# Learning Complex Bayesian Network Features for Classification

Péter Antal, András Gézsi, Gábor Hullám and András Millinghoffer
Department of Measurement and Information Systems
Budapest University of Technology and Economics

## Abstract

The increasing complexity of the models, the abundant electronic literature and the relative scarcity of the data make it necessary to use the Bayesian approach to complex queries based on prior knowledge and structural models. In the paper we discuss the probabilistic semantics of such statements, the computational challenges and possible solutions of Bayesian inference over complex Bayesian network features, particularly over features relevant in the conditional analysis. We introduce a special feature called Markov Blanket Graph. Next we present an application of the ordering-based Monte Carlo method over Markov Blanket Graphs and Markov Blanket sets.

In the Bayesian approach to a structural feature $F$ with values $F(G) \in \{f_i\}_{i=1}^{R}$ we are interested in the feature posterior induced by the model posterior given the observations $D_N$, where $G$ denotes the structure of the Bayesian network (BN)

$$p(f_i|D_N) = \sum_G 1(F(G) = f_i)p(G|D_N) \qquad (1)$$

The importance of such inference results from (1) the frequently impractically high sample and computational complexity of the complete model, (2) a subsequent Bayesian decision-theoretic phase, (3) the availability of stochastic methods for estimating such posteriors, and (4) the focusedness of the data and the prior on certain aspects (e.g. by pattern of missing values or better understood parts of the model). Correspondingly, there is a general expectation that for small amount of data some properties of complex models can be inferred with high confidence and relatively low computation cost preserving a model-based foundation.

The irregularity of the posterior over the discrete model space of the Directed Acyclic Graphs (DAGs) poses serious challenges when such feature posteriors are to be estimated. This induced the research on the application of Markov Chain Monte Carlo (MCMC) methods for elementary features (Madigan et al., 1996; Friedman and Koller, 2000). This paper extends these results by investigating Bayesian inference about BN features with high-cardinality, relevant in classification. In Section 1 we present a unified view of BN features enriched with free-text annotations as a probabilistic knowledge base (pKB) and discuss the corresponding probabilistic semantics. In Section 2 we overview the earlier approaches to feature learning. In Section 3 we discuss structural BN features and introduce a special feature called Markov Blanket Graph or Mechanism Boundary Graph. Section 4 discusses its relevance in conditional modeling. In Section 5 we report an algorithm using ordering-based MCMC methods to perform inference over Markov Blanket Graphs and Markov Blanket sets. Section 6 presents results for the ovarian tumor domain.

## 1 BN features in pKBs

Probabilistic and causal interpretations of BN ensure that structural features can express a wide range of relevant concepts based on conditional independence statements and causal assertions (Pearl, 1988; Pearl, 2000; Spirtes et al., 2001). To enrich this approach with subjective domain knowledge via free-text annotations, we introduce the concept of Probabilistic Annotated Bayesian Network knowledge base.

**Definition 1.** A Probabilistic Annotated Bayesian Network knowledge base $K$ for a fixed set $\boldsymbol{V}$ of discrete random variables is a first-order logical knowledge base including standard graph, string and BN related predicates, relations and functions. Let $G^w$ represent a target DAG structure including all the target random variables. It includes free-text descriptions for the subgraphs and for their subsets. We assume that the models $M$ of the knowledge base vary only in $G^w$ (i.e. there is a mapping $G \leftrightarrow M$) and a distribution $p(G^w|\xi)$ is available.

A sentence $\alpha$ is any well-formed first-order formula in $K$, the probability of which is defined as the expectation of its truth

$$E_{p(\mathcal{M}|K)}[\alpha^{\mathcal{M}}] = \sum_{G} \alpha^{\mathcal{M}(G)} p(G|K).$$

where $\alpha^{\mathcal{M}(G)}$ denotes its truth-value in the model $\mathcal{M}(G)$. This hybrid approach defines a distribution over models by combining a logical knowledge base with a probabilistic model. The logical knowledge base describes the certain knowledge in the domain defining a set of models (legal worlds) and the probabilistic part ($p(G^w|\xi)$) expresses the uncertain knowledge over these worlds.

