as well for growing sample size.

Next, we report results about the effects of syntactic and semantic aggregations. Fig. 3 and Fig. 4 report the maximal posteriors for k-MBS, also showing the posterior of the MAP MBS.

Fig. 5 reports the probability of relevance of SNPs at the aggregation level of genes.Fig. 6 indicates the decomposability of the MBS posteriors according to Section 6. Finally, Fig. 7 reports the sequential posteriors that a given SNP is relevant for Asthma, Allergy and Rhinitis, both separately and jointly. It also shows the approximation of the MBM posterior for the joint target set based on the MBM posteriors for individual targets according to Eq. 4.

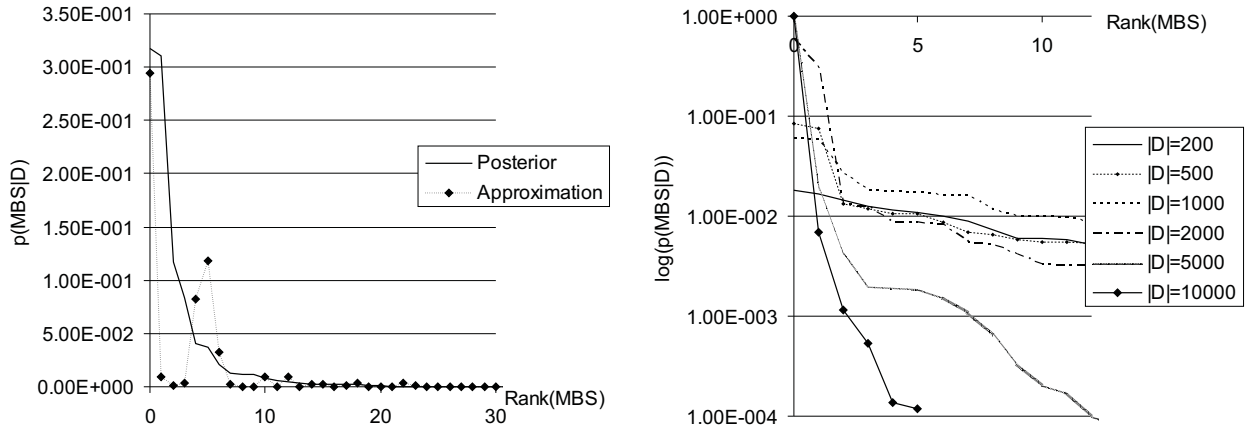

Figure 1: (Left) The peakness of the posteriors of the most probable MB sets using the D^-_{100} data set. The bold line denotes the estimated posteriors, the dots denote the corresponding MBM-based approximation. (Right) The ranked posteriors of the most probable MB sets using the D^+ data sets with growing sample size. The horizontal axis shows the posteriors of the MB sets; the vertical axis shows the ranks.

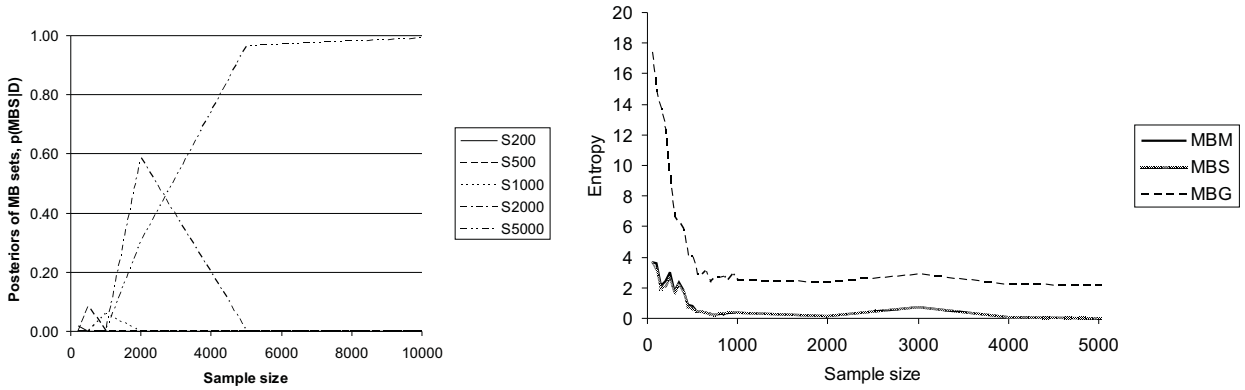

Figure 2: (Left) Sequential posteriors of relevant MB sets using data set D^+. The sets are the MAP sets for the 200, 500, 1000, 2000, 5000, and 10000 sample sizes. (Right) The summed entropies of the $p(\mathrm{MBM}(Y, X_i, G)|D^-)$ posteriors, and the entropy of the MBS and MBG posteriors for the data set D^-.

