Electronic Journal of Statistics
Vol. 11 (2017) 4000–4032
ISSN: 1935-7524
DOI: 10.1214/17-EJS1350

Constrained parameter estimation with uncertain priors for Bayesian networks*

Ali Karimnezhad

Department of Statistics and Computer Science
K. N. Toosi University of Technology, Tehran, Iran
e-mail: a.karimnezhad@kntu.ac.ir

Peter J. F. Lucas

Institute for Computing and Information Sciences
University of Nijmegen, Nijmegen, The Netherlands
e-mail: peterl@cs.ru.nl

and

Ahmad Parsian†

School of Mathematics, Statistics and Computer Science
University of Tehran, Tehran, Iran
e-mail: ahmad_p@khayam.ut.ac.ir

Abstract: In this paper we investigate the task of parameter learning of Bayesian networks and, in particular, we deal with the prior uncertainty of learning using a Bayesian framework. Parameter learning is explored in the context of Bayesian inference and we subsequently introduce Bayes, constrained Bayes and robust Bayes parameter learning methods. Bayes and constrained Bayes estimates of parameters are obtained to meet the twin objective of simultaneous estimation and closeness between the histogram of the estimates and the posterior estimates of the parameter histogram. Treating the prior uncertainty, we consider some classes of prior distributions and derive simultaneous Posterior Regret Gamma Minimax estimates of parameters. Evaluation of the merits of the various procedures was done using synthetic data and a real clinical dataset.

MSC 2010 subject classifications: Primary 62F15, 62C10; secondary 62F30, 62F35.
Keywords and phrases: Bayesian networks, constrained Bayes estimation, directed acyclic graph, posterior regret, robust Bayesian learning.

Received January 2017.

Contents

*The authors are grateful to the Editor, an anonymous Associate Editor and an anonymous referee for making valuable comments and suggestions on an earlier version of this article which led to substantial improvement.

†Ahmad Parsian's research supported by a grant of the Research Council of the University of Tehran.

1. Introduction

Bayesian networks (BNs) have become one of the most popular probabilistic models for representing joint probability distributions of a set of random variables [7, 28, 33]. Learning BNs from data is normally split into two different, although related steps: (1) learning the structure of the network and (2) learning the parameters [8, 18]. Sometimes the network structure is designed using expert knowledge. Once the structure of a network is obtained, parameter learning becomes possible.

Several methods are available to learn the structure of a BN (see [6, 10, 17, 44] among others), and there are many good software implementations of many of these [e.g. 30].

The focus of this paper is on the task of parameter learning only in BNs whose nodes represent discrete random variables. However, later in our final remarks, we refer to three common approaches of extending such BNs to BNs whose nodes are continuous random variables. Parameter learning has been studied also widely, giving rise to many different approaches. Most of the studies are based on the maximum likelihood (ML), the maximum a posteriori (MAP), or the posterior mean (PM) criterion. The ML estimation is a classical technique providing a parameter estimator by maximizing the joint probability density functions (pdfs), while the MAP and PM estimates, as Bayesian solutions, combine the information derived from the data with *a priori* knowledge concerning the parameter, see [4, 8, 9, 25, 35] among others.

Parameters of a BN possess an inherent symmetry as the sum of the parameters of a specific node is always equal to one, and thus, we expect their estimates satisfy this condition as well. Our main focus is to estimate parameters of a BN using Bayesian methods but it is very well-known that Bayes estimates highly depend on hyperparameters of a chosen prior and this may affect the corresponding results. Such a dependence in a learning procedure has been reported to be a serious problem [1, 42]. We adjust the task of Bayesian parameter learning using the idea of constrained Bayesian (CB) estimation of [29]. Further, we introduce and motivate the use of the simultaneous robust Bayes concept. The

notion of robustness used in this paper is different from the one explored by [34] in their description of the robust Bayes estimator, where they deal with missing data by means of probability intervals.

This paper is organized as follows: In Section 2, we introduce some preliminaries. Section 3 is devoted to simultaneous Bayes and the idea of CB learning. In addition, explicit forms for parameter estimates are derived. In Section 4, we introduce the idea of simultaneous posterior regret gamma minimax (SPRGM) learning in the presence of prior uncertainty and derive the corresponding estimates. In Section 5, we carry out an experimental study and compare performance of the proposed estimators using synthetic data from a well-known example network. Further, we study the impact of the proposed methods using real clinical data and a real-world BN. Finally, we conclude with some final remarks and a discussion. To keep readers in track, all the proofs along with some supplementary materials are provided in the Appendix.

2. Preliminaries

In this section we summarize the required basic material needed later. For more information see [8, 18, 24, 25, 31, 40].

2.1. Basic notions

A BN consists of a set of variables (or nodes) $V = \{X_1, \ldots, X_d\}$ and a subset of directed links E (also sometimes called edges or arcs) contained in the Cartesian product $V \times V$. We say the structure of a BN is known if the variables in the set V are connected to each other according to the links in E. Mathematically, the structure is called a directed graph. The directed graph is called acyclic, if it does not contain any directed cycle. We refer to such a directed acyclic graph (DAG) by $\mathcal{G} = (V, E)$. In the BNs context, a node is instantiated when its value is known through observing what it represents. We say we have a complete instantiation if all the nodes of a BN are simultaneously observed.

Suppose that for each $j = 1, \ldots, d$, the variable X_j takes values in the set $\mathcal{X}_j = \{x_j^{(1)}, \ldots, x_j^{(k_j)}\}$. The set of all possible outcomes for the experiment may be denoted by $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_d$. Hence, a sample of cases is given by $\boldsymbol{x} = (\boldsymbol{x}'_{(1)}, \ldots, \boldsymbol{x}'_{(n)})$, where $\boldsymbol{x}_{(i)} = (x_{i,1}^{(j_1)}, \ldots, x_{i,d}^{(j_d)})$ denotes the i-th complete instantiation and $\boldsymbol{x}'_{(i)}$ stands for the transpose of $\boldsymbol{x}_{(i)}$. For each variable X_j, denote all possible instantiations of the parent set Λ_j by the set $\{\lambda_j^{(1)}, \ldots, \lambda_j^{(q_j)}\}$. Thus, $\lambda_j^{(l)}$ implies that the parent configuration of variable X_j is in state $\lambda_j^{(l)}$ and there are q_j possible configurations of Λ_j.

For a given graph structure $\mathcal{G} = (V, E)$, let

$$n_{jilk} = \begin{cases} 1, & \text{if } (x_j^{(i)}, \lambda_j^{(l)}) \text{ is found in } \boldsymbol{x}_{(k)} \\ 0, & \text{otherwise,} \end{cases} \tag{2.1}$$

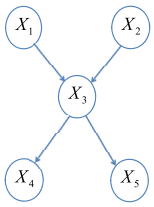

FIG 1. *A 5-node DAG.*

where $(x_j^{(i)}, \lambda_j^{(l)})$ is a configuration of the family (X_j, Λ_j). Let $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ denote the set of parameters defined by

$$\theta_{jil} = P\left(X_j = x_j^{(i)} | \Lambda_j = \lambda_j^{(l)}\right), \tag{2.2}$$

for $l = 1, \ldots, q_j$, $i = 1, \ldots, k_j$, $j = 1, \ldots, d$, with $\sum_{i=1}^{k_j} \theta_{jil} = 1$.

Using the decomposition of the probability distribution defined by the BN, the joint probability of a case $\boldsymbol{x}_{(k)}$ may be written as

$$p_{\boldsymbol{X}_{(k)}|\boldsymbol{\Theta}}(\boldsymbol{x}_{(k)}|\boldsymbol{\theta}, E) \propto \prod_{j=1}^{d} \prod_{l=1}^{q_j} \prod_{i=1}^{k_j} \theta_{jil}^{n_{jilk}}.$$

For independent observations $(\boldsymbol{x}_{(1)}, \ldots, \boldsymbol{x}_{(n)})$, the joint probability of the cases is

$$p_{\boldsymbol{X}|\boldsymbol{\Theta}}(\boldsymbol{x}|\boldsymbol{\theta}, \mathcal{G}) \quad \propto \quad \prod_{k=1}^{n} \prod_{j=1}^{d} \prod_{l=1}^{q_j} \prod_{i=1}^{k_j} \theta_{jil}^{n_{jilk}} = \prod_{j=1}^{d} \prod_{l=1}^{q_j} \prod_{i=1}^{k_j} \theta_{jil}^{n_{jil.}},$$

where $n_{jil.} = \sum_{k=1}^{n} n_{jilk}$, which is the likelihood function. One can observe that the ML estimate of θ_{jil} in Eq. (2.2) is given by

$$\delta_{jil}^{ML} = \frac{n_{jil.}}{n_{j.l.}}, \tag{2.3}$$

where $n_{j.l.} = \sum_{i=1}^{k_j} n_{jil.}$.

Observe that all parameters $(\theta_{j1l}, \ldots, \theta_{jk_jl})$ of a specific node X_j preserve the inherent symmetry of $\sum_{i=1}^{k_j} \theta_{jil} = 1$. We expect the corresponding estimates $(\delta_{j1l}, \ldots, \delta_{jk_jl})$ preserve this symmetry and satisfy the constraint $\sum_{i=1}^{k_j} \delta_{jil} = 1$. This constraint is automatically achieved by the ML estimates in Eq. (2.3) and $\sum_{i=1}^{k_j} \delta_{jil}^{ML} = 1$.

Example 2.1. *Consider the DAG depicted in Fig. 1 with five nodes* $X_1, \ldots, X_5$. *Suppose that all the nodes except* X_3 *are binary variables and* X_3 *takes values 0, 1 and 2 with the same probability. Hence,* $d = 5$, $k_i = 2$ *for* $i = 1, 2, 4, 5$, $k_3 = 3$,

and the parent set of X_5 *has three possible instantiations* $\lambda_5^{(1)}=0$, $\lambda_5^{(2)}=1$ *and* $\lambda_5^{(3)}=2$. *Suppose complete instantiations of 10 cases are available as below*

$$\boldsymbol{x}=\begin{pmatrix}\boldsymbol{x}_{(1)}\\ \vdots\\ \boldsymbol{x}_{(10)}\end{pmatrix}=\begin{pmatrix}(0,0,1,0,1)\\ (0,1,0,1,1)\\ (1,1,1,0,1)\\ (0,0,0,0,0)\\ (1,0,2,0,1)\\ (0,1,1,1,0)\\ (0,1,2,1,1)\\ (1,0,0,0,1)\\ (0,0,0,0,0)\\ (1,1,2,0,1)\end{pmatrix}$$

and we are interested in learning the parameters $\boldsymbol{\theta}_{5l}=(\theta_{51l},\theta_{52l})$, $l=1,2,3$. *The ML estimate of* $\boldsymbol{\theta}_{5l}$ *is given by* $\boldsymbol{\delta}_{5l}^{ML}=(\delta_{51l}^{ML},\delta_{52l}^{ML})$, *where* $\delta_{5il}^{ML}=\frac{n_{5il.}}{n_{5.l.}}$, $n_{5.l.}=\sum_{i=1}^{2}n_{5il.}$, $i=1,2$, $l=1,2,3$. *So, the ML estimate of* $\boldsymbol{\theta}_{52}$ *is given by* $\boldsymbol{\delta}_{52}^{ML}=(\frac{1}{3},\frac{2}{3})$.

2.2. Bayesian learning methods

In Example 2.1, one might believe that the sequence $(0,0,1,1,1)$ occurs in 80 percent of cases. If so, we could take *a priori* knowledge into account, assuming that some prior knowledge in forms of a prior distribution is available.

To derive the Bayes estimate of θ_{jil} in Eq. (2.2), consider the conjugate Dirichlet prior distribution $\mathrm{Dir}(\alpha_{j1l},\ldots,\alpha_{jk_jl})$, with pdf

$$\pi(\theta_{j1l},\ldots,\theta_{jk_jl})\ \propto\ \prod_{i=1}^{k_j}\theta_{jil}^{\alpha_{jil}-1}, \tag{2.4}$$

where $0<\theta_{jil}<1$, $\sum_{i=1}^{k_j}\theta_{jil}=1$ and $\alpha_{jil}>0$. Given the data $\boldsymbol{x}=(\boldsymbol{x}_{(1)},\ldots,\boldsymbol{x}_{(n)})$, it can be verified that $(\theta_{j1l},\ldots,\theta_{jk_jl})|\boldsymbol{x}\sim\mathrm{Dir}(n_{j1l.}+\alpha_{j1l},\ldots,n_{jk_jl.}+\alpha_{jk_jl})$. Obviously the marginal posteriors have Beta distributions, i.e., $\theta_{jil}|\boldsymbol{x}\sim Beta(n_{jil.}+\alpha_{jil},n_{j.l.}+\alpha_{j.l}-n_{jil.}-\alpha_{jil})$, where $\alpha_{j.l}=\sum_{i=1}^{k_j}\alpha_{jil}$.

It is easy to observe that the MAP and PM estimates of θ_{jil} are

$$\delta_{jil}^{MAP}=\arg\max_{\theta_{jil}}\pi(\theta_{jil}|\boldsymbol{X}=\boldsymbol{x})=\frac{n_{jil.}+\alpha_{jil}-1}{n_{j.l.}+\alpha_{j.l}-2}, \tag{2.5}$$

$$\delta_{jil}^{PM}=E[\theta_{jil}|\boldsymbol{X}=\boldsymbol{x}]=\frac{n_{jil.}+\alpha_{jil}}{n_{j.l.}+\alpha_{j.l}}. \tag{2.6}$$

Example 2.2. *(Example 2.1, cont.) To derive the MAP and PM estimates of* $\boldsymbol{\theta}_{5i2}=(\theta_{512},\theta_{522})$, *consider the conjugate* $Dir(\alpha_{512},\alpha_{522})$*-prior with* $\alpha_{512}=1$ *and* $\alpha_{522}=2$. *Then from (2.5) and (2.6),* $\boldsymbol{\delta}_{52}^{MAP}=(\frac{n_{512.}+\alpha_{512}-1}{n_{5.2.}+\alpha_{5.2}-2},\frac{n_{522.}+\alpha_{522}-1}{n_{5.2.}+\alpha_{5.2}-2})=(\frac{1}{4},\frac{3}{4})$ *and* $\boldsymbol{\delta}_{52}^{PM}=(\frac{n_{512.}+\alpha_{512}}{n_{5.2.}+\alpha_{5.2}},\frac{n_{522.}+\alpha_{522}}{n_{5.2.}+\alpha_{5.2}})=(\frac{1}{3},\frac{2}{3})$.

3. Constrained Bayesian learning

In the preceding section, assuming the squared error loss (SEL) function, we observed that if the only objective is simultaneous estimation of the BN parameters $\boldsymbol{\theta}_{jl} = (\theta_{j1l}, \ldots, \theta_{jk_jl})$ with θ_{jil} defined in Eq. (2.2), the Bayes estimate is a vector of posterior means, i.e., $\boldsymbol{\delta}_{jl}^{PM} = (\delta_{j1l}^{PM}, \ldots, \delta_{jk_jl}^{PM})$ with δ_{jil}^{PM} given by Eq. (2.6). In this section, following the idea of CB estimation of [29], we provide an adjusted version of δ_{jil}^{PM}.

