Discrete exponential bayesian networks structure learning for density estimation

Aida Jarraya, Philippe Leray, Afif Masmoudi

Discrete exponential Bayesian networks structure learning for density estimation

Aida Jarraya[1,2], Philippe Leray[2], and Afif Masmoudi[1]
[1] Laboratory of Probability and Statistics
Faculty of Sciences of Sfax, University of Sfax. Tunisia.

aidajarraya@yahoo.fr, afif.masmoudi@fss.rnu.tn

[2] LINA Computer Science Lab UMR 6241
Knowledge and Decision Team, University of Nantes, France.

philippe.leray@univ-nantes.fr

Abstract. Our work aims at developing or expliciting bridges between Bayesian Networks and Natural Exponential Families, by proposing discrete exponential Bayesian networks as a generalization of usual discrete ones. In this paper, we illustrate the use of discrete exponential Bayesian networks for Bayesian structure learning and density estimation. Our goal is to empirically determine in which contexts these models can be a good alternative to usual Bayesian networks for density estimation.

1 Introduction

Bayesian networks (BNs) are probabilistic graphical models used to model complex systems with variables of different natures. In the literature we find many works about discrete Bayesian network where the conditional distribution of each variable given its parents is a multinomial distribution. As initially proposed by [5], we are interested in extending this conditional distribution to natural exponential families (NEF) [1][9]. This idea has been used by [2] for instance with conjugate-exponential models, for Bayesian networks with latent variables. They concentrate their work on variational EM estimation needed because of latent variables, but they don't explicit the Bayesian estimators used and restrict their experiments to usual multinomial distributions. [12] also propose one great study of graphical models as exponential families, showing that very specific structures of directed or undirected probabilistic graphical models can be interpreted as an element of exponential family. Our work deals with the same general idea, developing or expliciting bridges between BNs and NEFs, dealing with discrete exponential BNs instead of usual discrete ones. We formally introduced in [6] discrete exponential Bayesian networks (deBNs) with a specific prior family proposed by [4] and demonstrated that this prior is a generalization of Dirichlet priors usually considered with discrete BNs. We illustrate now the use of deBNs for Bayesian structure learning and density estimation. Our goal is to empirically determ-

ine in which contexts deBNs can be a good alternative to usual BNs for density estimation. The present paper is structured as follows. In section 2, we summarized our theoretical results concerning structure and parameter learning for discrete exponential BNs. Section 3 then describes our experimental protocol, evaluation criteria, and finally gives interpretations of results obtained in this context. Section 4 concludes our paper by giving some perspectives for future work.

2 Discrete exponential Bayesian network

Usually, the statistical model of a discrete BN is a multinomial distribution [11]. We described in [6] how to use discrete exponential families in a more general way. We summarized here the main points concerning the definition of discrete exponential BNs (deBNs), our proposal of Bayesian scoring function and parameter estimator.

2.1 Notations

A Bayesian network (BN) is defined as a set of variables $X = \{X_1, \ldots, X_n\}$ with a network structure G that encodes a set of conditional independence assertions about variables in X, and a set P of local probability distributions associated with each variable. Together, these components define the joint probability distribution for X. The network structure G is a directed acyclic graph (DAG). Each X_i denotes both the variable and its corresponding node in G, and $Pa(X_i)$ the parents of node X_i in G. For BN learning, we assume that we have one dataset $d = \{x^{(1)}, \ldots, x^{(M)}\}$ of size M where $x^{(l)} = \{x_1^{(l)}, \ldots, x_n^{(l)}\}$ is the l^{th} sample and $x_i^{(l)}$ is the value of variable X_i for this sample.

2.2 DeBN definition

A discrete exponential Bayesian network is defined as a Bayesian network where conditional probability distributions are discrete natural exponential families (NEF). Let F be a NEF, usually described by its parameters $\upsilon, k_\upsilon = k$ and $\Psi_\upsilon = \Psi$. These general parameters allow us to describe any discrete exponential distribution (Poisson, Negative Binomial, ...).

