Structure learning of Bayesian networks with MCMC: extension to incomplete data and decision trees as local models

Péter Antal, András Millinghoffer, Gábor Hullám, Dirk Timmerman, Yves Moreau, Olivier Gevaert, and Bart De Moor, *Fellow, IEEE*

***Abstract*—Full Bayesian inference of structural properties remains a challenge for graphical models. Yet, it would be an attractive option, in particular in bioinformatics and biomedicine where the relative scarcity of the data compared to the number of variables, the possibility of prior knowledge incorporation, and the potential of posterior exploration and fusion from multiple sources make a full Bayesian treatment appealing. First, we overview recent results about exact model averaging and about Directed Acyclic Graph (DAG) MCMC and ordering MCMC for Bayesian networks, which allows Bayesian inference of conditional independencies, causal statements, and relevant features as well. Next, we describe the use of a novel Metropolis-in-Gibbs scheme to use incomplete data. Then we discuss the challenge of using sparser (and thus statistically more tractable) local models, such as decision trees, in DAG MCMC and ordering MCMC. Because of the high computational complexity of these methods, we consider the applicability of desktop grids, distributed shared-memory grids, and dedicated hardware as well. The proposed methods are illustrated on a real-world medical problem in the area of ovarian cancer.**

***Index Terms*—Bayesian networks, Bayesian inference, biomedicine, bioinformatics, MCMC, local models, incomplete data, parallel computation**

I. Introduction

PROBABILISTIC graphical models are widely used tools in biomedical data exploration both for the general characterization of the domain and also for the more focused analysis of feature subsets. In both cases, a main focus is the structural properties of the underlying model (i.e., graph properties). Because of the high sample complexity of model identification, the Bayesian statistical framework is attractive because it allows model averaging as an automated solution for the multiple testing problem and marginalization to detect relevant or simply identifiable aspects of the model. In Section II, we overview recent results about exact model averaging and MCMC methods for Bayesian networks. In Section III we present current developments in the Bayesian approach to the problem of feature subset selection (FSS). We also discuss the pros and contras of using domain models (e.g., Bayesian networks) versus conditional models (e.g., logistic regression, decision trees, SVMs) in FSS, especially in the case of incomplete data. In Section IV we introduce a novel Metropolis-in-Gibbs algorithm for structure learning, which performs a random walk in the joint space of DAGs and missing values. In Section V, we summarize approaches to decrease model complexity by parameter sharing (e.g., module networks) and by the application of sparse, domain specific local models (e.g., restricted decision trees) in Bayesian networks. Next, we discuss the extensions of DAG MCMC and ordering MCMC over Bayesian networks with decision trees as local models, in which we integrate existing MCMC methods for Bayesian networks with decision trees as building blocks. Finally in Section VI, we discuss computational complexity to understand the use of grids, distributed shared memory systems, FPGAs for various parallelization schemes to scale up the method for high-dimensional bioinformatics tasks. Methods will be illustrated on the modeling of clinical data for ovarian cancer.

P. Antal, A, Millonghoffer, G. Hullm is with the Department of Measurement and Information Systems, Budapest University of Technology and Economics, e-mail: (see http://www.mit.bme.hu/ antal).

D. Timmerman is with the Department of Obstetrics and Gynecology, University Hospitals Leuven, Katholieke Universiteit Leuven

O. Gevaert, Y. Moreau, and B. De Moor are with the Department of Electrical Engineering, Katholieke Universiteit Leuven

Manuscript received July 31, 2009;

II. Earlier work

Bayesian inference of structural properties of Bayesian networks started with the introduction of Dirichlet parameter priors with parameter independence by Spiegelhalter et al. [1]. In the beginning of the 1990's the full Bayesian approach was proposed in a seminal paper by Buntine et al. [2]. Cooper et al. [3] discussed the general use of the posterior over Bayesian network structures as an inductive probabilistic knowledge base. Later, Madigan et al. [4] proposed an MCMC scheme to approximate such Bayesian inference, while Heckerman [5] considered the application of this full Bayesian treatment to causal Bayesian networks. The Directed Acyclic Graph (DAG) MCMC method was improved by Castelo et al. [6]. Also, Dash et al. [7] reported a method to perform exact full Bayesian inference in a restricted case of naive Bayesian classifiers. Friedman et al. [8] reported the ordering MCMC scheme. Koivisto et al. [9], proposed a method to perform exact full Bayesian inference over modular features in $\mathcal{O}(n2^n)$ time. For the application of Metropolis Coupled MCMC, see [10], [11].

