Hierarchical Multinomial-Dirichlet model for the estimation of conditional probability tables

Laura Azzimonti
IDSIA - SUPSI/USI
Manno, Switzerland
laura@idsia.ch

Giorgio Corani
IDSIA - SUPSI/USI
Manno, Switzerland
giorgio@idsia.ch

Marco Zaffalon
IDSIA - SUPSI/USI
Manno, Switzerland
zaffalon@idsia.ch

Abstract—We present a novel approach for estimating conditional probability tables, based on a joint, rather than independent, estimate of the conditional distributions belonging to the same table. We derive exact analytical expressions for the estimators and we analyse their properties both analytically and via simulation. We then apply this method to the estimation of parameters in a Bayesian network. Given the structure of the network, the proposed approach better estimates the joint distribution and significantly improves the classification performance with respect to traditional approaches.

Keywords-hierarchical Bayesian model; Bayesian networks; conditional probability estimation.

I. INTRODUCTION

A Bayesian network is a probabilistic model constituted by a directed acyclic graph (DAG) and a set of *conditional probability tables* (CPTs), one for each node. The CPT of node X contains the conditional probability distributions of X given each possible configuration of its parents. Usually all variables are discrete and the conditional distributions are estimated adopting a Multinomial-Dirichlet model, where the Dirichlet prior is characterised by the vector of hyper-parameters α. Yet, Bayesian estimation of multinomials is sensitive to the choice of α and inappropriate values cause the estimator to perform poorly [1]. Mixtures of Dirichlet distributions have been recommended both in statistics [2], [3] and in machine learning [4] in order to obtain more robust estimates. Yet, mixtures of Dirichlet distributions are computationally expensive; this prevents them from being widely adopted. Another difficulty encountered in CPT estimation is the presence of rare events. Assuming that all variables have cardinality k and that the number of parents is q, we need to estimate k^q conditional distributions, one for each joint configuration of the parent variables. Frequently one or more of such configurations are rarely observed in the data, making their estimation challenging.

We propose to estimate the conditional distributions by adopting a novel approach, based on a hierarchical Multinomial-Dirichlet model. This model has two main characteristics. First, the prior of each conditional distribution is constituted by a mixture of Dirichlet distributions with parameter α; the mixing is attained by treating α as a

random variable with its own prior and posterior distribution. By estimating from data the posterior distribution of α, we need not to fix its value *a priori*. Instead we give more weight to the values of α that are more likely given the data. Secondly, the model is hierarchical since it assumes that the conditional distributions within the same CPT (but referring to different joint configurations of the parents) are drawn from the same mixture. The hierarchical model *jointly* estimates all the conditional distributions of the same CPT, called *columns* of the CPT. The joint estimates generate information flow between different columns of the CPT; thus the hierarchical model exploits the parameters learned for data-rich columns to improve the estimates of the parameters of data-poor columns. This is called *borrowing statistical strength* [5, Sec 6.3.3.2] and it is well-known within the literature of hierarchical models [6]. Also the literature of Bayesian networks acknowledges [7, Sec.17.5.4] that introducing dependencies between columns of the same CPT could improve the estimates, especially when dealing with sparse data. However, as far as we known, this work is the first practical application of joint estimation of the columns of the same CPT.

To tackle the problem of computational complexity we adopt a variational inference approach. Namely, we compute a factorised approximation of the posterior distribution that is highly efficient. Variational inference appears particularly well suited for hierarchical models; for instance the inference of Latent Dirichlet Allocation [8] is based on variational Bayes. By extensive experiments, we show that our novel approach considerably improves parameter estimation compared to the traditional approaches based on Multinomial-Dirichlet model. The experiments show large gains especially when dealing with small samples, while with large samples the effect of the prior vanishes as expected.

The paper is organised as follows. Section II introduces the novel hierarchical model. Section III provides an analytical study of the resulting estimator, proving that the novel hierarchical approach provides lower estimation error than the traditional approaches, under some mild assumptions on the generative model. Section IV presents some simulation studies showing that, given the same network structure,

hierarchical estimation yields both a better fit of the joint distribution and a consistent improvement in classification performance, with respect to the traditional estimation under parameter independence. Section V reports some concluding remarks.

II. ESTIMATION UNDER MULTINOMIAL-DIRICHLET MODEL

We want to induce a Bayesian network over the set of random variables $\boldsymbol{X} = \{X_1, \ldots, X_I\}$. We assume that each variable $X_i \in \boldsymbol{X}$ is discrete and has r_i possible values in the set $\mathcal{X}_i$. The parents of X_i are denoted by Pa_i and they have q_i possible joint states collected in the set $\mathcal{P}_i$.

