Structure Learning from Related Data Sets with a Hierarchical Bayesian Score

Laura Azzimonti LAURA@IDSIA.CH
Giorgio Corani GIORGIO@IDSIA.CH
Marco Scutari SCUTARI@IDSIA.CH
Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA), USI/SUPSI, Lugano, Switzerland

Abstract

Score functions for learning the structure of Bayesian networks in the literature assume that data are a homogeneous set of observations; whereas it is often the case that they comprise different related, but not homogeneous, data sets collected in different ways. In this paper we propose a new Bayesian Dirichlet score, which we call Bayesian Hierarchical Dirichlet (BHD). The proposed score is based on a hierarchical model that pools information across data sets to learn a single encompassing network structure, while taking into account the differences in their probabilistic structures. We derive a closed-form expression for BHD using a variational approximation of the marginal likelihood and we study its performance using simulated data. We find that, when data comprise multiple related data sets, BHD outperforms the Bayesian Dirichlet equivalent uniform (BDeu) score in terms of reconstruction accuracy as measured by the Structural Hamming distance, and that it is as accurate as BDeu when data are homogeneous. Moreover, the estimated networks are sparser and therefore more interpretable than those obtained with BDeu, thanks to a lower number of false positive arcs.

Keywords: Bayesian networks; structure learning; hierarchical priors; Dirichlet mixtures; network scores.

1. Introduction

Investigating challenging problems at the forefront of science increasingly requires large amounts of data that can only be gathered through collaborations between several institutions. This naturally leads to heterogeneous data sets that are in fact the collation of related, but not identical, subsets of data that will necessarily differ in the details of how they are collected. Examples can be found in multi-centre clinical trials, in which protocols are applied in slightly different ways to different patient populations; population genetics, which studies the architecture of phenotypic traits across populations and their evolution; ecology and environmental sciences, which produce different patterns of measurement errors and limitations in different environments. A common goal in analysing these complex data is to construct a mechanistic model that elucidates the interplay between different elements under investigation, either as a step towards building a causal model or to perform accurate prediction from a purely probabilistic perspective.

On the one hand, the task of efficiently modelling such related data sets can be tackled using hierarchical models (Gelman et al., 2014), which make it possible to pool the information common to the different subsets of the data while correctly encoding the information that is specific to each subset. On the other hand, Bayesian networks (BNs; Koller and Friedman, 2009) provide a rigorous approach for both causal and predictive modelling by representing variables as nodes and probabilistic dependencies as arcs in a graph. To the best of our knowledge, however, no method has

been proposed in the literature to combine these two approaches to learn a single BN structure from a set of related data sets and get the best of both worlds. Available methods focus on learning an ensemble of BNs that have similar structures by penalising differences in their arc sets (Niculescu-Mizil and Caruana, 2007; Oates et al., 2016). Parameter learning from related data sets has been investigated in De Michelis et al. (2006) for Gaussian BNs and in Malovini et al. (2012) for discrete BNs. However, they only consider a naive Bayes structure and they initialise their hyperprior with maximum likelihood point estimates.

In this paper, we show how to learn the structure of a BN from related data sets, containing the same variables, by building on our previous work on parameter learning in Azzimonti et al. (2019). In Section 2 we briefly introduce BNs and hierarchical models in the context of discrete data as well as prior work on parameter learning from related data sets. We then propose a score function for related data sets in Section 3, and we study its performance on simulated data in Section 4. Finally we discuss our results and possible future research directions in Section 5.

2. Background and Notation

Bayesian networks (BNs) are a class of graphical models that use a directed acyclic graph (DAG) $\mathcal{G}$ to model a set of random variables $\mathbf{X} = \{X_1, \ldots, X_N\}$: each node is associated with one $X_i \in \mathbf{X}$ and arcs represent direct dependence relationships. Graphical separation of two nodes implies the conditional independence of the corresponding random variables. In principle, there are many possible choices for the joint distribution of $\mathbf{X}$; literature has focused mostly on discrete BNs (Heckerman et al., 1995), in which both $\mathbf{X}$ and the X_i are categorical (multinomial) random variables. Other possibilities include Gaussian BNs and conditional linear Gaussian BNs (Lauritzen and Wermuth, 1989), which include both discrete and Gaussian BNs as particular cases.

The task of learning a BN from a data set $\mathcal{D}$ of n observations is performed in two steps in an inherently Bayesian fashion:

$$\underbrace{\mathrm{P}(\mathcal{G}, \Theta \,|\, \mathcal{D})}_{\text{learning}} = \underbrace{\mathrm{P}(\mathcal{G} \,|\, \mathcal{D})}_{\text{structure learning}} \cdot \underbrace{\mathrm{P}(\Theta \,|\, \mathcal{G}, \mathcal{D})}_{\text{parameter learning}}, \tag{1}$$

where Θ are the parameters of $\mathbf{X}$. *Structure learning* consists in finding the DAG $\mathcal{G}$ that encodes the dependence structure of the data. In this paper we will focus on score-based algorithms, which are typically heuristic search algorithms that use a goodness-of-fit score such as BIC (Schwarz, 1978) or the Bayesian Dirichlet equivalent uniform (BDeu) marginal likelihood (Heckerman et al., 1995) to find an optimal $\mathcal{G}$. *Parameter learning* involves the estimation of the parameters Θ given the DAG $\mathcal{G}$ learned in the first step. Thanks to the Markov property, this step is computationally efficient because if the data are complete the *global distribution* of $\mathbf{X}$ decomposes into

$$\mathrm{P}(\mathbf{X} \,|\, \mathcal{G}) = \prod_{i=1}^{N} \mathrm{P}(X_i \,|\, \Pi_{X_i}) \tag{2}$$

and the *local distribution* associated with each node X_i depends only on the configurations of its parents Π_{X_i}. Note that this decomposition does not uniquely identify a BN; different DAGs can encode the same global distribution, thus grouping BNs into equivalence classes (Chickering, 1995) characterised by the skeleton of $\mathcal{G}$ (its underlying undirected graph) and its v-structures (patterns of arcs of the type $X_j \to X_i \leftarrow X_k$, with no arc between X_j and X_k).