11. Discussion

At the pairwise level, the advantage of the MBM posteriors compared to standard pairwise statistical association tests is that despite its pairwise representation it correctly indicates whether the variables are part of a multivariate, direct relation. This is the consequence of the fact that it is still a model-based statistical relation (derived by Bayesian model

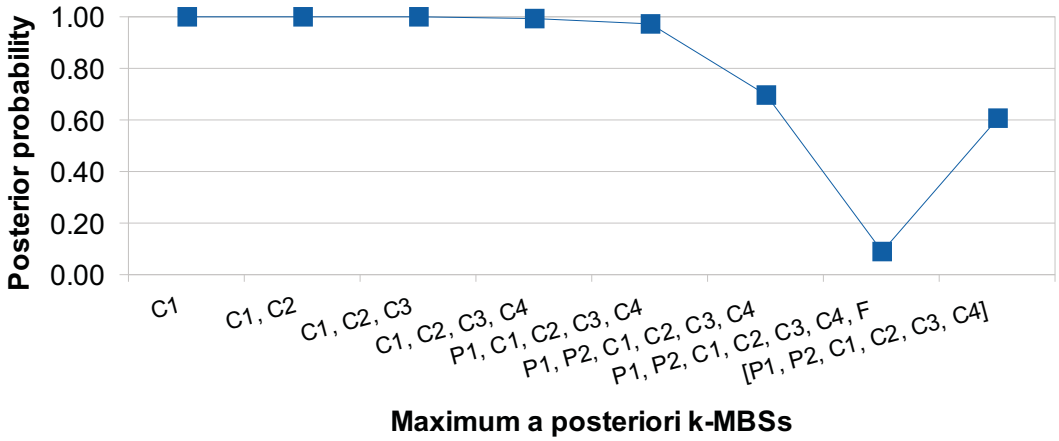

Figure 3: The maximal posteriors for k-MBS-s compatible with the MAP MBS for increasing $k = 1, \ldots, 6$ using the D^{-}_{300} data set. The last data point shows the posterior of the MAP MBS.

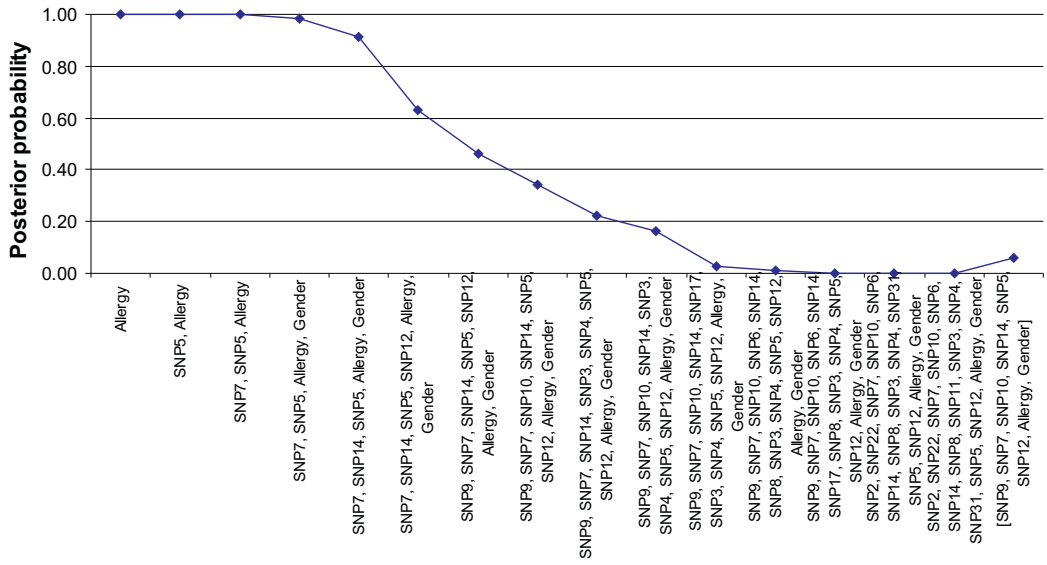

Figure 4: The maximal posteriors for k-MBS-s compatible with the MAP MBS for increasing $k = 1, \ldots, 15$ using the D^{+}_{1000} data set. The last data point shows the posterior of the MAP MBS.

averaging). Another general advantage of the MBM posteriors is that their multiple use is not hindered by the problem of multiple testing and they can be used in sequential analysis.