## 2 Earlier works

To avoid the statistical and computational burden of identifying complete models, related local algorithms for identifying causal relations were reported in (Silverstein et al., 2000) and in (Glymour and Cooper, 1999; Mani and Cooper, 2001). The majority of feature learning algorithms targets the learning of relevant variables for the conditional modeling of a central variable, i.e. they target the so-called *feature subset selection* (FSS) problem (Kohavi and John, 1997). Such examples are the Markov Blanket Approximating Algorithm (Koller and Sahami, 1996) and the Incremential Association Markov Blanket algorithm (Tsamardinos and Aliferis, 2003). The subgraphs of a BN as features were targeted in (Pe'er et al., 2001). The bootstrap approach inducing confidence measures for features such as compelled edges, Markov blanket membership and pairwise precedence was investigated in (Friedman et al., 1999).

On the contrary, the Bayesian framework offers many advantages such as the normative, model-based combination of prior and data allowing unconstrained application in the small sample region. Furthermore the feature posteriors can be embedded in a probabilistic knowledge base and they can be used to induce priors for other model spaces and for a subsequent learning. In (Buntine, 1991) proposed the concept of a posterior knowledge base conditioned on a given ordering for the analysis of BN models. In (Cooper and Herskovits, 1992) discussed the general use of the posterior for BN structures to compute the posterior of arbitrary features. In (Madigan et al., 1996) proposed an MCMC scheme over the space of DAGs and orderings of the variables to approximate Bayesian inference. In (Heckermann et al., 1997) considered the application of the full Bayesian approach to causal BNs. Another MCMC scheme, the ordering-based MCMC method utilizing the ordering of the variables were reported in (Friedman and Koller, 2000). They developed and used a closed form for the order conditional posteriors of Markov blanket membership, beside the earlier closed form of the parental sets.

## 3 BN features

The prevailing interpretation of BN feature learning assumes that the feature set is significantly simpler than the complete domain model providing an overall characterization as marginals and that the number of features and their values is tractable (e.g linear or quadratic in the number of variables). Another interpretation is to identify high-scoring arbitrary subgraphs or parental sets, Markov blanket subsets and estimate their posteriors. A set of simple features means a fragmentary representation for the distribution over the complete domain model from multiple, though simplified aspects, whereas using a given complex feature means a focused representation from a single, but complex point of view. A feature $F$ is complex if the number of its values is exponential in the number of domain variables. First we cite a central concept and a theorem about relevance for variables $V = \{X_1, \ldots, X_n\}$ (Pearl, 1988).

**Definition 2.** A set of variables $MB(X_i)$ is called the Markov Blanket of $X_i$ w.r.t. the distribution $P(V)$, if $(X_i \perp\!\!\!\perp V \setminus MB(X_i)|MB(X_i))$. A minimal Markov blanket is called a Markov boundary.

**Theorem 1.** *If a distribution $P(V)$ factorizes w.r.t DAG $G$, then*

$$\forall\, i = 1, \ldots, n : (X_i \perp\!\!\!\perp V \setminus bd(X_i)|bd(X_i, G))_P,$$

*where $bd(X_i, G)$ denotes the set of parents, children and the children's other parents for $X_i$.*

So the set $bd(X_i, G)$ is a Markov blanket of $X_i$. So we will also refer to $bd(X_i, G)$ as a Markov blanket of $X_i$ in $G$ using the notation $MB(X_i, G)$ implicitly assuming that $P$ factorizes w.r.t. $G$.

The induced, symmetric pairwise relation is the Markov Blanket Membership $MBM(X_i, X_j, G)$ w.r.t. $G$ between $X_i$ and $X_j$ (Friedman et al., 1999)

$$MBM(X_i, X_j, G) \leftrightarrow X_j \in bd(X_i, G) \qquad (2)$$

Finally, we define the Markov Blanket Graph.

**Definition 3.** A subgraph of $G$ is called the Markov Blanket Graph or Mechanism Boundary Graph $MBG(X_i, G)$ of a variable $X_i$ if it includes the nodes from $MB(X_i, G)$ and the incoming edges into $X_i$ and into its children.

It is easy to show, that the characteristic property of the MBG feature is that it completely defines the distribution $P(Y|V \setminus Y)$ by the local dependency models of $Y$ and its children in a BN model $G$, in case of point parameters $(G, \boldsymbol{\theta})$ and of parameter priors satisfying global parameter independence (Spiegelhalter and Lauritzen, 1990) and parameter modularity (Heckerman et al., 1995). This property offers two interpretations for the MBG feature. From a probabilistic point of view the $MBG(G)$ feature defines an equivalence relation over the DAGs w.r.t. $P(Y|V \setminus Y)$, but clearly the MBG feature is not a unique representative of a conditionally equivalent class of BNs. From a causal point of view, this feature uniquely represents the minimal set of mechanisms including $Y$. In short, under the conditions mentioned above, this structural feature of the causal BN domain model is necessary and sufficient to support the manual exploration and automated construction of a conditional dependency model.

