periodica polytechnica

Electrical Engineering and Computer Science
57/2 (2013) 35–48
doi: 10.3311/PPee.2088
http://periodicapolytechnica.org/ee

RESEARCH ARTICLE

The effect of parameter priors on Bayesian relevance and effect size measures

Gábor Hullám / *Péter* Antal

Received 2013-04-08, accepted 2013-05-07

Abstract

The application of Bayesian network based methods is increasingly popular in several research fields where the investigation of complex dependency patterns are of central importance. Bayesian networks provide a rich, graph-based language for the refined characterization of relevance types, and has a built-in mechanism for the correction of multiple testing. In the paper we discuss two main topics: the effects of priors and the applicability of Bayesian structure based odds ratio. The selection of an adequate prior is generally required by Bayesian methods and yet there is no general method for prior selection in the multivariate case. Here we analyze the effects of different priors and propose a method for prior selection based on expected effect size. In the second part of the paper we investigate structural and parametric aspects of relevance, and demonstrate a hybrid effect size measure that allows an integrated analysis of these aspects.

Keywords

Bayesian statistical framework · Bayesian networks · effect size · relevance measures

Acknowledgement

The work reported in the paper has been developed in the framework of the project "Talent care and cultivation in the scientific workshops of BME" project. This project is supported by the grant TÁMOP - 4.2.2.B-10/1–2010-0009.

Gábor Hullám

Department of Measurement and Information System, Faculty of Electrical Engineering and Informatics, Budapest University of Technology and Economics, Magyar tudósok krt. 2., H-1117 Budapest, Hungary
e-mail: hullam.gabor@mit.bme.hu

Péter Antal

Department of Measurement and Information System, Faculty of Electrical Engineering and Informatics, Budapest University of Techology and Economics, Magyar tudósok krt. 2., H-1117 Budapest, Hungary
e-mail: antal@mit.bme.hu

1 Introduction

Graphical models, especially Bayesian networks, became increasingly popular in several fields such as biomedicine, when the need for modeling potentially complex dependency structures between genomic, environmental, and clinical factors and disease state indicators emerged. Rapidly evolving measurement technologies led to new approaches unraveling the genetic background of multifactorial diseases. Genome-wide association studies (GWAS) for example, created a seemingly data rich environment with data sets of tens of thousands of measured factors, and of relatively large sample size. Using standard frequentist statistical methods however, the interpretation of the results was problematic in many cases, due to the strict thresholds on significance levels, which were used to cope with the multiple testing problem. Clearly, new methods were required to alleviate the problem, and to provide a meaningful analysis. The application of Bayesian network based Bayesian methods was motivated by two of its main properties: the 'built-in correction' for multiple testing, and the multivariate modeling capability of dependency relations.

First, the correction is expressed by the overall flatness of a posteriori probabilities (posteriors). By flatness we mean that even the posteriors of relevant factors (i.e. relevant with respect to a target e.g. a disease state) are moderate (or low), and are relatively closer to the posteriors of less relevant factors. Note, that in case of Bayesian methods, this is the usual symptom of insufficient or moderately sufficient sample size. The advantage of this approach is, that at least some characteristics of promising factors can be identified, whereas according to the principle of standard statistics, all factors with a p-value above a given threshold are discarded. Therefore, Bayesian methods may enhance the process of selecting candidates for further investigation, particularly when highly significant results are not present. On the other hand, one might argue that without a firm threshold, the selection of promising results rests solely in the hands of a subjective expert. The debate on this matter seems neverending, and can only be overcome by openly describing the of criteria of selection.

Second, the capability of modeling multivariate relationships

is essential, since mostly multifactorial diseases (i.e. illnesses with complex genetic background and related environmental factors, e.g. asthma, rheumatoid arthritis) are targeted by gene association studies. Although the assumption of independence of factors is highly improbable, in many cases the univariate approach is acceptable. Especially it is so, when the aim is to identify only the most significant factors that might lead to efficient biomarkers (i.e. indicators of the presence or the severity of a disease), and the discovery of interactions and other features is secondary. The relative simplicity and efficiency of univariate methods compared to more complex, computationally intensive multivariate methods frequently tip the balance in favor of the former. The application of univariate methods is usually justified, if highly significant results are found. However, when only weakly significant factors are present (and subsequently discarded due to a strict significance threshold), then multivariate methods seem much more appealing. Bayesian network based Bayesian methods provide a solution to discover the interactions among factors, and to identify their joint effects. Furthermore, under certain conditions even causal relationships can be identified [18].

Another characteristic of Bayesian methods is a hypothesis free analysis. In case of gene association studies this is an important aspect since most investigations are conducted using genetic models, such as additive, dominant or recessive models as alternative hypotheses. Frequently, the statistics are computed for several models and then the most significant one is selected. A Bayesian solution to this problem would be to average over possible models. In a univariate framework this can be achieved by using a weighted mixture of models as described in [24]. The assessment of weights, which reflect the beliefs of the investigator in certain possible models however, is not straightforward. In contrast, the model averaging in a Bayesian multivariate framework is done automatically.

Previously, we applied a full Bayesian approach, the Bayesian network based Bayesian Multilevel Analysis (BN-BMLA) in a candidate gene association study to analyze the relevance of single nucleotide polymorphisms (SNPs) at a structural level [27]. In this paper, we investigate two important topics: first, the effect of different priors on the overall learning process, and second, the extension of BN-BMLA to facilitate the Bayesian analysis of effect size parameters at a parametric level.

The selection of an appropriate prior (i.e. a suitable type and adequate hyperparameters) for Bayesian methods is a traditionally well known problem, which is relatively overlooked nowadays. There are no general methods for prior selection in the multivariate case, especially not for moderate or small sample size with respect to the number of variables. According to previously reported empirical results, some priors work well in the latter case, while others perform better in an asymptotic case. In this first main section, we overview the effects of different priors, including known analytic biases and anomalies we observed in finite sample cases.

The BN-BMLA method provides posteriors for structural properties at different abstraction levels, such as edges, so called Markov blanket sets (a variable X is part of the Markov blanket of the 'target variable' Y, which means, that the node that represents X is either a parent, a child or another parent of a child of node Y). However, in most real-world applications, particularly in genomics, apart from the information on structural properties, parametric information, such as effect size is also required. Although there are methods for the calculation of effect size for known causal structures, so far it was not used in a Bayesian context with posteriors of structural properties. In this section we demonstrate how a Bayesian version of odds ratio can be computed based on the structural properties.

2 Overview of priors

The basic paradigm of Bayesian methods is that using a prior probability distribution $P(A)$ and a likelihood $P(B|A)$ the posterior probability $P(A|B)$ can be computed according to the Bayes' theorem. In case of Bayesian network structure learning, given a data set D and a directed acyclic graph G that represents the joint probability distribution of discrete random variables $\mathbf{V} = X_1, ..., X_n$ having a multinomial distribution, the aim is to estimate the posterior of

$$p(G|D) = \frac{p(G)p(D|G)}{P(D)}. \quad (1)$$

Neglecting $P(D)$ as a modeling constant results in

$$p(G|D) \propto p(G)p(D|G), \quad (2)$$

that is in order to get the posterior of a certain structure given the data, a prior distribution of the possible structures $p(G)$ and a likelihood $p(D|G)$ is needed. Note, that the typical goal of structure learning is to find the maximum a posteriori (MAP) structure, i.e. the one with the highest posterior. This practical viewpoint somewhat contradicts the Bayesian philosophy of defining distributions instead of thresholds, i.e. a 'truly' Bayesian method would aim to identify the posterior distribution of possible structures instead of only finding the best. For practical reasons however, this is infeasible in most real-world applications.

The first term is responsible for the incorporation of a priori knowledge, i.e. an expert may judge some structures more probable than others. There are two opposing approaches towards this notion: permissive and restrictive. The former states that existing a priori knowledge is valuable and should be used to guide the search. In case of gene association studies this could mean the incorporation of results of previous studies. The latter approach argues the viability of this concept. What if previous experiments had unknown faults? The incorporation of false knowledge might influence the analysis in an improper way and thus hinder present efforts. Therefore, the restrictive approach advocates the use of uniform priors, i.e. that the prior probability should be the same for all possible entities. This way no harm

is done, although nothing is gained from previous expert knowledge. The counterargument is that Bayesian analysis should be data driven and not prior driven, and the effect of a prior should be overridden by the data if sufficient sample size is available. In real cases however, sample size is often close to the limit of sufficiency or even below, thus making the learning process highly sensitive to the selected prior.

The second term is the likelihood score, which measures that given the structure how probable the data is. There are several possible choices for this scoring metric such as Bayesian Dirichlet (BD), information criteria (e.g. Akaike information criterion [2], Bayesian information criterion [19], and minimum description length (MDL) based metrics [7]). Bouckaert defined these as quality measures, and investigated their asymptotic and finite sample based behavior [8]. While no difference was found in the former case, in the case of finite sample size the Bayesian Dirichlet prior allows structures with parental sets as large as $N/2$ to have the highest score, while MDL and information criteria measures limit this to $logN$, where N is the size of the database.

2.1 The Bayesian Dirichlet prior

The most popular prior is Bayesian Dirichlet (BD) prior, since it is a so called conjugate prior for multinomial sampling, and allows the finding of the maximum a posteriori (MAP) structure (i.e. assigns the highest score to the MAP structure) [6, 9, 12]. Being conjugate means that the distribution function of the prior and the posterior belongs to the same family under a given sampling model. This is a key property, because it enables analytical computation. The Bayesian Dirichlet prior takes the following form.

Definition 1. Given $S = \{X_1, X_2, ..., X_n\}$ a set of random discrete variables that take values from the set of possible states $1...r_i$ (e.g. $X_1 = k$ means that X_1 is in state k), let G be a Bayesian network structure that contains only variables from S. Each variable $X_i \in S$ has a set of parents PA_i with q_i possible configurations. Let pa_{ij} denote the jth instantiation of the parents. Then given a data set D let N_{ijk} be the number of cases in D in which $X_i = k$ and $PA_i = pa_{ij}$. Furthermore, let $N_{ij} = \prod_{k=1}^{r_i} N_{ijk}$ and let θ_{ijk} denote a corresponding conditional probability parameter of G for N_{ijk}. In case of a Dirichlet distribution θ_{ijk} can be given as

$$\theta_{ijk} = \frac{N_{ijk} + N'_{ijk}}{N_{ij} + N'_{ij}} \tag{3}$$

where N'_{ijk} is the *virtual sample size* with respect to N_{ijk} [9].