To evaluate the necessity of the multivariate approach, we have to investigate the relation of the MBM posteriors and the MBS posteriors. Note that because of the Bayesian foundation the MBM posteriors are marginal distributions of the MBS posterior. Fig. 1

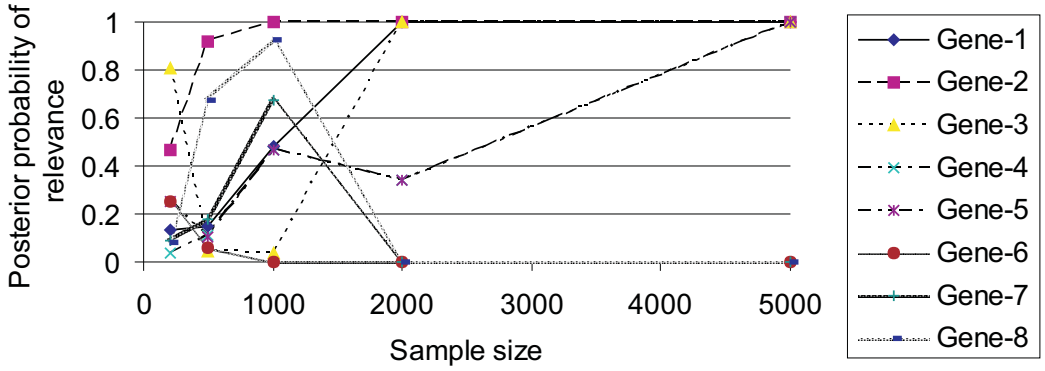

Figure 5: The sequential posteriors that a given gene contains a SNP relevant for asthma using the data set D^+. The probability of relevance is induced by the posterior $p(MBM(SNP_i, Asthma|D^+)$ for varying sample size.

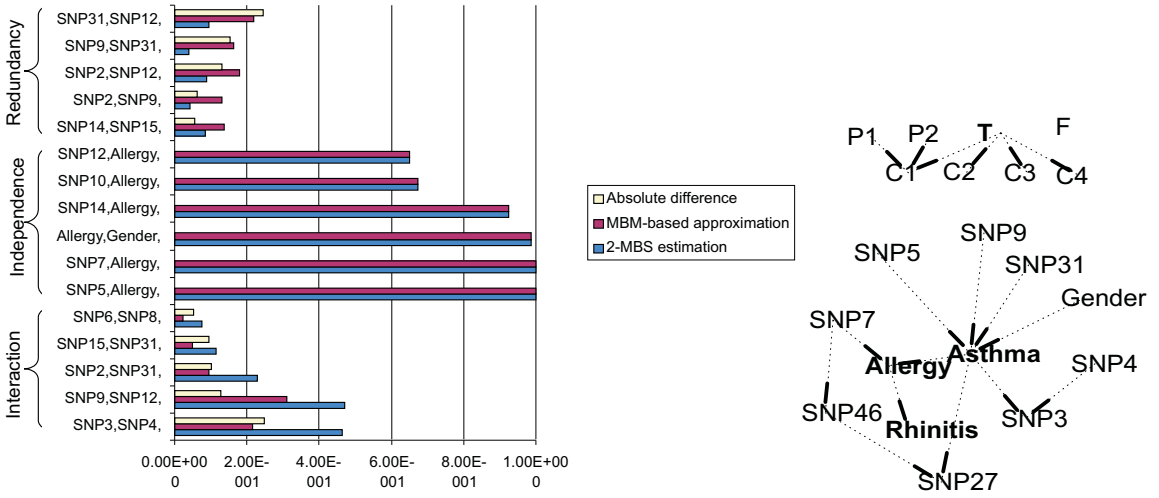

Figure 6: (Left) The quantification of interaction and redundancy by the decomposability of the posterior using the data set D^+_{1000}. The 2-MBS posteriors, their MBM) based approximation, and their difference are reported for the 5 most interacting, 5 most independent, and 5 most redundant pairs of variables (i.e., when the difference is minimal, closest to zero, and maximal). (Right) The Markov Blanket Graph for the target set *Asthma*, *Allergy*, *Rhinitis* in the reference SNP model G^+_0 and the complete artificial model G^-_0.