2.1 Classic Multinomial-Dirichlet Parameterisation

In the case of discrete BNs, we assume that each $X_i \,|\, \Pi_{X_i}$ follows a categorical distribution for each configuration of Π_{X_i}. Hence the parameters of $X_i \,|\, \Pi_{X_i}$ are the conditional probabilities $\boldsymbol{\theta}_{X_i \,|\, \Pi_{X_i}} = \{\boldsymbol{\theta}_{X_i \,|\, j}, j = 1, \ldots, |\Pi_{X_i}|\}$, whose kth element corresponds to $\mathrm{P}(X_i = k \,|\, \Pi_{X_i} = j)$, for which we assume a conjugate Dirichlet prior:

$$\begin{aligned} \boldsymbol{\theta}_{X_i \,|\, j} \,\big|\, \boldsymbol{\alpha}_i &\sim \text{Dirichlet}(\boldsymbol{\alpha}_i) & j &= 1, \ldots, |\Pi_{X_i}|, \\ X_i \,\big|\, \Pi_{X_i} = j, \boldsymbol{\theta}_{X_i \,|\, j} &\sim \text{Categorical}\left(\boldsymbol{\theta}_{X_i \,|\, j}\right) & j &= 1, \ldots, |\Pi_{X_i}|, \end{aligned} \tag{3}$$

where $\boldsymbol{\alpha}_i = \{\alpha_{ijk}, i = 1, \ldots N; j = 1, \ldots, |\Pi_{X_i}|; k = 1, \ldots, |X_i|\}$ is a hyperparameter vector defined over a simplex with sum $\sum_{jk} \alpha_{ijk} = s_i > 0$. The posterior estimator of $\boldsymbol{\theta}_{X_i \,|\, \Pi_{X_i}}$ is:

$$\left[\widehat{\boldsymbol{\theta}}_{X_i \,|\, j}\right]_k = \frac{\alpha_{ijk} + n_{ijk}}{\alpha_{ij} + n_{ij}}, \qquad n_{ij} = \sum\nolimits_k n_{ijk}, \;\; \alpha_{ij} = \sum\nolimits_k \alpha_{ijk}, \tag{4}$$

where n_{ijk} represents the number of observations for which $X_i = k$ and $\Pi_{X_i} = j$. It is common to set $\alpha_{ijk} = s_i/(|X_i||\Pi_{X_i}|)$ with the same *imaginary sample size* $s_i = s$ for all X_i.

In the context of structure learning, we have $\mathrm{P}(\mathcal{G} \,|\, \mathcal{D}) \propto \mathrm{P}(\mathcal{D} \,|\, \mathcal{G})\,\mathrm{P}(\mathcal{G})$ and we can use $\mathrm{P}(\mathcal{D} \,|\, \mathcal{G})$ as a score function. (Implicitly, we are saying that $\mathrm{P}(\mathcal{G}) \propto 1$ by disregarding it.) Assuming *positivity* ($\boldsymbol{\theta}_{X_i \,|\, \Pi_{X_i}} > 0$), *parameter independence* (columns of $\boldsymbol{\theta}_{X_i \,|\, \Pi_{X_i}}$ associated to different parent configurations are independent), *parameter modularity* ($\boldsymbol{\theta}_{X_i \,|\, \Pi_{X_i}}$ associated with different nodes are independent) and *complete data*, Heckerman et al. (1995) derived a closed form expression for $\mathrm{P}(\mathcal{D} \,|\, \mathcal{G})$ known as the *Bayesian Dirichlet* (BD) family of scores:

$$\mathrm{BD}(\mathcal{G}, \mathcal{D}; \boldsymbol{\alpha}) = \prod_{i=1}^{N} \mathrm{BD}\left(X_i \,|\, \Pi_{X_i}; \boldsymbol{\alpha}_i\right) = \prod_{i=1}^{N} \prod_{j=1}^{|\Pi_{X_i}|} \left[\frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + n_{ij})} \prod_{k=1}^{|X_i|} \frac{\Gamma(\alpha_{ijk} + n_{ijk})}{\Gamma(\alpha_{ijk})}\right]. \tag{5}$$

Choosing again $\alpha_{ijk} = s/(|X_i||\Pi_{X_i}|)$ gives the *Bayesian Dirichlet equivalent uniform* (BDeu) score. A default value of $s = 1$ has been recommended by Ueno (2010). Assuming a uniform prior for both $\mathcal{G}$ and $\boldsymbol{\theta}_{X_i \,|\, \Pi_{X_i}}$ is common in the literature, even if they can have serious impact on the accuracy of the learned structures (Scutari, 2016), especially for sparse data which are likely to lead to violations of the positivity assumption (Scutari, 2018). These assumptions are taken to represent lack of prior knowledge, and they result in BDeu giving the same score to BNs in the same equivalence class (*score-equivalence*). BDeu is the only BD score with this property.

2.2 Hierarchical Multinomial-Dirichlet Parameterisation for Related Data Sets

The classic Multinomial-Dirichlet model in (3) can be extended to handle related data sets by treating it as a particular case of the hierarchical Multinomial-Dirichlet (hierarchical MD) model presented in Azzimonti et al. (2019). For this purpose, we introduce an auxiliary variable F which identifies the $|F|$ related data sets. Assuming that the data sets contain the same variables and that F is always observed, we can learn a BN with a common structure $\mathcal{G}$ but with different parameter estimates for each related data set.

