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

P. Antal[1], A. Millinghoffer[1], G. Hullám[1], Cs. Szalai[2], and A. Falus[3]

[1] Dept. of Measurement and Information Systems, Budapest Univ. of Tech.
[2] Inflammation Biology and Immunogenomics Res. Group, Hung. Acad. of Sci.
[3] Dept. of Genetics, Cell- and Immunobiology, Semmelweis University, Hungary
antal@mit.bme.hu

Abstract. Earlier, we formulated a Bayesian approach to Feature Subset Selection using Bayesian networks, which jointly estimate the posteriors of Markov Blanket Memberships (MBMs), Markov Blanket Sets (MBSs), and Markov Blanket Subgraphs (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. In this paper we discuss applications of the Bayesian approach to new challenges in relevance analysis. First, we formulate refined levels in BMLA by introducing the concepts of k-MBSs and k-MBGs, which are intermediate, scalable model properties expressing relevance. Second, we consider the extension of BMLA to multiple targets. Third, 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 concepts and methods in the field of the 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 [5]. We presented a methodology of Bayesian, Multilevel Analysis (BMLA) of the relevance of input variables, which is capable of analyzing relevance at different abstraction levels (i.e., at the levels of Markov Blanket Memberships, Markov Blanket sets, and Markov Blanket graphs). BMLA can express the sufficiency of the data, and the uncertainty at the proposed multilevel representations.

However, there are many open issues in BMLA such as (1) more refined levels, (2) multiple target variables, (3) redundancy and interaction of features, (4) contextual relevance, and (5) predictive value of features. In this paper we discuss these extensions and experimentally investigate the first three issues. The paper is organized as follows. In Section 2 we overview the Bayesian approach to the feature subset selection problem (FSS), including the concept of Markov Blanket subgraph as a central Bayesian network property for the FSS.

In Section 3 we formulate the concepts of k-MBS and k-MBG with scalable, polynomial cardinality between pairwise relevance and complete subsets of relevant features, and between edges and subgraphs. In Section 4, we illustrate the concept of informative input feature aggregation. In Section 5, we discuss the concept of redundancy and interaction based on the decomposability of the structure posterior. In Section 6 we discuss the application of BMLA for multiple targets. In Section 7 we overview the concept of contextual relevance and its relation to relevance, conditional relevance, and interactions. In Section 9 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 tertiary genomic variables [18]. SNPs and genes are anonymized, because the biomedical publications of these results are still in progress.

2 Background

In the predictive approach to feature relevance, 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* [12]. In the *filter approach*, typically non-predictive methods approximate the following model-based definition of relevance [15].

Definition 1 (Markov boundary). *A set of variables* $\mathbf{X}'$ *is called a* Markov blanket *set of* X_i *w.r.t. the distribution* $p(X_1, \ldots, X_n)$*, 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 *[15]. Its indicator function is* $\mathrm{MBS}_p(X_i, \mathbf{X}')$.

Bayesian networks (BNs) and their properties offer a wide range of options for representing relevance, the discussion of which started with J.Pearl's seminal work [15]. The following theorem gives a sufficient condition for the unambiguous BN representation of the relevant features.

Theorem 1. *For a distribution* p *defined by Bayesian network* (G, θ) *the variables* $\mathrm{bd}(Y, G)$ *form a Markov blanket of* Y*, where* $\mathrm{bd}(X_i, G)$ *denotes the set of parents, children and the children's other parents for* X_i *[15]. If the distribution* p *is stable w.r.t. the DAG* G *[16], 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 [19].*

We also refer to $\mathrm{bd}(Y, G)$ as the Markov blanket for Y in G using the notation $\mathrm{MBS}(Y, G)$ by the implicit assumption that p is Markov compatible with G[4]. The induced (symmetric) pairwise relation $\mathrm{MBM}(Y, X_j, G)$ w.r.t. G between Y and X_j is called *Markov blanket membership.*

[4] 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 also the unique Markov Boundary with probability 1, because the possible parametrically encoded independencies have measure 0 [14].

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

To extend the FSS problem we proposed the use of Markov Blanket subGraph (MBG) feature (property), a.k.a. classification subgraph [5, 1] (see Fig. 3).