We denote by $\theta_{x|\mathrm{pa}}$ the probability of X_i being in state $x \in \mathcal{X}_i$ when its parent set is in state $\mathrm{pa} \in \mathcal{P}_i$, i.e., $\theta_{x|\mathrm{pa}} = p(X_i = x|\mathrm{Pa}_i = \mathrm{pa}) > 0$. We denote by $\boldsymbol{\theta}_{X_i|\mathrm{pa}}$ the parameters of the conditional distribution of X_i given $\mathrm{Pa}_i = \mathrm{pa}$. A common assumption [7, Sec.17] is that $\boldsymbol{\theta}_{X_i|\mathrm{pa}}$ is generated from a Dirichlet distribution with known parameters. The collection of the conditional probability distributions $\boldsymbol{\theta}_{X_i} = (\boldsymbol{\theta}_{X_i|\mathrm{pa}_1}, \ldots, \boldsymbol{\theta}_{X_i|\mathrm{pa}_{q_i}})$ constitutes the *conditional probability table* (CPT) of X_i. Each vector of type $\boldsymbol{\theta}_{X_i|\mathrm{pa}}$, with $\mathrm{pa} \in \mathcal{P}_i$, is a *column* of the CPT.

The assumption of *local parameter independence* [7, Sec. 17] allows to estimate each parameter vector $\boldsymbol{\theta}_{X_i|\mathrm{pa}}$ independently of the other parameter vectors. The assumed generative model, $\forall i \in 1, \ldots, I$, is:

$$p(\boldsymbol{\theta}_{X_i|\mathrm{pa}}) = \mathrm{Dir}(s\boldsymbol{\alpha}) \qquad \mathrm{pa} \in \mathcal{P}_i,$$
$$p(X_i|\mathrm{Pa}_i = \mathrm{pa}, \boldsymbol{\theta}_{X_i|\mathrm{pa}}) = \mathrm{Cat}(\boldsymbol{\theta}_{X_i|\mathrm{pa}}) \qquad \mathrm{pa} \in \mathcal{P}_i,$$

where $s \in \mathbb{R}$ denotes the prior strength, also called *equivalent sample size*, and $\boldsymbol{\alpha} \in \mathbf{R}^{r_i}$ is a parameter vector such that $\sum_{x \in \mathcal{X}} \alpha_x = 1$. The most common choice is to set $\alpha_x = 1/r_i$ and $s = 1/q_i$, which is called BDeu prior [7, Sec.17].

If there are no missing values in the data set D, the posterior expected value of $\theta_{x|\mathrm{pa}}$ is [6]:

$$E[\theta_{x|\mathrm{pa}}] = \frac{n_{x,\mathrm{pa}} + s\alpha_x}{n_{\mathrm{pa}} + s},$$

where $n_{x,\mathrm{pa}}$ is the number of observations in D characterised by $X_i = x$ and $\mathrm{Pa}_i = \mathrm{pa}$, while $n_{\mathrm{pa}} = \sum_{x \in \mathcal{X}_i} n_{x,\mathrm{pa}}$.

III. HIERARCHICAL MODEL

The proposed hierarchical model estimates the conditional probability tables by removing the local independence assumption. In order to simplify the notation we present the model on a node X with a single parent Y. X has r states in the set $\mathcal{X}$, while Y has q states in the set $\mathcal{Y}$. Lastly, we denote by n_{xy} the number of observations with $X = x$ and $Y = y$ and by $n_y = \sum_{x \in \mathcal{X}} n_{xy}$ the number of observations with $Y = y$, where $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.

As described in Section II, $\boldsymbol{\theta}_{X_i|\mathrm{pa}}$, for $i = 1, \ldots, I$ and $\mathrm{pa} \in \mathcal{P}_i$, are usually assumed to be independently drawn from a Dirichlet distribution, with known parameter $\boldsymbol{\alpha}$. On

the contrary, the hierarchical model treats $\boldsymbol{\alpha}$ as a hidden random vector, thus making different columns of the CPT dependent. Specifically, we assume $\boldsymbol{\alpha}$ to be drawn from a higher-level Dirichlet distribution with hyper-parameter $\boldsymbol{\alpha}_0$.

We assume that (x_k, y_k) for $k = 1, \ldots, n$ are n *i.i.d.* observations from the hierarchical Multinomial-Dirichlet model:

$$p(\boldsymbol{\alpha}|\boldsymbol{\alpha}_0) = \mathrm{Dirichlet}(\boldsymbol{\alpha}_0)$$
$$p(\boldsymbol{\theta}_{X|y}|s, \boldsymbol{\alpha}) = \mathrm{Dirichlet}(s\boldsymbol{\alpha}) \qquad y \in \mathcal{Y} \quad (1)$$
$$p(X|Y = y, \boldsymbol{\theta}_{X|y}) = \mathrm{Cat}(\boldsymbol{\theta}_{X|y}) \qquad y \in \mathcal{Y}$$

where $s \in \mathbb{R}$ is the equivalent sample size, and $\boldsymbol{\alpha}_0 \in \mathbb{R}^r$ is a vector of hyper-parameters.