For simplicity, we apply the hierarchical model independently to each local distribution to estimate the joint distribution of (X_i, Π_{X_i}) conditional on F, $\boldsymbol{\theta}_{X_i, \Pi_{X_i} \,|\, F} = \{\boldsymbol{\theta}^f_{X_i, \Pi_{X_i}}, f = 1, \ldots, |F|\}$,

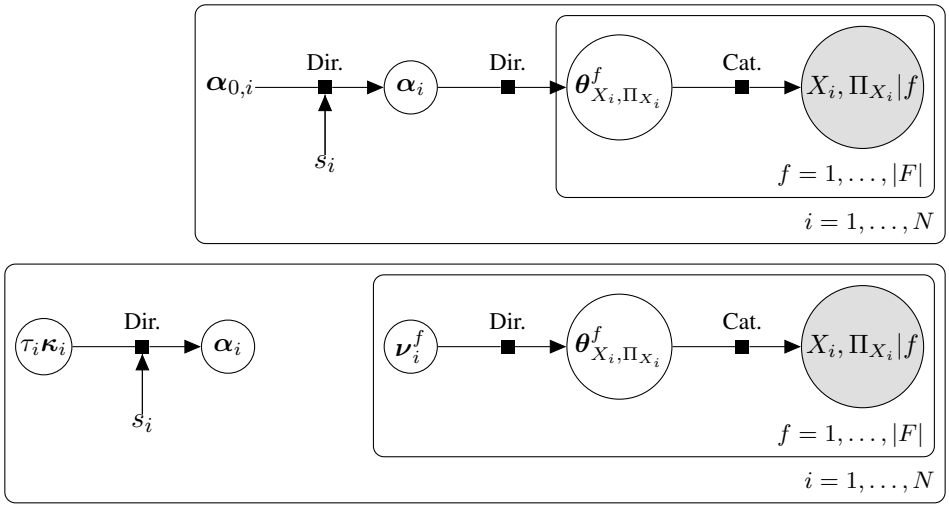

Figure 1: Directed factor graphs representing hierarchical Multinomial-Dirichlet model for related data sets (top panel) and its variational approximation (bottom panel). Cat. and Dir. represent respectively Categorical and Dirichlet distributions.

by pooling information between different data sets. The resulting hierarchical model is shown in the top panel of Figure 1. Specifically, for each node X_i we assume $\boldsymbol{\alpha}_i$ to be a latent random vector and we add a Dirichlet hyperprior to make $\boldsymbol{\theta}^f_{X_i,\Pi_{X_i}}$ a mixture of Dirichlet distributions:

$$\begin{aligned}
\boldsymbol{\alpha}_i \,|\, s_i, \boldsymbol{\alpha}_{0,i} &\sim s_i \cdot \text{Dirichlet}(\boldsymbol{\alpha}_{0,i}), \\
\boldsymbol{\theta}^f_{X_i,\Pi_{X_i}} \,\Big|\, \boldsymbol{\alpha}_i &\sim \text{Dirichlet}(\boldsymbol{\alpha}_i) && f = 1, \ldots, |F|, \\
X_i, \Pi_{X_i} \,\Big|\, F = f, \boldsymbol{\theta}^f_{X_i,\Pi_{X_i}} &\sim \text{Categorical}\left(\boldsymbol{\theta}^f_{X_i,\Pi_{X_i}}\right) && f = 1, \ldots, |F|,
\end{aligned} \tag{6}$$

where this time $\boldsymbol{\alpha}_i = \{\alpha_{ijk}\}$ is a latent random vector defined over a simplex with sum s_i. The new hyperparameters of this model are the imaginary sample size s_i and the parameter vector $\boldsymbol{\alpha}_{0,i}$, which in turn is defined over a simplex with sum $s_{0,i}$; in the following we will omit both for brevity.

The marginal posterior distribution for $\boldsymbol{\theta}^f_{X_i,\Pi_{X_i}}$ is not analytically tractable, as noted in Azzimonti et al. (2019). However, the posterior average can be compactly expressed as:

$$\left[\widehat{\boldsymbol{\theta}}^f_{X_i,\Pi_{X_i}}\right]_{jk} = \frac{\text{E}[\alpha_{ijk}] + n^f_{ijk}}{s_i + n^f_i}, \qquad n^f_i = \sum\nolimits_{jk} n^f_{ijk}. \tag{7}$$

$\text{E}[\alpha_{ijk}]$ represents the posterior average of α_{ijk}; it cannot be written in closed form but can be approximated using variational inference (Jordan et al., 1999; Wainwright and Jordan, 2008). The resulting $\widehat{\boldsymbol{\theta}}^f_{X_i,\Pi_{X_i}}$ are data-set-specific but depend on all the available data via the *partial pooling* (Gelman et al., 2014) of the information present in the $|F|$ related data sets, thanks to the shared $\text{E}[\alpha_{ijk}]$ term. On the one hand, this produces more reliable estimates for sparse data and for related data sets with unbalanced sample sizes (Casella and Moreno, 2009). On the other hand, the prior

in (6) violates the parameter independence assumption, leading to a marginal likelihood that does not decompose over parent configurations and that is not score-equivalent. The prior is specified on (X_i, Π_{X_i}), as opposed to $X_i \,|\, \Pi_{X_i}$, which is only later computed from the joint distribution. As a result, $\mathrm{P}(X_i, \Pi_{X_i} \,|\, F) \neq \mathrm{P}(X_i \,|\, \Pi_{X_i}, F)\,\mathrm{P}(\Pi_{X_i} \,|\, F)$ because $\mathrm{P}(X_i, \Pi_{X_i} \,|\, F)$ and $\mathrm{P}(\Pi_{X_i} \,|\, F)$ are estimated by applying the hierarchical model separately to two different sets of variables, thus pooling the available information differently.