Definition 2 (Markov Blanket subGraph). *A subgraph of Bayesian network structure G is called the Markov Blanket subGraph or Mechanism Boundary subGraph* $\mathrm{MBG}(Y, G)$ *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, a representation of observation equivalent MBGs, bounds for its cardinality, and its use in prediction see [1, 5]. An important property of the MBG feature is that it is sufficient for relevance analysis in case of complete data (which is the direct consequence of Thm. 1). Unfortunately, there is no closed-form for the posterior $p(mbg|D_N)$ in general, but it is easy to derive a closed-form for the ordering-conditional posterior, which can be exploited in ordering-MCMC methods [5].

Earlier works on using Bayesian network properties in relevance analysis include the Markov Blanket Approximating Algorithm [13], its recent extensions [21], the IAMB algorithm and its variants [19, 20, 2, 17].

In the Bayesian approach, we are interested in the posteriors for various model properties expressing relevance for a given target variable Y. E.g. assuming BNs as a model class the MBS posterior is defined as follows

$$p(\mathrm{MBS}(Y, \boldsymbol{X}')|D_N) = \sum_G 1(\mathrm{MBS}(Y, G) = \boldsymbol{X}')p(G|D_N).$$

where the indicator function $1(\mathrm{MBS}(Y, G) = \boldsymbol{X}')$ is true, if the set of variables $\boldsymbol{X}'$ is the MBS of the target variable Y, given a specific DAG structure G; and D_N denotes the data set. The goal of the Bayesian multilevel analysis of relevance is the joint analysis of posteriors corresponding to features X_i, sets of features, joint models of interactions of relevant features. Following our assumption in this paper about the underlying BN representation, it means the estimation of posteriors for the Markov Blanket Memberships ($\mathrm{MBM}(Y, X_i)$), Markov Blanket sets ($\mathrm{MBS}(Y, \mathbf{X}')$), and Markov Blanket graphs ($\mathrm{MBG}(Y, \mathrm{MBG})$).

3 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 [5]). Typically, —even in the "flat" posterior case— the most probable MBS and MBG feature values

often show a significant common part. As a response to this we define concepts focused on target variables with scalable complexity between MBMs and MBSs, and between edges and MBGs as follows.

Definition 3 (k-MBS). *For a distribution* $p(\boldsymbol{V})$ *(*$|\boldsymbol{V}| = n$*), if all the variables* X_i *in* $\boldsymbol{s} \subseteq \boldsymbol{V}$ *are members of a Markov Boundary set* mbs *and* $|\boldsymbol{s}| = k$*, then* $\boldsymbol{s}$ *is called a k-ary Markov Boundary subset*[5] $(\text{k-MBS}_p(\boldsymbol{s}, Y) \Leftrightarrow (\exists\, \text{mbs} : \text{MBS}_p(\text{mbs}, Y), \boldsymbol{s} \subseteq \text{mbs})$*. Its indicator is denoted by* $\text{k-MBS}_p(\boldsymbol{s}, Y)$*.*

Proposition 1. *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).*[5]

The concept of k-MBS-s covers the gap between the MBS and MBM features (MBM ≡"1-MBS").

Definition 4 (k-MBG). *A subgraph* g *of Bayesian network structure* G *is called the k-ary Markov Blanket subGraph* $\text{k-MBG}(g, Y, G)$ *of variable* Y *if it includes* k *edges of the* $\text{MBG}(Y, G)$[6]*.*

The k-MBS and k-MBG offer scalable features for the analysis of relevance, as their cardinalities are polynomial ($\mathcal{O}(n^k)$ and $\mathcal{O}(n^{2k})$). In practice this means, that we can analyze the most probable $\text{k-MBS}(Y)$ and $\text{k-MBG}(Y)$ feature values for all reasonable k values. The posteriors for k-MBS and k-MBG can be derived off-line from the estimates for the MBG and MBS 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).

4 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 noninformative or informative meaning.

[5] Because p is stable with probability 1 in case of Dirichlet distributions applied in the paper to specify $p(\theta|G)$ [14], 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 not stable cases with potential non-minimality of s, we call these sets in general k-ary Markov Blanket subsets.