We suppose that $X_i|Pa_i = j \sim P(\mu_{ij}, F)$. This conditional probability distribution can be expressed in an "exponential" way, where μ_{ij} parameters are mutually independent

$$P(x_i|Pa_i = j) = e^{\langle \psi(\mu_{ij}), x_i \rangle - k(\psi(\mu_{ij}))} \upsilon\{x_i\}. \quad (1)$$

For Bayesian estimation or structure learning, we also need to define a prior distribution for parameters μ_{ij}. In [6], we propose to choose the $\widetilde{\Pi}$ prior family introduced by [4] and demonstrated that this prior was a generalization of Dirichlet priors usually considered with discrete BNs. So $\mu_{ij} \sim \widetilde{\Pi}_{t_{ij}, m_{ij}}$ with

$$\tilde{\Pi}_{t_{ij},m_{ij}}(\mu_{ij}) = \tilde{K}_{t_{ij},m_{ij}} e^{t_{ij}<\psi(\mu_{ij}),m_{ij}> - t_{ij} k(\psi(\mu_{ij}))} \quad (2)$$

where $\tilde{K}_{t_{ij},m_{ij}}$ is a normalizing constant depending on the considered NEF.

2.3 DeBN learning

As for their usual counterpart, deBN structure learning can be performed by using any heuristic method whose objective is the optimization of the marginal likelihood or one of its approximations. In the Bayesian estimation framework, we described in [6] the computation of this marginal likelihood for discrete exponential BN and a generalized scoring function gBD extending the Bayesian Dirichlet (BD) score to any NEF distribution.

$$gBD(d,G) = P(G)\prod_{i=1}^{n}\prod_{j=1}^{q_i} \frac{\tilde{K}_{t_{ij},m_{ij}}}{\tilde{K}_{N_{ij}+t_{ij},\frac{\sum_{h\in M_{ij}} x_i^{(h)} + t_{ij}m_{ij}}{N_{ij}+t_{ij}}}} \quad (3)$$

where $M_{ij} = \left\{ h \in \{1,...,M\} \,\middle|\, Pa_i^{(h)} = j \right\}$ and $N_{ij} = \left| M_{ij} \right|$.

We also demonstrated that the Maximum a Posteriori (MAP) estimator of parameter μ_{ij} is given by the following closed form : $\hat{\mu}_{ij}^{MAP} = \left(\frac{\overline{X}_i + t_{ij}m_{ij}}{(t_{ij}/N_{ij}) + 1}\right)$ (4)

where $\overline{X}_i = \frac{1}{N_{ij}} \sum_{h\in M_{ij}} x_i^{(h)}$.

2.4 DeBN examples: Poisson and Negative Binomial BNs

Let us apply these previous results to Poisson and Negative Binomial distributions. For the Poisson distribution, the normalizing constant is

$$\tilde{K}_{t_{ij},m_{ij}} = \frac{t_{ij}^{t_{ij}m_{ij}+1}}{\Gamma(t_{ij}m_{ij}+1)}. \quad (5)$$

The score function gBD(d,G) is given by

$$P(G)\prod_{i=1}^{n}\prod_{j=1}^{q_i} \frac{t_{ij}^{t_{ij}m_{ij}+1}}{(N_{ij}+t_{ij})^{\sum_{h\in M_{ij}} x_i^{(h)} + t_{ij}m_{ij}+1}} \frac{\Gamma(t_{ij}m_{ij} + \sum_{h\in M_{ij}} x_i^{(h)} + 1)}{\Gamma(t_{ij}m_{ij}+1)} \quad (6)$$

For the Negative Binomial Model, we get the following normalizing constant:

$$\widetilde{K}_{t_{ij},m_{ij}} = \frac{\Gamma(t_{ij}m_{ij} + t_{ij})}{\Gamma(t_{ij}m_{ij} + 1)\Gamma(t_{ij} - 1)}, \qquad t_{ij} > 1 . \tag{7}$$