3. Structure Learning from Related Data Sets

In this section we derive the marginal likelihood score associated with the hierarchical model in (6) to implement structure learning from related data sets that contain the same variables. As the hierarchical model is not analytically tractable, we replace it with the approximate variational model shown in the bottom panel of Figure 1:

$$\begin{aligned}
\boldsymbol{\alpha}_i \,|\, s_i, \tau_i, \boldsymbol{\kappa}_i &\sim s_i \cdot \text{Dirichlet}\left(\tau_i \boldsymbol{\kappa}_i\right), \\
\boldsymbol{\theta}^f_{X_i,\Pi_{X_i}} \,\Big|\, \boldsymbol{\nu}^f_i &\sim \text{Dirichlet}\left(\boldsymbol{\nu}^f_i\right) && f = 1, \ldots, |F|, \\
\mathbf{X}_i, \Pi_{X_i} \,\Big|\, F = f, \boldsymbol{\theta}^f_{X_i,\Pi_{X_i}} &\sim \text{Categorical}\left(\boldsymbol{\theta}^f_{X_i,\Pi_{X_i}}\right) && f = 1, \ldots, |F|,
\end{aligned} \tag{8}$$

where $\boldsymbol{\nu}^f_i = \{\nu^f_{ijk}\}$, $\boldsymbol{\kappa}_i = \{\kappa_{ijk}\}$ with $i = 1, \ldots N$; $j = 1, \ldots, |\Pi_{X_i}|$; $k = 1, \ldots, |X_i|$ and $f = 1, \ldots, |F|$; $\sum_{jk} \kappa_{ijk} = 1$ and $\tau_i \in \mathbb{R}^+$ for $i = 1 \ldots N$. These parameters are estimated from the available data by minimising the Kullback-Leibler divergence between the exact posterior distribution p and its variational approximation q, as described in Azzimonti et al. (2019). Since F is assumed to be the parent of any node in the network and to be always observed, we treat it as an input variable in a conditional Bayesian network (Koller and Friedman, 2009, Section 5.6) and we do not explicitly assign it a distribution. Therefore, the auxiliary variable F will not influence the score.

The variational model (8) is similar to the original hierarchical MD model (6), but it removes the dependence between $\boldsymbol{\theta}^f_{X_i,\Pi_{X_i}}$ and $\boldsymbol{\alpha}_i$ thus making it possible to derive an approximation of the marginal likelihood $\mathrm{P}(\mathcal{D} \,|\, F, \mathcal{G})$.

Lemma 1 *Given $|F|$ complete and related data sets $\mathcal{D} = \{\mathcal{D}_f, f = 1, \ldots, |F|\}$, under the assumption that the related data sets have the same dependence structure $\mathcal{G}$ and that each local distribution follows the hierarchical MD (6) with positive parameters, the marginal likelihood $\mathrm{P}(\mathcal{D} \,|\, F, \mathcal{G})$ can be approximated with the quantity*

$$q(\mathcal{D} \,|\, F, \mathcal{G}) = \prod_{i=1}^{N} \prod_{f=1}^{|F|} \prod_{j=1}^{|\Pi_{X_i}|} \left[\frac{\Gamma(s_i \widehat{\kappa}_{ij})}{\Gamma(s_i \widehat{\kappa}_{ij} + n^f_{ij})} \prod_{k=1}^{|X_i|} \frac{\Gamma(s_i \widehat{\kappa}_{ijk} + n^f_{ijk})}{\Gamma(s_i \widehat{\kappa}_{ijk})} \right], \; n^f_{ij} = \textstyle\sum_k n^f_{ijk}, \widehat{\kappa}_{ij} = \sum_k \widehat{\kappa}_{ijk}, \tag{9}$$

where $s_i \widehat{\kappa}_{ijk}$ represents the posterior average of α_{ijk} under the variational model (8).

Proof Starting from the variational model (8), we can derive the conditional distribution of X_i given $\Pi_{X_i} = j$ and $F = f$, with $i = 1, \ldots N$; $j = 1, \ldots, |\Pi_{X_i}|$ and $f = 1, \ldots, |F|$, as a categorical distribution with parameters $\boldsymbol{\theta}^f_{X_i \,|\, j}$, whose kth element is $\left[\theta^f_{X_i,\Pi_{X_i}}\right]_{jk} / \sum_{\tilde{k}} \left[\theta^f_{X_i,\Pi_{X_i}}\right]_{j\tilde{k}}$

with $k = 1, \ldots, |X_i|$. Thanks to the properties of Dirichlet distributions, $\boldsymbol{\theta}^f_{X_i \,|\, j}$ and $\boldsymbol{\alpha}_{i \,|\, j}$, whose kth element is defined as $\alpha_{ijk} / \sum_{\tilde{k}} \alpha_{ij\tilde{k}}$, are still distributed as Dirichlet distributions respectively with parameters $\boldsymbol{\nu}^f_{ij}$, whose kth element is ν^f_{ijk}, and $\tau_i \boldsymbol{\kappa}_{ij}$, whose kth element is $\tau_i \kappa_{ijk}$.