In the paper we will summarize the advantages and inconvenients of DAG MCMC versus ordering MCMC, such as the effect of bias of the explicit prior over the orderings and the effect of analytical marginalization (a.k.a. Rao-Blackwellisation) [12], the effect of a high number of structural features on confidence estimation (i.e., the top K most probable feature problem [13]), and integrated MCMC

estimation and feature search in applying ordering MCMC for complex structural features [14].

III. Bayesian approach to the feature subset selection problem

Previously we presented the methodology of Bayesian Multilevel Analysis (BMLA) to detect the relevance of input variables [14], [13]. BMLA enables the analysis of relevance at different abstraction levels: model-based pairwise relevance, relevance of variable sets, and interaction models of relevant variables. In the Bayesian model averaging framework, each of these levels corresponds to a structural property of Bayesian networks (i.e., Markov Blanket Membership, Markov Blanket set, and Markov Blanket graph), and the essence of BMLA is that the estimated posteriors of these properties can be used to assess the relevance of input variables. Furthermore, Markov blanket graph posteriors provide principled confidence measures for multivariate variable selection and facilitates the identification of interaction models of relevant variables (for an extension with scalable structural properties see [15]).

In the paper we summarize the pros and contras of using domain models and conditional models in the Bayesian approach to Feature Subset Selection (FSS), because Bayesian conditional methods (e.g., logistic regression or multilayer perceptrons) are widely used in biomedicine (e.g., see [16], [17], [18], [19], [20]).

IV. Sampling the posterior using incomplete data

In case of incomplete data, a frequently made assumption is that unknown values are Missing at Random (MAR) [16]. In case of model identification, the Expectation-Maximization (EM) algorithm is one of the most well-known methods for imputation. In the context of graphical models, several EM variants were developed, such as Structural EM and Bayesian Structural EM by Friedman et al. [21], [22]. Another approach to learning parameters in Bayesian networks (BN) is the Bound and Collapse algorithm [23].

To sample from the posterior in case of incomplete data, Tanner and Wong proposed a two-phased iterative algorithm inspired by EM [24]. In its imputation step missing values are drawn from a predictive distribution, then in the posterior step a parameter value is drawn from the posterior distribution conditional on the data set completed in the previous step (e.g., see [16], [10]). For sampling from the parameter posterior in case of a fixed BN structure using Gibbs sampling, see [25]. Other methods applying the data augmentation principle for sampling both parameters and structure went one step further and combined Gibbs sampling and importance sampling [26], [27].

In the paper we propose a Metropolis-Hastings within Gibbs method to sample from the posterior over DAG structures in case of incomplete data. The Gibbs sampling steps are used to fill out missing values, then using the completed data set a DAG structure can be sampled in the M-H step. We explain the potential of such a combination, which is valid because the Gibbs sampler itself is a special case of M-H. Indeed, within the M-H method the kernels can be sampled sequentially without the need to reach convergence separately according to the product of kernels principle [28].

V. MCMC over Bayesian networks with decision tree local models

A serious challenge for graphical models as conditional independence models are the representation of contextual independencies (for a discussion and refinement of the concept, see [29]). For the explicit representation of contextual independencies, Boutilier et al. [30] proposed decision trees. From a statistical point of view this problem arises as a model complexity problem, for which special, sparse, local models, such as the classical noisy-OR were suggested (for parameter sharing, see for example module networks [31]). Decision trees and decision graphs as local models were investigated also from an inductive point of view [32], [33], [34] (for the management of continuous variables with hybrid BNs using ANNs as local models, see [35]).

However, the Bayesian treatment of such sparse BNs is currently unsolved. Even for decision trees, the Bayesian treatment is challenging (for Bayesian model averaging, see [36]; for Bayesian inference over decision trees, see [37], [38], [39], [40], [41]; for Bayesian phylogeny, see [42]).

The use of MCMC methods over decision trees as building blocks can be considered both in the DAG MCMC and in the ordering MCMC. In the DAG MCMC however the application of a M-H-in-Gibbs method is problematic, because the trees can not be sampled separately given the global DAG contraint. In the paper, we discuss the use of decision graphs and present an adaptation of ordering MCMC for BNs with decision trees.

VI. Advantages of parallelization

Finally, in the paper we will summarize the computational complexity of Bayesian inference methods, including theoretical and empirical bounds on the rate of MCMC convergence, and consider the use of shared memory systems, cluster/grid computing, and FPGAs [43], [44].

VII. Application to clinical modeling of ovarian cancer

All methods will be illustrated using clinical data on ovarian cancer domain [45]. More specifically, we analyze a multicentre data set containing more than 50 variables and 3500 patients to assess the malignancy of ovarian masses. Previous work on this data set focused on solving a binary classification problem and excluding cases with missing values. In the paper we will extend this to a multi-class classification problem corresponding to the subclassification of the tumor. In addition, an important biomarker for ovarian cancer, CA125, is often missing in this data set and thus provides an excellent case for methods to deal with incomplete data.