Note that $PA_i = pa_{ij}$, N_{ijk} and N'_{ijk} are the hyperparameters of the Dirichlet distribution. Given these assumptions the Bayesian Dirichlet metric is given as follows:

$$p(D, G) = p(G) \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(N'_{ij})}{\Gamma(N'_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(N'_{ijk} + N_{ijk})}{\Gamma(N'_{ijk})} \tag{4}$$

In addition, if the hyperparameters satisfy the following condition:

$$N'_{ijk} = N' \cdot p(X_i = k, PA_i = pa_{ij}|G') \tag{5}$$

then the metric ensures likelihood equivalence and thus it is called the Bayesian Dirichlet equivalence (BDe) metric [12]. G' is a hypothetical Bayesian network structure that encodes the prior knowledge and N' is the equivalent sample size (ESS), a free parameter determined by the user. A special case of BDe was described by Buntine [9], and termed as "BDeu" by Heckerman et al. [12], in which the same value $N'_{ijk} = \frac{N'}{r_i \cdot q_i}$ is applied for all N'_{ijk} hyperparamaters for a variable. A further variant, the Cooper-Herskovits (CH) prior uses a fixed value of virtual sample size $N'_{ijk} = 1$ for all variables [10].

Several papers were devoted to analyze BDeu and to investigate its properties. Bouckaert observed that when a database is relatively small compared to the number of variables, then the resulting network structures will likely contain a large number of excess arcs [7]. Steck and Jaakkola demonstrated that as the ESS asymptotically went to zero in case of a large sample, the deletion of arcs in a Bayesian network structure was favored by the metric, and in the opposite case when ESS became large, then the addition of arcs was favored resulting in structures with several extra arcs [23]. Silander and Myllymaki investigated the role of ESS, and compared the CH and BIC scores [21]. The results confirmed that increasing the ESS causes the addition of several arcs in the structures when the sample size is large. Subsequently, Silander et al. conducted a series of experiments to find the optimal ESS, and demonstrated that the results were highly sensitive to the selected ESS [20], with the conclusion that averaging out ESS (that is computing the BDeu metric for several ESS values and taking the average of the scores) would be a possible, though computationally expensive solution. Later, Steck showed that the optimal ESS value is approximately independent of sample size and of the number of variables in the domain. It was also demonstrated that if the data implies a skewed distribution or strong dependencies between the variables, then the optimal ESS-value is small [22]. Furthermore, Ueno provided an asymptotic analysis of both the general form of BDeu and the log-BDeu [26]. In the latter case, the score is decomposed into a log-posterior reflecting the non-uniformity of the distribution and a penalty term. The paper investigated the complex behavior of ESS, which participates simultaneously in blocking and adding arcs via the log-posterior and penalty terms respectively. Ueno stated that ESS should be small in case of skewed conditional distributions to prevent the overfitting of BDeu, while in the opposite case of non-skewed distributions a large ESS value is recommended to avoid underfitting.

To provide further insight into the application of BDeu prior and its other variants, such as CH, we conducted experiments with BN-BMLA on an artificial data set that was generated using real-world data. Results are explained in a subsequent section.

2.2 Priors for genetic association studies

Balding investigated the issue of priors for univariate Bayesian methods used in genetic association studies [24]. Two family of priors were investigated: priors with a normal distribution or a mixture of normal distributions, and priors with a normal exponential gamma (NEG) distributions. The priors were defined in terms of effect size (log odds ratio). The normal priors required a hyperparameter, the proportion π of SNPs having a non-zero effect size (e.g. Balding suggested $\pi = 10^{-4}$ that is 1 out of 10,000 [24]). These priors provided acceptable probabilities for SNPs with moderate and small effect size, but were overly strict in case of a large effect size. This was explained by the 'small tail' property of normal distributions. In order to alleviate this problem, [24] applied NEG priors of different shape and scale parameters that have 'fat tails', thus increasing the probability of larger effect size values. The use of NEG priors is a more flexible Bayesian approach, since it avoids the need of a null-hypothesis (i.e. defining π) as it is necessary in the case of normal priors under the hypothesis testing paradigm.

However, these priors follow a univariate approach, and are defined on the abstraction level of log-odds ratios. In contrast, in Bayesian networks usually such priors are used that are related to conditional probabilities in local, multinomial models (e.g. Dirichlet prior). Therefore, we investigated a prior for log-odds derived from Dirichlet priors defined at the level of conditional probabilities, i.e. derived from a lower abstraction level.

Definition 2. Let $X_1, X_2, ..., X_n$ denote discrete variables that encode SNP states 0,1,2 that refer to common homozygote, heterozygote, rare homozygote genotypes respectively. Then $X_i^{(s)}$ denotes SNP X_i in state s. In case of a disease indicator Y, the non-affected and the affected states (control and case) are denoted with $Y^{(0)}$ and $Y^{(1)}$ respectively. An *odds* is defined as

$$O_{X_i^{(s)}} = \frac{p(Y^{(1)}|X_i^{(s)})}{p(Y^{(0)}|X_i^{(s)})} \tag{6}$$

Consequently an *odds ratio* e.g. heterozygous (1) versus common homozygous (0) is given as

$$OR_{X_i^{(1,0)}} = \frac{O_{X_i^{(1)}}}{O_{X_i^{(0)}}} \tag{7}$$

Therefore, a log OR has the following form:

$$\log OR_{X_i^{(1,0)}} = \log O_{X_i^{(1)}} - \log O_{X_i^{(0)}} \tag{8}$$

$$= \log \frac{p(Y^{(1)}|X_i^{(1)})}{p(Y^{(0)}|X_i^{(1)})} - \log \frac{p(Y^{(1)}|X_i^{(0)})}{p(Y^{(0)}|X_i^{(0)})} \tag{9}$$

$$= \log \frac{\nu_{Y^{(1)}|X_i^{(1)}}}{\nu_{Y^{(0)}|X_i^{(1)}}} - \log \frac{\nu_{Y^{(1)}|X_i^{(0)}}}{\nu_{Y^{(0)}|X_i^{(0)}}},$$

where $\nu_{Y^{(i)}|X^{(i)}}$ denotes a specific conditional probability value. In case of a multinomial distribution the parameters defining the distribution correspond directly to the conditional probability values. Thus we need to apply the transformation $t(\nu) = \log \frac{\nu}{1-\nu}$ to the parameters $\nu_{(.)}$ defining the multinomial distribution of a categorical variable W.

The probability density function over these parameters for variable W having k different values is typically defined by a Dirichlet distribution which has the following general form

$$Dir(\nu_1, ..., \nu_{k-1}|\alpha_1, ..., \alpha_k) = \frac{1}{\text{Beta}(\alpha)} \cdot \prod_{i=1}^{k} \nu_i^{\alpha_i - 1}, \tag{10}$$

where ν_i denotes $p(W = w_i)$ (i.e. the probability that W is instantiated with value w_i). Note that in case of a finite data set ν_i is typically estimated by maximum likelihood estimates $\frac{N_i}{N}$, where N_i is the number of observations in which $W = w_i$ and N is the size of the data set. The Beta(α) function can be expressed in terms of the $\Gamma(.)$ function as

$$\text{Beta}(\alpha) = \frac{\prod_{i=1}^{k} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{k}(\alpha_i))}. \tag{11}$$

The transformed function $g(z)$ is as follows

$$g(z) = Dir(t^{-1}(z)) \cdot (t^{-1}(z))'. \tag{12}$$

The inverse function and its derivative is given by

$$t^{-1}(z) = \frac{1}{1 + e^{-z}} \tag{13}$$

$$(t^{-1}(z))' = \frac{e^{-z}}{(1 + e^{-z})^2} \tag{14}$$

The transformed distribution arises in the following form

$$g(z) = \frac{Dir(\frac{1}{1+e^{-z}}|\alpha_i) \cdot e^{-z}}{(1 + e^{-z})^2} \tag{15}$$

In a binary case of 2 hyperparameters (α and β) the transformation results in the following form:

$$g(z, \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \cdot \Gamma(\beta)} \cdot \frac{(\frac{1}{1+e^{-z}})^{\alpha-1} \cdot (1 - \frac{1}{1+e^{-z}})^{\beta-1} \cdot e^{-z}}{(1 + e^{-z})^2} \tag{16}$$

Then applying the uniform (structure) prior assumption (all structures are equally possible), implies that hyperparameters are equal ($\alpha = \beta$).

$$g(z, \alpha) = \frac{\Gamma(2\alpha)}{\Gamma(\alpha)^2} \cdot (\frac{1}{1 + e^{-z}})^{\alpha+1} \cdot (1 - \frac{1}{1 + e^{-z}})^{\alpha-1} \cdot e^{-z} \tag{17}$$

However, this formula is analytically intractable (e.g. estimation of the high probability density region), thus we sampled the distribution for the $\alpha_i = 1, 5$ and 10 case.

The importance of this formula is that it allows the investigation of the effect of the virtual sample size on the probability density function of odds ratios. In other words, it provides means to analyze the effect of a selected prior on the resulting probabilities for odds and odds ratios. Figure 1 indicates that as the virtual sample size increases the probability of a large odds decreases.