To avoid any unambiguity, we define the following key terms. Let $\boldsymbol{\delta}_{jl} = (\delta_{j1l}, \ldots, \delta_{jk_jl})$ be a vector of arbitrary estimates of elements of $\boldsymbol{\theta}_{jl} = (\theta_{j1l}, \ldots, \theta_{jk_jl})$ with θ_{jil} defined in Eq. (2.2). Define the sample mean and the sample variance of ensemble of the estimates δ_{jil} by $\bar{\delta}_{j.l} = \frac{1}{k_j}\sum_{i=1}^{k_j}\delta_{jil}$ and $\frac{1}{k_j}\sum_{i=1}^{k_j}\left(\delta_{jil} - \bar{\delta}_{j.l}\right)^2$, respectively. Also, define the posterior expected sample mean (PESM) and the posterior expected sample variance (PESV) of ensemble of the parameters θ_{jil} by $\frac{1}{k_j}E\Big[\sum_{i=1}^{k_j}\theta_{jil}|\boldsymbol{X} = \boldsymbol{x}\Big]$ and $\frac{1}{k_j}E\Big[\sum_{i=1}^{k_j}\left(\theta_{jil} - \bar{\theta}_{j.l}\right)^2 | \boldsymbol{X} = \boldsymbol{x}\Big]$ with $\bar{\theta}_{j.l} = \frac{1}{k_j}\sum_{i=1}^{k_j}\theta_{jil}$, respectively.

[29] suggested that problems with using posterior means as Bayes estimates might be dealt with by constructing a vector of CB estimators for which the sample mean and the sample variance of an ensemble of them are equal to the PESM and the PESV of an ensemble of parameters, respectively. Particularly, he proved that under normal likelihood with normal prior, the sampling variability of a collection of Bayes estimates is smaller than the posterior expectation of the corresponding population variability, see [16]. This property holds true in BNs, as provided in the following lemma. See the Appendix for a detailed verification of this inequality.

Lemma 3.1. *Let $\boldsymbol{\delta}_{jl}^{PM} = (\delta_{j1l}^{PM}, \ldots, \delta_{jk_jl}^{PM})$ be a vector of PM's of $\boldsymbol{\theta}_{jl} = (\theta_{j1l}, \ldots, \theta_{jk_jl})$ with θ_{jil} defined in Eq. (2.2), w.r.t. some prior π. Then, for a fixed j and l, the sample variance of ensemble of the Bayes estimates in $\boldsymbol{\delta}_{jl}^{PM}$ is smaller than the PESV of ensemble of parameters in $\boldsymbol{\theta}_{jl}$, i.e.,*

$$\frac{1}{k_j}\sum_{i=1}^{k_j}\left(\delta_{jil}^{PM} - \bar{\delta}_{j.l}^{PM}\right)^2 < \frac{1}{k_j}E\Big[\sum_{i=1}^{k_j}\left(\theta_{jil} - \bar{\theta}_{j.l}\right)^2 | \boldsymbol{X} = \boldsymbol{x}\Big]. \tag{3.1}$$

where $\bar{\delta}_{j.l}^{PM} = \frac{1}{k_j}\sum_{i=1}^{k_j}\delta_{jil}^{PM}$ and $\bar{\theta}_{j.l} = \frac{1}{k_j}\sum_{i=1}^{k_j}\theta_{jil}$.

By the CB approach of [29], the empirical distribution function of CB estimates becomes close to the empirical distribution function of the corresponding unknown parameters. This way, the sampling variability of a collection of estimates is a better estimate of the underlying variability among the population parameters. For more details, see [11, 13, 14, 15].

The idea of matching the first two moments from posterior distribution of parameters with the corresponding moments from distribution of estimates has been followed in a wide range of problems, mostly to derive adjusted empirical

Bayes estimators in problems such as disease mapping or environmental risk assessment. For example, in disease mapping it is supposed that there are k regions labeled with the indices $1, 2, \ldots, k$. By this setting, [11] follows a hierarchical model for disease counts and estimate the true disease rates θ_i, $i = 1, 2, \ldots, k$. For more information, see [11, 16] and papers cited therein.

Now, consider the problem of estimating θ_{jil} defined in (2.2) under the SEL function for a fixed j and l. If interest lies in both simultaneous estimation and closeness between the distribution of estimates and the posterior distribution of the parameters, the idea of deriving CB estimation might be helpful. To make a motivation, it is of interest to compare our estimation problem with the disease mapping problem considered in [11]. In the latter problem, some levels were considered for the parameter of interest (the true disease rates θ_i, $i = 1, 2, \ldots, k$) and in our estimation problem, in a specific node and for a specific parent, the parameter of interest, i.e., θ_{jil}, has different levels when changing i in the set $\{1, 2, \ldots, k_j\}$. Thus, CB estimation can be considered in order to meet the twin objective of simultaneous estimation and closeness between the distribution of the estimates and the posterior distribution of the parameters.

Here, we consider the problem of obtaining CB estimates of θ_{jil}, subject to the constraints considered by [29], and an additional constraint which is imposed due to the nature of parameter learning in BNs, i.e.,

(i) $\sum_{i=1}^{k_j} \delta_{jil} = \sum_{i=1}^{k_j} E[\theta_{jil}|\boldsymbol{X} = \boldsymbol{x}]$,
(ii) $\frac{1}{k_j}\sum_{i=1}^{k_j}(\delta_{jil} - \bar{\delta}_{j.l})^2 = \frac{1}{k_j} E\Big[\sum_{i=1}^{k_j}(\theta_{jil} - \bar{\theta}_{j.l})^2|\boldsymbol{X} = \boldsymbol{x}\Big]$,
(iii) $\sum_{i=1}^{k_j} \delta_{jil} = 1$,

where $\bar{\delta}_{j.l} = \frac{1}{k_j}\sum_{i=1}^{k_j} \delta_{jil}$ and $\bar{\theta}_{j.l} = \frac{1}{k_j}\sum_{i=1}^{k_j} \theta_{jil}$.

It is interesting to note that since $\sum_{i=1}^{k_j} \theta_{jil} = \sum_{i=1}^{k_j} \delta_{jil}^{PM} = 1$, the constraint (i) results in (iii). However, each one of the constraints (i) and (iii) plays its separate role and hence, we simultaneously consider both of these constraints for later use. The following theorem provides CB estimates of parameters in BNs. The main idea of this theorem comes from a proof that appeared in [29]. See the Appendix for a version of the proof compatible with the constraints considered in this paper.

Theorem 3.1. *Let $\boldsymbol{\delta}_{jl}^{PM} = (\delta_{j1l}^{PM}, \ldots, \delta_{jk_jl}^{PM})$ be a vector of PM's of $\boldsymbol{\theta}_{jl}$ w.r.t. some prior π. Then under the constraints (i)-(iii), the CB estimate of $\boldsymbol{\theta}_{jl}$ is given by $\boldsymbol{\delta}_{jl}^{CB} = (\delta_{j1l}^{CB}, \ldots, \delta_{jk_jl}^{CB})$, where $\delta_{jil}^{CB} = a_{jl}\delta_{jil}^{PM} + (1 - a_{jl})\frac{1}{k_j}$ and*

$$a_{jl} = \left\{ \frac{S_{jl}(\boldsymbol{x}) - \frac{1}{k_j}}{T_{jl}(\boldsymbol{x}) - \frac{1}{k_j}} \right\}^{\frac{1}{2}},$$

with $S_{jl}(\boldsymbol{x}) = E[\sum_{i=1}^{k_j} \theta_{jil}^2|\boldsymbol{X} = \boldsymbol{x}]$ and $T_{jl}(\boldsymbol{x}) = \sum_{i=1}^{k_j}(\delta_{jil}^{PM})^2$.

Example 3.1. *(Example 2.1, cont.) To derive the CB estimate of $\boldsymbol{\theta}_{52}$, w.r.t. the $Dir(1, 2)$-prior note that $\theta_{512}|\boldsymbol{x} \sim Beta(2, 4)$ and $\theta_{522}|\boldsymbol{x} \sim Beta(4, 2)$. Verify*

that $S_{52}(\boldsymbol{x}) = \frac{26}{42}$, $T_{jl}(\boldsymbol{x}) = \frac{5}{9}$ *and* $a_{52} = \sqrt{\frac{15}{7}}$. *Hence,* $\boldsymbol{\delta}_{52}^{CB} = (\delta_{512}^{CB}, \delta_{512}^{CB})$ *with* $\delta_{512}^{CB} = \sqrt{\frac{15}{7}}\delta_{512}^{PM} + (1 - \sqrt{\frac{15}{7}})\frac{1}{2} = 0.2560$ *and* $\delta_{522}^{CB} = 0.7440$ *is the CB estimate of* $\boldsymbol{\theta}_{52}$.

Bayes estimates generally depend on hyperparameters of a chosen prior and this can affect the relevant results. The following example clarifies this point.

Example 3.2. *(Example 2.2, cont.) In Examples 2.2 and 3.1, the PM and CB estimates of* $\boldsymbol{\theta}_{5i2} = (\theta_{512}, \theta_{522})$ *with the hyperparameter choices* $\alpha_{512} = 1$ *and* $\alpha_{522} = 2$ *reported as* $\boldsymbol{\delta}_{52}^{PM} = (0.3333, 0.6667)$ *and* $\boldsymbol{\delta}_{52}^{CB} = (0.2560, 0.7440)$, *respectively. Now, if one considers the hyperparameters as* $\alpha_{512} = 1$ *and* $\alpha_{522} = 4$, *it is easy to verify that the PM and CB estimates become* $\boldsymbol{\delta}_{52}^{PM} = (0.25, 0.75)$ *and* $\boldsymbol{\delta}_{52}^{CB} = (0.2113, 0.7887)$.

That the hyperparameters affect learning BN structures has been reported as a serious problem [41, 45]. In the next section, we consider this issue and explore robust Bayesian methods to overcome this problem.

4. Posterior regret Gamma minimax learning

When available, a particular prior distribution is usually somewhat arbitrary and there are good reasons to question the reliability of such a distribution. Usually, there is no way for a user to say that a particular prior is better than another one. Thus, in practice, prior knowledge is often vague. Alternatively, the expert may be unable to specify the prior completely. This situation may also occur when two or more experts do agree on the choice of a prior distribution arising in a decision making problem but differ in opinion w.r.t. the choice of the hyperparameters. A common solution to handle prior uncertainty in Bayesian statistical inference is to choose a class Γ of prior distributions and compute some quantity, such as the posterior risk, the Bayes risk or the posterior expected value, as the prior ranges over Γ. This is known as *robust Bayesian analysis*. This methodology is connected with studying the effect of changing a prior within a class Γ over some quantity, see [1, 2, 3]. In this section, we use the idea of SPRGM estimation in the parameter learning procedure. Readers may refer to the treatise by [19] for a detailed discussion of literature on various robust Bayes analysis problems. The book contains chapters on robust Bayes rules including many references dealing with various standard classes of priors (e.g., Chapters 8 and 13) as well as some applications provided in Chapters 17-21.

It is worth stressing that, in addition to the debate on being robust Bayesian, there are other strong arguments in the literature about incorporating prior knowledge into the task of data analysis of which [12] and [37] are excellent references. The relevant approach, known as hierarchical Bayes approach, robustifies the conjugate distribution, assuming a fully Bayesian model. The idea is that one may have structural and subjective prior information at the same time and would like to model this in stages. The attention is often on two stage priors and is used when the first stage of prior elicitation leads to a class Γ of

priors and then the statistician in the second stage, puts a prior on Γ. Thus, if $\Gamma = \{\pi_1 \text{ is of a given functional form and } \lambda \in \Lambda\}$, then the second stage would consist of putting a prior, $\pi_2(\lambda)$, on the hyperparameter λ. While specification of the hyperparameter is usually done based on subjective beliefs assuming that it reflects the best guess of statistician, it is difficult. The difficulty level of the hyperparameter specification is more tangible as number of hyparameters increases. BNs are a prime example of such a complicated specification and thus in this paper, we only emphasize on the robust Bayes approach. The difficulty of specifying the hyperprior has made common the use of noninformative priors at the second stage [e.g. 1, 37] but the noninformative priors might lead to inappropriate choices of priors. In contrast, not only the robust Bayes approach we consider in this paper obviates the complicatedness of prior elicitation, it leads to a global prevention against inappropriate choices of priors or their hyperparameters [22, 23]. See [12, 37] for more information on robust Bayes and hierarchical Bayes approaches, and [21, 22] for applications of these approaches as well as a quick list of some of their advantages and disadvantages.

Now, let $\rho(\pi, \delta_{jil})$ be posterior risk of the estimate δ_{jil} of θ_{jil} in Eq. (2.2) under the SEL function, i.e., $\rho(\pi, \delta_{jil}) = E[(\theta_{jil} - \delta_{jil})^2 | \boldsymbol{X} = \boldsymbol{x}]$. For a learning procedure of the parameters $\boldsymbol{\theta}_{jl}$ in a DAG under the SEL function and given a class of priors Γ, the posterior regret of choosing δ_{jil} instead of the Bayes estimate δ^{PM}_{jil} is $r_p(\delta_{jil}, \delta^{PM}_{jil}) = \rho(\pi, \delta_{jil}) - \rho(\pi, \delta^{PM}_{jil}) = \left(\delta_{jil} - \delta^{PM}_{jil}\right)^2$. With respect to simultaneous estimation, we define the posterior regret of choosing $\boldsymbol{\delta}_{jl}$ instead of $\boldsymbol{\delta}^{PM}_{jl}$ to be

$$r_p(\boldsymbol{\delta}_{jl}, \boldsymbol{\delta}^{PM}_{jl}) = \sum_{i=1}^{k_j} \sup_{\pi_i \in \Gamma} r_p(\delta_{jil}, \delta^{PM}_{jil}) = \sum_{i=1}^{k_j} \sup_{\pi_i \in \Gamma} \left(\delta_{jil} - \delta^{PM}_{jil}\right)^2,$$

with the constraint $\sum_{i=1}^{k_j} \delta_{jil} = 1$. Then we define $\boldsymbol{\delta}^{SPR}_{jl} = (\delta^{SPR}_{j1l}, \ldots, \delta^{SPR}_{jk_jl})$ to be the SPRGM value over the class Γ of priors if

$$r_p(\boldsymbol{\delta}^{SPR}_{jl,\Gamma}, \boldsymbol{\delta}^{PM}_{jl}) = \inf_{\boldsymbol{\delta}_{jl} \in \mathcal{D}} \sum_{i=1}^{k_j} \sup_{\pi \in \Gamma} r_p(\delta_{jil}, \delta^{PM}_{jil}) = \inf_{\boldsymbol{\delta}_{jl} \in \mathcal{D}} \sum_{i=1}^{k_j} \sup_{\pi \in \Gamma} \left(\delta_{jil} - \delta^{PM}_{jil}\right)^2, \quad (4.1)$$

where $\mathcal{D}$ is the class of all possible estimates of $\boldsymbol{\theta}_{jl}$.