reports the posteriors of the MAP MB sets and their pairwise MBM-based approximated values. The monotone decreasing curve corresponds to the ranked estimated MBS posteriors (the posteriors are insignificant for sets with ranks larger than 100). A good approximation indicates the conditional independence of the features. As Fig. 1 shows this is not the case, as the MBM-based approximation performs badly both w.r.t. estimation and rank-

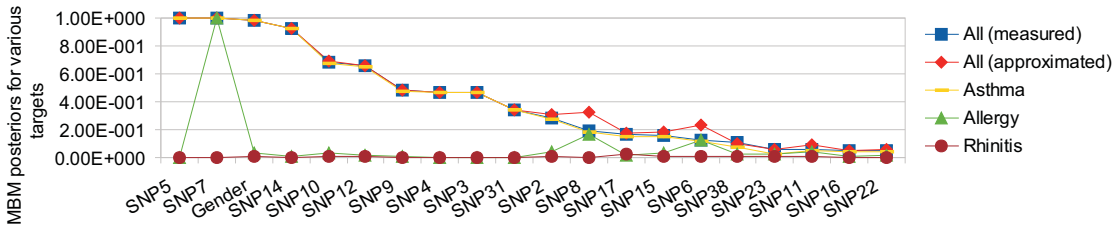

Figure 7: The MBM posteriors that a SNP is relevant for asthma, allergy and rhinitis, both separately and jointly using the data set D^{+}_{1000}. It also shows the approximation of the posterior for the joint target set based on the posteriors for individual targets.

ing. Specifically, the MAP MB set and the set defined by the pairwise MAP MBM values are different. Indeed, there are no trivial thresholds for the pairwise MBM posteriors to identify the certain and uncertain members of the MB sets with high posteriors, for which the k-MBS concept was proposed. Furthermore, Fig. 1 indicates the lack of a dominating MB set, i.e. the lack of single set of variables with an outstanding posterior. The joint evaluation of these indicators at multiple levels can support a detailed evaluation of this phenomenon as follows. Flat posteriors at the MBM and MBS levels can be the consequence either of insufficient data or the redundancy of the features, which can be decided by the evaluation of the posterior at the MBS level, because it explicitly shows the alternative sets of features. Because such alternating sets are not present in the most probable MB sets, the data set probably does not define a posterior at the MBS level that is peaked enough. We also investigated the change of the entropy at the MBM and MBS levels for amount of data D^{-}_{0} in the small domain. This shows a rapid convergence for 1000 samples for all levels; the non-zero value of the entropy at the MBG level is the consequence of the 4 MBGs in the observationally equivalent class of G^{-}_{0}. Note that the sum of the entropies for the MBM posteriors approximates closely the entropy of the underlying MBS posterior, which shows that the MBM features are relatively independent of each other (equality holds only for total independence, otherwise the sum of decomposed entropies is strictly larger then the original, see (10)).

The reference models contain several interactions and features of conditional relevance, see Fig. 6. Their MBS and MBG properties can be correctly identified by the MAP MBS and MBG above 10^4 samples (see Table 1 and 2), which asymptotic observation holds for the newly introduced k-MBS and k-MBG features, gene level aggregation, and for multiple target variables as well. Furthermore, the sequential analysis in Table 1 is also consistent with strength of dependencies encoded in $(G^{-}_{0}, \boldsymbol{\theta}^{-})$. To illustrate the effect of input aggregation and multiple outputs we used 10^3 samples, which is moderate, typical sample size w.r.t. this set of variables. This sample size was also used for the investigation of the decomposability of the posterior to infer interaction and redundancy.

As for the proposed k-MBS and k-MBG features with intermediate complexity, Fig. 3 indicates that for the G^{-} domain 300 samples are enough to ensure a high posterior ($0.9 <$) for $k < 6$. In the G^{+} domain Fig. 4 indicates that 10^3 samples are enough to ensure a

high posterior (0.9 <) for $k < 5$. It is also noteworthy, that the posterior of the MAP MBS (0.0592) is significantly lower than the corresponding 8-MBS defined by its members (0.345). Despite the restriction in the use of maximal posteriors, these results clearly justify that the proposed k-MBS feature can fulfill its intended role to fill the gap between MBM and MBS features.

Fig. 5 shows the result of feature aggregation based on domain knowledge using a sequential approach, which illustrates the easy transformation and fusion of Bayesian results to support interpretation.