Proposition 1. *In a practical approach, this notion can be used to define the prior according to an expected a priori distribution of odds ratios.*

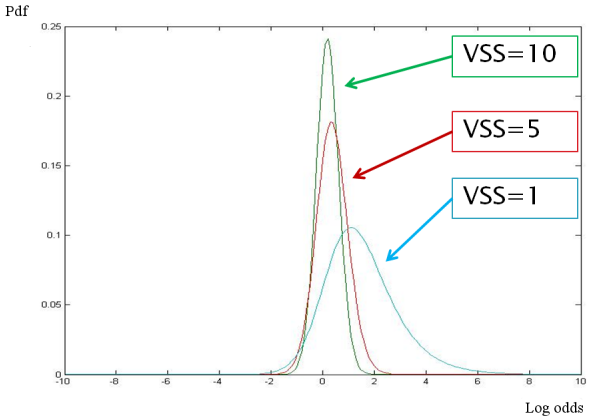


Fig. 1. Probability density function of log odds given various prior settings. VSS denotes virtual sample size, log odds are shown on the horizontal axis, and the probability density function (pdf) is displayed on the vertical axis.

2.3 An empirical study on the effect of priors

We investigated the effect of priors using BN-BMLA on a semi artificial data set containing 115 variables and 10,000 samples. The data set was generated using a model learned from real-world data of a candidate gene association study of asthma [14]. We used *BDeu* and *CH* for priors, since they are popular choices in Bayesian learning. We applied the BN-BMLA method on the data set using different parameter settings, such as varying virtual sample size (VSS), and varying sample size ($100 \ldots 10000$). Data sets of various sizes (100, 300, 500, 800, 1000, 2000, 5000) were created by truncating the original data set of 10,000 samples in order to investigate the small sample size case. In case of *BDeu* prior the equivalent sample size (ESS) α parameter was set to $\alpha = 1, 5, 10, 20, 50$ resulting in a virtual sample size of $N'_{ijk} = \frac{\alpha}{r_i \cdot q_i}$. In case of the *CH* prior the value of the VSS was set directly ($N'_{ijk} = 1, 5, 10, 20, 50$). Every combination of parameters was considered as a setting, and for each setting five parallel runs of BN-BMLA were executed. The results were averaged out (per setting) in order to provide a robust base for the analysis. Although the BN-BMLA method is capable of estimating posteriors for several types of structural features, in this analysis only Markov Blanket Memberships (MBM) were investigated.

Definition 3. A set of variables $\mathbf{X}'$ is called a Markov Blanket Set (MBS) of Y if conditioned on $\mathbf{X}'$ the 'target' variable Y is independent of all other variables. Formally, given the set $\mathbf{V}$ containing all variables $X_1, \ldots, X_n$, a subset of variables $\mathbf{X}' \subset \mathbf{V}$ is a Markov blanket set of Y with respect to the distribution $p(X_1, \ldots, X_n, Y)$ if $(Y \perp\!\!\!\perp \mathbf{V} \setminus \mathbf{X}' | \mathbf{X}')_p$, where $\perp\!\!\!\perp$ denotes conditional independence [17].

Note that the reason for learning MBS of a target variable Y (denoted as MBS_Y) is that given the stability condition and the faithfulness condition [18] the elements of the MBS_Y are strongly relevant with respect to Y (see subsequent sections). In other words, the MBS provides a multivariate characterization indicating the strong relevance of a set of variables.

In order to characterize the individual relevance of a variable, the Markov Blanket Membership (MBM) concept was proposed [11].

Definition 4. The Markov Blanket Membership $\mathrm{MBM}(X_i, Y)$ holds if $X_i \in MBS_Y$.

In practice the identification of the unique minimal Markov blanket is frequently not feasible due to data insufficiency and to the inherent noise in the data. Therefore we developed a Bayesian approach, the BN-BMLA method, which provides posteriors for these relations [5]. The limitation of the data can be seen by the low or moderate maximum a posteriori (MAP) value for MBSs. Even in case when there is only a small number of MBSs (e.g.: 3) that have a relatively high posterior, by investigating only the MAP MBS valuable information contained in the other MBSs with relatively high posteriors would be neglected. Furthermore, in case there are hundreds of MBSs with similarly low posteriors, analyzing only the MAP MBS would lead to improper results. Therefore, in accordance with the Bayesian principle, instead of considering only the MAP MBS, the whole distribution of MBSs should be analyzed.

The Markov blanket set of the model, which the data set was generated from, was identified and was used as a reference. In order to assess the effect of the priors standard statistical measures of performance, such as sensitivity, specificity, accuracy and AUC score, were computed based on the comparison of MBM posteriors and the reference. For the sake of simplicity we treated $p(MBM_Y(X_i)) \geq 0.5$ as a positive result, that is every X_i having a posterior above 0.5 was identified as strongly relevant. This threshold was an appropriate choice for this data set, but in general this threshold has to be chosen based on the distribution of posteriors. For example in case of a relatively small sample size with respect to the number of variables the MAP MBM posterior can be lower than 0.5, and yet indicate that certain variables are relevant.

Performance measures for the basic case of CH prior (VSS=1) and the BDeu prior (ESS=1) are summarized in Table 1 and in Table 2 respectively. Data sets with sample sizes below 1000 resemble the small sample size cases with respect to the number of variables (115). That is the sample size should be at least an order of magnitude higher than the number of variables. Depending on the strength of dependencies between the variables within the data set, that is the skewedness of the underlying distribution [26], the practical limit may be lower. On the other hand, the complete data set with 10,000 samples can be regarded as an asymptotic case. Figure 2 compares the sensitivity, specificity and area under the ROC curve (AUC) respectively for CH and BDeu priors.

For the CH prior, the case of 100 samples is definitely a small sample size scenario. Only the third of the relevant variables are identified correctly and the AUC score (0.71) is relatively low. The case with 300 samples is at the limit of data sufficiency, since approximately half of the relevant variables are correctly

Tab. 1. Performance measures for CH prior (VSS=1) for various sample sizes. AUC refers to the area under the ROC curve.

Sample size	Sensitivity	Specificity	Accuracy	AUC
100	0.33	0.85	0.78	0.71
300	0.53	0.90	0.85	0.92
500	0.67	0.94	0.90	0.89
800	0.60	0.95	0.90	0.91
1000	0.53	0.99	0.93	0.97
2000	0.87	0.99	0.97	0.99
5000	1.00	0.99	0.99	0.99
10000	1.00	0.99	0.99	0.99

identified (0.53) and the number of false positives is relatively high (specificity: 0.9). In a practical case, when the ratio of relevant variables (with respect to all variables) is typically lower than 0.1, this would mean that the majority of the variables identified as relevant would be false positives. The next case, the data set containing 500 samples, could be considered as a moderate sample case. The specificity (0.94) is acceptable and a significant portion (0.67) of the relevant variables is identified correctly.

Tab. 2. Performance measures for BDeu prior (ESS=1) for various sample sizes. AUC refers to the area under the ROC curve.

Sample size	Sensitivity	Specificity	Accuracy	AUC
100	0.07	0.95	0.83	0.59
300	0.13	1.00	0.89	0.58
500	0.27	1.00	0.90	0.65
800	0.40	1.00	0.92	0.77
1000	0.47	1.00	0.93	0.79
2000	0.80	1.00	0.97	0.95
5000	0.87	1.00	0.98	0.99
10000	0.93	1.00	0.99	0.99

For more than 500 samples one would expect a gradual increase in both the sensitivity and specificity measures. Instead, as Table 1 and Figure 2 shows, an interesting anomaly is encountered. Though the specificity increases (from 0.90 to 0.93) as expected, the sensitivity decreases (from 0.67 to 0.53). This is due to a characteristic feature of the data set, in which not all relevant variables in the reference MBS have a direct relationship with the target, i.e. there are variables that are in pure interaction with the target. This means that it is possible, that by excluding false positives some true positives (that form a dependency structure with them) are also excluded temporarily. For 2000 samples this effect vanishes and the sensitivity significantly increases (0.81), and a nearly perfect specificity (0.99) is achieved. The last two cases are almost ideal having AUC scores above 0.99.

In case of the BDeu prior, the results in Table 2 for sample sizes ranging from 100 to 1000 indicate a poor performance in terms of sensitivity, which is also reflected by the AUC scores. Even in case of 800 samples only the 40% of the relevant variables are identified correctly, having an AUC score of 0.77. An

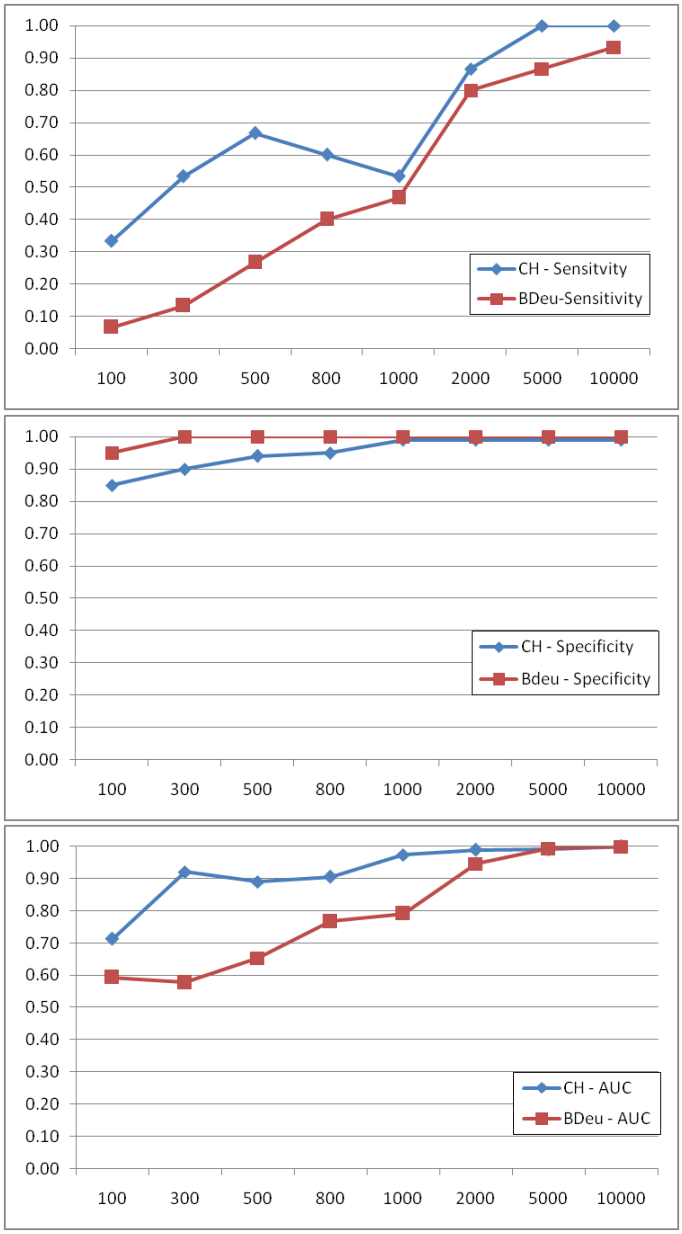


Fig. 2. The comparison of sensitivity (top), specificity (middle) and AUC measures (bottom) in case of CH and BDeu priors for different sample sizes.

acceptable sensitivity is achieved only at the sample size of 2000 (0.80), which increases as high as 0.93 at 10,000 samples, thus it does not reach the ideal performance. In contrast, an ideal specificity is reached above 300 samples, which means that the result is free of false positives.