[6] 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.

Besides 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. A straightforward way to augment the SNP space is to introduce the level of genes. Many SNPs are related to a given gene, therefore genes can be regarded as aggregations of SNPs. 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)

$$p(MBM(Y, g|D)) = \sum_G p(G|D) \max_s 1(onGene(g, s)) \times 1(MBM(Y, s, G)). \quad (2)$$

5 Interaction, redundancy based on posterior decomposition

Typically, we focus on high-scoring subfeatures, although low probabilities may also indicate important relations, because composite measures representing high-level semantic properties can be constructed, e.g. for redundancy and for interactions. The discovery of interacting features and the redundancy of features have great significance. To construct such a score supporting their discovery, note that the k-MBS posterior can be approximated by the corresponding MBM posteriors as follows.

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

That is the product of the Markov Blanket Membership probabilities of each member variable X_i of k-MBS, as if their occurrence were independent. The exact k-MBS posterior can be calculated by summing up the probabilities of the MBSs containing the examined set $\mathbf{X}'$. This enables a direct Bayesian approach to the concept of redundancy and interaction based on the decomposability of the structure posterior. 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.

Definition 5 (Interaction and redundancy). *The features* $\mathbf{X}' = \{X_{i_1}, \ldots, X_{i_k}\}$ *are 1,k-product independent/redundant/interacting if* $p(\text{k-MBS}(\mathbf{X}', Y, G)|D_N)$ *is equal/less/larger 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.

6 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 that are relevant to the target set. Note, that this is similar to the aggregation of input features in Section 4, but in this case the target variables are "aggregated". Fortunately, the basic concepts of relevance discussed earlier, such as the probabilistic concept of Markov blankets in Definition 1, the concept of relevance in Definition 6, and the graph-theoretic concept of neighbourhood in Theorem 1 can easily be extended to use target sets instead of a single target node.

Definition 6 (Multiple target Relevance). *A feature (stochastic variable) X_i is strongly/weakly relevant to $\boldsymbol{Y}$, if it is strongly 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 2. *If* $\mathrm{MBS}(\mathbf{Y}) = (\bigcup_{Y_i \in \mathbf{Y}} \mathrm{MBS}_p(Y_i)) \backslash \mathbf{Y}$*, 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 filtered appropriately. Note, that the posterior for a given target set $\boldsymbol{Y}$ cannot be calculated from the posteriors of a partition of $\boldsymbol{Y} = \bigcup_i \boldsymbol{Y}_i$, because of the interdependencies. However posteriors corresponding to subsets of the target set can be used for an approximation. In case of MBMs, e.g.

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

Furthermore 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 $Y_i \in \boldsymbol{Y}$.

In case of multiple targets, the posteriors of canonic features (MBM, MBS, and MBG) of target sets can be estimated with standard DAG-based Monte Carlo methods. However, the ordering-based Monte Carlo methods need serious modifications, as they contain special search methods within the estimation to find high-scoring MBS and MBG features [5]. Exceptionally, for the MBM feature, the ordering conditional MBM posterior $p(\mathrm{MBM}(\boldsymbol{Y}, G)| \prec, D_N)$, where $\prec$ denotes ordering, can still be computed in polynomial time by the proper adaptation of the single target formula [8].

7 Conditional and contextual relevance

The fundamental definitions of relevance in Def. 1 and 6 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 extensions for relevance are as follows.

Definition 7 (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 8 (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 $(\boldsymbol{X'} \not\perp\!\!\!\perp \boldsymbol{Y} | \boldsymbol{C'})$ as assumed.*

This definition applies for 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 is able to perform a model-based Bayesian inference about conditional relevance (see Section 10). However, to handle contextual relevances, a Bayesian network representing contextual dependencies is necessary, e.g. using decision trees as local dependency models [6].

8 Posteriors for the predictive power of input features

Since the wrapper approach in practice is based on predictive power and the filter approach is based on some model-based relevance, their relation is an open issue and their joint usage needs caution, just as using filter approaches to support predictive model construction (e.g., see [7] for the bias of the model-based approach on BN classifiers). As a special case, consider the asymmetricity of the approaches: a variable can be identified as relevant by having a high MBM posterior, yet its predictive power can be negligible.