The score function gBD(d,G) is given by

$$P(G)\prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{\Gamma(t_{ij}m_{ij} + t_{ij})\Gamma(t_{ij}m_{ij} + \sum_{h\in M_{ij}} x_i^{(h)} + 1)\Gamma(N_{ij} + t_{ij} - 1)}{\Gamma(t_{ij}m_{ij} + 1)\Gamma(t_{ij}m_{ij} + \sum_{h\in M_{ij}} x_i^{(h)})\Gamma(t_{ij} - 1)} \tag{8}$$

3 Experimentations

3.1 Data

In order to evaluate the interest of using deBNs instead of usual BNs for density estimation, we carried out repetitive experiments in several contexts. In the first context, data are generated from distributions described by usual BNs (dist=multi). In the second context, data are generated from distributions described by Poisson deBNs (dist=poisson). In these contexts, we are able to control several parameters such as the number n of variables (n = 10, 20, 50) and the size M of generated datasets (M = 100, 1.000, 10.000). The maximal cardinality K of our discrete variables is also controlled for usual BNs (K = 2, 3, 5) but measured in the generated samples for Poisson deBNs.

Every dataset generation in such conditions is iterated 10x10 times, with 10 randomly generated DAGs, and 10 random parameter values for each of these DAGs.

3.2 Models and algorithms used

Our goal is comparing performances of usual discrete BN models (model=multi) versus Poisson deBN (model=poisson) learnt with the previous datasets. Prior parameters are chosen in their simplest form, $\alpha_{ij} = 1$, uniform Dirichlet coefficient, for discrete BNs and $t_{ij} = 1$, $m_{ij} = 0$ for Poisson deBNs.

Structure learning procedure used for optimizing the Bayesian scoring function is an usual greedy search procedure as proposed in [3]. In order to obtain more robust results, this greedy search is performed 10 times with different random initializations and the best result of the 10 runs is kept. Maximum A Posteriori estimation is used for parameter learning. Our various models and algorithms have been implemented in Matlab with BNT [10] and BNT Structure Learning Package [8].

3.3 Evaluation criteria

Accuracy evaluation of each model is estimated by the Kullback-Leibler (KL) divergence between the "original" distribution described by the model used for generating a given dataset and the "final" distribution obtained by the model learnt from this dataset. For large numbers of variable configurations (greater than 10^5), an MCMC approximation is used with 10^5 random configurations.

Comparison of both models is illustrated by plotting absolute values of KL obtained by deBNs versus usual BNs for the same datasets. The fact that one model is better than the other can be observed with respect to the first diagonal (upper triangle : deBN is better, versus lower triangle : usual BN is better). In order to determine whether the observed differences are statistically significant, we use the Wilcoxon paired signed rank test, with a significance level equal to 0.05, for the 100 experiments performed for one given context (dist, n, M, K).

3.4 Results and interpretations

Our preliminary results described in Figure 1 concern n = 10. As we can see, when data are generated from Poisson distributions (results in magenta), our Poisson deBNs are logically better models than usual BNs. When data are generated from multinomial distributions (results in blue), results depend on the sample size M. When M is high (M = 10.000, third figure on the right), usual BNs are better models than Poisson deBNs. When the sample size decreases (M = 1000), usual BNs and deBNs give similar results. With a small sample size (M = 100), deBNs are better than usual BNs. All these results are confirmed by Wilcoxon tests. By comparing results for M = 1.000 with respect to the maximum variable cardinality K (not described right now in the figure), we can observe than deBNs and usual BNs give similar results for K = 3 but the situation changes if K increases. For K = 6, deBN give better results than BNs.
These intermediate results need to be completed with other values of n and K, but we can already observe than deBNs seems to be a good alternative to usual BNs in several contexts. If we compare Poisson deBNs and usual BNs, the first ones have less free parameters than the others and this number of parameters is less dependent of the value of K. So when the sample size M is low or when the value of K is high, deBN are a good compromise for density estimation.