Comparing the performance of the two priors (see Figure 2) BDeu performs poorly in the small sample region, when the sample size is not an order of magnitude higher than the number of variables. In terms of sensitivity, CH definitely performs better than BDeu both in the small sample case, and interestingly it also outperforms BDeu in the asymptotic case. Considering the specificity, BDeu is less prone to false positives than CH, particularly in the small sample region. Furthermore, CH seems to be more sensitive to the strength of dependencies between variables defined by the data. Based on AUC scores, CH prior should be used in domains with small and moderate sample size. For a

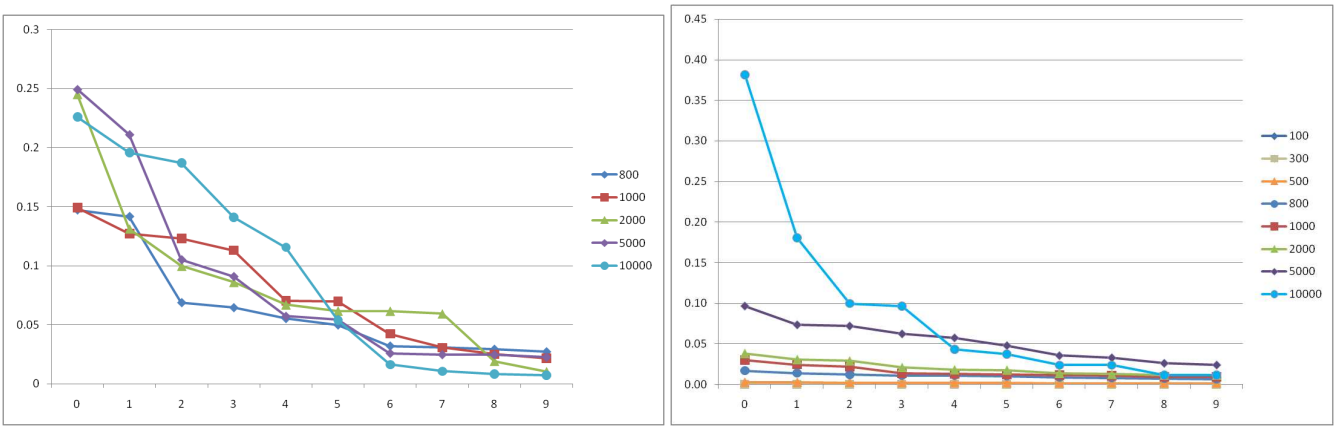


Fig. 3. Posteriors of the ten highest ranking MBSs in case of BDeu (left) and CH (right) priors for different sample sizes.

relatively large data set both the CH and the BDeu priors are appropriate choices, however, the parallel use of CH and BDeu priors could also be beneficial. Note that this notion is consistent with the Bayesian approach since it means averaging over priors.

In order to illustrate the necessity of Bayesian model averaging and the validity of the MBM based approach Figure 3 displays the posteriors of the ten highest ranking MBSs. Based on the evaluation of MBM posteriors, the CH prior performed well in small and moderate sample sized domains. In terms of MBS posteriors however, the MAP MBS is lower than 0.05 even for 2000 samples. In these cases, which we called as the 'flat posterior' case, a partial multivariate aggregation is a viable solution, apart from the aggregation into univariate MBM posteriors. The partial multivariate aggregation aims to find the most probable k size subsets of MBSs [4]. This enables the identification of the common elements of MBSs, that is relevant variable patterns. Note that an MBM is the $k = 1$ size subset of an MBS. The posterior distribution becomes peaked only in the nearly asymptotic case of the complete data set. In contrast, the MBS posterior in case of BDeu prior is relatively peaked for all sample sizes, although the MAP posterior is still relatively low for large samples. However, the relatively high MBS posteriors of the BDeu case, compared to the MBS posteriors in case of CH prior, are misleading in terms of sensitivity, as there are fewer relevant variables in these MBSs. This also confirms the necessity to apply the partial multivariate approach.

Apart from the basic setting of CH prior, other cases with different virtual sample size (VSS = 1,5,10,20,50) were examined. Figure 4 shows the sensitivity and the specificity measures for different sample sizes in case of CH prior with various VSS values. Interestingly, a distinct effect of VSS parameters is not observable in performance measures. A possible explanation is that the effect of varying sample size is far greater than that of VSS settings, dominating these measures. On a numerical level however, the larger the VSS is, the closer the posteriors get to extremes, i.e. the posterior of irrelevant variables is shifted towards 0, whereas the posterior of relevant variables approaches 1. This can be partially seen for the CH prior case in Figure 6 for larger sample sizes.

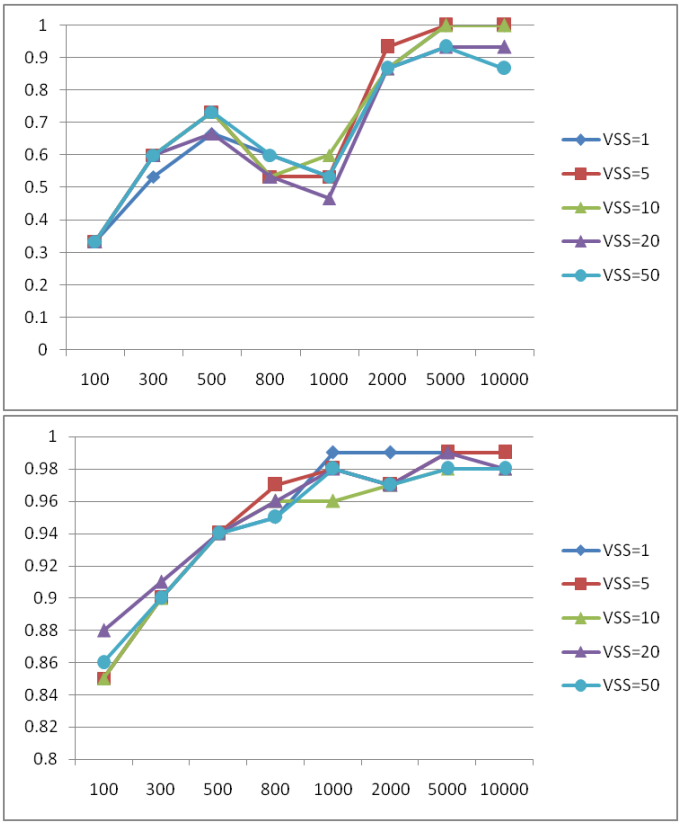


Fig. 4. Sensitivity (top) and specificity (bottom) measures based on MBM posteriors in case of CH prior for different sample sizes. VSS denotes the virtual sample size parameter for CH.

Figure 5 shows the sensitivity and the specificity measures for different sample sizes in case of BDEu prior with various ESS values. Similarly to the CH case, the effect of varying sample size seems to dominate the effect of ESS in the performance measures.

The average of MBM posteriors shown in Figure 6 is a rough approximation to characterize the MBM posterior distribution. In case of greater uncertainty, as in case of small sample size, the posterior values are relatively close to 0.5, thus the average is relatively high. As the data provides evidence for either relevance or irrelevance the posteriors get differentiated. Relevant

elements contribute the most to the average, as their posterior gets close to 1. Above 1000 samples, settings with $VSS \geq 1$ increase the average posterior as they 'push' the posteriors towards extremes.

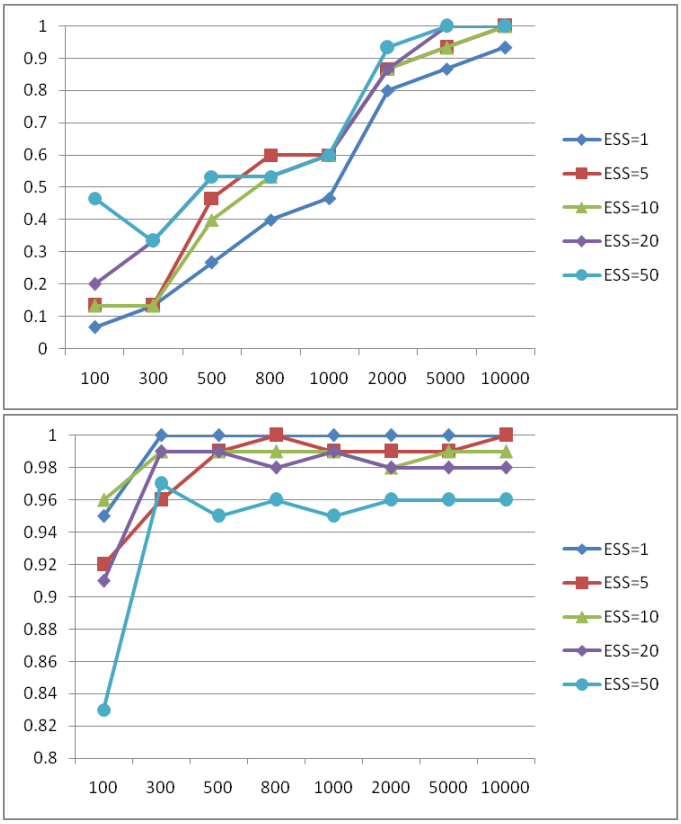


Fig. 5. Sensitivity (top) and specificity (bottom) measures based on MBM posteriors in case of BDeu prior for different sample sizes. ESS denotes the equivalent sample size parameter for BDeu.