There is no closed formula for the posterior $p(MBG(Y, G))$, which excludes the direct use of the MBG space in optimization or in Monte Carlo methods. However, there exist a formula for the order conditional posterior with polynomial time complexity if the size of the parental sets is bounded by $k$

$$\begin{aligned} & p(MBG(Y,G) = mbg|D_N) = \\ & \quad p(pa(Y, mbg)|D_N) \\ & \quad \prod_{\substack{Y \prec X_i \\ Y \in pa(X_i, mbg)}} p(pa(X_i, mbg)|D_N) \\ & \quad \prod_{\substack{Y \prec X_i \\ Y \notin pa(X_i, mbg)}} \sum_{Y \notin pa(X_i)} p(pa(X_i)|D_N). \end{aligned} \qquad (3)$$

The cardinality of the $MBG(Y)$ space is still super-exponential (even if the number of parents is bounded by $k$). Consider an ordering of the variables such that $Y$ is the first and all the other variables are children of it, then the parental sets can be selected independently, so the number of alternatives is in the order of $(n-1)^{n^2}$ (or $(n-1)^{(k-1)(n-1)}$). However, if $Y$ is the last in the ordering, then the number of alternatives is of the order $2^{n-1}$ or $(n-1)^{(k)}$). In case of $MBG(Y, G)$, variable $X_i$ can be (1) non-occurring in the MBG, (2) a parent of $Y$ ($X_i \in pa(Y, G)$), (3) a child of $Y$ ($X_i \in ch(Y, G)$) and (4) (pure) other parent $((X_i \notin pa(Y, G) \wedge (X_i \in pa(ch(Y)_j))))$. These types correspond to the irrelevant (1) and strongly relevant (2,3,4) categories (see, Def. 4). The number of DAG models $G(n)$ compatible with a given MBG and ordering $\prec$ can be computed as follows: the contribution of the variables $X_i \prec Y$ without any constraint and the contribution of the variables $Y \prec X_i$ that are not children of Y, which is still $2^{\mathcal{O}((k)(n)log(n))}$ (note certain sparse graphs are compatible with many orderings).

## 4 Features in conditional modeling

In the conditional Bayesian approach the relevance of predictor variables (features in this context) can be defined in an asymptotic, algorithm-, model- and loss-free way as follows

**Definition 4.** A feature $X_i$ is strongly relevant iff there exists some $x_i, y$ and $s_i = x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$ for which $p(x_i, s_i) > 0$ such that $p(y|x_i, s_i) \neq p(y|s_i)$. A feature $X_i$ is weakly relevant iff it is not strongly relevant, and there exists a subset of features $S'_i$ of $S_i$ for which there exists some $x_i, y$ and $s'_i$ for which $p(x_i, s'_i) > 0$ such that $p(y|x_i, s'_i) \neq p(y|s'_i)$. A feature is relevant if it is either weakly or strongly relevant; otherwise it is irrelevant (Kohavi and John, 1997).

In the so-called filter approach to feature selection we have to select a minimal subset $X'$ which fully determines the conditional distribution of the target ($p(Y|X) = p(Y|X')$). If the conditional modeling is not applicable and a domain model-based approach is necessary then the Markov boundary property (feature) seems to be an ideal candidate for identifying relevance. The following theorem gives a sufficient condition for uniqueness and minimality (Tsamardinos and Aliferis, 2003).

**Theorem 2.** *If the distribution $P$ is stable w.r.t. the DAG $G$, then the variables $bd(Y,G)$ form a unique and minimal Markov blanket of $Y$, $MB(Y)$. Furthermore, $X_i \in MB(Y)$ iff $X_i$ is strongly relevant.*

However, the MBG feature provides a more detailed description about relevancies. As an example, consider that a logistic regression (LR) model without interaction terms and a Naive BN model can be made conditionally equivalent using a local and transparent parameter transformation. If the distribution contains additional dependencies, then the induced conditional distribution has to be represented by a LR model with interaction terms.