The Bayesian analysis of relevance based on Bayesian networks in general corresponds to the model-based approaches, but it is specialized as much as possible towards the predictive approach by collapsing the structure posterior into a

simpler space of complex structural features representing exactly the predictive aspects (e.g. the MBG feature is a sufficient and necessary feature for prediction under broad conditions [5, 3]). Although the relation of the model-based and predictive approaches is outside the scope of the paper, we shortly summarize a parallel Bayesian view for quantitative, prediction oriented model properties.

Using Bayesian networks as conditional predictors allows the definition of quantitative model properties (features) for a given input-output relation expressing the predictive power of the input features on a given data set. Such features (assuming a binary target variable) are the following: the Misclassification Rate $MR(G, \boldsymbol{\theta}|Y, \boldsymbol{X}, D'_N)$, the Odds Ratio $OR(G, \boldsymbol{\theta}|Y, \boldsymbol{X}, D'_N)$, and the Area Under the (ROC) Curve $AUC(G, \boldsymbol{\theta}|Y, \boldsymbol{X}, D'_N)$ [4]. These random variables are defined by the Bayesian network $(G, \boldsymbol{\theta})$ for a given input-output relation $(\boldsymbol{X}, Y)$ and on a given external data set D'_N (the posterior $p(G, \boldsymbol{\theta}|D_N)$ is typically defined by a different training data set). Note, that by having a fully specified Bayesian network $(G, \boldsymbol{\theta})$ (i.e. the input distribution as well) we can define and use these random variables exclusively based on data D_N.

9 Results

We demonstrate the newly proposed general concepts in the discrete domain using a realistic reference model (G_0, θ_0) containing three clinical variables (*Asthma*, *Allergy*, *Rhinitis*) and forty-six SNPs selected from the asthma susceptibility region of chromosome 11q13 [18]. The structure and parameterization of the reference model was learned from a real data set containing 1117 samples, which was slightly modified to test special cases of relevance. We generated 10,000 complete random samples from this reference model.

To estimate the posteriors we applied both a DAG-based and an ordering-based Markov Chain Monte Carlo(MCMC) method [11, 8, 5]. Because of their correspondence, space constraints, and the more direct applicability of the DAG-MCMC in the proposed extensions, we report results only from this method. The length of the burn-in and MCMC simulation was 10^6 and $5 \cdot 10^6$, the probability of the operators from [11] is uniform. The Cooper-Herskovits (CH) prior was used as parameter prior and the structure prior was uniform prior. The maximum number of parents was 4. The lengths of the burn-in was selected using Geweke's z-score test and the R value of the multiple-chain method [9, 10]. The length of the MCMC simulation was selected to decrease the variances of the MCMC estimates below 10^{-2}.

First, we report results about the effects of syntactic and semantic aggregations. Fig. 1 reports the maximal posteriors for k-MBS compatible with the maximum a posteriori (MAP) MBS for increasing $k = 1, \ldots, 7$ (the MAP MBS contains seven variables). It also shows the posterior of the MAP MBS. Fig. 2 reports the probability of relevance of SNPs at the aggregation level of genes.Fig. 3 indicates the decomposability of the MBS posteriors according to Section 5. Finally, Fig. 4 reports the sequential posteriors that a given SNP is relevant for asthma, allergy and rhinitis, both separately and jointly. It also shows the

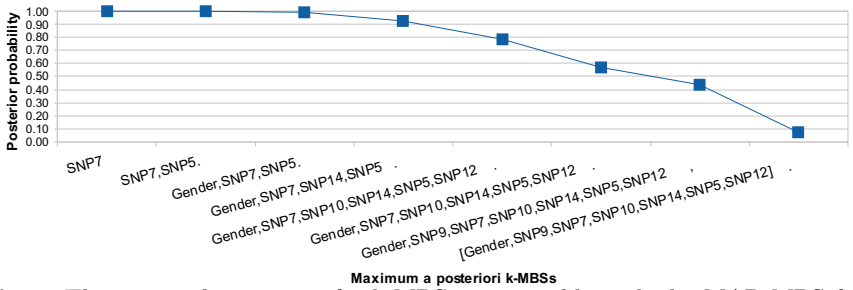

Fig. 1. The maximal posteriors for k-MBS-s compatible with the MAP MBS for increasing $k = 1, \ldots, 7$ (the MAP MBS contains seven variables). The last columns shows the posterior of the MAP MBS given a sample size of 1000.