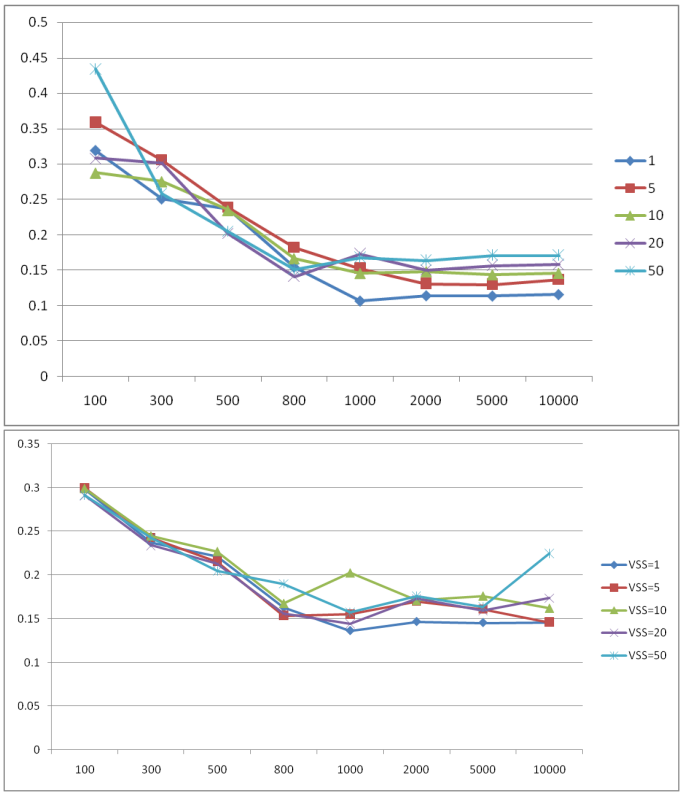


Fig. 6. Average of MBM posteriors in case of BDeu (top) and CH (bottom) priors for different sample size and virtual sample size parameters. ESS denotes the equivalent sample size parameter for BDeu, VSS denotes the virtual sample size parameter for CH.

3 Effect size and relevance

Effect size measures are a group of descriptors that aim to characterize the relevance of variables. Although there are many possible methods for assessing relevance, we distinguished three main approaches [13]:

- the *association based approach* neglecting structural aspects,
- the *causal approach* assuming a fixed structure,
- the *existential approach* assuming structural uncertainty.

The widespread association based approach uses effect size measures that do not take structural, multivariate relationships into account. Odds ratio is the most widely used such measure, especially in case-control studies.

The causal approach measures the effect of a variable X on another Y given a structure describing causal relationships, i.e. the nature and the strength of a relationship between variables X and Y. Structural equation modeling and the average causal effect measure [18] are two related methods providing a measure of effect size assuming a known causal structure. The obvious drawback of these methods is the lack of learning structures.

The structural (existential) uncertainty based approach uses Bayesian networks, which provide a graph based language for encoding relevance and representing dependency relationships. Each of these approaches focus on a different aspect of relevance, which we call respectively as

- *parametric* relevance,
- *causal* relevance,
- *structural (existential)* relevance.

These aspects appear to be separate dimensions of relevance, that is parametric relevance does not imply structural relevance (e.g. strong relevance explained below), and vice versa. For example an odds ratio is a quantitative measure of parametric relevance, that is a variable X can be relevant with respect to a selected target Y just by being over a certain threshold e.g. $OR(X,Y) \geq 2.5$, even though the structural relation between X and Y is unknown and has no influence on this aspect of relevance.

An advantage of the Bayesian network (BN) based Bayesian framework is that it allows to connect these two aspects by a structure based Bayesian effect size measure $OR(X,Y|\theta,G)$ based on $p(\theta,G|D_N)$, where θ denotes the parametrization and G the structure of an underlying BN, and D_N denotes data (i.e. $OR(X,Y|\theta,G)$ is a random variable with distribution $p(\theta,G|D_N)$).

A further reason for using a structure based Bayesian effect size measure is that there is no closed form for the posterior of effect size measures, such as odds ratio. Thus, the posterior has to be estimated, e.g. by sampling the log-odds ratio of a Dirichlet distribution based on the data. Another possible solution is to use the structural properties of BNs to guide the estimation process. Though instead of learning a whole BN from the data, which is computationally unfeasible in practical scenarios, the learning of relevant variables with respect to a target is sufficient.

Although relevance is a central concept, its relation to structural properties and to the concepts of association and effect size, seemed previously unclarified. We investigated various structural properties of BNs related to relevance and demonstrated the Bayesian application of BNs in relevance analysis [3].

Relevance can be defined formally in multiple ways. On one hand, it can be defined using conditional probability distributions without being specific to the applied model class used as a predictor, the optimization algorithm, the data set, and the loss function [15].

Definition 5 (Strong and weak relevance)**.** A feature X_i is strongly relevant to Y, if there exist some $X_i = x_i, Y = y$ 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, if it is not strongly relevant, and there exists a subset of features S'_i of S_i for which there exist 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.

On the other hand, relevance can also be defined by using Markov blankets as structural properties of BNs. The following theorem gives a sufficient condition for the unambiguous BN representation of the relevant structural properties [25].

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 $\mathrm{bd}(Y, G)$ denotes the set of parents, children and the children's other parents of Y [17]. If the distribution p is stable with respect to the DAG G, then $\mathrm{bd}(Y, G)$ forms a unique and minimal Markov blanket of Y, denoted as $\mathrm{MBS}_p(Y)$. Furthermore, $X_i \in \mathrm{MBS}_p(Y)$ iff X_i is strongly relevant.

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 distribution p is *Markov compatible* with graph structure G [18]. This theorem means that an $\mathrm{MBS}(Y, G)$ contains all the strongly relevant variables X_j with respect to Y, thus we can rely on this set of variables and corresponding parameters instead of taking the whole BN structure into consideration (under the conditions of multinomial sampling and global parameter independence).

3.1 Effect size in known causal structures

In case of a new domain that needs to be explored, most research studies aim to identify the mechanisms, which define relationships between the entities of that domain, and the overall behavior. In terms of Bayesian network learning, the first step is the learning of structures with high probability, that is identifying the direct and indirect relationships between entities of a domain, i.e. between the variables of a data set. In order to characterize the relationships, a second step, the learning of conditional distribution parameters is needed. The identification of this parametric level is necessary to calculate effect size, which describes the strength of a relationship between two variables. More specifically, in a mechanism centered approach, one is interested in the causal relationships that form the mechanism. Given a causal relationship $A \to B$ the effect size descriptor defines the amount of effect a value of variable A has on specific values of B. Note, that the Bayesian network representation under the Causal Markov Condition (for details see [18]) can be interpreted in a causal context, in which an edge between vertices X_i and X_j denotes a causal relationship $X_i \to X_j$.

In case of a known causal structure the structural equation modeling (SEM) methodology [18] provides a straightforward way of defining cause-effect relationships.

Despite the fact that the whole methodology of SEM was devised in order to quantitatively describe causal relationships and to assess effect size, the constraints on its applicability prohibits its wide-spread usage. In most practical cases (e.g. in gene association studies) there is no a priori causal structure, or the number of possible a priori structures makes this approach infeasible.

Due to the difficulties of applying methods based on causality, most researchers turn towards other methods, in which the causal interpretation is either omitted or prohibited. Interestingly, in case of gene association studies one of the usual long term goals is to identify causal biomarkers or to link a disease to a causal pathway, in terms of applied methods however the causal interpretation seems forbidden.

3.2 Bayesian effect size estimation

A possible multivariate Bayesian approach to effect size estimation is to utilize the underlying $\mathrm{BN}(G, \theta)$ that is a graph structure G and its parametrization θ for odds ratio computation $\mathrm{OR}(X_i, Y, \theta, G)$. More specifically, we are interested in such structures G_j where X_i is strongly relevant with respect to Y. This is the basic step towards the integration of structural and parametric aspects of relevance.

Definition 6. The Bayesian structure based odds ratio is defined as

$$p(\mathrm{OR}(X_i, Y)|I_{\mathrm{MBM}(X_i,Y|G)}) = \frac{p(\mathrm{OR}(X_i, Y), I_{\mathrm{MBM}(X_i,Y|G)})}{p(I_{\mathrm{MBM}(X_i,Y|G)})}, \quad (18)$$

where $I_{\mathrm{MBM}(X_i,Y|G_j)}$ means whether X_i is a member of $\mathrm{MBS}(Y, G_j)$, in other words whether X_i is strongly relevant or not.

By applying Bayesian model averaging over structures and parameters this can be estimated using computationally inten-

sive Markov chain Monte Carlo simulation, see e.g. [16]. However, given the number of possible structures G_j and their parameterizations (assuming a Dirichlet prior $\theta_k \sim Dir(\nu_k|\alpha_k)$), this computation is highly redundant. Therefore, we propose to sample the parameters in the 'relevant part' of the BN, instead of the whole structure. The structural property in which the relevant structural elements (with respect to the target) are encoded is called the *Markov blanket graph* [1, 3].

Definition 7. A Markov blanket graph MBG(Y, G) of a variable Y is a subgraph of a Bayesian network structure G, which contains the nodes of the Markov blanket set of Y, that is MBS(Y, G) and the incoming edges into Y and its children. Given a target node, which corresponds to the target variable Y, MBG(Y, G) as a (sub)graph structure consists of nodes that are (1) parents of Y, (2) children of Y or (3) "other parents" of the children of Y.