The approximated conditional distribution satisfies *independence between related data sets*. Moreover, given a data set $F = f$, both *parameter modularity* and *parameter independence* are satisfied. Thus,

$$q(\mathcal{D} \,|\, F, \mathcal{G}) = \iint \prod_{f=1}^{|F|} \prod_{i=1}^{N} \prod_{j=1}^{|\Pi_{X_i}|} q(X_i \,|\, \Pi_{X_i} = j, F = f, \boldsymbol{\theta}^f_{X_i \,|\, j}, \boldsymbol{\alpha}_{i \,|\, j}, \mathcal{G}) \\ q(\boldsymbol{\theta}^f_{X_i \,|\, j}, \boldsymbol{\alpha}_{i \,|\, j} \,|\, \mathcal{G}) \, d\boldsymbol{\theta}^f_{X_i \,|\, j} d\boldsymbol{\alpha}_{i \,|\, j}.$$

Thanks to the independence between $\boldsymbol{\theta}^f_{X_i \,|\, j}$ and $\boldsymbol{\alpha}_{i \,|\, j}$ induced by the variational model and the fact that $\int q(\boldsymbol{\alpha}_{i \,|\, j} \,|\, \mathcal{G}) \, d\boldsymbol{\alpha}_{i \,|\, j} = 1$, we obtain

$$q(\mathcal{D} \,|\, F, \mathcal{G}) = \prod_{f=1}^{|F|} \prod_{i=1}^{N} \prod_{j=1}^{|\Pi_{X_i}|} \int q(X_i \,|\, \Pi_{X_i} = j, F = f, \boldsymbol{\theta}^f_{X_i \,|\, j}, \mathcal{G}) q(\boldsymbol{\theta}^f_{X_i \,|\, j} \,|\, \mathcal{G}) \, d\boldsymbol{\theta}^f_{X_i \,|\, j}.$$

Since $q(\boldsymbol{\theta}^f_{X_i \,|\, j} \,|\, \mathcal{G}) = \text{Dirichlet}(\boldsymbol{\nu}^f_{ij})$ we derive

$$q(\mathcal{D} \,|\, F, \mathcal{G}) = \prod_{i=1}^{N} \prod_{f=1}^{|F|} \prod_{j=1}^{|\Pi_{X_i}|} \left[\frac{\Gamma(\nu^f_{ij})}{\Gamma(\nu^f_{ij} + n^f_{ij})} \prod_{k=1}^{|X_i|} \frac{\Gamma(\nu^f_{ijk} + n^f_{ijk})}{\Gamma(\nu^f_{ijk})} \right], \quad \nu^f_{ij} = \sum_k \nu^f_{ijk}. \tag{10}$$

The parameter ν^f_{ijk} is not known *a priori*, but it is estimated *a posteriori* as $\widehat{\nu}^f_{ijk} = s_i \widehat{\kappa}_{ijk} + n^f_{ijk}$, see Azzimonti et al. (2019) for details on the derivation. Hence *a posteriori* $\boldsymbol{\theta}^f_{X_i \,|\, j}$ is distributed as a $\text{Dirichlet}(s_i \widehat{\boldsymbol{\kappa}}_{ij} + \boldsymbol{n}^f_{ij})$. Thanks to this relationship and the conjugacy between Categorical and Dirichlet distributions, we can replace ν^f_{ijk} with $s_i \widehat{\kappa}_{ijk}$ in (10), thus obtaining (9). ■

Note that the marginal likelihood (9) has the same form as the classic BD score (5), with α_{ijk} replaced by $s_i \widehat{\kappa}_{ijk}$, which represents the posterior average of α_{ijk} under the hierarchical variational model. The posterior average is shared between different related data sets, thus inducing a pooling effect that makes $\boldsymbol{\theta}^f_{X_i, \Pi_{X_i}}$ and $\boldsymbol{\alpha}_i$ dependent once more.

From Lemma 1, we define the approximated Bayesian hierarchical Dirichlet score as

$$\text{BHD}(\mathcal{G}, \mathcal{D} \,|\, F) = q(\mathcal{D} \,|\, F, \mathcal{G}).$$

The proposed BHD score can be factorised over the nodes, *i.e.*,

$$\text{BHD}(\mathcal{G}, \mathcal{D} \,|\, F) = \prod_{i=1}^{N} \text{BHD}(X_i \,|\, \Pi_{X_i}, F),$$

and can be used to learn a common structure for all related data sets, taking into account potential differences in the probabilistic relationships between variables.

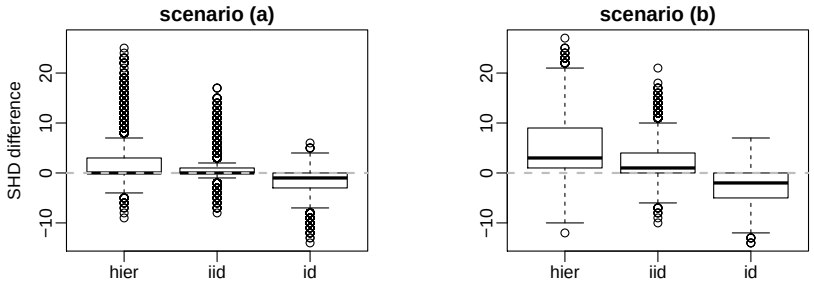

Figure 2: Boxplots of SHD difference between BHD and BDeu score for scenario (a) (left panel) and (b) (right panel). Positive values favour the hierarchical score.

4. Simulation Study

We now compare the empirical performance of BHD to that of BDeu and BIC with a simulation study; for brevity, we will not discuss the results for BIC in detail since they are fundamentally the same as those for BDeu. In particular, we are interested in structure learning in the following two scenarios:

(a) the true underlying network is the same for all the related data sets;

(b) the true underlying network is the same for all the related data sets, apart from N_F data sets in which N_A randomly selected arcs have been removed.