## 5 Estimating complex features

The basic task is the estimation of the expectation of a given random variable over the space of DAGs with a specified confidence level in Eq. 1. We assume complete data, discrete domain variables, multinomial local conditional distributions and Dirichlet parameter priors. It ensures efficiently computable closed formulas for the (unnormalized) posteriors of DAGs. As this posterior cannot be sampled directly and the construction of an approximating distribution is frequently not feasible, the standard approach is to use MCMC methods such as the Metropolis-Hastings over the DAG space (see e.g. (Gamerman, 1997; Gelman et al., 1995)).

The DAG-based MCMC method for estimating a given expectation is generally applicable, but for certain types of features such as Markov blanket membership an improved method, the so-called ordering-based MCMC method can be applied, which utilizes closed-forms of the order conditional feature posteriors computable in $\mathcal{O}(n^{k+1})$ time, where $k$ denotes the maximum number of parents (Friedman and Koller, 2000).

In these approaches the problem is simplified to the estimation of separate posteriors. However, the number of target features can be as high as $10^4 - 10^6$ even for a given type of pairwise features and moderate domain complexity. This calls for a decision-theoretic report of selection and estimation of the features, but here we use a simplified approach targeting the selection-estimation of the $K$ most probable feature values. Because of the exponential number of feature values a search method has to be applied either iteratively or in an integrated fashion. The first approach requires the offline storage of orderings and corresponding common factors, so we investigated this latter option. The integrated feature selection-estimation is particularly relevant for the ordering-based MC methods, because it does not generate implicitly "high-scoring" features and features that are not part of the solution cause extra computational costs in estimation.

The goal of search within the MC cycle at step $l$ is the generation of MBGs with high order conditional posterior, potentially using the already generated MBGs and the posteriors $p(MBG| \prec_l, D_N)$. To facilitate the search we define an order conditional MBG state space based on the observation that the order conditionally MAP MBG can be found in $\mathcal{O}(n^{k+1})$ time with a negligible constant increase only. An MBG state is represented by an $n'$ dimensional vector $\underline{s}$, where $n'$ is the number of variables not preceding the target variable $Y$ in the ordering $\prec$:

$$n' = \sum_{i=1}^{n} 1(Y \preceq X_i) \tag{4}$$

The range of the values are integers $s_i = 0, \ldots, r_i$ representing either separate parental sets or (in case of $X_i$ where $Y \prec_l X_i$) a special set of parental sets not including the target variable. The product of the order conditional posteriors of the represented sets of parental sets gives the order conditional posterior of the represented MBG state in Eq. 3. We ensure that the conditional posteriors of the represented sets of parental sets are monotone decreasing w.r.t. their indices:

$$\forall s_i < s_i' : p(s_i|D_N, \prec) \geq p(s_i'|D_N, \prec) \tag{5}$$

which can be constructed in $\mathcal{O}(n^{k+1} \log(\max_i r_i))$ time, where $k$ is the maximum parental set size.

This MBG space allows the application of efficient search methods. We experimented with a direct sampling, top-sampling and a deterministic search. The direct sampling was used as a baseline, because it does need the MBG space. The top-sampling method is biased towards sampling MBGs with high order conditional posterior, by sampling only from the L most probable sets of parental sets for each $Y \precsim X_i$. The uniform-cost search constructs the MBG space, then performs a uniform-cost search to a maximum number of MBGs or to threshold $p(MBG^{MAP,\prec}| \prec, D_N)/\rho^S$.

The pseudo code of searching and estimating MAP values for the MBG feature of a given variable is shown in Alg. 1 (for simplicity the estimation of simple classification features such as edge and MBM relations, and the estimation of the MB features of a given variable using the estimated MAP MBG collection is not shown).