References

[1] D. J. Spiegelhalter, A. Dawid, S. Lauritzen, and R. Cowell, "Bayesian analysis in expert systems," *Statistical Science*, vol. 8, no. 3, pp. 219–283, 1993.

[2] W. L. Buntine, "Theory refinement of Bayesian networks," in *Proc. of the 7th Conf. on Uncertainty in Artificial Intelligence (UAI-1991)*. Morgan Kaufmann, 1991, pp. 52–60.
[3] G. F. Cooper and E. Herskovits, "A Bayesian method for the induction of probabilistic networks from data," *Machine Learning*, vol. 9, pp. 309–347, 1992.
[4] D. Madigan, S. A. Andersson, M. Perlman, and C. T. Volinsky, "Bayesian model averaging and model selection for Markov equivalence classes of acyclic digraphs," *Comm.Statist. Theory Methods*, vol. 25, pp. 2493–2520, 1996.
[5] D. Heckermann, C. Meek, and G. Cooper, "A bayesian aproach to causal discovery," Technical Report, MSR-TR-97-05, 1997.
[6] P. Giudici and R. Castelo, "Improving Markov Chain Monte Carlo model search for data mining," *Machine Learning*, vol. 50, pp. 127–158, 2003.
[7] D. Dash and G. F. Cooper, "Exact model averaging with naive bayesian classifiers," in *Proceedings of the Nineteenth International Conference on Machine Learning*, 2002, pp. 91–98.
[8] N. Friedman and D. Koller, "Being Bayesian about network structure," *Machine Learning*, vol. 50, pp. 95–125, 2003.
[9] M. Koivisto and K. Sood, "Exact bayesian structure discovery in bayesian networks," *Journal of Machine Learning Research*, vol. 5, pp. 549–573, 2004.
[10] J. S. Liu, *Monte Carlo Strategies in Scientific Computing*. Springer-Verlag, 2004.
[11] G. Altekar, S. Dwarkadas, J. P. Huelsenbeck, and F.Ronquist, "Parallel metropolis coupled markov chain monte carlo for bayesian phylogenetic inference," *Bioinformatics*, vol. 20, no. 3, pp. 407–415, 2004.
[12] P. Antal, *Integrative Analysis of Data, Literature, and Expert Knowledge*. Ph.D. dissertation, K.U.Leuven, D/2007/7515/99, 2007.
[13] A. Millinghoffer, G. Hullám, and P. Antal, "On inferring the most probable sentences in bayesian logic," in *Workshop notes on Intelligent Data Analysis in bioMedicine And Pharmacology (IDAMAP-2007)*, 2007, pp. 13–18.
[14] P. Antal, G. Hullám, A. Gézsi, and A. Millinghoffer, "Learning complex bayesian network features for classification," in *Proc. of third European Workshop on Probabilistic Graphical Models*, 2006, pp. 9–16.
[15] P. Antal, A. Millinghoffer, G. Hullám, C. Szalai, and A. Falus, "A bayesian view of challenges in feature selection: Feature aggregation, multiple targets, redundancy and interaction," *JMLR Proceeding*, vol. 4, pp. 74–89, 2008.
[16] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*. London: Chapman & Hall, 1995.
[17] P. Antal, G. Fannes, D. Timmerman, Y. Moreau, and B. D. Moor, "Bayesian applications of belief networks and multilayer perceptrons for ovarian tumor classification with rejection," *Artificial Intelligence in Medicine*, vol. 29, pp. 39–60, 2003.
[18] C. Kooperberg and I. Ruczinski, "Identifying interacting snps using monte carlo logic regression," *Genet Epidemiol*, vol. 28, no. 2, pp. 157–170, 2005.
[19] D. J. Balding, "A tutorial on statistical methods for population association studies," *Nature*, vol. 7, pp. 781–91, 2006.
[20] M. A. Province and I. B. Borecki, "Gathering the gold dust: Methods for assessing the aggregate impact of small effect genes in genomic scans," in *Proc. of the Pacific Symposium on Biocomputing (PSB08)*, vol. 13, 2008, pp. 190–200.
[21] N. Friedman, "Learning belief networks in the presence of missing values and hidden variables," in *Proceedings of the Fourteenth International Conference on Machine Learning*, 1997, pp. 125–133.
[22] ——, "The bayesian structural EM algorithm," in *Proc. of the 14th Conf. on Uncertainty in Artificial Intelligence(UAI-1998)*. Morgan Kaufmann, 1998, pp. 129–138.
[23] M. Ramoni and P. Sebastiani, "Robust learning with missing data," *Machine Learning*, vol. 45, no. 2, pp. 147–170, 2001.