The evaluation of interactions and redundancy according to Def. 7 indicated mostly independence, but in Fig. 6 we report the five-five pairs with maximal difference between their estimated posteriors and MBM based product approximations according to Eq. 3 in the reference model G_0^+. The pair *SNP3,SNP4* with "highest" difference indicating interaction really does form an interaction in the reference model (*SNP3* is a child of *Asthma* and *SNP4* is another parent of *SNP3*). The members of the pair *SNP9,SNP31* with "highest" difference indicating redundancy are potentially redundant, multiple parents of *Asthma*.

Finally Fig. 7 demonstrates the joint use of multiple target variables, although in this case one of the target variables (*Asthma*) nearly determines the MBM posteriors for the whole set. Furthermore, the relevant variables for the target variables in the reference model are mostly different, thus the approximation in Eq. 4 gives close values.

12. Conclusion

The Bayesian multilevel methodology using Bayesian network features together with the optional sequential analysis using growing sample size provides deeper insight to the sufficiency of the data. It is also capable for incorporating wide range of prior knowledge, thus it is especially applicable for the analysis of data with small sample size. The exact modeling of interactions by the MBG features using Bayesian networks and the Bayesian approach to the feature subset "selection" problem offered a principled solution for quantifying the uncertainty in inferring relevant features and their joint interactions. In the paper we treated the following issues.

1. *Joint management of interactions.* We discussed an exact generalization of the FSS problem with interactions using the concept of MBGs. It allows the identification of interactions of relevant variables, and conditionally relevant features.

2. *Bayesian MultiLevel Analysis of relevance.* The joint usage of different feature levels has multiple advantages: we can better understand the types of the relevance relations, and the necessity and possibility of the multivariate, and the multivariate-interactionist analysis. Using Bayesian networks in the Bayesian framework, we showed that there is a significant amount of uncertainty at the level of feature subsets and their joint interaction in typical scenarios with less than 10000 samples. That is the size of the data is typically not enough for a dominant MAP solution, which means that there are many MBSs with high posteriors, see Fig. 1.

3. *Concepts of* k-MBS *and* k-MBG. We introduced new prediction-oriented (focused) Bayesian network properties for relevance analysis with scalable $\mathcal{O}(n^k)$ polynomial complexity.

4. *Multiple target variables.* We formulated concepts for the relevance analysis in case of multiple targets and showed that it is a distinct problem in the Bayesian approach in general, because the posterior for the target set cannot be reconstructed from the posteriors for the partitions of the target set.

5. *Interaction and redundancy discovery.* We introduced a direct Bayesian approach to evaluate interaction and redundancy, which is based on the decomposability of the structure posterior.

Note, that these extensions, e.g. the concepts of k-MBS or k-MBG and "multi-target" relevance can be useful in frequentist methods as well.

Acknowledgements

This study was supported by grants: OTKA (National Scientific Research Fund): T046372 (C. Szalai); TS/2 044707 (A. Falus); and ETT (Ministry of Health) 451/2006 (C. Szalai), OTKA-PD (Hungarian Scientific Research Fund):76348 (P.Antal), Bolyai Grant (Hungarian Academy of Science): (P.Antal).

References

[1] S. Acid, L. M. de Campos, and J. G. Castellano. Learning bayesian network classifiers: searching in a space of partially directed acyclic graphs. *Machine Learning*, 59:213–235, 2005.

[2] C.F. Aliferis, I. Tsamardinos, and A. Statnikov. Large-scale feature selection using markov blanket induction for the prediction of protein-drug binding, 2003.

[3] 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.

[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] P. Antal, A. Millinghoffer, G. Hullám, G. Hajós, Cs. Szalai, and A. Falus. A bioinformatic platform for a Bayesian, multiphased, multilevel analysis in immunogenomics. In D.R.Flower M.N.Davies, S.Ranganathan, editor, *Bioinformatics for Immunomics*, pages –. Springer.

[6] C. Boutilier, N. Friedman, M. Goldszmidt, and D. Koller. Context-specific independence in bayesian networks, 1996.

[7] W. L. Buntine. Theory refinement of Bayesian networks. In *Proc. of the 7th Conf. on Uncertainty in Art.Int.(UAI-1991)*, pages 52–60. Morgan Kaufmann, 1991.

[8] The International HapMap Consortium. A second generation human haplotype map of over 3.1 million snps. *Nature*, 449:851–862, 2007.

[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] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley & Sons, 2001.

[11] N. Friedman and D. Koller. Being Bayesian about network structure. *Machine Learning*, 50:95–125, 2003.