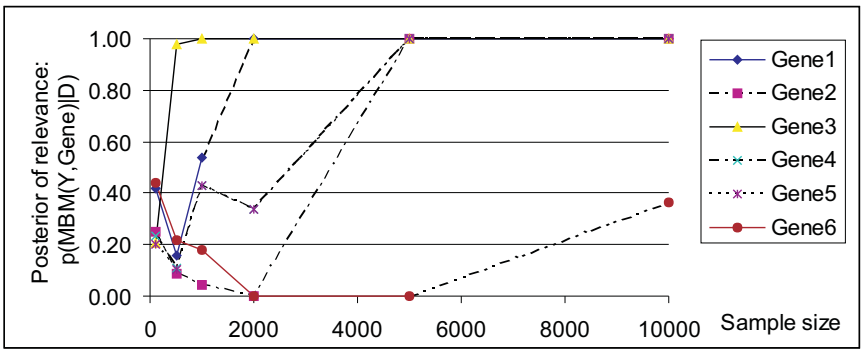

Fig. 2. The sequential posteriors that a given gene contains a SNP relevant for asthma. The probability of relevance is induced by the posterior $p(MBM(SNPi, Asthma|D)$.

approximation of the MBM posterior for the joint target set based on the MBM posteriors for individual targets according to Eq. 4.

10 Discussion

The reference model G_0 for the three target variables contains several interactions and features of conditional relevance, see Fig. 3. The MBS(Asthma, G_0) and MBG(Asthma, G_0) can be correctly identified by the MAP MBS and MBG above 10^4 samples. Consequently, this asymptotic observation holds for the newly introduced k-MBS and k-MBG features, gene level aggregation, and for multiple target variables as well. To illustrate the effect of input aggregation and multiple outputs we used 10^3 samples, which is moderate sample size w.r.t. this set of variables, and typical in practice. 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. 1 indicates that 10^3 samples are enough to ensure a high posterior

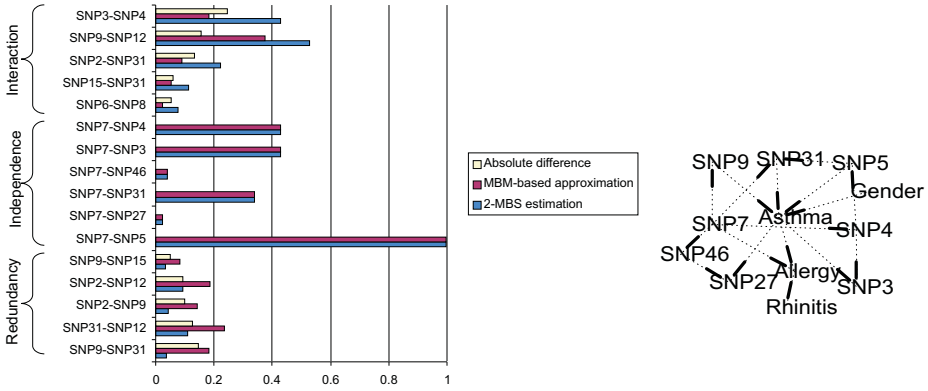

Fig. 3. (Left) The quantification of interaction and redundancy by the decomposability of the posterior. The 2-MBS posteriors, their MBM) based approximation, and their difference are reported for the 5 most interacting, 5 most independent, and 5 most redundant pairs of variables (i.e., when the difference is minimal, closest to zero, and maximal). (Right) The Markov Blanket Graph for the target set *Asthma*, *Allergy*, *Rhinitis* in the reference model G_0.