For each scenario, we consider three different models for the local distributions of each node:

hier: the parameters associated with each of the related data sets are sampled from a hierarchical Dirichlet distribution with imaginary sample size equal to 10 and parameter $\boldsymbol{\alpha}$ that is in turn sampled from a Dirichlet distribution with all $\alpha_{0,ijk} = 1$;

iid: the parameters associated with each of the related data sets are independently sampled from the same Dirichlet distribution with imaginary sample size equal to 10 and all $\alpha_{ijk} = 1$;

id: the parameters are identical for all data sets.

The first approach follows the distributional assumptions of the hierarchical model underlying BHD and may favour the proposed score. The last approach may favour methods not considering that data may comprise related data sets, which pool all the data and assume they are generated from the same distribution. The second approach is a middle ground between the first and the third, since parameters associated to the related data sets are different but they are not generated from the hierarchical model.

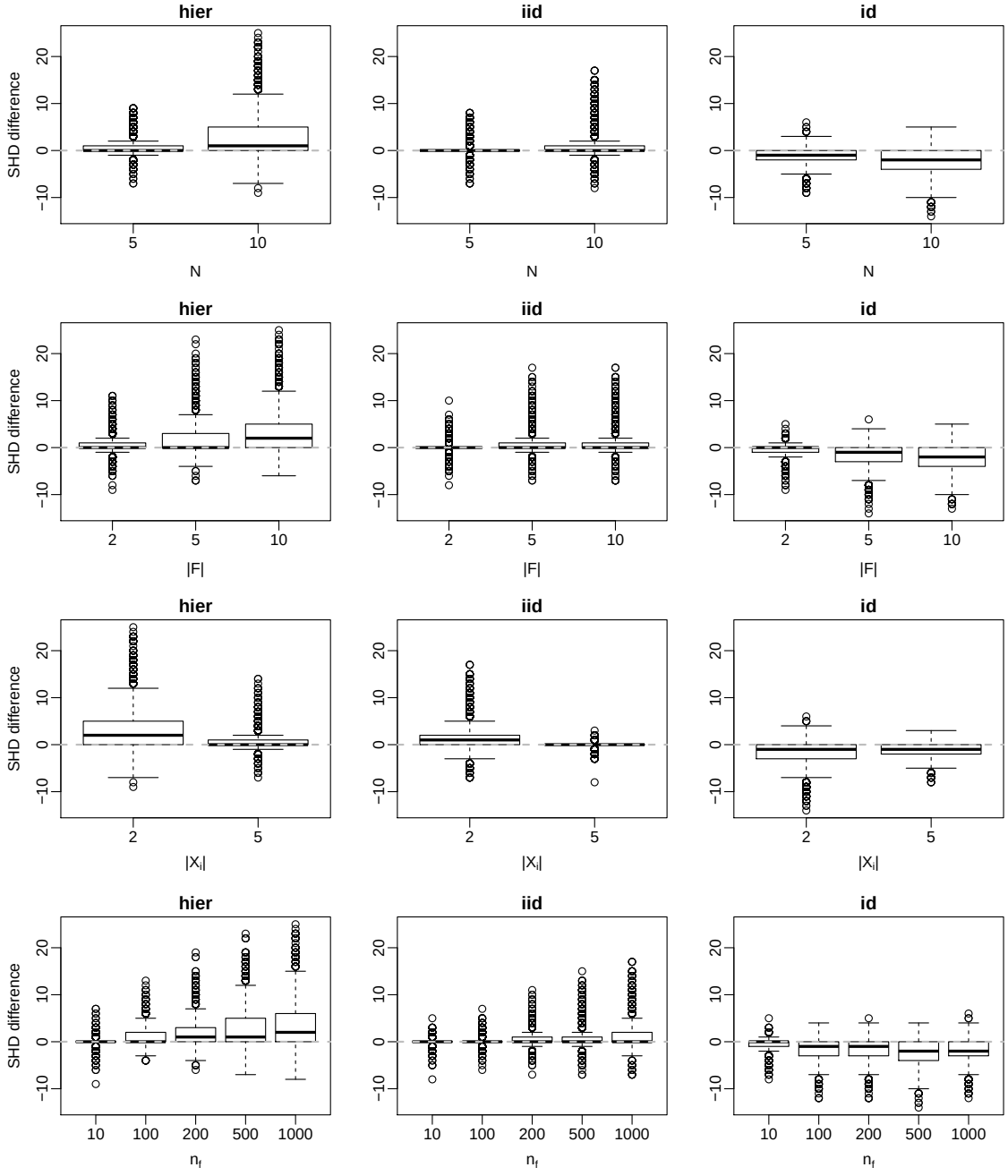

Figure 3: Boxplots of SHD difference between BHD and BDeu score for scenario (a) (equal structures) with different values of number of variables N, number of related groups $|F|$, number of states $|X_i|$ and number of observations n_f, with parameters sampled with the hierarchical (left panels), i.i.d (central panels) or identical distribution (right panel) approach. Positive values favour the hierarchical score.

To evaluate the performance of BHD, we first sample 3 network structures for each of three different levels of sparsity, such that the proportion of the number of arcs per node is equal to $c \in \{1, 1.2, 1.5\}$, and each combination of:

- $N = \{5, 10\}$, where N represents the number of nodes;
- $|F| \in \{2, 5, 10\}$, where $|F|$ represents the number of related data sets;
- $|X_i| \in \{2, 5\}$, where $|X_i|$ represents the number of states for each variable.

Then, for both scenario (a) and (b), we replicate the same structure for all the $|F|$ related data sets. In scenario (b), for each of the N_F data sets differing from the others, we randomly remove N_A arcs from the network, with $N_F \in \{1, 2\}$ and $N_A \in \{1, 2\}$. Thus, in scenario (b) we deal with N_F structures that differ from one another and from the main structure by N_A arcs.