As it is obvious from Eq. (4.1), deriving SPRGM would be possible by determining the supremum of $r_p(\delta_{jil}, \delta^{PM}_{jil})$, where the prior varies over all priors in the class Γ. As δ_{jil} does not depend on prior information, one way to obtain insight into the supremum of $r_p(\delta_{jil}, \delta^{PM}_{jil})$ is to look at the behavior of the Bayes estimate δ^{PM}_{jil} in Eq. (2.6). For fixed data and fixed j and l, variation of the hyperparameters $\alpha_{j1l}, \alpha_{j2l}, \ldots, \alpha_{jk_jl}$ in some given intervals determines the behavior of the PM estimate δ^{PM}_{jil} and thus, the supremum of $r_p(\delta_{jil}, \delta^{PM}_{jil})$ can be analyzed. To make it clear, we recall Example 2.2 where $\delta^{PM}_{512} = \frac{n_{512.} + \alpha_{512.}}{n_{5.2.} + \alpha_{512} + \alpha_{522}}$. Obviously, δ^{PM}_{512} is increasing in α_{512} and decreasing in α_{522}. Now, if the hyperparameters α_{512} and α_{522} (which in fact reflect prior beliefs) vary over some

intervals, δ_{512}^{PM} can take some minimum and maximum values and thus, the supremum of $r_p(\delta_{jil}, \delta_{jil}^{PM})$ can be analyzed in order to determine the SPRGM estimate.

To derive SPRGM estimates of θ_{jil}, $i = 1, \ldots, k_j$, once again, consider the conjugate $Dir(\alpha_{j1l}, \ldots, \alpha_{jk_jl})$ prior and let $K_j = \{\alpha_{j1l}, \ldots, \alpha_{jk_jl}\}$. Also, to adopt prior information in the robust learning methodology, our prior knowledge about the Dirichlet hyperparameters may cluster them at three disjoint sets, i.e., the prior information may indicate that it would be better to consider some elements of K_j, say α_{jul}, are fixed known constants and some other elements, say α_{jvl}, are varied over some fixed known intervals. We refer to these cases as U_j and V_j, respectively. Thus, α_{jul} is a fixed hyperparamer if $u \in U_j$ and similarly, α_{jvl} is a varying hyperparameter if $v \in V_j$. To cover all the possible cases of hyperparameter variations, let $W_j = K_j - U_j - V_j$ consist of all the other cases. The set W_j is not necessarily empty, since prior knowledge may suggest letting the sum of all the hyperparameters vary in a fixed known interval. This clustering leads to different classes of priors. The following are examples of such classes of Dirichlet priors $\Pi_j = Dir(\alpha_{j1l}, \ldots, \alpha_{jk_jl})$

$$\Gamma^{\dagger} = \left\{\Pi_j : \alpha_{jul} = \alpha_{jul}^{*}, \underline{\alpha}_{jvl} \leq \alpha_{jvl} \leq \overline{\alpha}_{jvl}, u \in U_j, v \in K_j - U_j, V_j = \emptyset\right\}, \quad (4.2)$$

$$\Gamma^{\ddagger} = \left\{\Pi_j : \underline{\alpha}_{jvl} \leq \alpha_{jvl} \leq \overline{\alpha}_{jvl}, \underline{\alpha}_{w} \leq \sum_{w \in K_j - V_j} \alpha_{jwl} \leq \overline{\alpha}_{w}, v \in V_j, w \in K_j - V_j, U_j = \emptyset\right\}, \quad (4.3)$$

where α_{jul}^{*}, $\underline{\alpha}_{jvl}$, $\overline{\alpha}_{jvl}$, $\underline{\alpha}_{w}$ and $\overline{\alpha}_{w}$ are known constants. The classes in (4.2) and (4.3) are very general. A special case occurs when either $U_j = \emptyset$ in $\Gamma^{\dagger}$ or $V_j = K_j$ in $\Gamma^{\ddagger}$. The resulting class of priors is

$$\Gamma^{\dagger\ddagger} = \left\{\Pi_j : \underline{\alpha}_{jvl} \leq \alpha_{jvl} \leq \overline{\alpha}_{jvl}, v \in K_j\right\}, \quad (4.4)$$

where $\underline{\alpha}_{jvl}$ and $\overline{\alpha}_{jvl}$ are fixed known constants. As seen above, there can be a wide variety of classes of Dirichlet priors for a specific problem. We emphasize that each of the possible classes of priors reflect the prior knowledge behind the choice of such a class of prior and this does not mean at all that a chosen class is superior to many alternatives. In fact, when choosing a class of priors, we only decide based on our experience.

Although SPRGM estimates of $\boldsymbol{\theta}_{jl} = (\theta_{j1l}, \ldots, \theta_{jk_jl})$ can be derived for different values of k_j, we provide two most promising cases with $k_j = 2$ and $k_j = 3$. The following theorem provides one SPRGM estimator of $\boldsymbol{\theta}_{jl}$ under the sum of SEL function when $k_j = 2$. For the proof, see the Appendix.

Theorem 4.1. *Let Γ be a class of priors and suppose that, for $i = 1, 2$, $\underline{\delta}_{jil}(\boldsymbol{X}) \equiv \underline{\delta}_{jil} = \inf_{\pi \in \Gamma} \delta_{jil}^{PM}$ and $\overline{\delta}_{jil}(\boldsymbol{X}) \equiv \overline{\delta}_{jil} = \sup_{\pi \in \Gamma} \delta_{jil}^{PM}$ are finite. Then, the SPRGM estimate of $(\theta_{j1l}, \theta_{j2l})$ over the class Γ subject to the constraint $\delta_{j1l} + \delta_{j2l} = 1$, is given by $\boldsymbol{\delta}_{jl,\Gamma}^{SPR} = (\delta_{j1l,\Gamma}^{SPR}, \delta_{j2l,\Gamma}^{SPR})$ with*

$$\delta^{SPR}_{j1l,\Gamma}=\begin{cases}\frac{1}{2}\left(1+\overline{\delta}_{j1l}-\overline{\delta}_{j2l}\right), & \text{if } \overline{\delta}_{j1l}+\underline{\delta}_{j2l}\geq 1 \ \&\ \underline{\delta}_{j1l}+\overline{\delta}_{j2l}\geq 1\\ \frac{1}{2}\left(1+\underline{\delta}_{j1l}-\underline{\delta}_{j2l}\right), & \text{if } \overline{\delta}_{j1l}+\underline{\delta}_{j2l}\leq 1 \ \&\ \underline{\delta}_{j1l}+\overline{\delta}_{j2l}\leq 1\\ \text{does not exist,} & \text{Otherwise,}\end{cases}$$

and $\delta^{SPR}_{j2l,\Gamma}=1-\delta^{SPR}_{j1l,\Gamma}$.

The following example illustrates how to derive SPRGM estimates in practice.

Example 4.1. *(Example 3.1, cont.) To derive the SPRGM estimates of* $\boldsymbol{\theta}_{52}$*, consider the following classes of priors*

$$\begin{aligned}\Gamma^{\dagger} &= \Big\{Dir(\alpha_{512},\alpha_{522}): 0.5\leq\alpha_{512}\leq 1.5,\ \alpha_{522}=2\Big\},\\ \Gamma^{\ddagger} &= \Big\{Dir(\alpha_{512},\alpha_{522}): 2\leq\alpha_{522}\leq 3,\ \alpha_{512}=1\Big\},\\ \Gamma^{\dagger\ddagger} &= \Big\{Dir(\alpha_{512},\alpha_{522}): 0.5\leq\alpha_{512}\leq 1.5,\ 2\leq\alpha_{522}\leq 3\Big\}.\end{aligned}$$

Notice that $\delta^{PM}_{5i2}=\frac{n_{5i2.}+\alpha_{5i2}}{n_{5.2.}+\alpha_{5.2}}$*, for a fixed* i*, is increasing in* α_{5i2} *and decreasing in* α_{jml}*,* $m\neq i$*. Thus over* $\Gamma^{\dagger}$*,* $\underline{\delta}_{512}=\frac{3}{11}$*,* $\overline{\delta}_{512}=\frac{5}{13}$*,* $\underline{\delta}_{522}=\frac{8}{13}$ *and* $\overline{\delta}_{522}=\frac{8}{11}$*. Obviously,* $\overline{\delta}_{512}+\underline{\delta}_{522}\leq 1$*,* $\underline{\delta}_{512}+\overline{\delta}_{522}\leq 1$ *and hence,* $\delta^{SPR}_{512,\Gamma^{\dagger}}=\frac{1}{2}\left(1+\underline{\delta}_{512}-\underline{\delta}_{522}\right)=\frac{47}{143}$ *and* $\delta^{SPR}_{522,\Gamma^{\dagger}}=1-\delta^{PR}_{512,\Gamma^{\dagger}}=\frac{96}{143}$*. Also over* $\Gamma^{\ddagger}$*,* $\underline{\delta}_{512}=\frac{2}{7}$*,* $\overline{\delta}_{512}=\frac{1}{3}$*,* $\underline{\delta}_{522}=\frac{2}{3}$ *and* $\overline{\delta}_{522}=\frac{5}{7}$*, and since* $\overline{\delta}_{512}+\underline{\delta}_{522}\leq 1$ *and* $\underline{\delta}_{512}+\overline{\delta}_{522}\leq 1$*, thus* $\delta^{SPR}_{512,\Gamma^{\ddagger}}=\frac{1}{2}\left(1+\underline{\delta}_{512}-\underline{\delta}_{522}\right)=\frac{13}{42}$ *and* $\delta^{SPR}_{522,\Gamma^{\ddagger}}=1-\delta^{SPR}_{522,\Gamma^{\ddagger}}=\frac{29}{42}$*. Similarly,* $\delta^{SPR}_{512,\Gamma^{\dagger\ddagger}}=\frac{1}{2}\left(1+\underline{\delta}_{512}-\underline{\delta}_{522}\right)=\frac{4}{13}$ *and* $\delta^{SPR}_{522,\Gamma^{\dagger\ddagger}}=1-\delta^{SPR}_{522,\Gamma^{\ddagger}}=\frac{9}{13}$.

In the next theorem, we provide one SPRGM estimator of $\boldsymbol{\theta}_{jl}$ under the sum of SEL function when $k_j=3$. For the proof, see the Appendix.

Theorem 4.2. *Let* Γ *be a class of priors and suppose that, for* $i=1,2,3$*,* $\underline{\delta}_{jil}(\boldsymbol{X})\equiv\underline{\delta}_{jil}=\inf_{\pi\in\Gamma}\delta^{PM}_{jil}$ *and* $\overline{\delta}_{jil}(\boldsymbol{X})\equiv\overline{\delta}_{jil}=\sup_{\pi\in\Gamma}\delta^{PM}_{jil}$ *are finite. Then, the SPRGM estimate of* $(\theta_{j1l},\theta_{j2l},\theta_{j2l})$ *over the class* Γ *subject to the constraint* $\delta_{j1l}+\delta_{j2l}+\delta_{j3l}=1$*, is given by* $\boldsymbol{\delta}^{SPR}_{jl,\Gamma}=(\delta^{SPR}_{j1l,\Gamma},\delta^{SPR}_{j2l,\Gamma},\delta^{SPR}_{j3l,\Gamma})$ *in which* $\delta^{SPR}_{j3l,\Gamma}=1-\delta^{SPR}_{j1l,\Gamma}-\delta^{SPR}_{j2l,\Gamma}$ *and* $\delta^{SPR}_{jil,\Gamma}$*,* $i=1,2$*, are determined by one of the following conditions:*

(i) $\delta^{SPR}_{j1l,\Gamma}=\frac{1}{3}(1+2\overline{\delta}_{j1l}-\overline{\delta}_{j2l}-\overline{\delta}_{j3l})$ *and* $\delta^{SPR}_{j2l,\Gamma}=\frac{1}{3}(1+2\overline{\delta}_{j2l}-\overline{\delta}_{j1l}-\overline{\delta}_{j3l})$ *provided that* $\delta^{SPR}_{j1l,\Gamma}\leq\frac{1}{2}(\underline{\delta}_{jil}+\overline{\delta}_{jil})$*,* $i=1,2,3$*,*

(ii) $\delta^{SPR}_{j1l,\Gamma}=\frac{1}{3}(1+2\overline{\delta}_{j1l}-\overline{\delta}_{j2l}-\underline{\delta}_{j3l})$ *and* $\delta^{SPR}_{j2l,\Gamma}=\frac{1}{3}(1+2\overline{\delta}_{j2l}-\overline{\delta}_{j1l}-\underline{\delta}_{j3l})$*, provided that* $\delta^{SPR}_{jil,\Gamma}\leq\frac{1}{2}(\underline{\delta}_{jil}+\overline{\delta}_{jil})$*,* $i=1,2$ *and* $\delta^{SPR}_{j3l,\Gamma}>\frac{1}{2}(\underline{\delta}_{j3l}+\overline{\delta}_{j3l})$*,*

(iii) $\delta^{SPR}_{j1l,\Gamma}=\frac{1}{3}(1+2\overline{\delta}_{j1l}-\underline{\delta}_{j2l}-\overline{\delta}_{j3l})$ *and* $\delta^{SPR}_{j2l,\Gamma}=\frac{1}{3}(1+2\underline{\delta}_{j2l}-\overline{\delta}_{j1l}-\overline{\delta}_{j3l})$*, provided that* $\delta^{SPR}_{jil,\Gamma}\leq\frac{1}{2}(\underline{\delta}_{jil}+\overline{\delta}_{jil})$*,* $i=1,3$ *and* $\delta^{PM}_{j2l}>\frac{1}{2}(\underline{\delta}_{j2l}+\overline{\delta}_{j2l})$*,*

(iv) $\delta^{SPR}_{j1l,\Gamma}=\frac{1}{3}(1+2\underline{\delta}_{j1l}-\overline{\delta}_{j2l}-\overline{\delta}_{j3l})$ *and* $\delta^{SPR}_{j2l,\Gamma}=\frac{1}{3}(1+2\overline{\delta}_{j2l}-\underline{\delta}_{j1l}-\overline{\delta}_{j3l})$*, provided that* $\delta^{SPR}_{jil,\Gamma}\leq\frac{1}{2}(\underline{\delta}_{jil}+\overline{\delta}_{jil})$*,* $i=2,3$ *and* $\delta^{PM}_{j1l}>\frac{1}{2}(\underline{\delta}_{j1l}+\overline{\delta}_{j1l})$*,*

(v) $\delta^{SPR}_{j1l,\Gamma}=\frac{1}{3}(1+2\overline{\delta}_{j1l}-\underline{\delta}_{j2l}-\underline{\delta}_{j3l})$ *and* $\delta^{SPR}_{j2l,\Gamma}=\frac{1}{3}(1+2\underline{\delta}_{j2l}-\overline{\delta}_{j1l}-\underline{\delta}_{j3l})$*, provided that* $\delta^{SPR}_{j1l,\Gamma}\leq\frac{1}{2}(\underline{\delta}_{j1l}+\overline{\delta}_{jil})$ *and* $\delta^{SPR}_{jil,\Gamma}>\frac{1}{2}(\underline{\delta}_{jil}+\overline{\delta}_{jil})$ *,* $i=2,3$*,*