In contrast with a graph structure G containing all the dependency relationships of variables, a Markov blanket graph MBG(Y, G) includes only the mechanisms, in which Y is involved. This property makes the MBG(Y, G) an ideal candidate to serve as a base for measuring effect size. Since the goal is to measure the effect of a factor X_i (e.g. SNPs) on a target variable Y (e.g. disease susceptibility) the fact that X_i is a member of MBG(Y, G) is a relevant information. The following proposition allows the derivation of an efficient sampling scheme using MBGs.

Proposition 2. *The Bayesian structure based odds ratio (see Definition 6) can be computed using the posterior of MBGs parameterized by the data set.*

Proof: The first step is to expand Eq. 18 by averaging over all possible structures G.

$$\frac{1}{p(\mathrm{MBM}(X_i, Y))} \cdot \sum_{\forall G} p(\mathrm{OR}(X_i, Y), G, I_{\mathrm{MBM}(X_i,Y|G)})), \qquad (19)$$

where the first term serves as a normalization factor. The joint distribution of the odds ratio, the structure and the indicator function of strong relevance can be factorized according to the chain rule

$$\begin{aligned} &p(\mathrm{OR}(X_i, Y), G, I_{\mathrm{MBM}(X_i,Y|G)}) = \\ &= p(\mathrm{OR}(X_i, Y)|G, I_{\mathrm{MBM}(X_i,Y|G)}) \cdot p(I_{\mathrm{MBM}(X_i,Y|G)}|G) \cdot p(G), \end{aligned} \qquad (20)$$

where $p(G)$ is the prior probability of a given structure G, and $p(I_{\mathrm{MBM}(X_i,Y|G)})|G)$ is 1 if $X_i \in \mathrm{MBS}(Y, G)$ for a given G and 0 otherwise. This means that all those structures can be omitted, in which X_i is non-relevant.

$$\frac{1}{p(\mathrm{MBM}(X_i, Y))} \cdot \sum_{\forall G_j} p(\mathrm{OR}(X_i, Y)|G_j) \cdot p(G_j), \qquad (21)$$

where G_j denotes all those structures for which $I_{\mathrm{MBM}(X_i,Y|G_j)} = 1$.

If X_i is a member of the Markov blanket, then the probability of a certain value of the target Y (e.g. in case of Y as a disease state, the values are: "case" and "control") can be estimated based on the MBG(Y, G) and a specific instantiation of X_i [17]. This in turn allows the estimation of the structure based odds ratio by substituting graph structures G with MBGs in Eq. 21

$$\begin{aligned} p(\mathrm{OR}(X_i, Y)|\theta, G) \sim \frac{1}{p(\mathrm{MBM}(X_i, Y))} \cdot \\ \sum_{\forall \mathrm{MBG}_j(Y)} p(\mathrm{OR}(X_i, Y|\,\mathrm{MBG}_j(Y))) \cdot p(\mathrm{MBG}_j(Y)), \end{aligned} \qquad (22)$$

where $\mathrm{MBG}_j(Y)$ denotes all those MBG(Y, G) for which $I_{\mathrm{MBM}(X_i,Y|G)} = 1$, that is X_i is a member of a given Markov blanket of Y □.

Note, that from all the possible edges between these nodes, MBG(Y, G) only contains those that end in Y or one of its children. *Mechanism boundary graph* is another term for Markov blanket graphs due to its significant role in describing causal relationships (i.e. mechanisms). From this perspective given that the causal Markov assumption holds [18], MBG(Y, G) describes the direct causal relationships of Y (i.e. edges from nodes of type (1) and edges to nodes of type (2)), and also some of the indirect relationships of Y that form a special dependency pattern, called a *v-structure* [18]. This pattern $X \rightarrow Z \leftarrow Y$ consist of X a node of type (3) having an edge to Z a node of type (2), and the target Y also having an edge to Z.

Even though it is practical to use Markov blanket graphs to estimate structure based odds ratios instead of whole graph structures, their cardinality is even greater than that of Markov blanket sets, which is super exponential in the number of variables. This means that the sufficient sample size for the estimation of posteriors of Markov blanket graphs is also higher. Therefore, in a practical scenario the distribution of Markov blanket graph posteriors is even flatter than that of Markov blanket sets. Typically, in such a practical case there are thousands of Markov blanket graphs with relatively low posteriors. Therefore, selecting the MAP MBG is typically not the best solution, thus we resort to model averaging (see Algorithm 1).

Algorithm 1 Calculation of $BOR(X_i, Y)$ and its credible interval

Require: $n, m, \mathrm{MBG}(Y, G), D$
  for $\mathrm{MBG}_{1\ldots n}$ **do**
    for $\theta_{1\ldots m}$ **do**
      draw parametrization $\theta_k = (X_{k1} = x_{k1}, \ldots, X_{kr} = x_{kr})$
      for all $X_k \in \mathrm{MBG}_j$, so that $X_k \neq X_i$.
      estimate $P(Y = 0|X_i = x_i, \theta_k)$
      compute $\mathrm{Odds}(X_i = x_i, \theta_k) = \frac{P(Y=1|X_i=x_i,\theta_k)}{P(Y=0|X_i=x_i,\theta_k)}$
      compute $\mathrm{OR}(X_i, \theta_k) = \frac{\mathrm{Odds}(X_i=x_i^1,\theta_k)}{\mathrm{Odds}(X_i=x_i^0,\theta_k)}$
    end for
    compute $OR_{MBG}(X_i|\,\mathrm{MBG}_j) = \sum_{\theta_k=1}^{m} \mathrm{OR}(X_i, \theta_k)$
    update $\mathrm{OR}_{histogram}(X_i)$
  end for
  $BOR(X_i, Y) = \sum_{\mathrm{MBG}_j=1}^{n} OR_{MBG}(X_i|\,\mathrm{MBG}_j) \cdot p(\mathrm{MBG}_j)$
  calculate credible interval for $BOR(X_i, Y)$ based on $\mathrm{OR}_{histogram}(X_i)$

Definition 8. The Bayesian MBG-based odds ratio (BOR) is computed by averaging over the estimates of odds ratios based on possible MBGs as follows

$$BOR(X_i, Y) = \sum_{j=1}^{m} OR_{MBG}(X_i | \mathrm{MBG}_j(Y,G)) \cdot p(\mathrm{MBG}_j(Y,G)) \cdot I(X_i, \mathrm{MBG}_j(Y,G)), \quad (23)$$

where m is the number of MBGs with a posterior $p(\mathrm{MBG}_j(Y,G)) > 0$. The indicator function $I(X_i, \mathrm{MBG}_j(Y,G))$ is 1 if $X_i \in \mathrm{MBG}_j(Y,G)$ and 0 otherwise.

Assuming a binary target variable Y the MBG-based odds can be computed as

$$Odds_{MBG_{j(Y,G)}}(X_i = x_k, Y) = \frac{p(Y=1|\mathrm{MBG}_j(Y,G), X_i = x_k)}{p(Y=0|\mathrm{MBG}_j(Y,G), X_i = x_k)}, \quad (24)$$

where x_k is an instantiation of X_i. The computation of MBG-based odds ratio is based on these odds using instantiation values x_k depending on the used genetic model for inheritance (e.g.: dominant). Though averaging over genetic models is a possibility, comparing the calculated odds ratios and their credible intervals (i.e. Bayesian analogue of confidence intervals in standard statistical hypothesis testing framework) is generally preferred by experts.

This Bayesian measure of effect size can be extended to a set of variables, allowing the assessment of the joint effect size of multiple factors.

Definition 9. Given a set of predictors $\mathbf{V} = \{X_1, X_2, \ldots, X_n\}$ the multivariate *BOR* is calculated as

$$BOR^*(\mathbf{V}, Y) = \sum_{j=1}^{m} OR_{MBG}(\mathbf{V}|\mathrm{MBG}_j)(Y,G) \cdot p(\mathrm{MBG}_j(Y,G)) \cdot I^*(\mathbf{V}, \mathrm{MBG}_j(Y,G)), \quad (25)$$

where the indicator function $I^*(\mathbf{V}, \mathrm{MBG}_j(Y,G))$ is 1 if for any $X_i \in \mathbf{V}$ it is true that $X_i \in \mathrm{MBG}_j(Y,G)$, and 0 otherwise.

Correspondingly, the MBG-based odds for a set of variables $\mathbf{V}$ is given as

$$Odds^*_{MBG_j(Y,G)}(\mathbf{V}, Y) = = \frac{p(Y=1|\mathrm{MBG}_j(Y,G), X_{n1} = x_{n1}, \ldots, X_{nr} = x_{nr})}{p(Y=0|\mathrm{MBG}_j(Y,G), X_{n1} = x_{n1}, \ldots, X_{nr} = x_{nr})}, \quad (26)$$

where $x_{n1} \ldots x_{nr}$ are instantiations of variables $X_{ni} \in \mathbf{V}$ that are in $\mathrm{MBG}_j(Y,G)$. Note that if only one variable X_{ni} out of $|\mathbf{V}| = n$ elements of the set $\mathbf{V}$ is in $\mathrm{MBG}_j(Y,G)$ then Eq. 26 reverts to Eq. 24. Though the computation of odds for a given value configuration of variables of the set $\mathbf{V}$ is straightforward, the calculation of odds ratios presents a problem. In case of a univariate odds ratio the number of possible denominators is limited by the cardinality of one variable. In the multivariate case however, the odds for any subset of value configurations can serve as a denominator, theoretically. Practically, there are two basic choices: (1) the odds in case of "zeros" ($X_1 = 0, X_2 = 0, \ldots, X_n = 0$) and (2) the odds for all configurations except the one used in the numerator ($X_{n1} \neq x_{n1}, \ldots, X_{nr} \neq x_{nr}$).

3.3 Properties of the Bayesian odds ratio

We investigated the properties of Bayesian odds ratio on an artificial data set similar to the data set used for the prior case study (5000 samples, 115 variables).

Tab. 3. Comparison of Bayesian credible intervals (95% CR) and confidence intervals (95% CI) in case of a data set of 1000 and 5000 samples. Suffixes L and U denote the Lower and the Upper half of the interval respectively.