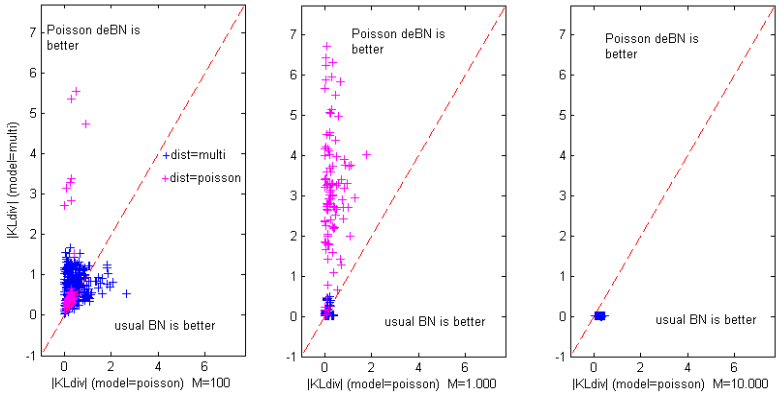

Fig. 1. Comparison of KL divergence obtained by Poisson deBNs versus usual BNs for the same datasets (upper triangle : deBN is better, versus lower triangle : usual BN is better) with respect to dataset size (M = 100, 1.000, 10.000) and data original distribution (dist=poisson vs. multinomial) for n = 10.

4 Conclusion and perspectives

In this paper, we have developed the concept of discrete exponential Bayesian network (deBN) previously, described in [6], for Bayesian structure learning and density estimation. Experiments described here show us that Poisson deBNs can be a good alternative to usual BNs for density estimation when the sample size is low or when the maximum variable cardinality is high, because of the reduced number of parameters used by deBNs. These experiments could be extended to other discrete NEF distributions such as the Negative Binomial one, or continuous distributions. For each distribution, we need to propose a better way to deal with a priori parameters such as t_{ij} and m_{ij} for Poisson distribution, in order to obtain Bayesian scoring functions verifying the Markov equivalence property (like BDe scoring function for usual discrete BNs). Probabilistic inference algorithms also have to be extended for these distributions, which seems to be not so difficult for any exponential distribution as shown in [7] for hybrid BNs with conditional Gaussian distributions.

References

1. Barndorff-Nielsen, O.: Information and Exponential families in Statistical Theory. John Wiley (1978)
2. Beal, M., Ghahramani, Z.: The variational bayesian EM algorithm for incomplete data:with application to scoring graphical model structures. Bayesian Statistics 7, 453-464 (2003)
3. Chickering, D., Geiger, D., Heckerman, D.: Learning bayesian networks: Search methods and experimental results. In: Proceedings of Fifth Conference on Artificial Intelligence and Statistics. pp. 112-128 (1995)
4. Consonni, G., Veronese, P.: Conjugate priors for exponential families having quadratic variance functions. J. Amer. Statist. Assoc 87, 1123-1127 (1992)
5. Geiger, D., Heckerman, D., King, H., Meek, C.: Stratified exponential families: graphical models and model selection. Annals of Statistics 29, 505-529 (2001)
6. Jarraya, A., Leray, P., Masmoudi, A.: Discrete exponential bayesian networks: an extension of bayesian networks to discrete natural exponential families. In: 23rd IEEE International Conference on Tools with Artificial Intelligence (ICTAI'2011). pp. 205-208. Boca Raton, Florida, USA (2011)
7. Lauritzen, S.L., Jensen, F.: Stable local computation with conditional Gaussian distributions. Statistics and Computing 11(2), 191-203 (Apr 2001)
8. Leray, P., Francois, O.: BNT structure learning package: Documentation and experiments. Tech. rep., Laboratoire PSI (2004)
9. Letac, G.: Lectures on natural exponential families and their variance functions. No. 50 in Monograph. Math., Inst. Mat. Pura Aplic. Rio (1992)
10. Murphy, K.: The bayesnet toolbox for matlab. In: Computing Science and Statistics: Proceedings of Interface. vol. 33 (2001)
11. Studeny, M.: Mathematical aspects of learning bayesian networks: Bayesian quality criteria. Research Report 2234, Institute of Information Theory and Automation, Prague (December 2008)
12. Wainwright, M.J., Jordan, M.I.: Graphical models, exponential families, and variational inference. Foundations and Trends in Machine Learning 1(1-2), 1-305 (2008)