**Algorithm 1** Search and estimation of classification features using the MBG-ordering spaces

```
Require: $p(\prec)$,p($pa(X_i)| \prec$),k,R,$\rho$,$L^S$,$\rho^S$,$L^T$,M;
Ensure: K MAP feature value with estimates
  Cache order-free parental posteriors $\Pi = \{\forall\, i, |pa(X_i)| \leq k : p(pa(X_i)|D_N)\}$
  Initialize MCMC, the MBG-tree $\mathcal{T}$, MBM and edge posterior matrices $\mathcal{R}$, $\mathcal{E}$;
  Insert a priori specified MBGs in $\mathcal{T}$;
  for $l = 0$ to $M$ do {the sampling cycle}
    Draw next ordering;
    Cache order specific common factors $\Psi$ for $|pa(X_i)| \leq k$:
    $p(pa(X_i)| \prec_l)$ for all $X_i$
    $p(Y \notin pa(X_i)| \prec_l)$ for $Y \prec_l X_i$;
    Compute $p(\prec_l |D_N)$;
    Construct order conditional MBG-Subspace($\Pi, \Psi, R, \rho$)=$\Phi$
    $S^S$=Search($\Phi$,$L^S$, $\rho^S$);
    for all $mbg \in S^S$ do
      if $mbg \notin \mathcal{T}$ then
        Insert($\mathcal{T}, mbg$)
    if $L^T < |\mathcal{T}|$ then
      $\mathcal{T}$=PruneToHPD($\mathcal{T}$,$L^T$);
    for all $mbg \in \mathcal{T}$ do
      $\hat{p}(mbg|D_N)+ = p(mbg| \prec_l, D_N)$;
  Report K MAP MBGs from $\mathcal{T}$;
  Report K' MAP MBs using the MBGs in $\mathcal{T}$;
```

Parameters $R, \rho$ allow the restriction of the MBG subspace separately for each dimension to a less than $R$ values by requiring that the corresponding posteriors are above the $\exp(-\rho)$ ratio of the respective MAP value. The uniform-cost search starts from the order conditional MAP MBG, and stops after the expansion of $L^S$ number of states or if the most probable MBG in its search list drops below $\exp(\rho^S)$ ratio of the order conditional posterior of the starting MBG.

Generally, the expansion phase has high computational costs, but for large $L^T$ the update of the MBGs in $\mathcal{T}$ is high as well. In order to maintain tractability the usage of more refined methods such as partial updating are required. Within the explored OC domain however the full, exact update has acceptable costs if the size of the estimated MBG set is $L^T \in [10^5, 10^6]$. This $L^T$ ensures that the newly inserted MBGs are not pruned before their estimates can reliable indicate their high-scoring potential and still allows an exact update. In larger domains this balance can be different and the analysis of this question in general is for future research.

The analysis of the MBG space showed that the conditional posteriors of the ranked parental sets after rank 10 are negligible, so subsequently we will report results using values $R = 20, \rho = 4$ and $L^S = 10^4, \rho^S = 10^{-6}$. Note that the expansion with the $L^S$ conditionally most probable MBGs in each step does not guarantee that the $L^S$ most probable MBGs are estimated, not even the MAP MBG.

## 6 Results

We used a data set consisting of 35 discrete variables and 782 complete cases related to the preoperative diagnosis of ovarian cancer (see (Antal et al., 2004)).

First we report the estimation-selection of MB features for the central variable *Pathology*. We applied the heuristic deterministic search-estimation method in the inner cycle of the MCMC method. The length of the burn-in and MCMC simulation was 10000, the probability of the pairwise replace operator was 0.8, the parameter prior was the $BD_{eu}$ and the structure prior was uniform prior for the parental set sizes (Friedman and Koller, 2000). The maximum number of parents was 4 (the posteriors of larger sets are insignificant). For preselected high-scoring MB values after 10000 burn-in the single-chain convergence test from Geweke comparing averages has z-score approximately 0.5, the R value of the multiple-chain method of Gelman-Rubin with 5 chains drops below 1.05 (Gamerman, 1997; Gelman et al., 1995). The variances of the MCMC estimates of these preselected test feature values drop below $10^{-2}$. We also applied the deterministic search-estimation method for a single ordering, because a total ordering of the variables was available from an expert. Fig. 1 reports the estimated posteriors of the MAP MB sets for *Pathology* with their MBM-based approximated values assuming the independence of the MBM values and Table 1 shows the members of the MB sets. Note that the two monotone decreasing curves correspond to independent rankings, one for the expert's total ordering and one for the unconstrained case. It also reports the MB set spanned by a prior BN specified by the expert ($E$), the MB set spanned by the MAP BN ($BN$) and the set spanned by the MAP MBG ($MBG$) (see Eq. 6). Furthermore we generated another reference set ($LR$) from a conditional standpoint using the logistic regression model class and the SPSS 14.0 software with the default setting for forward model construction (Hosmer and Lemeshow, 2000). $MB_p$ reports the result of the deterministic select-estimate method using the total ordering of the expert and the $MB_1, MB_2, MB_3$ report the result of the unconstrained ordering-based MCMC with deterministic select-estimate method. Variables $FamHist$,$CycleDay$, $HormTherapy$, $Hysterectomy$, $Parity$ PMenoAge are never selected and the variables $Volume$, $Ascites$, $Papillation$, $PapFlow$, $CA125$, $WallRegularity$ are always selected, so they are not reported.