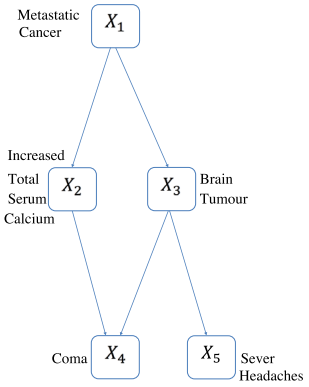

FIG 2. *A BN for the lung cancer problem.*

(vi) $\delta^{SPR}_{j1l,\Gamma} = \frac{1}{3}(1 + 2\underline{\delta}_{j1l} - \overline{\delta}_{j2l} - \underline{\delta}_{j3l})$ *and* $\delta^{SPR}_{j2l,\Gamma} = \frac{1}{3}(1 + 2\overline{\delta}_{j2l} - \underline{\delta}_{j1l} - \underline{\delta}_{j3l})$, *provided that* $\delta^{SPR}_{j2l,\Gamma} \leq \frac{1}{2}(\underline{\delta}_{j2l} + \overline{\delta}_{j2l})$ *and* $\delta^{SPR}_{jil,\Gamma} > \frac{1}{2}(\underline{\delta}_{jil} + \overline{\delta}_{jil})$, $i = 1, 3$,

(vii) $\delta^{SPR}_{j1l,\Gamma} = \frac{1}{3}(1 + 2\underline{\delta}_{j1l} - \underline{\delta}_{j2l} - \overline{\delta}_{j3l})$ *and* $\delta^{SPR}_{j2l,\Gamma} = \frac{1}{3}(1 + 2\underline{\delta}_{j2l} - \underline{\delta}_{j1l} - \overline{\delta}_{j3l})$, *If* $\delta^{SPR}_{j3l,\Gamma} \leq \frac{1}{2}(\underline{\delta}_{j3l} + \overline{\delta}_{j3l})$ *and* $\delta^{SPR}_{jil,\Gamma} > \frac{1}{2}(\underline{\delta}_{jil} + \overline{\delta}_{jil})$, $i = 1, 2$,

(viii) $\delta^{SPR}_{j1l,\Gamma} = \frac{1}{3}(1 + 2\underline{\delta}_{j1l} - \underline{\delta}_{j2l} - \underline{\delta}_{j3l})$ *and* $\delta^{SPR}_{j2l,\Gamma} = \frac{1}{3}(1 + 2\underline{\delta}_{j2l} - \underline{\delta}_{j1l} - \underline{\delta}_{j3l})$, *provided that* $\delta^{SPR}_{jil,\Gamma} > \frac{1}{2}(\underline{\delta}_{jil} + \overline{\delta}_{jil})$, $i = 1, 2, 3$.

5. Experiments

5.1. Synthetic data

In this section, we provide a simulation study to compare performance of the ML, MAP, PM, CB and SPRGM estimates. For this purpose, we use the well-known metastatic lung cancer BN shown in Fig. 2. This network appeared in the early literature on BNs, see [24, 43] among others.

For our simulation study, let X_1 be distributed according to $B(1, 0.2)$, where $B(1, p)$ stands for a Bernoulli distribution with success probability p. To generate values for the variables X_2 and X_3, note that their possible parent sets are $\lambda_j^{(1)} = 0$ and $\lambda_j^{(2)} = 1$, $j = 2, 3$. Now, suppose $\theta_{211} = 0.8$, $\theta_{212} = 0.2$, $\theta_{311} = 0.95$, $\theta_{312} = 0.80$, and generate the variables $X_j|\lambda_j^{(l)} \sim B(1, \theta_{jil})$ for the specified indices. To generate values for X_4, the possible parent sets are $\lambda_4^{(1)} = (0, 0)$, $\lambda_4^{(2)} = (0, 1)$, $\lambda_4^{(3)} = (1, 0)$ and $\lambda_4^{(4)} = (1, 1)$, we generate the variables $X_4|\lambda_4^{(l)} \sim B(1, \theta_{4il})$ for the specified indices with $\theta_{411} = 0.95$, $\theta_{412} = \theta_{413} = \theta_{414} = 0.2$. Finally, we define the variable X_5 to be zero with probability $\theta_{511} = 0.4$ if the output of X_3 is zero. Otherwise, X_5 takes one with probability $\theta_{522} = 0.8$, indicating that a patient who has Brain tumour will suffer from severe headaches.

To draw conclusions about performance of the different estimates provided earlier, we consider estimates of the conditional probabilities $\boldsymbol{\theta}_{5l} = (\theta_{51l}, \theta_{52l})$,

$l = 1, 2$. To obtain the MAP, PM and CB estimates of θ_{52l}, $l = 1, 2$, we use the conjugate $Dir(\alpha_{52l}, \alpha_{51l})$-prior distribution. Notice that in Bayes estimation of θ_{51l}, the conjugate prior is $Dir(\alpha_{51l}, \alpha_{52l})$. To make a choice in estimating θ_{522} w.r.t. the hyperparameters, suppose that three experts have provided information about having a brain tumour and subsequently estimated chance of being affected by severe headaches. Assume that one of the experts based on some prior knowledge assumes the conjugate $Dir(\alpha_{522}, \alpha_{512})$-prior with $\alpha_{512} = 5$ and $\alpha_{522} = 35$, implying that the mean chance is about 0.875. Suppose that this expert opinion does not attract consensus of opinion from the two other experts. Rather, they believe in different hyperparameters. They attribute $Dir(40, 10)$ and $Dir(45, 15)$-priors, respectively, reflecting that they believe that the prior mean is about 0.80 and 0.75. We shall refer to these three chosen priors by π_1, π_2 and π_3, respectively. Clearly, the three experts attributed priors with means around the real parameter 0.8, but the resulting Bayes estimates can still be quite different. To deal with this issue, we consider the following class of priors incorporating the three experts' beliefs:

$$\Gamma \quad = \quad \Big\{ Dir(\alpha_{522}, \alpha_{512}) : 5 \leq \alpha_{512} \leq 15,\ 35 \leq \alpha_{522} \leq 45 \Big\}. \tag{5.1}$$

We rely on this class to derive the SPRGM estimate of $\boldsymbol{\theta}_{52} = (\theta_{512}, \theta_{522})$.

Now, to estimate θ_{521}, the probability that a patient has severe headaches in the absence of a Brain tumor, suppose similar to the above situation, that three experts have provided estimates of this conditional probability. The opinion of the three experts is expressed by the $Dir(\alpha_{521}, \alpha_{511})$-prior with $(\alpha_{521}, \alpha_{511}) = (40, 25), (45, 25), (35, 30)$, implying that the mean chance is around 0.60. We shall refer to these priors by π_1^*, π_2^* and π_3^*, respectively. To obtain the SPRGM estimate of $\boldsymbol{\theta}_{51} = (\theta_{511}, \theta_{521})$, we consider the following class of priors incorporating the three experts' beliefs:

$$\Gamma^* \quad = \quad \Big\{ Dir(\alpha_{521}, \alpha_{511}) : 25 \leq \alpha_{511} \leq 30,\ 35 \leq \alpha_{521} \leq 45 \Big\} \tag{5.2}$$

Consider the three priors π_j, π_j^*, $j = 1, 2, 3$, and the classes of priors Γ and Γ^*, as defined above. The following steps in the simulation study are taken:

Step 1. Complete instantiations $(x_1, \ldots, x_5)$ of n cases with $n = 25$, 50, 100, 200 are generated.

Step 2. For each $i = 1, 2$, taking each of the priors π_j, $j = 1, 2, 3$, and the class Γ into account, the estimates δ_{5i2}^{ML}, δ_{5i2}^{MAP,π_j}, δ_{5i2}^{PM,π_j}, δ_{5i2}^{CB,π_j} and $\delta_{5i2,\Gamma}^{SPR}$ of θ_{5i2} are computed. For each fixed $i = 1, 2$, these computations result in 11 estimates of θ_{5i2} denoted by $d[k, i]$, $k = 1, \ldots, 11$, and $i = 1, 2$. Similarly, taking each of the priors π_j^*, $j = 1, 2, 3$, and the class Γ^* into account, the estimates δ_{5i1}^{ML}, $\delta_{5i1}^{MAP,\pi_j^*}$, $\delta_{5i1}^{PM,\pi_j^*}$, $\delta_{5i1}^{CB,\pi_j^*}$ and $\delta_{5i1,\Gamma^*}^{SPR}$ of θ_{5i1} are computed. Again, these computations result in 11 estimates of θ_{5i1} denoted by $d^*[k, i]$, $k = 1, \ldots, 11$, and $i = 1, 2$.

Step 3. Steps 1 and 2 are run $N = 10,000$ times. Based on the generated data, mean, average of Kullback-Leibler divergence (AKLD) and average of sample variance (ASV) of ensemble of the estimates $(d[k,1], d[k,2])$ of $(\theta_{512}, \theta_{522})$, $k = 1, \ldots, 11$, are computed as follows:

$$\begin{aligned}
\text{Mean}\,(d[k,i]) &= \frac{1}{N}\sum_{m=1}^{N} d_m[k,i],\\
\text{AKLD}\,(d[k,1], d[k,2]) &= \frac{1}{N}\sum_{m=1}^{N}\left(\theta_{512}\log_2\left(\frac{\theta_{512}}{d_m[k,1]}\right) + \theta_{522}\log_2\left(\frac{\theta_{522}}{d_m[k,2]}\right)\right),\\
\text{ASV}\,(d[k,1], d[k,2]) &= \frac{1}{N}\sum_{m=1}^{N}\frac{1}{2}\left(\left(d_m[k,1]-\frac{1}{2}\right)^2 + \left(d_m[k,2]-\frac{1}{2}\right)^2\right), \quad (5.3)
\end{aligned}$$

where $d_m[k,i]$ stands for the estimate $d[k,i]$ in the m-th repetition. The mean, AKLD and ASV of ensemble of the estimates $(d^*[k,1], d^*[k,2])$ of $(\theta_{511}, \theta_{521})$, $k = 1, \ldots, 11$, are similarly computed.

The quantitative results for different values of n are summarized in Table 1 and Table A.1 of the Appendix. Before drawing any conclusion, we would like to restate that the true value of the parameters θ_{511}, θ_{521}, θ_{512} and θ_{522} are 0.4, 0.6, 0.2 and 0.8 respectively. Thus, based on the mean criterion in Step 3, any of the proposed estimates which has a mean close to the corresponding true value would be preferred to the alternatives. By the AKLD criterion, any estimate with lowest AKLD value would be preferred to the other alternatives. We introduced the ASV criterion based on the condition (ii) in Theorem 3.1. By this criterion, sample variance of ensemble of the corresponding CB estimates $(d[k,1], d[k,2])$ of $\boldsymbol{\theta}_{52} = (\theta_{512}, \theta_{522})$ is equal to the PESV of ensemble of the parameters in $\boldsymbol{\theta}_{52}$.

From Table 1, we observe that the simulation process failed to compute the ML estimate for $n = 25$, 50, 100. However, for $n = 200$ in Table 1 and all sample sizes in Table A.1 of the Appendix, the ML estimates perform quite well, although one should notice that in practice, we use them when there is no source of prior knowledge.

The three different priors in Table 1 have led to the different prior-based estimates MAP, PM and CB estimates. When considering π_2, i.e., $Dir(\alpha_{522}, \alpha_{512})$-prior with $\alpha_{512} = 10$ and $\alpha_{522} = 40$ (in this case the prior mean is equal to the true parameter 0.8), the corresponding MAP, PM and CB estimates, i.e., δ_{5i2}^{MAP,π_2}, δ_{5i2}^{PM,π_2}, δ_{5i2}^{CB,π_2}, perform better than the other Bayes and CB estimates. Similarly, in Table A.1 of Appendix the MAP, PM and CB estimates w.r.t. the prior π_1^* (which has a mean closer than the mean of other priors to the true parameter 0.6), outperform the other Bayes and CB estimates. As noted earlier, it is not possible to decide only relying on one source of prior information. Rather one should respect the knowledge of all the experts. Thinking in this way, we observe that the SPRGM estimates computed over Γ and Γ^* perform better than the other prior-based estimates. In other words, the SPRGM estimates are better in most cases because the case with correct prior actually yields equally good estimates, although correct prior knowledge may be rare.

TABLE 1. *Quantitative statistics for different values of n.*

	n	i	δ^{ML}_{5i2}	δ^{MAP,π_1}_{5i2}	δ^{MAP,π_2}_{5i2}	δ^{MAP,π_3}_{5i2}	δ^{PM,π_1}_{5i2}	δ^{PM,π_2}_{5i2}	δ^{PM,π_3}_{5i2}	δ^{CB,π_1}_{5i2}	δ^{CB,π_2}_{5i2}	δ^{CB,π_3}_{5i2}	$\delta^{SPR}_{5i2,\Gamma}$
Mean	25	1	†na	0.1099	0.1880	0.2400	0.1285	0.2000	0.2484	0.1250	0.1950	0.2426	0.2000
		2	na	0.8901	0.8120	0.7600	0.8715	0.8000	0.7516	0.8750	0.8050	0.7574	0.8000
AKLD			na	0.0519	0.0013	0.0069	0.0304	0.0005	0.0098	0.0337	0.0006	0.0077	0.0005
ASV			na	0.1523	0.0975	0.0677	0.1382	0.0901	0.0634	0.1408	0.0931	0.0663	0.0901
Mean	50	1	na	0.1143	0.1886	0.2389	0.1318	0.2001	0.2470	0.1284	0.1953	0.2414	0.2001
		2	na	0.8857	0.8114	0.7611	0.8682	0.7999	0.7530	0.8716	0.8047	0.7586	0.7999
AKLD			na	0.0482	0.0018	0.0068	0.0288	0.0010	0.0095	0.0318	0.0010	0.0076	0.0010
ASV			na	0.1492	0.0972	0.0684	0.1359	0.0901	0.0642	0.1384	0.0931	0.0670	0.0901
Mean	100	1	na	0.1216	0.1894	0.2366	0.1375	0.2002	0.2443	0.1341	0.1956	0.2391	0.2002
		2	na	0.8784	0.8106	0.7634	0.8625	0.7998	0.7557	0.8659	0.8044	0.7609	0.7998
AKLD			na	0.0416	0.0024	0.0065	0.0257	0.0016	0.0089	0.0283	0.0016	0.0072	0.0016
ASV			na	0.1437	0.0968	0.0697	0.1320	0.0903	0.0656	0.1344	0.0930	0.0683	0.0903
Mean	200	1	0.1994	0.1328	0.1905	0.2324	0.1460	0.1999	0.2395	0.1429	0.1959	0.2349	0.1999
		2	0.8006	0.8672	0.8095	0.7676	0.8540	0.8001	0.7605	0.8571	0.8041	0.7651	0.8001
AKLD			0.0454	0.0330	0.0034	0.0061	0.0216	0.0026	0.0080	0.0235	0.0026	0.0066	0.0026
ASV			0.1011	0.1357	0.0964	0.0720	0.1261	0.0906	0.0683	0.1283	0.0930	0.0707	0.0906

†The simulation process failed to compute the ML estimates. The three priors π_1, π_2 and π_3 stand for $Dir(35,5)$, $Dir(40,10)$ and $Dir(45,15)$-priors, respectively. Γ stands for the class of conjugate Dirichlet priors in (5.1).