S-1000	OR	CI-L	CI-U	CR-L	CR-U
S1 1 vs 0	6.71	3.79	11.87	4.42	5.08
S2 1 vs 0	3.23	1.76	5.91	2.93	3.49
S-5000	**OR**	**CI-L**	**CI-U**	**CR-L**	**CR-U**
S1 1 vs 0	5.60	4.41	7.12	5.05	5.73
S2 1 vs 0	3.38	2.54	4.51	3.32	3.46

Bayesian odds ratios and related credible intervals (CR) were estimated based on Markov blanket graphs learned by the BN-BMLA method using 300, 500, 1000 and 5000 samples. We also computed odds ratios and corresponding confidence intervals (CI), which were corrected for multiple hypothesis testing using the number of a priori known, strongly relevant variables (11). Table 3 compares the properties of two selected variables ($S1$ and $S2$) which are both strongly relevant and have high odds ratios. In case of 5000 samples the length of Bayesian credible intervals is smaller and are contained within the corrected confidence intervals (see Figure 7), e.g. $OR^{(1,0)}_{S1}$ 95% CI: 4.41-7.12, whereas 95% CR: 5.05-5.73. As expected, in case of a smaller data set of 1000 samples the length of the corrected confidence intervals increase, since the data provides less evidence (e.g. $OR^{(1,0)}_{S2}$ 95% CI for 5000 samples is 2.54-4.51, while it is 1.76-5.91 for 1000 samples). A similar effect can be observed in case of the Bayesian credible interval for variable $S2$, but not for $S1$. Apart from the change of interval length, another significant effect is the drifting towards the neutral odds ratio of 1 (e.g. $OR^{(1,0)}_{S1}$ 95% CR : 5.05-5.73, and 4.42-5.08 for 5000 and 1000 samples respectively). A possible explanation is related to the 'flat posterior' case of relevance posteriors. As the sample size decreases, the sufficiency of the data decreases as well, which results in the vanishing difference between the posteriors of relevant and non-relevant variables. The increasing structural uncertainty may cause the 'degradation' of the credible interval of the Bayesian odds ratio.

Furthermore, Bayesian odds ratio has its own limitation for small sample sizes. This is due to the insufficiency of the data for learning Markov blanket graphs. This results in remarkably different Markov blankets, which may imply significantly different structure based odds ratios and corresponding credible intervals. Figure 8 illustrates the case when the Bayesian odds ratio is computed separately for the 10 most relevant Markov blan-

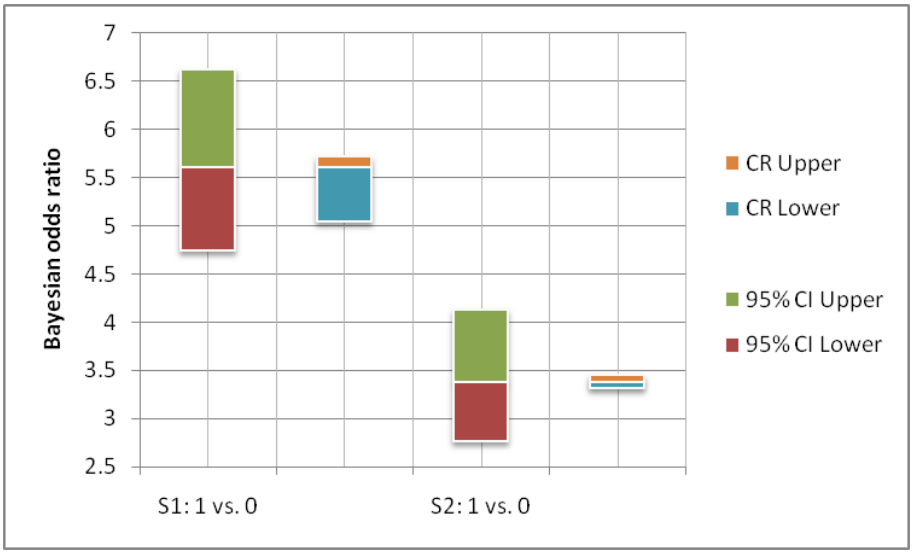


Fig. 7. Comparison of confidence intervals (CI) and credible intervals (CR) of selected variables based on the data set of 5000 samples. S1:1 vs.0 and S2: 1 vs. 0 denote odds ratios of heterozygous (1) versus common homozygous (0) cases of variable S1 and S2 respectively.

ket graphs for 300 samples. A joint (i.e. based on all MBGs) estimation of a Bayesian credible interval is rather problematic in this case, although a possible solution is to cover the whole concerned region.

In addition, the posterior distribution curves of Bayesian odds ratios provide a further tool for the characterization of effect size. Figure 9 shows such distribution curves for Bayesian odds ratios of three trichotomous variables ($V3, V10, V11$) with values $0, 1, 2$. The Bayesian odds ratios were computed with respect to the target variable T, which is a binary disease state descriptor. The odds corresponding to value 0 was used as a basis for odds ratio calculation. Each plot shows the posteriors within the 95% credible interval of a Bayesian odds ratio. Each curve depicts the outline of a histogram of possible odds ratio values (only those parts of the curve are shown that are within the 95% credible interval). A highly peaked curve indicates a distinct value with high certainty, while a flat curve indicates several possible values with moderate or low certainty. Curves may have different forms that reflect possible dependency models of analyzed factors (i.e. MBGs). On one hand, multiple peaks indicate that the possible models entail remarkably different odds ratio values (e.g. a model emphasizing synergistic effects versus a model focusing on main effects). On the other hand, large plateaus indicate that possible models entail similar odds ratio values.

In case of $V3$ both variants 1 and 2 highly increased the risk of the disease with credible intervals of (3.32-3.47) and (7.81-8.38) respectively. The posterior distribution curve related to variant 1 is highly peaked, which means that all possible dependency models of factors support that $V3$ has a strong effect on the susceptibility to the disease. The curve corresponding to variant 2 is relatively flat indicating that different dependency models entail somewhat different odds ratios, although the two local maxima (7.81 and 8.0) are close to each other.

Variable $V11$ is also relevant (such as $V3$) with respect to the target, though its effect size characteristics are remarkably different. The posterior distribution curve of the Bayesian odds ratio of variant 1 ($V11$) is similarly peaked as in case of variant 1 ($V3$), its effect size however is less significant (1.34-1.40). In contrast with this narrow credible interval, variant 2 ($V11$) has an extremely large credible interval of (4.0-12.0), and its distribution curve is flat. This means that possible dependency models entail remarkably different odds ratio values. Furthermore, the flatness of the posterior distribution curve of variant 2 indicates insufficient sample size, that is the ratio of variant 2 ($V11$) is low within the sample. Note that the flatness of the posterior distribution is a proper Bayesian response to low sample size.

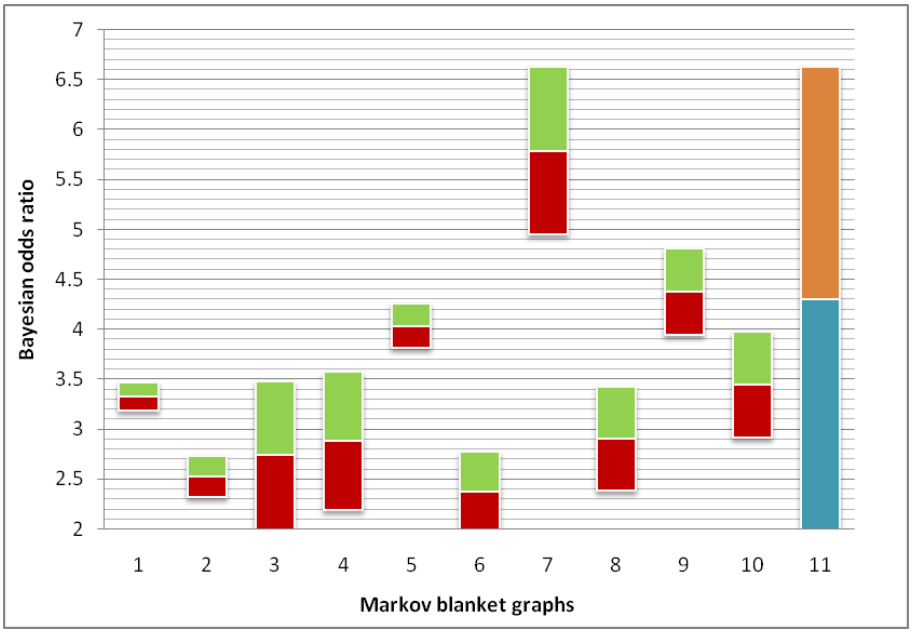
The Bayesian odds ratios of variable $V10$ show a protective effect, i.e. odds ratios less than 1 with respect to the target. Both in case of variants 1 and 2 of $V10$ the posterior distribution curves are relatively flat, whereas the credible intervals are narrow, (0.22-0.37) and (0.039-0.065) respectively. This indicates that the ratio of both variants are relatively low in the data set.

4 Conclusion

Bayesian networks and the Bayesian statistical framework provide multiple advantages in feature subset analysis, particularly in the exploration of interactions and in the characterization of relevance types [27]. Beside the multivariate aspect, Bayesian networks in the Bayesian statistical framework can also be used to explore the quantitative aspect of relevance by estimating the distribution of effect size parameters. This full Bayesian approach allows the joint modeling of structural and parametric aspects of relevance. The combination of these aspects offer refined concepts of relevance such as the proposed Bayesian structure based odds ratio.

In the paper we discussed the effects of priors and the applicability of Bayesian structure based odds ratio. Preliminary results confirm the general opinion that the theoretically sound BDeu parameter prior is more sensitive to virtual sample size selection, whereas the CH prior is proved to be surprisingly ro-

Fig. 8. Bayesian Markov blanket based odds ratios of a selected variable in case of a data set of 300 samples. The 10 credible intervals shown on the horizontal axis correspond to the 10 highest ranking Markov blankets. The additional 11th item illustrates a hypothetical joint credible interval.