A. Posterior moments for $\boldsymbol{\theta}_{X|y}$

We now study the hierarchical model, deriving an analytical expression for the posterior average of $\theta_{x|y}$, which is the element x of vector $\boldsymbol{\theta}_{X|y}$, and for the posterior covariance between $\theta_{x|y}$ and $\theta_{x'|y'}$. To keep notation simple, in the following we will not write explicitly the conditioning with respect to the fixed parameters s and $\boldsymbol{\alpha}_0$. We introduce the notation $\mathbb{E}^D[\cdot] = \mathbb{E}[\cdot|D]$ to represent the posterior average and $\mathbb{C}\mathrm{ov}^D(\cdot, \cdot) = \mathbb{C}\mathrm{ov}(\cdot, \cdot|D)$ to represent the posterior covariance.

Definition 1. We define the pointwise estimator $\hat{\theta}_{x|y}$ for the parameter $\theta_{x|y}$ as its posterior average, i.e., $\hat{\theta}_{x|y} = \mathbb{E}^D[\theta_{x|y}]$, and the pointwise estimator $\hat{\alpha}_x$ for the element x of the parameter vector $\boldsymbol{\alpha}$ as its posterior average, i.e., $\hat{\alpha}_x = \mathbb{E}^D[\alpha_x]$.

Theorem 1. *Under model* (1)*, the posterior average of $\theta_{x|y}$ is*

$$\hat{\theta}_{x|y} = \mathbb{E}^D[\theta_{x|y}] = \frac{n_{xy} + s\hat{\alpha}_x}{n_y + s}, \qquad (2)$$

while the posterior covariance between $\theta_{x|y}$ and $\theta_{x'|y'}$ is

$$\mathbb{C}\mathrm{ov}^D(\theta_{x|y}, \theta_{x'|y'}) = \delta_{yy'} \frac{\hat{\theta}_{x|y}\delta_{xx'} - \hat{\theta}_{x|y}\hat{\theta}_{x'|y}}{n_y + s + 1} + \frac{s^2 \mathbb{C}\mathrm{ov}^D(\alpha_x, \alpha_{x'})}{C_{yy'}},$$

where $C_{yy'}$ is defined as

$$C_{yy'} = \begin{cases} (n_y + s)(n_{y'} + s) & \text{if } y \neq y' \\ (n_y + s)(n_y + s + 1) & \text{if } y = y'. \end{cases}$$

The posterior average and posterior covariance of $\boldsymbol{\alpha}$ cannot be computed analytically. Some results concerning their expression and numerical computation, together with the complete proof of Theorem 1, are detailed in [9]. A sketch of the proof is reported in Appendix.

Notice that the pointwise estimator $\hat{\theta}_{x|y}$ is a mixture of traditional Bayesian estimators obtained under (non-hierarchical) Multinomial-Dirichlet models with $\boldsymbol{\alpha}$ fixed, i.e, $\frac{n_{xy} + s\alpha_x}{n_y + s}$. Indeed, thanks to the linearity in α_x, we obtain

$$\hat{\theta}_{x|y} = \frac{n_{xy} + s\hat{\alpha}_x}{n_y + s} = \int \frac{n_{xy} + s\alpha_x}{n_y + s} p(\boldsymbol{\alpha}|D) d\boldsymbol{\alpha}.$$

This mixture gives more weight to the values of $\boldsymbol{\alpha}$ that are more likely given the observations.

B. Properties of the estimator $\hat{\boldsymbol{\theta}}_{X|y}$

We study now the mean-squared error (MSE) of $\hat{\theta}_{x|y}$ and we compare it to the MSE of other traditional estimators. In order to study the MSE of $\hat{\theta}_{x|y}$ we need to assume the generative model

$$p(\boldsymbol{\theta}_{X|y}|s, \tilde{\boldsymbol{\alpha}}) = \text{Dirichlet}(s\tilde{\boldsymbol{\alpha}}) \qquad y \in \mathcal{Y},$$
$$p(X|Y = y, \boldsymbol{\theta}_{X|y}) = \text{Cat}(\boldsymbol{\theta}_{X|y}) \qquad y \in \mathcal{Y}, \quad (3)$$

where s and $\tilde{\boldsymbol{\alpha}}$ are the true underlying parameters. Moreover, since $\boldsymbol{\theta}_{X|y}$ is a random vector, we define the MSE for an estimator $\bar{\theta}_{x|y}$ of the single component $\theta_{x|y}$ as

$$\text{MSE}\left(\bar{\theta}_{x|y}\right) = \mathbb{E}_\theta\left[\mathbb{E}_n\left[\left(\bar{\theta}_{x|y} - \theta_{x|y}\right)^2\right]\right], \quad (4)$$

where $\mathbb{E}_\theta[\cdot]$ and $\mathbb{E}_n[\cdot]$ represent respectively the expected value with respect to $\theta_{x|y}$ and n_{xy}, and the MSE for the estimator $\bar{\boldsymbol{\theta}}_{X|y}$ of the vector $\boldsymbol{\theta}_{X|y}$ as $\text{MSE}\left(\bar{\boldsymbol{\theta}}_{X|y}\right) = \sum_{x \in \mathcal{X}} \text{MSE}\left(\bar{\theta}_{x|y}\right)$.