Conducting a more precise investigation, Fig. 3 provides side-by-side histograms of sample variance of ensemble of the estimates in $\boldsymbol{\delta}_{52} = (\delta_{512}, \delta_{522})$ of the parameters $\boldsymbol{\theta}_{52} = (\theta_{512}, \theta_{522})$ with δ_{5i2} replaced by one of the estimates δ_{5i2}^{MAP,π_j}, δ_{5i2}^{PM,π_j}, δ_{5i2}^{CB,π_j} and $\delta_{5i2,\Gamma}^{SPR}$, $i = 1, 2$, $j = 1, 2, 3$ (the simulation process failed to compute the ML estimates). For these simulations we took $n = 50$ but our investigation led to similar results for other values of n. Associated with each of the priors π_j, $j = 1, 2, 3$, in each row of Fig. 3, we provide histograms of PESV of ensemble of the parameters in $\boldsymbol{\theta}_{52}$ to show how similarly they behave, compared to the sample variance of ensemble of the estimates in $\boldsymbol{\delta}_{52}$. We observe that the histograms of PESV and the sample variance of ensemble of the CB estimates w.r.t. all the priors π_j, $j = 1, 2, 3$ coincide. This is in fact an illustration of Theorem 3.1. It is also of interest to note that as we observe from Fig. 3, the CB estimator is the only estimator with the same distribution of the posterior distribution of the parameters (the corresponding histogram and histogram of PESV fall on each other). Further, Fig. 4 provides averages of PESV (APESV) of ensemble of the parameters in $\boldsymbol{\theta}_{52}$ and ASV of ensemble of the different estimates w.r.t. all the priors π_j, $j = 1, 2, 3$ given by Eq. (5.3). This figure also confirms that PESV associated with each of the priors π_j, $j = 1, 2, 3$, is always greater than sample variance of the corresponding PM estimates (as provided by Lemma 3.1), and the CB estimator is the only estimator of which the corresponding sample variance is equal to the PESV of ensemble of the parameters in $\boldsymbol{\theta}_{52}$ (as an illustration of Theorem 3.1).

On the other hand, if $\boldsymbol{\delta}_{5l}$ estimates $\boldsymbol{\theta}_{5l}$ very well, the corresponding ASV is expected to be close to $\frac{1}{2}\left((\theta_{51l} - \frac{1}{2})^2 + (\theta_{52l} - \frac{1}{2})^2\right)$, which is equal to 0.01 for $l = 1$ and 0.09 for $l = 2$. From Fig. 4 we observe that the ASV of the SPRGM estimates of $\boldsymbol{\theta}_{52}$ is not close to the APESV but its ASV is very close to 0.09. Also, this is clearly observed from Fig. 3 in which the histogram of SPRGM estimates is centred about 0.09. Comparing the PM and the CB estimates, we observe that ASV of the CB estimates w.r.t. the prior π_3 is closer to 0.09 than the corresponding PM estimates and thus, their performance is better than the PM estimates w.r.t. the priors π_3. This also can be confirmed from Table 1. Thus, in some situations, the CB estimates act better than the PM ones.

The same conclusions are deduced when estimating the parameters θ_{5i1}, $i = 1, 2$, w.r.t the priors π_j^*, $j = 1, 2, 3$, and the class of priors Γ^*, see Table A.1, Fig. A.1 and Fig. A.2 of the Appendix.

5.2. Real clinical data

In this section, we analyze a clinical dataset using an associated, expert-designed BN and compare performance of ML, MAP, PM, CB and SPRGM estimates. For this purpose, we consider the Hepar BN model [32], which is a causal BN concerning a subset of the domain of hepatology: 11 liver diseases (described by 9 disorder nodes), 18 risk factors, and 44 symptoms and laboratory tests results. Fig. 5 shows a simplified fragment of the Hepar BN model.

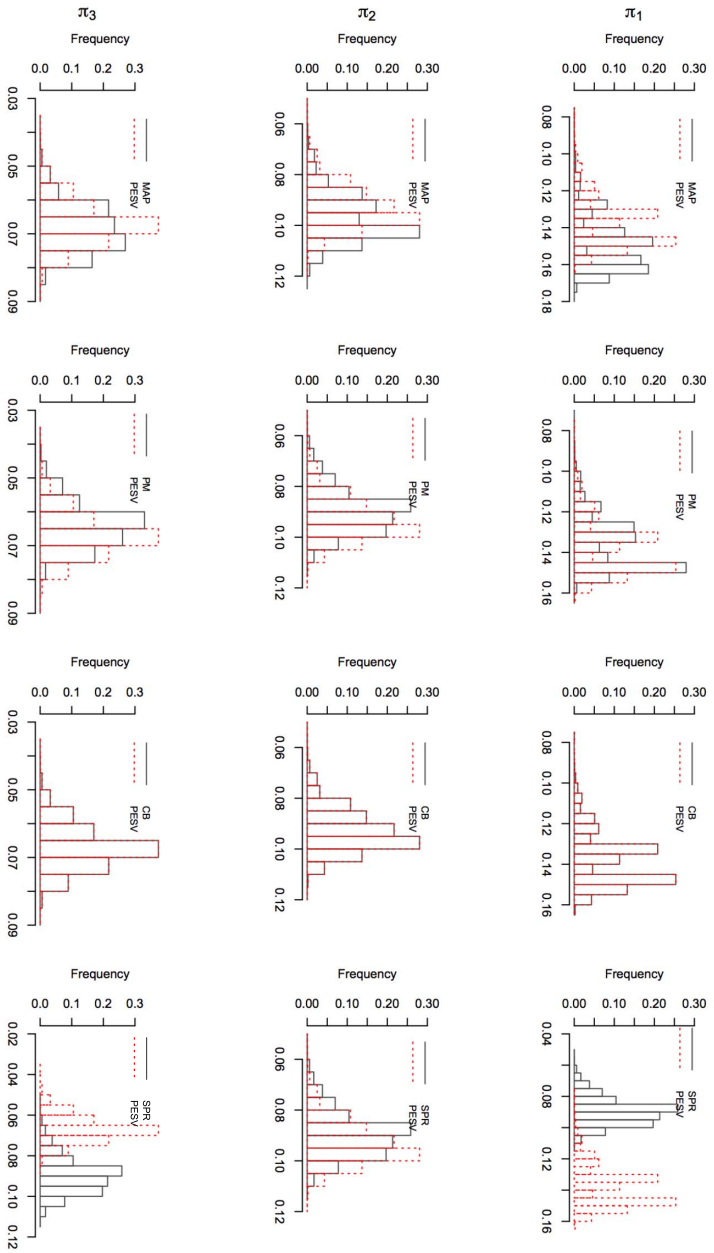

FIG 3. *Histograms of ASV of ensemble of the MAP, PM, CB estimates w.r.t. the priors* π_j, $j = 1, 2, 3$, *and SPRGM estimates w.r.t. the class of priors* Γ *along with histograms of the PESV of ensemble of the parameters in* $\boldsymbol{\theta}_{52}$. *Each row is associated with one of the priors* π_j, *as indicated on the y-axis of the first histograms.*

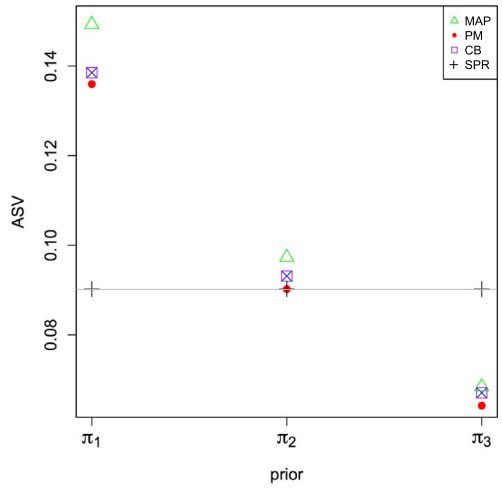

FIG 4. *Plots of ASV of the MAP, PM, CB estimates w.r.t. the priors* π_j, $j = 1, 2, 3$, *and SPRGM estimates w.r.t. the class of priors* Γ *along with the APESV of ensemble of the parameters in* $\boldsymbol{\theta}_{52}$. *In the figure,* $\times$ *represents PSEV. Also, green triangle corresponds to ASV of the MAP estimates, red dot refers to ASV of the PM estimates, purple square represents ASV of the CB estimates, and black plus sign corresponds to ASV of the SPRGM estimates.*

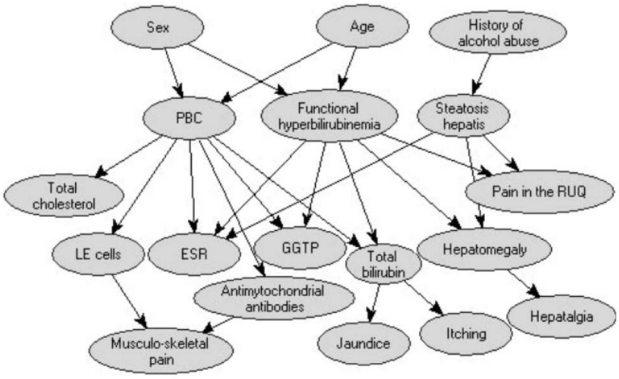

FIG 5. *The structure of the Hepar BN.*

The network models 18 variables related to diagnosis of a small set of hepatic disorders: three risk factors, 12 symptoms and test results, and three disorder nodes. To give the reader an idea of the number of numerical parameters needed to quantify a BN, let us assume for simplicity that each variable in the model in Fig. 5 is Binary.

We are interested in computing the probability $P(\text{PBC} \mid \text{Evidence})$, where 'PBC' stands for primary biliary cirrhosis, one of the possible liver diseases modelled in the network, and 'Evidence' would be a set of variables with their

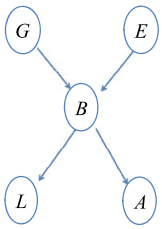

FIG 6. *Simplified Hepar BN model.*

values that pertain to PBC in some way. We will use this as an example to examine the effects of different parameter estimation methods. For example, LE_cells = 0 and Antimytochondrial AB = 1, Gender = female will already give a high probability if the Age is above 40 and these are indeed part of the characteristics of the disease according to the clinical literature.

For simplicity, we refer to PBC, LE_cells, Antimytochondrial AB, Gender and Age by B, L, A, G and E, respectively. The variables A, G and L are assumed to be binary. For the variable Age, we consider that E takes value 0 if the patient's age is under 40 and takes value 1, otherwise. Also, in our clinical data, B takes either the value zero (disease is absent) or one (disease is present). Thus, our goal is to compute

$$P(B = 1 \mid G = 0, E = 1, L = 0, A = 1). \tag{5.4}$$

Fig. 6 shows a simplified Hepar BN network with only these variables included. The following lemma restates the probability in (5.4) in terms of θ_{jil} defined in Eq. (2.2).

Lemma 5.1. *If we replace G, E, B, L, A by the variables $X_1, \ldots, X_5$ and their associated probabilistic parameters, the desired probability in (5.4) can be expressed as follows*

$$P(B = 1 \mid G = 0, E = 1, L = 0, A = 1) = \frac{\theta_{412}\theta_{522}\theta_{322}}{\theta_{411}\theta_{521}\theta_{312} + \theta_{412}\theta_{522}\theta_{322}} = \theta_D. \tag{5.5}$$

Suppose we know that the probability $\theta_{312} = P(B = 1|E > 40, A = 0)$ has a high value (from our prior experience), but we are unable to determine its exact value reliably based on the data available. From the data we first determine point estimates for the parameters in Eq. (5.5), i.e., we can at least propose a prior distribution by looking at possible estimates of θ_{312}, e.g. its ML estimate, which is 0.883. Based on this value, one may consider using the $Dir(\alpha_{312}, \alpha_{322})$-prior with $\alpha_{312} = 50$ and $\alpha_{322} = 5$, which gives a prior mean of 50/55=0.9091. However, this specific estimate might not be the same if we change the sample while it is obvious that a change in the available sample data would make a change in the point estimates. To make sure that the proposed prior is rich enough to include some other possible cases, one may consider the class Γ_1

below, which reflects the possibility of getting some estimates in the interval (40/48, 60/62).

By expressing the uncertainty of the parameters in terms of some classes of conjugate distributions as listed below, we make sure that a wider range of probabilities is covered.

$$\begin{aligned}
\Gamma_1 &= \Big\{Dir(\alpha_{312},\alpha_{322}) : 40 \le \alpha_{312} \le 60, 2 \le \alpha_{322} \le 8\Big\},\\
\Gamma_2 &= \Big\{Dir(\alpha_{411},\alpha_{421}) : 5 \le \alpha_{411} \le 25, 90 \le \alpha_{421} \le 110\Big\},\\
\Gamma_3 &= \Big\{Dir(\alpha_{412},\alpha_{422}) : 2 \le \alpha_{412} \le 8, 5 \le \alpha_{422} \le 15\Big\},\\
\Gamma_4 &= \Big\{Dir(\alpha_{511},\alpha_{521}) : 90 \le \alpha_{511} \le 110, 2 \le \alpha_{522} \le 4\Big\},\\
\Gamma_5 &= \Big\{Dir(\alpha_{512},\alpha_{522}) : 3 \le \alpha_{512} \le 8, 3 \le \alpha_{522} \le 8\Big\}.
\end{aligned}$$

One way to derive the SPRGM estimate of θ_D, is to compute SPRGM estimate for each of θ_{jil} as appeared in Eq. (5.5). The relevant computed estimates are shown in Table 2. It can be observed that the SPRGM estimate of the desired parameter θ_D is high enough, as somehow expected. For comparison, we also report on the corresponding ML estimates in Table 2. It should be emphasized that since ML estimates do not depend on the prior knowledge, comparing ML estimates and Bayesian estimates is not fair and we should rely on the ML estimates only in situations in which we do not have access to any source of prior information.

TABLE 2
Computed SPRGM estimates of the parameters appeared in the Eq. (5.5).

Estimates	θ_{312}	θ_{322}	θ_{411}	θ_{412}	θ_{521}	θ_{522}	θ_D
ML	0.8830	0.1170	0.1316	0.3071	0.0120	0.5679	0.9362
SPRGM	0.8889	0.1111	0.1317	0.3086	0.0200	0.0246	0.9150

6. Final remarks

In this paper we focused on discrete random variables. We would like to stress that this is common in BNs. For example, the well-known bnlearn software [39] is based on discrete random variables. One main reason is the fact that many BN learning algorithms are unable to treat efficiently continuous variables. However, as [5] reports, there are three common approaches of extending BNs to continuous variables introduced in the literature: one approach, introduced by [47], is to model the conditional pdf of each continuous random variable based on certain family of distribution first, and redesign the corresponding BN inference based on the parameterizations, next. Another approach is to use non-parametric distributions such as Gaussian processes [20], and the third approach would be to discretize the continuous variables based on some criteria such as the minimum description length. The third approach has been extensively used

in the literature and new developments have been introduced, see for example [17] among many others. Thus assuming the random variables are discrete is not a restrictive assumption.

We also highlight that in our developments we assumed a BN with a complete instantiation is available, meaning that no missing values are present. But we would like to emphasize that in the presence of incomplete/missing data it can be handled with one of the available methods in the literature. [8] provided a theoretical approach to handle the problem of learning with missing data. They show that one can solve this problem by taking a sum of the conditional probabilities over all posible values for each missing data point. [27] studied the parameter learning task in presence of some missing data based on the Expectation-Maximization (EM) technique. [36] applied the important sampling technique into solving such a problem.