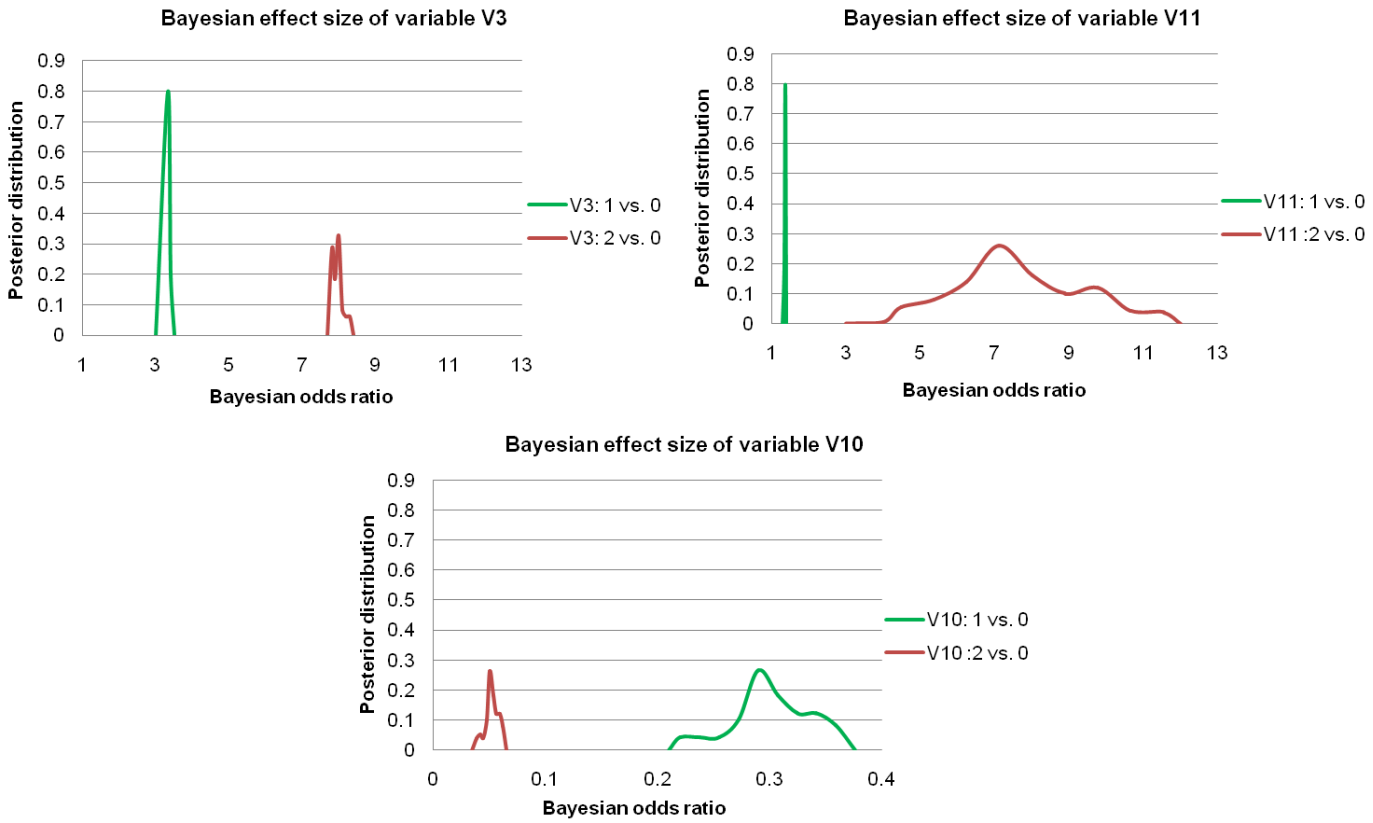


Fig. 9. Bayesian Markov blanket based odds ratios of selected variables. Bayesian odds ratio values are shown on the horizontal axis, whereas corresponding probability values are displayed on the vertical axis. Each curve depicts the outline of a histogram of possible odds ratio values (only those parts of the curve are shown that are within the 95% credible interval)

bust. The posterior distribution of odds ratio in the Bayesian statistical framework provides an easily interpretable overview, and allows the formulation of arbitrary probabilistic statements related to effect size, which is in sharp contrast with the limited usage of confidence intervals provided by the frequentist framework. Furthermore, the probabilistic statements can be further enriched by structural aspects. The proposed Bayesian structure based odds ratio can be seen as a first step along this line.

References

1 **Acid S, de Campos LM, Castellano JG**, *Learning Bayesian network classifiers: searching in a space of partially directed acyclic graphs.*, Machine Learning, **59**, (2005), 213–235, DOI 10.1007/s10994-005-0473-4.

2 **Akaike H**, *A new look at the statistical model identification.*, IEEE Trans. Auto. Cont., **19**, (1974), 716–723, DOI 10.1109/tac.1974.1100705.

3 **Antal P, Hullam G, Gezsi A, Millinghoffer A**, *Learning complex Bayesian network features for classification.*, In: Proc. of third European Workshop on Probabilistic Graphical Models, 2006, pp. 9–16, DOI 10.1.1.127.240.

4 **Antal P, Millinghoffer A, Hullam G, Szalai C, Falus A**, *A Bayesian multilevel analysis of feature relevance.*, In: Workshop on New challenges for feature selection in data mining and knowledge discovery (FSDM 2008) at The 19th European Conference on Machine Learning, Vol. 4, 2008, pp. 74–89.

5 **Antal P, Millinghoffer A, Hullam G, Szalai C, Falus A**, *A Bayesian view of challenges in feature selection: Feature aggregation, multiple targets, redundancy and interaction.*, JMLR Proceeding, **4**, (2008), 74–89.

6 **Bernardo JM**, *Bayesian Theory*, Wiley & Sons; Chichester, 1995.

7 **Bouckaert RR**, *Properties of Bayesian belief network learning algorithms*, In: Proceedings of the Tenth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-94), Morgan Kaufmann.; San Francisco, USA, 1994, pp. 102–109.

8 **Bouckaert RR**, *Bayesian belief networks: From construction to inference*, Ph.D. Thesis, Utrecht University, 1995.

9 **Buntine WL**, *Theory refinement of Bayesian networks.*, In: Proc. of the 7th Conf. on Uncertainty in Artificial Intelligence (UAI-1991), Morgan Kaufmann, 1991, pp. 52–60, DOI 10.1.1.52.1068.

10 **Cooper GF, Herskovits E**, *A Bayesian method for the induction of probabilistic networks from data.*, Machine Learning, **9**, (1992), 309–347, DOI 10.1.1.88.4436.

11 **Friedman N, Koller D**, *Being Bayesian about network structure.*, Machine Learning, **50**, (2003), 95-125, DOI 10.1.1.111.3429.

12 **Heckerman D, Geiger D**, *Likelihoods and parameter priors for Bayesian networks*, MicroSoft Research., 1995. Tech. Rep. MSR-TR-95-54.

13 **Hullam G, Antal P**, *Estimation of effect size posterior using model averaging over Bayesian network structures and parameters.*, In: Proceedings of the 6th European Workshop on Probabilistic Graphical Models.; Granada, Spain, 2012, pp. 147–154.

14 **Hullam G, Antal P, Millinghoffer A, Szalai C, Falus A**, *Evaluation of a Bayesian model-based approach in GA studies.*, JMLR Workshop and Conference Proceedings, **8**, (2010.), 30–43.

15 **Kohavi R, John GH**, *Wrappers for feature subset selection.*, Artificial Intelligence, **97**, (1997), 273–324, DOI 10.1016/S0004-3702(97)00043-X.

16 **Madigan D, York J**, *Bayesian graphical models for discrete data*, Internat. Statist. Rev., **63**, (1995), 215–232.

17 **Pearl J**, *Probabilistic Reasoning in Intelligent Systems.*, Morgan Kaufmann, San Francisco, CA, 1988, ISBN 978-1558604797.

18 **Pearl J**, *Causality: Models, Reasoning, and Inference.*, Cambridge University Press, 2000, ISBN 978-0521773621.

19 **Schwartz G**, *Estimating the dimension of a model*, Annals of Statistics, **6**, (1978), 461–464, DOI 10.1214/aos/1176344136.

20 **Silander T, Kontkanen P, Myllymaki P**, *On sensitivity of the map Bayesian network structure to the equivalent sample size parameter.*, In: Proceedings of the Twenty-Third Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-07), AUAI Press,; Corvallis, Oregon, USA, 2007, pp. 360–367.

21 **Silander T, Myllymaki P**, *A simple approach for finding the globally optimal Bayesian network structure.*, Proceedings of the Twenty-Second Annual Conference on Uncertainty in Artificial Intelligence (UAI-06), (2006), 445-452.

22 **Steck H**, *Learning the Bayesian network structure: Dirichlet prior versus data.*, Proceedings of the Twenty-Fourth Annual Conference on Uncertainty in Artificial Intelligence (UAI-08), (2008), 511–518.

23 **Steck H, Jaakkola TS**, *On the dirichlet prior and Bayesian regularization.*, Advances in Neural Information Processing Systems, Vol. 15, MIT Press, 2002, DOI 10.1.1.19.5502.

24 **Stephens M, Balding DJ**, *Bayesian statistical methods for genetic association studies.*, Nature Review Genetics, **10**(10), (2009), 681–690, DOI 10.1038/nrg2615.

25 **Tsamardinos I, Aliferis C**, *Towards principled feature selection: Relevancy,filters, and wrappers.*, Proc. of the Artificial Intelligence and Statistics, (2003), 334-342, DOI 10.1.1.12.4290.

26 **Ueno M**, *Learning networks determined by the ratio of prior and data.*, In: Proceedings of the Twenty-Sixth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-10), AUAI Press., 2010, pp. 598–605, `http://arxiv.org/abs/1203.3521`.

27 **Ungvari I, Hullam G, Antal P, Kiszel PS, Gezsi A**, *Evaluation of a partial genome screening of two asthma susceptibility regions using Bayesian network based Bayesian multilevel analysis of relevance*, PLoS ONE, **7**(2), (2012), DOI 10.1371/journal.pone.0033573.