The MBM-based approximation performs relatively well, particularly w.r.t. ranking in the case of the expert's ordering $\prec_0$, but it performs poorly in the unconstrained case both w.r.t. estimations and ranks (see the difference of $M_p$ set to $M_1$ w.r.t. variables $Age$, $Meno$, $PI$, $TAMX$, $Solid$).

We compared the $MBG(Y, G^{MAP})$ and $MB(Y, G^{MAP})$ feature values defined by the MAP BN structure $G^{MAP}$ against the MAP MBG feature value $MBG(Y)^{MAP}$ and the MAP MB feature value $MB(Y)^{MAP}$ including the MB feature value defined by the MAP MBG feature value $MB(Y, MBG(Y)^{MAP})$

Table 1: Markov blanket sets of Pathology among the thirty-five variables.

| | E | LR | $BN$ | $MG$ | $MB_p$ | $MB_1$ | $MB_2$ | $MB_3$ |
|---|---|---|---|---|---|---|---|---|
| Age | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| Meno | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 1 |
| PMenoY | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| PillUse | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Bilateral | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| Pain | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| Fluid | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| Septum | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| ISeptum | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| PSmooth | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Loc. | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 |
| Shadows | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| Echog. | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| ColScore | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| PI | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| RI | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| PSV | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| TAMX | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 |
| Solid | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 |
| FHBrCa | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| FHOvCa | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |

$$\begin{aligned} G^{MAP} &= \arg\max_{G} p(G|D_N) \qquad (6)\\ MBG(Y)^{MAP} &= \arg\max_{mbg(Y)} p(mbg(Y)|D_N)\\ MB(Y)^{MAP} &= \arg\max_{mb(Y)} p(mb(Y)|D_N) \end{aligned}$$

We performed the comparison using the best BN structure found in the MCMC simulation. The MAP MBG feature value $MBG(Y)^{MAP}$ differed significantly from the MAP domain model, because of an additional $Age$ and $Fluid$ variables in the domain model. The MAP MB feature value $MB(Y)^{MAP}$ similarly differs from the MB sets defined by the MAP domain models for example w.r.t. the vascularization variables such as $PI$. Interestingly, the MAP MB feature value also differs from the MB feature value defined by the MAP MBG feature value $MB(Y, MBG(Y)^{MAP})$, for example w.r.t. $TAMX$, $Solid$ variables. In conclusion these results together with the comparison against the simple feature-based analysis such as MBM-based analysis reported in Fig. 1, show the relevance of the complex feature-based analysis.

We also constructed an offline probabilistic knowledge base containing $10^4$ MAP MBGs. It is connected with the annotated BN knowledge base defined in Def. 1, which allows an offline exploration of the domain from the point of view of conditional modeling. The histogram of the number of parameters and inputs for the MBGs using only the fourteen most relevant variables are reported in Fig. 2.

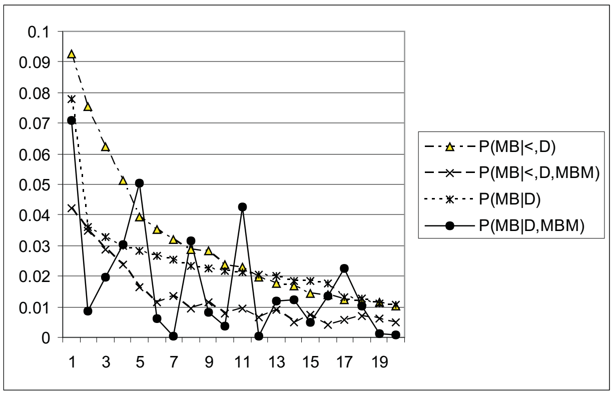

Figure 1: The ranked posteriors and their MBM-based approximations of the 20 most probable $MB(Pathology)$ sets for the single/unconstrained orderings.