Among the existing methods, we suggest using the EM algorithm due to its advantage of being easy to implement and having the property of converging relatively quickly [38].

Now, to apply the EM algorithm, suppose that in the kth sample, $k = 1, 2, \ldots, n$, of the variables in the set $\boldsymbol{x}_{(k)}$, X_m is the variable whose value is missing. The EM algorithm starts with an initial estimation θ_0 and at each iteration t, the data set is completed based on θ_t and then the parameters are re-estimated using the completed data set, obtaining θ_{t+1}. The E-step finds the conditional expectation of the complete data log-likelihood, given the observed component of the data and the current values of the parameters. In fact, the E-step computes the current expected log-likelihood of $\boldsymbol{\theta}$ given the data $\boldsymbol{x}$, as denoted by $Q(\boldsymbol{\theta}|\theta_t)$ for simplicity below

$$\begin{aligned} Q(\boldsymbol{\theta}|\theta_t) &= \sum_k \sum_{x_m} P(X_m = x_m|\boldsymbol{X}_{(k)} = \boldsymbol{x}_{(k)}, \theta_t) \log P(\boldsymbol{X}_{(k)} = \boldsymbol{x}_{(k)}, X_m = i|\boldsymbol{\theta}) \\ &= \sum_k \sum_{x_m} \sum_i \sum_j \sum_l P(X_m = x_m|\boldsymbol{X}_{(k)} = \boldsymbol{x}_{(k)}, \theta_t) n_{jilk} \log \theta_{jil} \\ &= \sum_i \sum_j \sum_l m^t_{jil} \log \theta_{jil}, \end{aligned}$$

where $m^t_{jil} = \sum_k \sum_{x_m} P(X_m = x_m|\boldsymbol{X}_{(k)} = \boldsymbol{x}_{(k)}, \theta_t)\, n_{jilk}$ in which n_{jilk} is given by the Eq. (2.1).

The M-step then computes the next estimate θ_t by maximizing the current expected log-likelihood of the data, i.e., $\theta_{t+1} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\theta_t)$. After some algebraic manipulations, for all i, j and k, we will get

$$\theta^{t+1}_{jil} = \frac{m^t_{jil}}{\sum_i m^t_{jil}}.$$

Here m^t_{jil} is interpreted as the number of cases where $X_j = i$ when its parent configuration is in the state λ_l in the completed data set. Thus, θ^{t+1}_{jil} is interpreted as the expected proportion of cases where $X_j = i$ among all possibilities when its parent configuration is in the state λ_l.

Since the EM algorithm converges [26, 38], this iterative approach leads to a replacement of θ_{jil} by θ^s_{jil}, where s is the time thereafter θ^t_{jil} is constant. Once this replacement is done, δ^{ML}_{jil} in Eq. (2.3) is derived. Since $n_{j.l.}$ will be known, we get $n_{jil.} = \delta^{ML}_{jil} n_{j.l.}$. Now, replacing the new value for $n_{jil.}$ in Eq. (2.5) and Eq. (2.6), as well as Theorems 3.1, 4.1 and 4.2 leads to MAP, PM, CB, and SPRGM estimates of parameters associated with the variable whose value is missing.

7. Conclusions and discussion

In this paper we considered the task of parameter learning in BNs. Improvements of Bayesian methods were provided, leading to the extension and application of the simultaneous estimation and robust Bayesian methodology to the context of parameter learning in BNs.

Assuming accessibility of some prior knowledge, we dealt with different approaches to incorporate prior knowledge and derived explicit forms of Bayes (MAP and PM), adjusted Bayes (CB) and robust Bayes (SPRGM) estimates. From the Bayesian estimation literature it is understood that, in presence of crisp prior knowledge, one can reach a reliable Bayes estimate for the desired parameter. Prior knowledge can be specified by determining hyperparameters of the underlying prior distribution, but in many situations there may be a lack of consensus among experts or decision-makers concerning these hyperparameters. In such situations, one sensible approach, as adopted in this paper, would be to define a class of priors to ensure that the existing knowledge fall within the proposed class. The corresponding rule, which we referred to as the 'robust Bayes rule', can be used in the hope of arriving at a rule consistent with the real world.

Our simulation study emphasizes that if the crisp prior is present, Bayes and CB rules are reliable methods. This was obvious from the choice $Dir(40, 10)$ and $Dir(35, 10)$-priors and the corresponding Bayes and CB estimates in Table A.1 of the Appendix, as the true parameter was 0.2 and 0.4, respectively. But it is seen that for the other specified priors, the resulting Bayes and CB estimates are quite far from the true parameters and thus, these selected priors are bad choices. However, as noted earlier in the simulation study, in practice, the availability of exact prior knowledge in terms of specific prior hyperparameters is rare. The overall class (5.1) was rich enough to ensure that it includes all the prior information attributed by the three experts. In addition to prevention of selecting bad choices of priors, quantitative statistics show that the SPRGM estimates perform quite well.

We emphasize that when the values of hyperparameters are not justifiably chosen, or when the exact prior knowledge is not available, SPRGM estimates outperform Bayes rules, as we should expect due to the fact that robust rules are aimed at global prevention of bad choices in a single prior. Obviously, for a justified choice of a single prior the results may reverse in the sense that for such a prior, the Bayes estimate outperforms robust Bayes rules. When specific hyperparameters of a prior are available, we encourage the use of MAP and

PM estimates. We encourage using the CB estimates only if the interest lies in both simultaneous estimation and closeness between distribution of estimates and posterior distribution of the parameters. When there is a lack of consensus of opinion about the prior hyperparameters, we encourage using the SPRGM estimate(s), with the hope of reaching an optimal estimate.

We would like to wrap up this work by addressing the main interest of Bayesian analysis considered in this paper. Although different prior-based point estimates of the desired parameters have been provided in this paper, the points estimates have been driven by recovering the posterior distribution. The MAP and PM rules are the points that minimize the posterior function which is informally averages of losses of choosing an estimator of the desired parameter w.r.t. the posterior distribution, the CB estimates adjust the PM estimates according to the additional constraints (i)-(iii) of Section 3 and the SPRGM estimates minimize the difference between posterior risk of any arbitrary estimator and the posterior risk of the Bayes estimator.

Appendix

Proof of Lemma 3.1.

$$
\begin{aligned}
& E\Big[\sum_{i=1}^{k_j}\big(\theta_{jil}-\bar{\theta}_{j.l}\big)^2\,|\boldsymbol{X}=\boldsymbol{x}\Big] \\
= & \sum_{i=1}^{k_j} E[\theta_{jil}^2|\boldsymbol{X}=\boldsymbol{x}]-k_jE[\bar{\theta}_{j.l}^2|\boldsymbol{X}=\boldsymbol{x}] \\
= & \sum_{i=1}^{k_j} E[\theta_{jil}^2|\boldsymbol{X}=\boldsymbol{x}]-k_jVar[\bar{\theta}_{j.l}|\boldsymbol{X}=\boldsymbol{x}]-k_jE^2[\bar{\theta}_{j.l}|\boldsymbol{X}=\boldsymbol{x}] \\
= & \sum_{i=1}^{k_j} E[\theta_{jil}^2|\boldsymbol{X}=\boldsymbol{x}]-\frac{1}{k_j} \qquad (\text{since } \bar{\theta}_{j.l}=\frac{1}{k_j}) \\
> & \sum_{i=1}^{k_j} E^2[\theta_{jil}|\boldsymbol{X}=\boldsymbol{x}]-\frac{1}{k_j} \qquad (\text{by Jensen inequality}) \\
= & \sum_{i=1}^{k_j} E^2[\theta_{jil}|\boldsymbol{X}=\boldsymbol{x}]-k_j\bar{\delta}_{jl}^{PM^2} \qquad (\text{since } \bar{\delta}_{jl}^{PM}=\frac{1}{k_j}\sum_{i=1}^{k_j}\delta_{jil}^{PM}=\frac{1}{k_j}) \\
= & \sum_{i=1}^{k_j} \delta_{jil}^{PM^2}-k_j\bar{\delta}_{j.l}^{PM^2} \\
= & \sum_{i=1}^{k_j}\big(\delta_{jil}^{PM}-\bar{\delta}_{j.l}^{PM}\big)^2 .
\end{aligned}
$$

Hence,

$$E\Big[\sum_{i=1}^{k_j}\left(\theta_{jil}-\bar{\theta}_{j.l}\right)^2|\boldsymbol{X}=\boldsymbol{x}\Big]>\sum_{i=1}^{k_j}\left(\delta_{jil}^{PM}-\bar{\delta}_{jl}^{PM}\right)^2. \qquad \square$$

Proof of Theorem 3.1. To derive CB estimates of the elements of $\boldsymbol{\theta}_{jl}$, we minimize

$$E\Big[\sum_{i=1}^{k_j}\left(\theta_{jil}-\delta_{jil}\right)^2|\boldsymbol{X}=\boldsymbol{x}\Big],$$

w.r.t. δ_{jil} subject to (i)-(iii). First note that

$$\begin{aligned}E\Big[\sum_{i=1}^{k_j}\left(\theta_{jil}-\delta_{jil}\right)^2|\boldsymbol{X}=\boldsymbol{x}\Big] &= E\Big[\sum_{i=1}^{k_j}\left(\theta_{jil}+\delta_{jil}^{PM}-\delta_{jil}^{PM}-\delta_{jil}\right)^2|\boldsymbol{X}=\boldsymbol{x}\Big]\\ = E\Big[\sum_{i=1}^{k_j}\left(\theta_{jil}-\delta_{jil}^{PM}\right)^2|\boldsymbol{X}=\boldsymbol{x}\Big]&+\sum_{i=1}^{k_j}\left(\delta_{jil}-\delta_{jil}^{PM}\right)^2. \qquad (A.1)\end{aligned}$$

The first term in the RHS of (A.1) does not depend on the estimates δ_{jil}. Hence, minimizing $E[\sum_{i=1}^{k_j}\left(\theta_{jil}-\delta_{jil}\right)^2|\boldsymbol{X}=\boldsymbol{x}]$ subject to the constraints (i)-(iii) is equivalent to minimizing $\sum_{i=1}^{k_j}\left(\delta_{jil}-\delta_{jil}^{PM}\right)^2$ subject to the conditions (i)-(iii). From the constraint (i), $\sum_{i=1}^{k_j}\delta_{jil}=\sum_{i=1}^{k_j}\delta_{jil}^{PM}$, we observe that

$$\begin{aligned}\sum_{i=1}^{k_j}\left(\delta_{jil}-\delta_{jil}^{PM}\right)^2 &= \sum_{i=1}^{k_j}\left(\delta_{jil}-\bar{\delta}_{j.l}+\bar{\delta}_{j.l}^{PM}-\delta_{jil}^{PM}\right)^2\\ &= \sum_{i=1}^{k_j}\left(\delta_{jil}-\bar{\delta}_{j.l}\right)^2+\sum_{i=1}^{k_j}\left(\delta_{jil}^{PM}-\bar{\delta}_{j.l}^{PM}\right)^2\\ &\quad -2\sum_{i=1}^{k_j}\left(\delta_{jil}^{PM}-\bar{\delta}_{j.l}^{PM}\right)\left(\delta_{jil}-\bar{\delta}_{j.l}\right)\\ &= k_j\left(Var(Z_{jl})+Var(W_{jl})-2Cov(Z_{jl},W_{jl})\right), \qquad (A.2)\end{aligned}$$

where for a fixed $j=1,\ldots,d$ and $l=1,\ldots,q_j$,

$$P(Z_{jl}=\delta_{jil},W_{jl}=\delta_{jil}^{PM})=\frac{1}{k_j},\qquad i=1,\ldots,k_j.$$

Due to the constraint (ii), for a fixed $j=1,\ldots,d$ and $l=1,\ldots,q_j$, $Var(Z_{jl})$ is constant. It is obvious that $Var(W_{jl})$ does not depend on δ_{jil} values. Thus, the right side of (A.2) is minimized when $Cov(Z_{jl},W_{jl})=\sqrt{Var(Z_{jl})}\sqrt{Var(W_{jl})}$ or equivalently the corresponding correlation is equal to one, i.e., $\rho(Z_{jl},W_{jl})=1$. This implies that $W_{jl}=a_{jl}Z_{jl}+b_{jl}$ with probability 1 for some $a_{jl}>0$ and $b_{jl}\in\Re$. Thus,

$$\delta_{jil} = a_{jl}\delta_{jil}^{PM}+b_{jl}. \qquad (A.3)$$

Hence by taking sum over i from both sides we have

$$\sum_{i=1}^{k_j}\delta_{jil} \quad = \quad a_{jl}\sum_{i=1}^{k_j}\delta_{jil}^{PM}+b_{jl}k_j$$

which using the constraints (i) and (iii) leads to

$$b_{jl}=(1-a_{jl})\frac{1}{k_j}. \tag{A.4}$$

Substituting (A.4) in (A.3) and the fact that $\sum_{i=1}^{k_j}\delta_{jil}=\sum_{i=1}^{k_j}\delta_{jil}^{PM}$, leads to

$$\delta_{jil}=a_{jl}\delta_{jil}^{PM}+(1-a_{jl})\frac{1}{k_j}. \tag{A.5}$$

or

$$\sum_{i=1}^{k_j}(\delta_{jil}-\bar{\delta}_{j.l})^2 \quad = \quad a_{jl}^2\sum_{i=1}^{k_j}(\delta_{jil}^{PM}-\bar{\delta}_{j.l}^{PM})^2. \tag{A.6}$$

Now, combining (A.6) and the constraint (ii), we set

$$a_{jl}=\left\{\frac{G_{jl}(\boldsymbol{x})}{H_{jl}(\boldsymbol{x})}\right\}^{\frac{1}{2}}, \tag{A.7}$$

where $G_{jl}(\boldsymbol{x})=E[\sum_{i=1}^{k_j}(\theta_{jil}-\bar{\theta}_{j.l})^2|\boldsymbol{X}=\boldsymbol{x}]$ and $H_{jl}(\boldsymbol{x})=\sum_{i=1}^{k_j}(\delta_{jil}^{PM}-\bar{\delta}_{j.l}^{PM})^2$. Notice that due to inherent symmetry in BNs, i.e., the fact $\sum_{i=1}^{k_j}\theta_{jil}=1$ or equivalently $\bar{\theta}_{j.l}=\frac{1}{k_j}$, $G_{jl}(\boldsymbol{x})$ can be simplified as follows

$$\begin{aligned} G_{jl}(\boldsymbol{x}) &= E\Big[\sum_{i=1}^{k_j}(\theta_{jil}-\frac{1}{k_j})^2|\boldsymbol{X}=\boldsymbol{x}\Big] \\ &= \sum_{i=1}^{k_j}E[\theta_{jil}^2|\boldsymbol{X}=\boldsymbol{x}]+\frac{1}{k_j}-\frac{2}{k_j}\sum_{i=1}^{k_j}\theta_{jil} \\ &= S_{jl}(\boldsymbol{x})-\frac{1}{k_j}, \end{aligned} \tag{A.8}$$

where $S_{jl}(\boldsymbol{x})=\sum_{i=1}^{k_j}E[\theta_{jil}^2|\boldsymbol{X}=\boldsymbol{x}]$.
To simplify $H_{jl}(\boldsymbol{x})$, note that due to the constraint (iii), we have $\bar{\delta}_{jil}^{PM}=\frac{1}{k_j}\sum_{i=1}^{k_j}\delta_{jil}^{PM}=\frac{1}{k_j}$. Thus,

$$\begin{aligned} H_{jl}(\boldsymbol{x}) &= E\Big[\sum_{i=1}^{k_j}(\delta_{jil}^{PM}-\frac{1}{k_j})^2|\boldsymbol{X}=\boldsymbol{x}\Big] \\ &= \sum_{i=1}^{k_j}E[(\delta_{jil}^{PM})^2|\boldsymbol{X}=\boldsymbol{x}]+\frac{1}{k_j}-\frac{2}{k_j}\sum_{i=1}^{k_j}\delta_{jil}^{PM} \\ &= T_{jl}(\boldsymbol{x})-\frac{1}{k_j}. \end{aligned} \tag{A.9}$$

where $T_{jl}(\boldsymbol{x}) = \sum_{i=1}^{k_j} E[\theta_{jil}^2 | \boldsymbol{X} = \boldsymbol{x}]$. Substituting (A.8) and (A.9) in (A.7) and combining (A.7) and (A.5) the proof is complete. □