Notice that the generative model (3) is the traditional, thus non-hierarchical, Multinomial-Dirichlet model, which implies parameter independence. Hence, the traditional Bayesian estimator satisfies exactly the assumptions of this model. The Bayesian estimator is usually adopted by assuming $\tilde{\boldsymbol{\alpha}}$ to be fixed to the values of a uniform distribution on $\mathcal{X}$, i.e., $\boldsymbol{\alpha}^{\text{B}} = 1/r \cdot \mathbf{1}_{1 \times r}$, see e.g. [6]. However, since in general $\tilde{\boldsymbol{\alpha}} \neq 1/r \cdot \mathbf{1}_{1 \times r}$, the traditional Bayesian approach generates biased estimates in small samples. On the contrary, the novel hierarchical approach estimates the unknown parameter vector $\tilde{\boldsymbol{\alpha}}$ basing on its posterior distribution. For this reason the proposed approach can provide estimates that are closer to the true underlying parameters, with a particular advantage in small samples, with respect to other traditional approaches.

In order to study the MSE of different estimators, we first consider an *ideal* shrinkage estimator

$$\boldsymbol{\theta}^*_{X|y} = \tilde{\omega}_y \boldsymbol{\theta}^{\text{ML}}_{X|y} + (1 - \tilde{\omega}_y)\tilde{\boldsymbol{\alpha}}, \quad (5)$$

where $\tilde{\omega}_y \in (0, 1)$ and $\theta^{\text{ML}}_{x|y} = \frac{n_{xy}}{n_y}$ is the maximum-likelihood (ML) estimator, obtained estimating from the observations each vector $\boldsymbol{\theta}_{X|y}$ independently of other vectors. This convex combination shrinks the ML estimator towards the true underlying parameter $\tilde{\boldsymbol{\alpha}}$. Setting $\tilde{\omega}_y = \frac{n_y}{n_y + s}$, the ideal estimator corresponds to a Bayesian estimator with known parameter $\tilde{\boldsymbol{\alpha}}$, i.e., $\theta^*_{x|y} = \frac{n_{xy} + \tilde{\alpha}_x}{n_y + s}$. However, since $\tilde{\boldsymbol{\alpha}}$ represents the true underlying parameter that is usually unknown, the ideal estimator (5) is unfeasible. Yet it is useful as it allows to study the MSE.

The main result concerns the comparison, in terms of MSE, of the ideal estimator with respect to the traditional (non hierarchical) Bayesian estimator, which estimates $\tilde{\alpha}_x$ by means of the uniform distribution $1/r$, i.e., $\theta^{\text{B}}_{x|y} = \frac{n_{xy} + s/r}{n_y + s}$, see e.g. [6]. The traditional Bayesian estimator can be written as

$$\boldsymbol{\theta}^{\text{B}}_{X|y} = \omega_y \boldsymbol{\theta}^{\text{ML}}_{X|y} + (1 - \omega_y)\frac{1}{r}, \quad (6)$$

where $\omega_y = \frac{n_y}{n_y + s}$.

Theorem 2. *Under the assumption that the true generative model is* (3)*, the MSE for the ideal estimator is*

$$MSE\left(\theta^*_{x|y}\right) = \left(\tilde{\omega}^2_y \frac{s}{n_y} + (1 - \tilde{\omega}_y)^2\right)\frac{\tilde{\alpha}_x - \tilde{\alpha}^2_x}{s + 1},$$

while the MSE for the traditional Bayesian estimator is

$$MSE\left(\theta^{B}_{x|y}\right) = MSE\left(\theta^*_{x|y}\right) + (1 - \omega^2_y)\left(\tilde{\alpha}_x - \frac{1}{r}\right)^2.$$

If $\tilde{\omega}_y = \omega_y$, $MSE\left(\boldsymbol{\theta}^*_{X|y}\right) \leq MSE\left(\boldsymbol{\theta}^{B}_{X|y}\right)$.

A sketch of the proof is reported in Appendix, the complete proof is reported in [9].

Since in general $\tilde{\alpha}_x \neq \frac{1}{r}$, the second term in (2) is positive and the ideal estimator achieves smaller MSE with respect to traditional Bayesian estimator. To improve the estimates of the traditional Bayesian model in terms of MSE, we propose to act exactly on the second term of (2). Specifically, we can achieve this purpose by estimating the parameter vector $\tilde{\boldsymbol{\alpha}}$ from data, instead of considering it fixed.

The proposed hierarchical estimator defined in (2) has the same structure of the ideal estimator (5):

$$\hat{\boldsymbol{\theta}}_{X|y} = \omega_y \boldsymbol{\theta}^{\text{ML}}_{X|y} + (1 - \omega_y)\hat{\boldsymbol{\alpha}},$$

where $\omega_y = \frac{n_y}{n_y + s}$. This convex combination of $\boldsymbol{\theta}^{\text{ML}}_{X|y}$ and $\hat{\boldsymbol{\alpha}}$ shrinks the ML estimator towards the posterior average of $\boldsymbol{\alpha}$, with a strength that is inversely proportional to n_y.