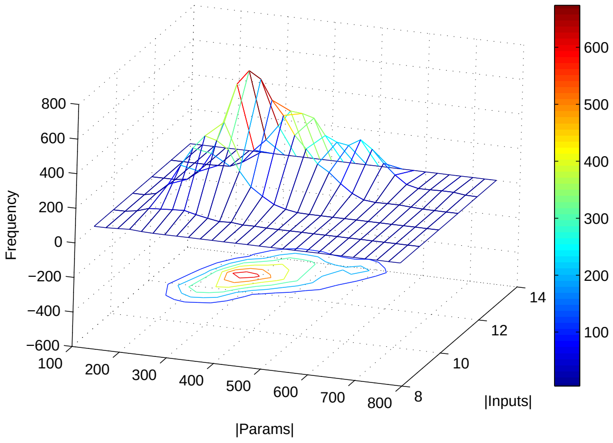

Figure 2: The histogram of the number of parameters and inputs of the MAP MBGs.

## 7 Conclusion

In the paper we presented a Bayesian approach for complex BN features, for the so-called Markov Blanket set and the Markov Blanket

Graph features. We developed and applied a select-estimate algorithm using ordering-based MCMC, which uses the efficiently computable order conditional posterior of the MBG feature and the proposed MBG space. The comparison of the most probable MB and MBG feature values with simple feature based approximations and with complete domain modeling showed the separate significance of the analysis based on these complex BN features in the investigated medical domain. The proposed algorithm and the offline knowledge base in the introduced probabilistic annotated BN knowledge base context allows new types of analysis and fusion of expertise, data and literature.

## Acknowledgements

We thank T. Dobrowiecki for his helpful comments. Our research was supported by grants from Research Council KUL: IDO (IOTA Oncology, Genetic networks).

## References

P. Antal, G. Fannes, Y. Moreau, D. Timmerman, and B. De Moor. 2004. Using literature and data to learn Bayesian networks as clinical models of ovarian tumors. *AI in Med.*, 30:257–281. Special issue on Bayesian Models in Med.

W. L. Buntine. 1991. Theory refinement of Bayesian networks. In *Proc. of the 7th Conf. on Uncertainty in Artificial Intelligence* , pages 52–60. Morgan Kaufmann.

G. F. Cooper and E. Herskovits. 1992. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347.

N. Friedman and D. Koller. 2000. Being Bayesian about network structure. In *Proc. of the 16th Conf. on Uncertainty in Artificial Intelligence*, pages 201–211. Morgan Kaufmann.

N. Friedman, M. Goldszmidt, and A. Wyner. 1999. Data analysis with bayesian networks: A Bootstrap approach. In *Proc. of the 15th Conf. on Uncertainty in Artificial Intelligence*, pages 196–205. Morgan Kaufmann.

D. Gamerman. 1997. *Markov Chain Monte Carlo.* Chapman & Hall, London.

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. 1995. *Bayesian Data Analysis.* Chapman & Hall, London.

C. Glymour and G. F. Cooper. 1999. *Computation, Causation, and Discovery.* AAAI Press.

D. Heckerman, D. Geiger, and D. Chickering. 1995. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20:197–243.

D. Heckermann, C. Meek, and G. Cooper. 1997. A bayesian aproach to causal discovery. *Technical Report*, MSR-TR-97-05.

D. W. Hosmer and S. Lemeshow. 2000. *Applied Logistic Regression.* Wiley & Sons, Chichester.

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

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

D. Madigan, S. A. Andersson, M. Perlman, and C. T. Volinsky. 1996. Bayesian model averaging and model selection for markov equivalence classes of acyclic digraphs. *Comm.Statist. Theory Methods*, 25:2493–2520.

S. Mani and G. F. Cooper. 2001. A simulation study of three related causal data mining algorithms. In *International Workshop on Artificial Intelligence and Statistics*, pages 73–80. Morgan Kaufmann, San Francisco, CA.

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

J. Pearl. 2000. *Causality: Models, Reasoning, and Inference.* Cambridge University Press.

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

C. Silverstein, S. Brin, R. Motwani, and J. D. Ullman. 2000. Scalable techniques for mining causal structures. *Data Mining and Knowledge Discovery*, 4(2/3):163–192.

D. J. Spiegelhalter and S. L. Lauritzen. 1990. Sequential updating of conditional probabilities on directed acyclic graphical structures. *Networks*, 20(.):579–605.

P. Spirtes, C. Glymour, and R. Scheines. 2001. *Causation, Prediction, and Search.* MIT Press.

I. Tsamardinos and C. Aliferis. 2003. Towards principled feature selection: Relevancy,filters, and wrappers. In *Proc. of the Artificial Intelligence and Statistics*, pages 334–342.