Proof of Theorem 4.1. For each $i = 1, 2$, $r_p(\delta_{jil}, \delta_{jil}^{PM})$ is a convex function of δ_{jil}^{PM} and attains its maximum at either $\delta_{jil}^{PM} = \underline{\delta}_{jil}$ or $\delta_{jil}^{PM} = \overline{\delta}_{jil}$. Following the four possible cases, we obtain the SPRGM estimates subject to the constraint $\delta_{j1l} + \delta_{j2l} = 1$.

i) $r_p(\delta_{j1l}, \overline{\delta}_{j1l}) \geq r_p(\delta_{j1l}, \underline{\delta}_{j1l})$ and $r_p(\delta_{j2l}, \overline{\delta}_{j2l}) \geq r_p(\delta_{j2l}, \underline{\delta}_{j2l})$. Then, $(\delta_{j1l}, \delta_{j2l})$ belongs to the following class of estimates

$$\mathcal{D}_1 = \left\{ (\delta_{j1l}, \delta_{j2l}) : \delta_{j1l} \leq \frac{\underline{\delta}_{j1l} + \overline{\delta}_{j1l}}{2}, \delta_{j2l} \leq \frac{\underline{\delta}_{j2l} + \overline{\delta}_{j2l}}{2} \right\}.$$

Using the constraint $\delta_{j1l} + \delta_{j2l} = 1$, we observe that $r_p(\boldsymbol{\delta}_{jl}, \boldsymbol{\delta}_{jl}^{PM}) = (\delta_{j1l} - \overline{\delta}_{j1l})^2 + (\delta_{j1l} + \overline{\delta}_{j2l} - 1)^2$, which obviously is convex in δ_{j1l} and has a unique minimum at $\delta_{j1l} = \frac{1}{2}\left(1 + \overline{\delta}_{j1l} - \overline{\delta}_{j2l}\right)$. It is easy to observe that this solution satisfies the conditions in $\mathcal{D}_1$ if $\underline{\delta}_{j1l} + \overline{\delta}_{j2l} \geq 1$ and $\overline{\delta}_{j1l} + \underline{\delta}_{j2l} \geq 1$. So, $\delta_{j1l,\Gamma}^{SPR} = \frac{1}{2}\left(1 + \overline{\delta}_{j1l} - \overline{\delta}_{j2l}\right)$ and $\delta_{j2l,\Gamma}^{SPR} = 1 - \delta_{j1l,\Gamma}^{SPR}$, provided $\underline{\delta}_{j1l} + \overline{\delta}_{j2l} \geq 1$ and $\overline{\delta}_{j1l} + \underline{\delta}_{j2l} \geq 1$.

ii) $r_p(\delta_{j1l}, \overline{\delta}_{j1l}) \geq r_p(\delta_{j1l}, \underline{\delta}_{j1l})$ and $r_p(\delta_{j2l}, \underline{\delta}_{j2l}) \geq r_p(\delta_{j2l}, \overline{\delta}_{j2l})$. Then, $(\delta_{j1l}, \delta_{j2l})$ belongs to the following class of estimates

$$\mathcal{D}_2 = \left\{ (\delta_{j1l}, \delta_{j2l}) : \delta_{j1l} \leq \frac{\underline{\delta}_{j1l} + \overline{\delta}_{j1l}}{2}, \delta_{j2l} \geq \frac{\underline{\delta}_{j2l} + \overline{\delta}_{j2l}}{2} \right\}.$$

Use the constraint $\delta_{j1l} + \delta_{j2l} = 1$, $r_p(\boldsymbol{\delta}_{jl}, \boldsymbol{\delta}_{jl}^{PM}) = (\delta_{j1l} - \overline{\delta}_{j1l})^2 + (\delta_{j1l} + \underline{\delta}_{j2l} - 1)^2$. $r_p(\boldsymbol{\delta}_{jl}, \boldsymbol{\delta}_{jl}^{PM})$ is convex in δ_{j1l} and has a unique minimum at $\delta_{j1l} = \frac{1}{2}\left(1 + \overline{\delta}_{j1l} - \underline{\delta}_{j2l}\right)$. This solution would be SPRGM estimate of θ_{j1l} if it belongs to $\mathcal{D}_2$. It is easy to verify that this is not possible and hence, this case does not lead to any SPRGM estimate.

iii) $r_p(\delta_{j1l}, \underline{\delta}_{j1l}) \geq r_p(\delta_{j1l}, \overline{\delta}_{j1l})$ and $r_p(\delta_{j2l}, \overline{\delta}_{j2l}) \geq r_p(\delta_{j2l}, \underline{\delta}_{j2l})$. Then, $(\delta_{j1l}, \delta_{j2l})$ belong to the following class of estimates

$$\mathcal{D}_3 = \left\{ (\delta_{j1l}, \delta_{j2l}) : \delta_{j1l} \geq \frac{\underline{\delta}_{j1l} + \overline{\delta}_{j1l}}{2}, \delta_{j2l} \leq \frac{\underline{\delta}_{j2l} + \overline{\delta}_{j2l}}{2} \right\}.$$

Using the constraint $\delta_{j1l} + \delta_{j2l} = 1$, $r_p(\boldsymbol{\delta}_{jl}, \boldsymbol{\delta}_{jl}^{PM}) = (\delta_{j1l} - \underline{\delta}_{j1l})^2 + (\delta_{j1l} + \overline{\delta}_{j2l} - 1)^2$, which is convex in δ_{j1l} and has a unique minimum at $\delta_{j1l} = \frac{1}{2}\left(1 + \underline{\delta}_{j1l} - \overline{\delta}_{j2l}\right)$. Similar to the case (ii), it is easy to verify that this case does not lead to any SPRGM solution.

iv) $r_p(\delta_{j1l}, \underline{\delta}_{j1l}) \geq r_p(\delta_{j1l}, \overline{\delta}_{j1l})$ and $r_p(\delta_{j2l}, \underline{\delta}_{j2l}) \geq r_p(\delta_{j2l}, \overline{\delta}_{j2l})$. Then, $(\delta_{j1l}, \delta_{j2l})$ belongs to the following class of estimates

$$\mathcal{D}_4 = \left\{ (\delta_{j1l}, \delta_{j2l}) : \delta_{j1l} \geq \frac{\underline{\delta}_{j1l} + \overline{\delta}_{j1l}}{2}, \delta_{j2l} \geq \frac{\underline{\delta}_{j2l} + \overline{\delta}_{j2l}}{2} \right\}.$$

Using the constraint $\delta_{j1l}+\delta_{j2l}=1$, we observe that $r_p(\boldsymbol{\delta}_{jl},\boldsymbol{\delta}_{jl}^{PM})=(\delta_{j1l}-\underline{\delta}_{j1l})^2+(\delta_{j1l}+\underline{\delta}_{j2l}-1)^2$, which is convex in δ_{j1l} and has a unique minimum at $\delta_{j1l}=\frac{1}{2}\left(1+\underline{\delta}_{j1l}-\underline{\delta}_{j2l}\right)$. It is easy to observe that this solution satisfies the conditions in $\mathcal{D}_4$ if $\underline{\delta}_{j1l}+\overline{\delta}_{j2l}\leq 1$ and $\overline{\delta}_{j1l}+\underline{\delta}_{j2l}\leq 1$. So, $\delta_{j1l,\Gamma}^{SPR}=\frac{1}{2}\left(1+\underline{\delta}_{j1l}-\underline{\delta}_{j2l}\right)$ and $\delta_{j2l,\Gamma}^{SPR}=1-\delta_{j1l,\Gamma}^{SPR}$, provided that $\underline{\delta}_{j1l}+\overline{\delta}_{j2l}\leq 1$ and $\overline{\delta}_{j1l}+\underline{\delta}_{j2l}\leq 1$. □

Proof of Theorem 4.2. For each $i=1,2,3$, $r_p(\delta_{jil},\delta_{jil}^{PM})$ is a convex function of δ_{jil}^{PM} and attains its maximum at either $\delta_{jil}^{PM}=\underline{\delta}_{jil}$ or $\delta_{jil}^{PM}=\overline{\delta}_{jil}$. Following the eight possible cases, we obtain the SPRGM estimates subject to the constraint $\delta_{j1l}+\delta_{j2l}+\delta_{j3l}=1$. We only prove (i), the proof of (ii)-(viii) is similar to (i).
Suppose $r_p(\delta_{jil},\overline{\delta}_{j1l})\geq r_p(\delta_{jil},\underline{\delta}_{j1l})$, $i=1,2,3$, Then, $(\delta_{j1l},\delta_{j2l},\delta_{j3l})$ belongs to the following class of estimates

$$\mathcal{D}_1^*=\left\{(\delta_{j1l},\delta_{j2l},\delta_{j3l}):\delta_{j1l}\leq\frac{\underline{\delta}_{j1l}+\overline{\delta}_{j1l}}{2},\delta_{j2l}\leq\frac{\underline{\delta}_{j2l}+\overline{\delta}_{j2l}}{2},\delta_{j3l}\leq\frac{\underline{\delta}_{j3l}+\overline{\delta}_{j3l}}{2}\right\}.$$

Using the constraint $\delta_{j1l}+\delta_{j2l}+\delta_{j3l}=1$, we observe that $r_p(\boldsymbol{\delta}_{jl},\boldsymbol{\delta}_{jl}^{PM})=(\delta_{j1l}-\overline{\delta}_{j1l})^2+(\delta_{j2l}-\overline{\delta}_{j2l})^2+(1-\delta_{j1l}-\delta_{j2l}-\overline{\delta}_{j3l})^2$. Based on the second partials test [46], one can verify that infimum of $r_p(\boldsymbol{\delta}_{jl},\boldsymbol{\delta}_{jl}^{PM})$ is achieved at $\delta_{j1l}=\frac{1}{3}(1+2\overline{\delta}_{j1l}-\overline{\delta}_{j2l}-\overline{\delta}_{j3l})$ and $\delta_{j2l}=\frac{1}{3}(1+2\overline{\delta}_{j2l}-\overline{\delta}_{j1l}-\overline{\delta}_{j3l})$. These solutions are the SPRGM estimates if they belong to $\mathcal{D}_1^*$. Thus, $\delta_{j1l,\Gamma}^{SPR}=\frac{1}{3}(1+2\overline{\delta}_{j1l}-\overline{\delta}_{j2l}-\overline{\delta}_{j3l})$,$\delta_{j2l,\Gamma}^{SPR}=\frac{1}{3}(1+2\overline{\delta}_{j2l}-\overline{\delta}_{j1l}-\overline{\delta}_{j3l})$ and $\delta_{j3l,\Gamma}^{SPR}=1-\delta_{j1l,\Gamma}^{SPR}-\delta_{j2l,\Gamma}^{SPR}$ provided that $\delta_{j1l,\Gamma}^{SPR}\leq\frac{1}{2}(\underline{\delta}_{jil}+\overline{\delta}_{jil})$, $i=1,2,3$. □

Proof of Lemma 5.1. First notice that using the Bayes' rule we have

$$P(B=1\,|\,G=0,E=1,L=0,A=1)=\frac{P(B=1,G=0,E=1,L=0,A=1)}{P(G=0,E=1,L=0,A=1)}. \tag{A.10}$$

The numerator of (A.10) can be written as

$$\begin{aligned}&P(B=1,G=0,E=1,L=0,A=1)\\&\quad=P(L=0,A=1\mid B=1,G=0,E=1)P(B=1\mid G=0,E=1)P(G=0,E=1)\\&\quad=P(L=0\mid B=1)P(A=1\mid B=1)P(B=1\mid G=0,E=1)P(G=0,E=1).\end{aligned} \tag{A.11}$$

Similar to the derivation of the numerator, one can easily derive that

$$\begin{aligned}&P(G=0,E=1,L=0,A=1)\\&=\sum_{b\in\{0,1\}}P(L=0\mid B=b)P(A=1\mid B=b)P(B=b\mid G=0,E=1)P(G=0,E=1)\end{aligned} \tag{A.12}$$

Hence, substituting (A.11) and (A.12) in (A.10), we obtain that

$$P(B=1\mid G=0,E=1,L=0,A=1)$$

$$= \frac{P(L=0 \mid B=1)P(A=1 \mid B=1)P(B=1 \mid G=0,E=1)P(G=0,E=1)}{\sum_{b\in\{0,1\}} P(L=0 \mid B=b)P(A=1 \mid B=b)P(B=b \mid G=0,E=1)P(G=0,E=1)}$$
$$= \frac{P(L=0 \mid B=1)P(A=1 \mid B=1)P(B=1 \mid G=0,E=1)}{\sum_{b\in\{0,1\}} P(L=0 \mid B=b)P(A=1 \mid B=b)P(B=b \mid G=0,E=1)}.$$

Now, if we replace G, E, B, L, A by the variables $X_1, \ldots, X_5$ and their associated probabilistic parameters, the proof is complete. □