Contrary to the traditional Bayesian estimator (6), the hierarchical estimator provides an estimate of $\tilde{\boldsymbol{\alpha}}$ that converges to the true underlying parameter as n increases. Indeed, it is well known that the posterior average $\mathbb{E}^D[\boldsymbol{\alpha}]$ converges to the true underlying parameter $\tilde{\boldsymbol{\alpha}}$ as n goes to infinity, i.e., $\hat{\boldsymbol{\alpha}} \to \tilde{\boldsymbol{\alpha}}$ as $n \to +\infty$, [6]. As a consequence, $\hat{\theta}_{x|y}$ converges to $\theta^*_{x|y}$ and the MSE of $\hat{\theta}_{x|y}$ converges to the MSE of $\theta^*_{x|y}$, i.e., $\text{MSE}\left(\hat{\theta}_{x|y}\right) \to \text{MSE}\left(\theta^*_{x|y}\right)$ as $n \to +\infty$.

In the finite sample assumption $\text{MSE}\left(\hat{\theta}_{x|y}\right)$ differs from $\text{MSE}\left(\theta^*_{x|y}\right)$, since $\hat{\theta}_{x|y}$ includes an estimator of $\boldsymbol{\alpha}$. Since in the hierarchical model we cannot compute this quantity analytically, we verify by simulation that the hierarchical estimator provides good performances in terms of MSE with respect to the traditional Bayesian estimators.

In conclusion, as we will show in the numerical experiments, the hierarchical model can achieve a smaller MSE than the traditional Bayesian estimators, in spite of the unfavourable conditions of the generative model (3). This gain is obtained thanks to the estimation of the parameter $\tilde{\boldsymbol{\alpha}}$, rather than considering it fixed as in the traditional approaches. In more general conditions with respect to (3), the true generating model could not satisfy parameter independence and the MSE gain of the hierarchical approach would further increase.

IV. EXPERIMENTS

In the experiments we compute the proposed hierarchical estimator using variational Bayes inference in R by means of the *rstan* package [10]. The variational Bayes estimates are practically equivalent to those yielded by Markov Chain Monte Carlo (MCMC), though being the variational inference much more efficient than MCMC (less than a second for estimating a CPT compared to a couple of minutes for the MCMC method). In the following we report the results obtained via variational inference. The code is available at http://ipg.idsia.ch/software.php?id=139.

A. MSE analysis

In the first study we assess the performances of the hierarchical estimator in terms of MSE. We consider two different settings, in which we generate observations from model (3), where $\tilde{\alpha}$ is fixed. In the first setting (test 1) we sample $\tilde{\alpha}$ from a Dirichlet distribution with parameter $\mathbf{1}_{1 \times r}$, while in the second setting (test 2) we sample it from a Dirichlet distribution with parameter $10^6 \cdot \mathbf{1}_{1 \times r}$. Under test 2 the parameters of the sampling distribution for $\tilde{\alpha}$ are very large and equal to each other, implying $\tilde{\alpha}_x \simeq 1/r$, $\forall x \in \mathcal{X}$. For this reason, test 2 is the ideal setting for the traditional Bayesian estimator, while test 1 is the ideal setting for the hierarchical estimator.

In both test 1 and test 2 we consider all the possible combinations of r (the number of states of X) and q (the number of conditioning states), with $r \in \{2, 4, 6, 8\}$ and $q \in \{2, 4, 6, 8\}$. For each combination of r and q, for both test 1 and test 2, we generate data sets with size $n \in \{20, 40, 80, 160, 320, 640\}$. We repeat the data sampling and the estimation procedure 10 times for each combination of r, q and n. Then we compare the estimates yielded by the hierarchical estimator, with $s = r$ and $\alpha_0 = \mathbf{1}_{1 \times r}$, and by the traditional Bayesian estimator assuming parameter independence, with $s = r$. We compare the performance of different estimators by computing the difference in terms of average MSE, defined as $\mathrm{MSE}(\bar{\theta}_{X|Y}) = \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} \mathrm{MSE}(\bar{\theta}_{x|y})/rq$ for an estimator $\bar{\theta}_{X|Y}$. In every repetition of the experiment $\{1, \ldots, 10\}$, we then compute $\mathrm{MSE}(\theta^{\mathrm{B}}_{X|Y}) - \mathrm{MSE}(\hat{\theta}_{X|Y})$ and we represent it in Figure 1.

The results show that the hierarchical estimator mostly provides better or equivalent results in comparison to $\theta^{\mathrm{B}}_{X|Y}$, especially for small n and/or large q. In a Bayesian network it is usual to have large values of q, since q represents the cardinality of the parents' joint states set. In test 1 (light blue boxplots) the advantage of the hierarchical estimator over the Bayesian one is generally large, as expected. The advantage of the hierarchical model steadily increases as q increases, becoming relevant for $q = 6$ or $q = 8$. For large n the gap between the two estimators vanishes, although it is more persistent when dealing with large q. Interestingly, in test 2 (green boxplots), the traditional Bayesian estimator is just slightly better than the hierarchical one, even though

the former is derived exactly from the true generative model. The traditional estimator has a small advantage only for $q = 4$ and small values of n, and this advantage quickly decreases if either q or n increase.