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

[13] R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97:273–324, 1997.

[14] M. Koivisto and K. Sood. Exact bayesian structure discovery in bayesian networks. *Journal of Machine Learning Research*, 5:549–573, 2004.

[15] D. Koller and M. Sahami. Toward optimal feature selection. In *International Conference on Machine Learning*, pages 284–292, 1996.

[16] 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.

[17] C. Meek. Causal inference and causal explanation with background knowledge. In Philippe Besnard, Steve Hanks, Philippe Besnard, and Steve Hanks, editors, *Proc. of the 11th Conf. on Uncertainty in Artificial Intelligence (UAI-1995)*, pages 403–410. Morgan Kaufmann, 1995.

[18] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Francisco, CA, 1988.

[19] D. Pe'er, A. Regev, G. Elidan, and N. Friedman. Inferring subnetworks from perturbed expression profiles. *Bioinformatics, Proc. of ISMB 2001*, 17(Suppl. 1):215–224, 2001.

[20] 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.

[21] C. Szalai. Genomic investigation of asthma in human and animal models. In A. Falus, editor, *Immunogenomics and Human Disease*, pages 419–441. Wiley, London, 2005.

[22] 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.

[23] I. Tsamardinos, C.F. Aliferis, and A. Statnikov. Algorithms for large-scale local causal discovery and feature selection in the presence of limited sample or large causal neighbourhoods. In *The 16th International FLAIRS Conference*, 2003.

[24] Lei Yu and Huan Liu. Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research*, 5:1205–1224, 2004.

Table 1: The MAP summary of the BMLA of relevance with target T for the G_0^- reference model using a small (10^2), medium (10^3), and large (10^4) data set from D^-. The first row shows the types of the nodes in G_0^- w.r.t. List 10, then the MBM posteriors, the MAP Markov blanket sets, and the MAP Markov Blanket Graphs are shown for varying sample size. For the MBG type codes see Section 10.

	P 1	P 2	C 1	C 2	C 3	C 4	F
G_0^-	op	op	ic	sc	sc	sc	
MBM 100	87	76	95	100	98	69	58
MBS 100	1	1	1	1	1	1	1
MBG 100			sp	sc	sc		
MBM 1000	100	4	100	100	100	100	1
MBS 1000	1		1	1	1	1	
MBG 1000	op		ic	sp	sc	sc	
MBM 10000	100	100	100	100	100	100	0
MBS 10000	1	1	1	1	1	1	
MBG 10000	op	op	ic	sp	sc	sc	

Table 2: The MAP summary of the BMLA of relevance with *Asthma* target for the G_0^+ reference model using a (s)mall ($2 \cdot 10^2$), (m)edium (10^3), and (l)arge (10^4) data set from D^+. The first row shows the types of the nodes in G_0^+ w.r.t. List 10, then the MB(M) posteriors, the MAP Markov Blanket (S)ets, and the MAP Markov Blanket (G)raphs are shown for varying sample size (denoted by the combinations of s/m/l and M/S/G). Only the first two digits of the posterior values are shown and "+" denotes 1.00. The variables (*Gender*, *Allergy*, *Rhinitis*) are denoted by "G.", "A.", and "R.", for the SNP variables only their indices are shown (the following SNP variables are omitted as their MBM posterior is always less then 0.1 and never enter a MAP MBS and MBG: 1, 16, 18, 19, 21, 20, 22, 23, 24, 25, 26, 28, 32, 34, 35, 36, 37, 40, 41, 42, 43, 44. For the MBG type codes see Section 10.

	A.	G.	R.	2	3	4	5	6	7	8	9	10	11	12	13	14	15	17	27	29	30	31	33	38	39	45	46
G_0^+	ic	mp			ic	op	mp		op		mp								ic			mp					op
s-M	65	4	22	14	24	3	46	5	13	13	13	8	31	13	17	8	15	15	1	53	61	24	25	86	24	30	80
s-S	1						1	1							1						1			1			1
s-G	op						mp	mp							mp						op			ic			mp
m-M	+	98	0	27	46	46	99	11	99	18	47	67	5	64	1	92	14	14	1	0	0	33	0	7	0	0	3
m-S	1	1					1		1		1	1		1		1											
m-G	ic	ic					mp		op		mp	op		mp		mp											
l-M	+	+		0	+	+	+		+	0	+								+			+		0	0		+
l-S	1	1			1	1	1		1		1								1			1					1
l-G	ic	mp			ic	op	mp		op		mp								ic			mp					op