Once the structures have been generated, we sample 10 different parameter sets for each of **hier**, **iid** and **id**; and, for each of these parameter sets, we sample 10 times $|F|$ related data sets, each of them composed of the same number of observations $n_f \in \{10, 100, 200, 500, 1000\}$. We then perform structure learning on the resulting data by means of the hill-climbing implementation in bnlearn (Scutari, 2010) with BHD and BDeu scores. For both scores we use the imaginary sample $s = 1$. In the case of BDeu we pool all the available data from different related data sets.

We evaluate the accuracy in reconstructing the network with the Structural Hamming Distance (SHD) between the estimated and the true underlying structure, True Positive (TP), False Positive (FP) and False Negative (FN) arcs.

The difference between BDeu and BHD in terms of SHD for scenarios (a) and (b) is shown in Figure 2, left and right panel respectively. Positive values favour the proposed BHD score. When parameters are sampled from a hierarchical distribution (**hier**), BHD outperforms BDeu in both scenarios, with a larger improvement in scenario (b). In the **iid** case, BHD is competitive with BDeu when the underlying network structures are homogeneous, and it outperforms BDeu when the underlying network structures are different. On the other hand, in the **id** case BDeu has better accuracy than BHD because it correctly assumes that all the data are generated form the same distribution, while BHD has a large number of redundant parameters that would model the non-existing related data sets.

Figure 3 shows how the SHD difference between BHD and BDeu varies for different simulation parameters in scenario (a). Specifically, as the number of variables N or the number of related data sets $|F|$ increase, the differences between BHD and BDeu (positive for **hier** and **iid**, negative for **id**), become increasingly large in magnitude. On the other hand, as the number of states $|X_i|$ increases the differences in accuracy between BHD and BDeu gradually decrease. For what concerns the sample size, BHD increasingly outperforms BDeu in both **hier** and **iid** as n_f increases. In the **id** case we expect the two scores to be asymptotically equivalent; the values we consider for n_f are not large enough to clearly show this empirically.

Figure 4 shows the relationship between the difference in SHD and some key simulation parameters in scenario (b). The effect of both the number of related data sets $|F|$ and the number of observations n_f is more marked than in scenario (a). For the same $|F|$ and n_f, BHD outperforms BDeu by a larger margin when some network structures are different (scenario (b)) compared to when they are all identical (scenario (a)).

Figure 5 compares the difference in TP (left), FP (center) and FN (right panel) between BDeu and BHD for scenario (b). Positive values favour the proposed BHD score. While the two methods

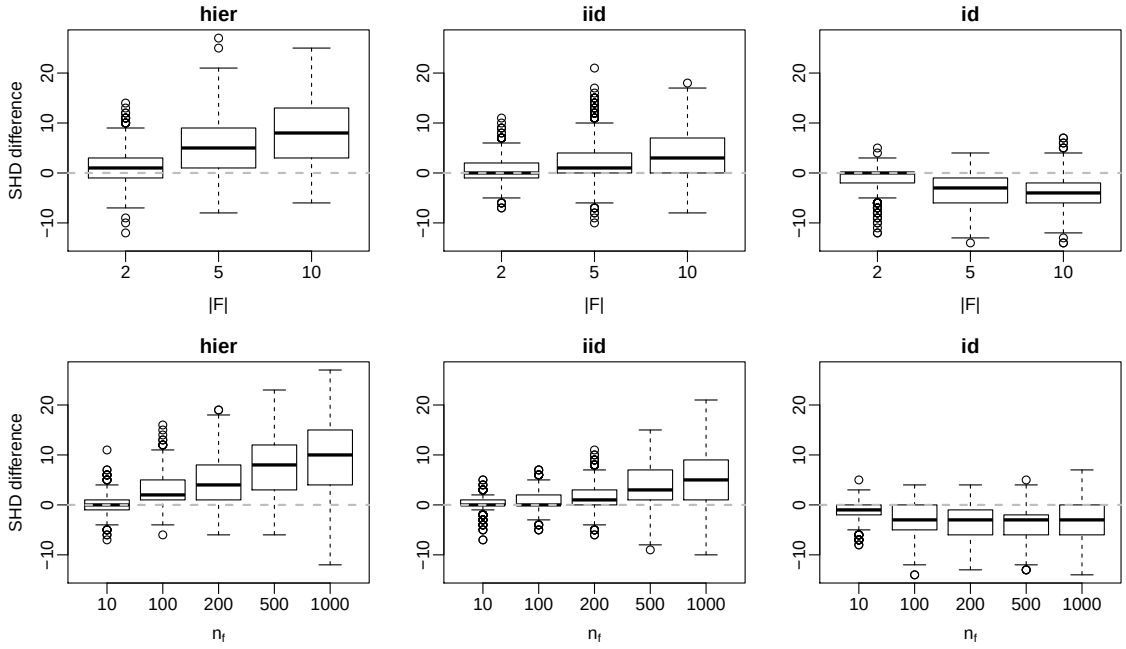

Figure 4: Boxplots of SHD difference between BHD and BDeu score for scenario (b) (different structures) with different values of number of related groups $|F|$ and number of observations n_f, with parameters sampled with the hierarchical (left panels), i.i.d (central panels) or identical distribution (right panel) approach. Positive values favour the hierarchical score.

perform similarly in terms of TP and FN, BHD outperforms BDeu in terms of FP in the **hier** case. The structures learned by BHD are thus sparser and more interpretable than those learned by BDeu.

We also perform some experiments with different values $s \in \{1, 2, 5, 10\}$ of the imaginary sample size. As s increases, BHD achieves marginally lower SHDs. However, its average SHD is not significantly different from that of BDeu for the same s (figures not shown for reasons of space).