B. Joint distribution fitting

In the second study we assess the performance of the hierarchical estimator in the recovery of the joint distribution of a given Bayesian network.

We consider 5 data sets from UCI Machine Learning Repository: *Adult*, *Letter Recognition*, *Nursery*, *Pen-Based Recognition of Handwritten Digits* and *Spambase*. We discretise all numerical variables into five equal-frequency bins and we consider only instances without missing values. For each dataset we first learn, from all the available data, the associated directed acyclic graph (DAG) by means of a hill-climbing greedy search, as implemented in the *bnlearn* package [11]. We then keep such structure as fixed for all the experiments referring to the same data set, since our focus is not structural learning. Then, for each data set and for each $n \in \{20, 40, 80, 160, 320, 640, 1280\}$ we repeat 10 times the procedure of 1) sampling n observations from the data set and 2) estimating the CPTs. We perform estimation using the proposed hierarchical approach, with $s = r$ and $\alpha_0 = \mathbf{1}_{1 \times r}$, and the traditional BDeu prior (Bayesian estimation under parameter independence) with $s = 1$ and $s = 10$. The choice of $s = 1$ is the most commonly adopted in practice, while $s = 10$ is the default value proposed by the *bnlearn* package. Conversely, we did not offer any choice to the hierarchical model. Indeed, we set the smoothing factor s in the proposed model to the number of states of the child variable, which has the same order of magnitude of the smoothing factors used in the traditional Bayesian approach. In spite of the more limited choice for the parameter s, the hierarchical estimator consistently outperforms the traditional Bayesian estimator, regardless whether the latter adopts a smoothing factor 1 or 10.

We then measure the log-likelihood of all the instances included in the test set, where the test contains all the instances not present in training set. We report in the top panels of Figure 2 the difference between the log-likelihood of the hierarchical approach and the log-likelihood of Bayesian estimation under local parameter independence, i.e., the log-likelihood ratio. The log-likelihood ratio approximates the ratio between the Kullback-Leibler (KL) divergences of the two models. The KL of a given model measures the distance between the estimated and the true underlying joint distribution.

The log-likelihood ratios obtained in the experiments are extremely large on small sample sizes, being larger than *one thousand* on all data sets (Figure 2, top panels). This shows the huge gain delivered by the hierarchical approach when dealing with small data sets. This happens regardless of the equivalent sample size used to perform Bayesian estimation
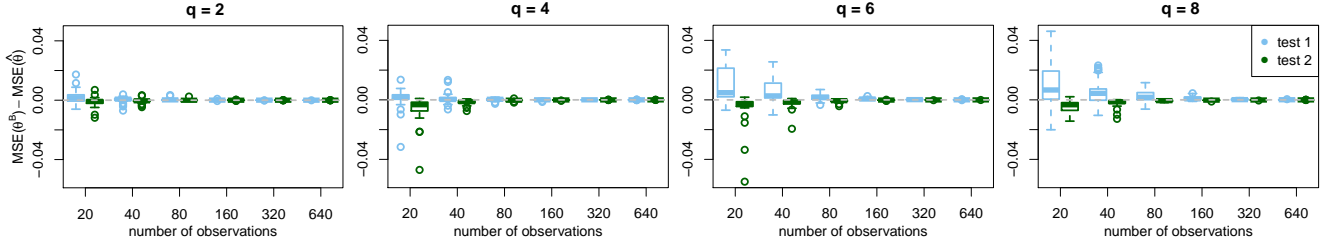
Figure 1: Boxplots of MSE difference between the Bayesian ($s = r$) and the hierarchical estimator in test 1 (light blue) and test 2 (green) with different dimension of the conditioning set ($q = 2, 4, 6, 8$). Positive values favour the hierarchical model.

under parameter independence: we note however that in general $s = 10$ yields better results than $s = 1$. The lowest gains are obtained on the data set *Nursery*; the reason is that the DAG of *Nursery* has the lowest number of parents per node (0.9 on average, compared to about twice as much for the other DAGs). Thus, this data set is the less challenging from the parameter estimation viewpoint, with respect to the others. We point out that significant likelihood ratios are obtained in general even for large samples, even though they are not apparent from the figure due to the scale. For instance, for $n = 320$ the log-likelihood ratios range from 50 (*Nursery*) to 85000 (*Letter*).

C. Classification

In the third study we assess the performance of the hierarchical estimator in terms of classification. We consider the same datasets of the previous experiment, discretised in the same way.