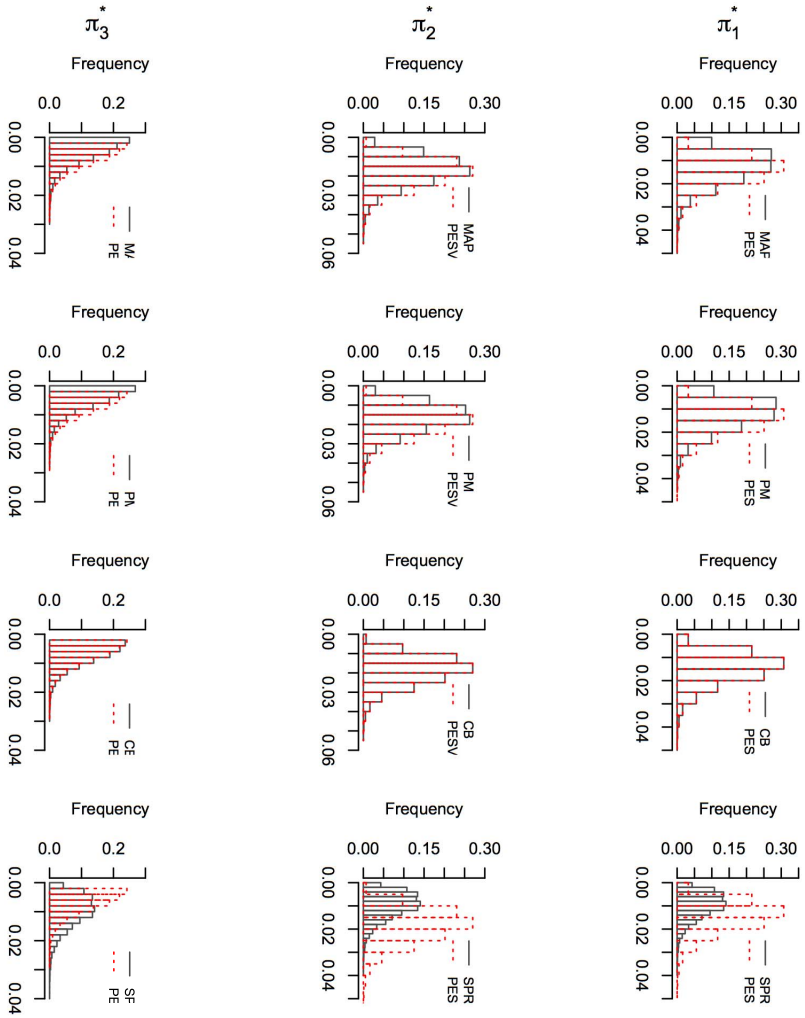

FIG A.1. *Histograms of ASV of ensemble of the MAP, PM, CB estimates w.r.t. the priors* π_j^*, $j = 1, 2, 3$, *and SPRGM estimates w.r.t. the class of priors* Γ^* *along with histograms of the PESV of ensemble of the parameters in* $\boldsymbol{\theta}_{51}$. *Each row is associated with one of the priors* π_j^*, *as indicated on the y-axis of the first histograms.*

TABLE A.1. *Quantitative statistics for different values of n.*

	n	i	δ_{5i1}^{ML}	$\delta_{5i1}^{MAP,\pi_1^*}$	$\delta_{5i1}^{MAP,\pi_2^*}$	$\delta_{5i1}^{MAP,\pi_3^*}$	$\delta_{5i1}^{PM,\pi_1^*}$	$\delta_{5i1}^{PM,\pi_2^*}$	$\delta_{5i1}^{PM,\pi_3^*}$	$\delta_{5i1}^{CB,\pi_1^*}$	$\delta_{5i1}^{CB,\pi_2^*}$	$\delta_{5i1}^{CB,\pi_3^*}$	$\delta_{5i1,\Gamma^*}^{SPRGM}$
Mean	25	1	0.4008	0.3862	0.3650	0.4443	0.3888	0.3679	0.4456	0.3768	0.3585	0.4291	0.4067
		2	0.5992	0.6138	0.6350	0.5557	0.6f112	0.6321	0.5544	0.6232	0.6415	0.5709	0.5933
AKLD			0.0340	0.0029	0.0059	0.0079	0.0026	0.0052	0.0082	0.0032	0.0071	0.0034	0.0022
ASV			0.0203	0.0137	0.0189	0.0038	0.0131	0.0181	0.0037	0.0157	0.0206	0.0064	0.0094
Mean	50	1	0.3993	0.3887	0.3716	0.4346	0.3907	0.3739	0.4358	0.3806	0.3657	0.4192	0.4048
		2	0.6007	0.6113	0.6284	0.5654	0.6093	0.6261	0.5642	0.6194	0.6343	0.5808	0.5952
AKLD			0.0163	0.0032	0.0052	0.0062	0.0030	0.0047	0.0064	0.0034	0.0059	0.0032	0.0026
ASV			0.0154	0.0133	0.0173	0.0052	0.0128	0.0167	0.0050	0.0150	0.0187	0.0072	0.0099
Mean	100	1	0.4002	0.3924	0.3801	0.4247	0.3938	0.3816	0.4256	0.3863	0.3752	0.4146	0.4036
		2	0.5998	0.6076	0.6199	0.5753	0.6062	0.6184	0.5744	0.6137	0.6248	0.5854	0.5964
AKLD			0.0081	0.0030	0.0039	0.0045	0.0029	0.0037	0.0046	0.0029	0.0042	0.0027	0.0027
ASV			0.0126	0.0125	0.0152	0.0066	0.0122	0.0149	0.0064	0.0137	0.0163	0.0080	0.0102
Mean	200	1	0.4006	0.3956	0.3877	0.4158	0.3964	0.3886	0.4165	0.3916	0.3843	0.4102	0.4026
		2	0.5994	0.6044	0.6123	0.5842	0.6036	0.6114	0.5835	0.6084	0.6157	0.5898	0.5974
AKLD			0.0040	0.0023	0.0026	0.0029	0.0022	0.0025	0.0029	0.0022	0.0027	0.0021	0.0021
ASV			0.0112	0.0116	0.0133	0.0078	0.0114	0.0131	0.0077	0.0124	0.0140	0.0087	0.0102

The priors π_1^*, π_2^* and π_3^* stand for $Dir(40, 25)$, $Dir(45, 25)$ and $Dir(35, 30)$-priors and Γ^* stands for the class of priors in (5.2).

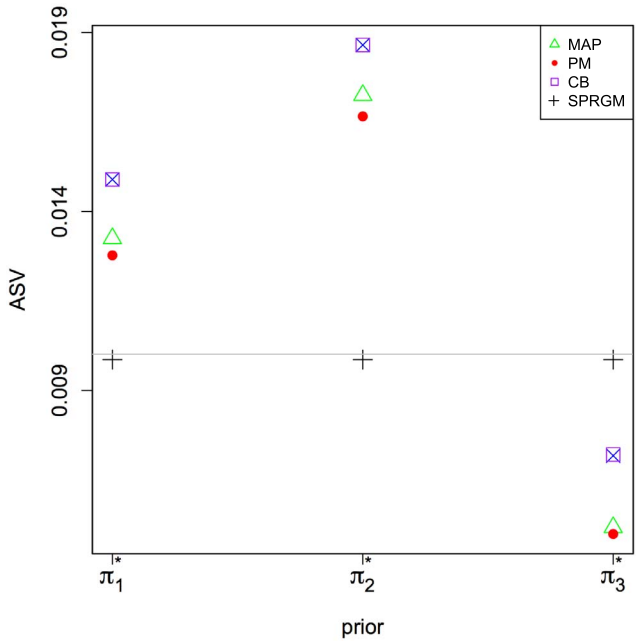

FIG A.2. *Plots of ASV of the MAP, PM, CB estimates w.r.t. the priors* π_j^*, $j = 1, 2, 3$, *and SPRGM estimates w.r.t. the class of priors* Γ^* *along with the APESV of ensemble of the parameters in* $\boldsymbol{\theta}_{51}$. *In the figure,* $\times$ *represents PSEV. Also, green triangle corresponds to ASV of the MAP estimates, red dot refers to ASV of the PM estimates, purple square represents ASV of the CB estimates, and black plus sign corresponds to ASV of the SPRGM estimates.*

References

[1] BERGER, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis.* Springer Science and Business Media. MR0804611

[2] BERGER, J. O. (1990). Robust Bayesian analysis: sensitivity to the prior. *J. Stat. Plan. Infer.* **25**(3) 303-328. MR1064429

[3] BERGER, J. O. (1994). An overview of robust Bayesian analysis. *Test* **3**(1) 5-124. MR1293110

[4] BUNTINE, W. (1991). Theory refinement on Bayesian networks. *In: Proceedings of the Seventh conference on Uncertainty in Artificial Intelligence*, 52-60. Morgan Kaufmann Publishers Inc.

[5] CHEN, Y. C., WHEELER, T. A. AND KOCHENDERFER, M. J. (2017). Learning discrete Bayesian networks from continuous data. *Artif. Intell. Res.* **59** 103-132 MR3670488

[6] CHENG, J., GREINER, R., KELLY, J., BELL, D. AND LIU, W. (2002). Learning Bayesian networks from data: an information-theory based approach. *Artif. Intell.* **137**(1-2) 43-90. MR1906473

[7] COOPER, G. F. (1989). Current research directions in the development of expert systems based on belief networks. *Appl. Stoch. Model. Data Anal.*

5(1) 39-52.

[8] Cooper, G. F. and Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Mach. Learn.* **9**(4) 309-347.

[9] de Campos, C. and Qiang, J. (2008). Improving Bayesian network parameter learning using constraints. *In: Proceedings of the 19th International Conference on Pattern Recognition*, 1-4.

[10] de Campos, L. M. (2006). A scoring function for learning Bayesian networks based on mutual information and conditional independence tests. *J. Mach. Learn. Res.* **7** 2149-2187. MR2274436

[11] Frey, J. and Cressie, N. (2003). Some results on constrained Bayes estimators. *Stat. Prob. Lett.* **65**(4) 389-399. MR2039883

[12] Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2014). *Bayesian Data Analysis*. Vol. 2. Chapman and Hall/CRC Boca Raton, FL, USA. MR3235677

[13] Ghosh, M.(1992). Constrained Bayes estimation with applications. *J. Am. Stat. Assoc.* **87**(418) 533-540. MR1173817

[14] Ghosh, M., Joon Kim, M. and Ho Kim, D. (2008). Constrained Bayes and empirical Bayes estimation under random effects normal ANOVA model with balanced loss function. *J. Stat. Plan. Infer.* **138**(7) 2017-2028. MR2406422

[15] Ghosh, M., Kim, M. J. and Kim, D. (2007). Constrained Bayes and empirical Bayes estimation with balanced loss functions. *Commun. Stat. Theory Methods* **36**(8) 1527-1542. MR2396441

[16] Ghosh, M. and Maiti, T. (1999). Adjusted Bayes estimators with applications to small area estimation. *Sankhya: Indian J. Stat., Series B* 71-90. MR1720722

[17] Grünwald, P. D. (2007). *The Minimum Description Length Principle*. MIT press.

[18] Heckerman, D., Geiger, D. and Chickering, D. M. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Mach. Learn.* **20**(3) 197-243. MR1615024

[19] Insua, D. R. and Ruggeri, F. (2000). *Robust Bayesian Analysis*. Vol. 152. Springer. MR1795206

[20] Ickstadt, K., Bornkamp, B., Grzegorczyk, M., Wieczorek, J., Sheriff, M. R., Grecco, H. E. and Zamir, E. (2011). Nonparametric Bayesian networks. *Bayesian Stat.* **9** 283. MR3204010

[21] Karimnezhad, A. and Moradi, F. (2016). Bayes, E-Bayes and robust Bayes prediction of a future observation under precautionary prediction loss functions with applications. *Appl. Math. Model.* **40**(15) 7051-7061. MR3508029

[22] Karimnezhad, A., Niazi, S., and Parsian, A. (2014). Bayes and robust Bayes prediction with an application to a rainfall prediction problem. *J. Korean Stat. Soci.* **43**(2) 275-291. MR3188368

[23] Karimnezhad, A. and Parsian, A. (2014). Robust Bayesian methodology with applications in credibility premium derivation and future claim size prediction. *AStA Adv. Stat. Anal.* **98**(3) 287-303. MR3227607

[24] Korb, K. B. and Nicholson, A. E. (2010). *Bayesian Artificial Intelligence*. 2nd Ed. Boca Raton, FL, USA: CRC Press, Inc. MR2130189

[25] Koski, T. and Noble, J. (2011). *Bayesian networks: an introduction*. Vol. 924. John Wiley and Sons. MR2641352

[26] Krishnan, T. and McLachlan, G. (1997). *The EM Algorithm and Extensions*. John Wiley and Sons. MR2392878

[27] Lauritzen, S. L. (1995). The EM algorithm for graphical association models with missing data. *Comp. Stat. Data Anal.* **19**(2) 191-201.

[28] Lauritzen, S. L. and Spiegelhalter, D.J. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *J. Royal Stat. Soci. Series B (Methodological)* 157-224. MR0964177

[29] Louis, T. A. (1984). Estimating a population of parameter values using Bayes and empirical Bayes methods. *J. Am. Stat. Assoc.* **79**(386) 393-398. MR0755093

[30] Nagarajan, R., Scutari, M. and Lèbre, S. (2013). *Bayesian Networks in R with Applications in Systems Biology*. New York. Springer. MR3059206

[31] Nielsen, T. D. and Jensen, F. V. (2009). *Bayesian Networks and Decision Graphs*. Springer Science and Business Media. MR2344166

[32] Oniésko, A., Lucas, P. and Druzdzel, M. J. (2001). Comparison of rule-based and Bayesian network approaches in medical diagnostic systems. *In: Artificial Intelligence in Medicine* 283-292. Springer.

[33] Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann. MR0965765

[34] Ramoni, M. and Sebastiani, P. (2001). Robust learning with missing data. *Mach. Learn.* **45** 147-170.

[35] Ramoni, M. and Sebastiani, P. (2003). Bayesian methods. *In: Intelligent Data Analysis*, 131-168. Springer.

[36] Riggelsen, C. and Feelders, A. (2005). Learning Bayesian network models from incomplete data using importance sampling. *In: Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, 301-308. MR2337929

[37] Robert, C. (2007). *The Bayesian Choice: from Decision-Theoretic Foundations to Computational Implementation*. Springer Science and Business Media. MR2723361

[38] Singh, M. (1997). Learning Bayesian networks from incomplete data. *In: Proceedings of Fourteenth National Conference on Artificial Intelligence, Providence, RI*, 534-539.

[39] Scutari, M. (2010). bnlearn: Bayesian network structure learning. *R package*.

[40] Sebastiani, P., Abad, M. M., and Ramoni, M. F. (2010). Bayesian networks. *In: Data Mining and Knowledge Discovery Handbook*, 175-208. Springer.

[41] Silander, T., Kontkanen, P. and Myllymaki, P. (2012). On sensitivity of the MAP Bayesian network structure to the equivalent sample size parameter. *arXiv preprint arXiv:1206.5293*.

[42] Singh, P., Singh, S. and Singh, U. (2008). Bayes estimator of inverse Gaussian parameters under general entropy loss function using Lindley's approximation. *Commun. Stat. Simul. Comput.* **37**(9) 1750-1762. MR2542431

[43] Spiegelhalter, D. J. (1989). Probabilistic reasoning in expert systems. *Am. J. Math. Management Sci.* **9**(3-4) 191-210. MR1082797

[44] Spirtes, P., Glymour, C. N. and Scheines, R. (2000). *Causation, Prediction, and Search*. Vol. 81. MIT Press. MR1815675

[45] Steck, H. and Jaakkola, T. S. (2002). On the Dirichlet prior and Bayesian regularization. *In: Advances in Neural Information Processing Systems*, 697-704.

[46] Thomas, G. B., Finney, R. L., Weir, M. D. and Giordano, F. R. (2001). *Thomas' Calculus*. Addison-Wesley.

[47] Weiss, Y. and Freeman, W. T. (2001). Correctness of belief propagation in Gaussian graphical models of arbitrary topology. *Neural Comput.* **13**(10) 2173-2200.