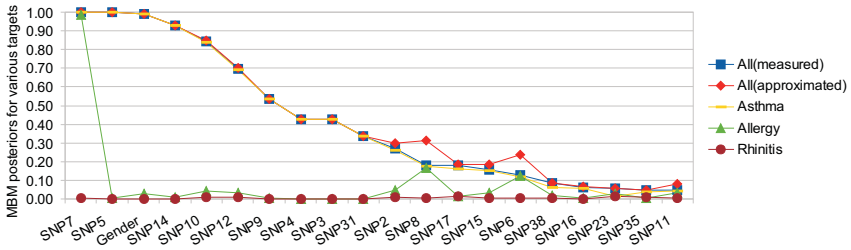

Fig. 4. The posteriors that a given SNP is relevant for asthma, allergy and rhinitis, both separately and jointly. It also shows the approximation of the MBM posterior for the joint target set based on the MBM posteriors for individual targets.

(0.9 <), though only for k-MBS $k < 5$. It is also noteworthy, that the posterior of the MAP MBS (0.0760) is significantly lower than the corresponding 7-MBS defined by its members (0.4374). Despite the restriction in the use of maximal posteriors and only k-MBS-s compatible with the MAP MBS, this result clearly justifies that the proposed k-MBS feature can fulfill its intended role to fill the gap between MBM and MBS features (the posterior for the MBM is 0.9989). Fig. 2 shows a similar but semantic aggregation using a sequential approach, which illustrates the easy transformation and fusion of Bayesian results to support interpretation.

Fig. 3 in general indicates mostly independence, but we report the five-five pairs with maximal difference between their estimated posteriors and MBM based product approximations according to Eq. 3. The reported independencies are compatible with the MBG($Asthma, Allergy, Rhinitis, G_0$) in the reference

model G_0, e.g. *SNP7* is the parent of *Allergy*, whereas the other variables are related to *Asthma*. 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 pair *SNP9,SNP31* with "highest" difference indicating redundancy are potentially redundant, multiple parents of *Asthma*.

Finally Fig. 4 demonstrates the joint use of multiple target variables, although in this case one of the target variables (*Asthma*) nearly determines the posterior MBM 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.

11 Conclusion

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 this paper, motivated by the Bayesian approach and the shortcomings of the Bayesian multilevel analysis of relevance, we introduced and investigated the following concepts and methods.

1. k-MBS, k-MBG the use of new prediction-oriented Bayesian network properties (features) with intermediate, scalable complexity.
2. *Multiple target variables*, which is a distinct problem, e.g. in the Bayesian approach in general the posterior for the target set cannot be reconstructed from the posteriors for the partitions of the target set.
3. *Interaction and redundancy discovery*, based on the decomposability of the structure posterior.

Furthermore, we overviewed open issues in FSS from the aspect of the Bayesian approach and Bayesian networks, such as contextual relevance, and the relation of relevance and predictive measures. Note, that these extensions, e.g. the concepts of k-MBS or k-MBG and multivariate 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), János Bolyai Research Scholarship of the 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. P. Antal. *Integrative Analysis of Data, Literature, and Expert Knowledge*. Ph.D. dissertation, K.U.Leuven, D/2007/7515/99, 2007.
4. 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.
5. 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.
6. C. Boutilier, N. Friedman, M. Goldszmidt, and D. Koller. Context-specific independence in bayesian networks, 1996.
7. N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian networks classifiers. *Machine Learning*, 29:131–163, 1997.
8. N. Friedman and D. Koller. Being Bayesian about network structure. *Machine Learning*, 50:95–125, 2003.
9. D. Gamerman. *Markov Chain Monte Carlo*. Chapman & Hall, London, 1997.
10. A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall, London, 1995.
11. P. Giudici and R. Castelo. Improving Markov Chain Monte Carlo model search for data mining. *Machine Learning*, 50:127–158, 2003.
12. R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97:273–324, 1997.
13. D. Koller and M. Sahami. Toward optimal feature selection. In *International Conference on Machine Learning*, pages 284–292, 1996.
14. 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.
15. J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Francisco, CA, 1988.
16. J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.
17. 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.
18. 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.
19. 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.
20. 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.
21. Lei Yu and Huan Liu. Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research*, 5:1205–1224, 2004.