For each dataset we first learn the Tree-Augmented Naive Bayes (TAN) structure by means of the *bnlearn* package. The networks are estimated on the basis of all the available samples and are kept fixed for all the experiments referring to the same data set, since our focus is not structural learning. Then, for each dataset and for each $n \in \{20, 40, 80, 160, 320, 640, 1280\}$, we sample n observations. We then estimate the CPTs of the Bayesian network from the sampled data by means of the hierarchical estimator ($s = r$ and $\boldsymbol{\alpha}_0 = \mathbf{1}_{1 \times r}$) and the traditional Bayesian estimators obtained under a BDeu prior ($s = 1$ and $s = 10$). We repeat the sampling and the estimating steps 10 times for each value of n and each data set. We then classify each instance of the test set, which contains 1000 instances sampled uniformly from all the instances not included in the training set. We assess the classification performance by measuring accuracy and area under the ROC (ROC AUC) of the classifiers. In the central panels of Figure 2 we report the difference in accuracy between the hierarchical estimator and the traditional Bayesian ones, while in the bottom panels of the same figure we report the difference in ROC AUC between the same classifiers.

The area under the ROC is a more sensitive indicator for the correctness of the estimated posterior probabilities

with respect to accuracy. According to Figure 2 (bottom panels), the hierarchical approach yields consistently higher ROC AUC than both the BDeu classifiers. The increase of ROC AUC in small samples ($n = 20$, $n = 40$) ranges between *2* and *20* points compared to both the BDeu priors. As n increases this gain in ROC AUC tends to vanish. However, for $n > 320$ the gain in ROC AUC for the datasets *Adult* and *Letter* ranges between 1 and 5 points.

Figure 2 (central panels) shows also improvements in accuracy, even if this indicator is less sensitive to the estimated posterior probability than the area under ROC. Indeed, in computing the accuracy, the most probable class is compared to the actual class, without paying further attention to its posterior probability. In small samples ($n = 20$, $n = 40$) there is an average increase of accuracy of about *5* points compared to the BDeu prior with $s = 10$ and of about *10* points compared to the BDeu prior with $s = 1$. The accuracy improvements tends to decrease as n increases; yet on both *Adult* and *Letter* data sets an accuracy improvement of about 1-2 points is shown also for $n = 1280$.

V. CONCLUSIONS

We have presented a novel approach for estimating the conditional probability tables by relaxing the local independence assumption. Given the same network structure, the novel approach yields a consistently better fit to the joint distribution than the traditional Bayesian estimation under parameter independence; it also improves classification performance. Moreover, the introduction of variational inference makes the proposed method competitive in terms of computational cost with respect to the traditional Bayesian estimation.

APPENDIX

Sketch of proof for Theorem 1: The posterior average of $\theta_{x|y}$ can be computed by means of the law of total expectation as $\mathbb{E}^D[\theta_{x|y}] = \mathbb{E}^D[\mathbb{E}^{\boldsymbol{\alpha},D}[\theta_{x|y}]]$. The posterior covariance between $\theta_{x|y}$ and $\theta_{x'|y'}$ is computed by means of the law of total covariance, deriving under model (1) $\mathbb{C}\text{ov}^D(\mathbb{E}^{\boldsymbol{\alpha},D}[\theta_{x|y}], \mathbb{E}^{\boldsymbol{\alpha},D}[\theta_{x'|y'}])$ and $\mathbb{E}^D[\mathbb{C}\text{ov}^{\boldsymbol{\alpha},D}(\theta_{x|y}, \theta_{x'|y})]$. ∎
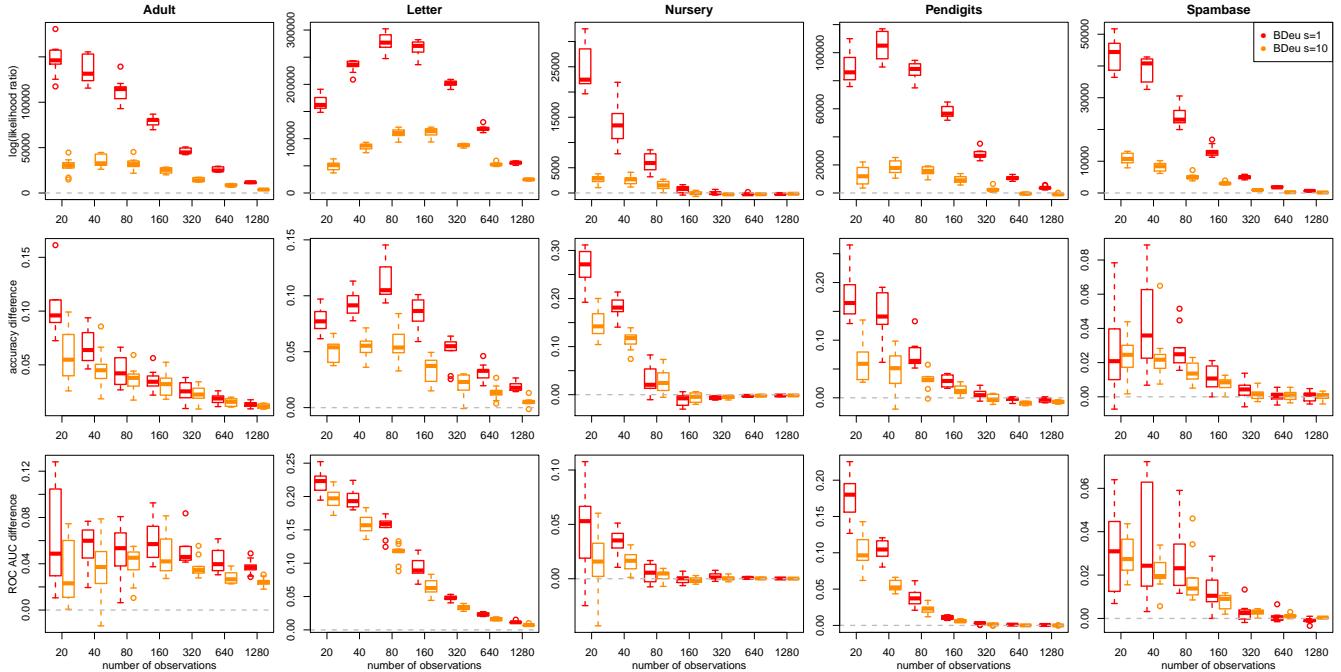
Figure 2: Boxplots of the logarithm of the likelihood ratio (top panels), accuracy gain (central panels) and area under the ROC gain (bottom panels) obtained comparing the hierarchical method with respect to BDeu ($s = 1$ in orange and $s = 10$ in red) for the five machine learning datasets analysed. Positive values favour the hierarchical model.