5. Conclusions and future work

In this work we propose a new Bayesian score, BHD, to learn a common BN structure from related data sets. BHD assumes that the joint distribution in each node of the network follows a mixture of Dirichlet distributions, thus pooling information between the data sets. We find that BHD outperforms both BDeu and BIC when applied to data that are composed by related data sets; and that it has comparable performance to BDeu and BIC when data are a single, homogeneous data set.

Learning a common BN structure with BHD builds on and complements our previous work on parameter learning from related data sets, described in Azzimonti et al. (2019). We can use the latter to learn the parameters associated with a network structure learned using BHD, thus obtaining different BNs (one for each related data set) with the same structure and related parameters. Com-

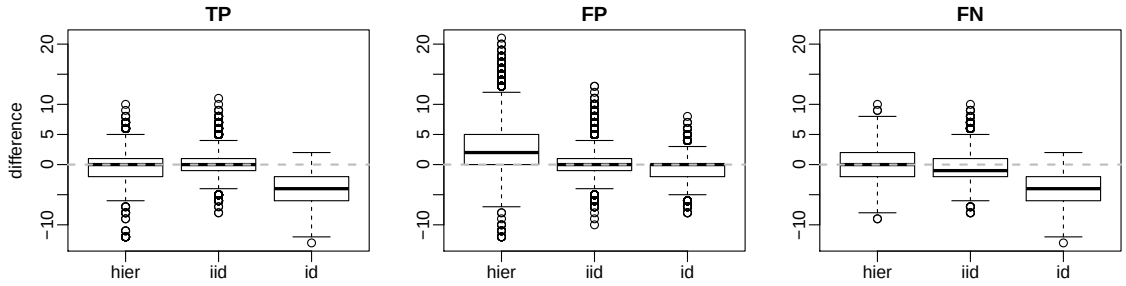

Figure 5: Boxplots of TP (left panel), FP (central panel) and FN (right panel) difference between BHD and BDeu score for scenario (b) (different structures). Positive values always favour the hierarchical score.

bining the two approaches may increase the performance of the BN models such as BN classifiers when dealing with related data sets.

The assumptions underlying BHD can be relaxed in several ways to extend its applicability to more complex scenarios. For instance, by relaxing the assumption that related data sets share the same dependence structure, it would be possible to detect independencies that hold only in certain contexts, similarly to Boutilier et al. (1996). In this case, the context-specific independence would be directly modelled by learning different but related network structures for each data set. Another interesting development would be to derive a conditional independence test from BHD to learn BNs from related data sets with constraint-based algorithm similarly to, *e.g.*, Tillman (2009).

Acknowledgments

We would like to acknowledge support for this project from the Swiss National Science Foundation (NSF, Grant No. IZKSZ2_162188).

References

L. Azzimonti, G. Corani, and M. Zaffalon. Hierarchical Estimation of Parameters in Bayesian Networks. *Computational Statistics & Data Analysis*, 137:67–91, 2019.

C. Boutilier, N. Friedman, M. Goldszmidt, and D. Koller. Context-Specific Independence in Bayesian Networks. In *Proceedings of the 12th International Conference on Uncertainty in Artificial Intelligence*, UAI '96, pages 115–123, 1996.

G. Casella and E. Moreno. Assessing Robustness of Intrinsic Tests of Independence in Two-Way Contingency Tables. *Journal of the American Statistical Association*, 104(487):1261–1271, 2009.

D. M. Chickering. A Transformational Characterization of Equivalent Bayesian Network Structures. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, UAI '95, pages 87–98, 1995.

F. De Michelis, P. Magni, P. Piergiorgi, M. A. Rubin, and R. Bellazzi. A Hierarchical Naive Bayes Model for Handling Sample Heterogeneity in Classification Problems: an Application to Tissue Microarrays. *BMC bioinformatics*, 7(1):514, 2006.

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis*. CRC press, 3rd edition, 2014.

D. Heckerman, D. Geiger, and D. M. Chickering. Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. *Machine Learning*, 20(3):197–243, 1995. Available as Technical Report MSR-TR-94-09.

M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An Introduction to Variational Methods for Graphical Models. *Machine Learning*, 37(2):183–233, 1999.

D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT press, 2009.

S. L. Lauritzen and N. Wermuth. Graphical Models for Associations Between Variables, Some of Which are Qualitative and Some Quantitative. *The Annals of Statistics*, 17(1):31–57, 1989.

A. Malovini, N. Barbarini, R. Bellazzi, and F. De Michelis. Hierarchical Naive Bayes for Genetic Association Studies. *BMC bioinformatics*, 13(14):S6, 2012.

A. Niculescu-Mizil and R. Caruana. Inductive Transfer for Bayesian Network Structure Learning. In *Proceedings of Artificial Intelligence and Statistics*, pages 339–346, 2007.

C. J. Oates, J. Q. Smith, S. Mukherjee, and J. Cussens. Exact Estimation of Multiple Directed Acyclic Graphs. *Statistics and Computing*, 26(4):797–811, 2016.

G. Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464, 1978.

M. Scutari. Learning Bayesian Networks with the bnlearn R Package. *Journal of Statistical Software*, 35(3):1–22, 2010.

M. Scutari. An Empirical-Bayes Score for Discrete Bayesian Networks. *Journal of Machine Learning Research (Proceedings Track, PGM 2016)*, 52:438–448, 2016.

M. Scutari. Dirichlet Bayesian Network Scores and the Maximum Relative Entropy Principle. *Behaviormetrika*, 45(2):337–362, 2018.

R. E. Tillman. Structure Learning with Independent Non-Identically Distributed Data. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 1041–1048, 2009.

M. Ueno. Learning Networks Determined by the Ratio of Prior and Data. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence*, pages 598–605, 2010.

M. J. Wainwright and M. I. Jordan. Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.