Sketch of proof for Theorem 2: Exploiting the linearity of the ideal estimator we obtain that $\theta^*_{x|y} - \theta_{x|y} = \tilde\omega_y\big(\theta^{\mathrm{ML}}_{x|y} - \theta_{x|y}\big) + (1 - \tilde\omega_y)\big(\tilde\alpha_x - \theta_{x|y}\big)$. Using the definition of MSE (4) and the assumptions of model (3), i.e., $\mathbb{E}_n\left[n_{xy}\right] = n_y\theta_{x|y}$, $\mathbb{E}_\theta\left[\theta_{x|y}\right] = \tilde\alpha_x$, $\mathbb{V}\mathrm{ar}_\theta\left(\theta_{x|y}\right) = \frac{\tilde\alpha_x - \tilde\alpha_x^2}{s+1}$ and $n_{xy} \perp\!\!\!\perp \tilde\alpha - \theta_{x|y}$, we obtain $\mathrm{MSE}\big(\theta^*_{x|y}\big) = \tilde\omega_y^2\mathrm{MSE}\big(\theta^{\mathrm{ML}}_{x|y}\big) + (1 - \tilde\omega_y)^2\mathrm{MSE}\left(\tilde\alpha_x\right)$, where $\mathrm{MSE}\big(\theta^{\mathrm{ML}}_{x|y}\big) = \frac{s}{n_y}\frac{\tilde\alpha_x - \tilde\alpha_x^2}{s+1}$ and $\mathrm{MSE}\left(\tilde\alpha_x\right) = \mathbb{V}\mathrm{ar}_\theta\left(\theta_{x|y}\right)$. Analogously, we compute the MSE for the Bayesian estimator (6) as $\mathrm{MSE}\big(\theta^{\mathrm{B}}_{x|y}\big) = \omega_y^2\mathrm{MSE}\big(\theta^{\mathrm{ML}}_{x|y}\big) + (1 - \omega_y^2)\mathrm{MSE}\left(1/r\right)$, where $\mathrm{MSE}\left(1/r\right) = \big(\tilde\alpha_x - \frac{1}{r}\big)^2 + \mathbb{V}\mathrm{ar}_\theta\left(\theta_{x|y}\right)$. Combining the terms, we obtain (2) and (2). If $\tilde\omega_y = \omega_y$, the inequality $\mathrm{MSE}\big(\theta^*_{x|y}\big) \le \mathrm{MSE}\big(\theta^{\mathrm{B}}_{x|y}\big)$ follows directly. ∎

Acknowledgment

References

[1] J. Hausser and K. Strimmer, "Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks," *J. Mach. Learn. Res.*, vol. 10, no. Jul, pp. 1469–1484, 2009.

[2] G. Casella and E. Moreno, "Assessing robustness of intrinsic tests of independence in two-way contingency tables," *J. Am. Stat. Assoc.*, vol. 104, no. 487, pp. 1261–1271, 2009.

[3] I. Good and J. Crook, "The robustness and sensitivity of the mixed-Dirichlet Bayesian test for independence in contingency tables," *Ann. Stat.*, pp. 670–693, 1987.

[4] I. Nemenman, F. Shafee, and W. Bialek, "Entropy and inference, revisited," in *Adv. Neural Inf. Process. Syst.*, 2002, pp. 471–478.

[5] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[6] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian data analysis*. CRC Press, 2013.

[7] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[8] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[9] L. Azzimonti, G. Corani, and M. Zaffalon, "Hierarchical Multinomial-Dirichlet model for the estimation of conditional probability tables," *ArXiv:1708.06935*, 2017.

[10] B. Carpenter, A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell, "Stan: A probabilistic programming language," *J. Stat. Softw.*, vol. 76, no. 1, pp. 1–32, 2017.

[11] M. Scutari, "Learning Bayesian networks with the bnlearn R package," *J. Stat. Softw.*, vol. 35, no. 3, 2010.