[24] M. A. Tanner and W. H. Wong, "The calculation of posterior distributions by data augmentation," *Journal of the American Statistical Association*, vol. 82, no. 398, pp. 528–540, 1987.
[25] D. Heckerman, *Learning in graphical models*. Cambridge, MA: MIT Press, 1999, ch. A Tutorial on Learning with Bayesian Networks.
[26] C. Riggelsen, "Learning bayesian networks from incomplete data: An efficient method for generating approximate predictive distributions," in *Proceedings of the SIAM International Conference on Data Mining, SDM'06*, 2006, p. 00.
[27] C. Riggelsen and A. Feelders, "Learning bayesian network models from incomplete data using importance sampling," in *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics*, 2005, p. 301308.
[28] S. Chib and E. Greenberg, "Understanding the metropolis-hastings algorithm," *The American Statistician*, vol. 49, no. 4, pp. 327–335, 1995.
[29] S. K. M. Wong, C. Ssa, and C. J. Butz, "Contextual weak independence in bayesian networks," in *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, 670–679*, 1999, pp. 670–679.
[30] C. Boutilier, N. Friedman, M. Goldszmidt, and D. Koller, "Context-specific independence in bayesian networks," in *Proc. of the 20th Conf. on Uncertainty in Artificial Intelligence (UAI-1996)*, E. Horvitz and F. V. Jensen, Eds. Morgan Kaufmann, 1996, pp. 115–123.
[31] E. S. D. Peer, A. Regev, D. Koller, and N. Friedman, "Learning module networks," in *Proc. of the 19th Conf. on Uncertainty in Artificial Intelligence (UAI-2003)*. Morgan Kaufmann, 2003, pp. 525–534.
[32] N. Friedman and M. Goldszmidt, "Learning Bayesian networks with local structure," in *Proc. of the 20th Conf. on Uncertainty in Artificial Intelligence (UAI-1996)*, E. Horvitz and F. V. Jensen, Eds. Morgan Kaufmann, 1996, pp. 252–262.
[33] D. M. Chickering, D. Heckerman, and N. Friedman, "A Bayesian approach to learning Bayesian networks with local structure," in *Proc. of the 20th Conf. on Uncertainty in Artificial Intelligence (UAI-1997)*. Morgan Kaufmann, 1997, pp. 80–89.
[34] J. Su and H. Zhang, "Full bayesian network classifiers," in *In Proc. ICML06*. ACM Press, 2006, pp. 897–904.
[35] S. Monti and G. Cooper, *Learning in graphical models*. Cambridge, MA: MIT Press, 1999, ch. Learning hybrid Bayesian networks from data.
[36] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
[37] H. A. Chipman, E. George, and R. McCulloch, "Bayesian cart model search," *Journal of the American Statistical Association*, vol. 93, no. 443, pp. 935–947, 1998.
[38] H. Chipman, E. I. George, R. E. McCulloch, and R. E, "Bayesian treed models," in *Machine Learning*, 2000, pp. 299–320.
[39] D. Denison, B. Mallick, and A. Smith, "Bayesian cart," *Biometrika*, vol. 85, no. 2, p. 363377, 1998.
[40] Y. Wu, *Bayesian Tree Models*. Ph.D. thesis, 20046.
[41] Y. Wu, H. Tjelmeland, and M. West, "Bayesian cart: Prior specification and posterior simulation." *Journal of Computational and Graphical Statistics*, vol. 16, no. 1, p. 4466, 2007.
[42] F. Ronquist and J. Huelsenbeck., "Mrbayes3: Bayesian phylogenetic inference under mixed models," *Bioinformatics*, vol. 19, pp. 1572–1574, 2003.
[43] T. Ramdas and G. Egan, "A survey of fpgas for acceleration of high performance computing and their application to computational molecular biology," in *TENCON 2005*, 2005, pp. 1–6.
[44] I. Pournara, C. Bouganis, and G. Constantinides, "Fpga-accelerated bayesian learning for reconstruction of gene regulatory networks," in *International Conference on Field Programmable Logic and Applications, 2005*, 2005, pp. 323–328.
[45] D. Timmerman, A. C. Testa, T. Bourne, E. Ferrazzi, L. Ameye, M. L. Konstantinovic, B. Van Calster, W. P. Collins, I. Vergote, S. Van Huffel, and L. a. Valentin, "Logistic regression model to distinguish between the benign and malignant adnexal mass before surgery: a multicenter study by the international ovarian tumor analysis group." *J Clin Oncol*, vol. 23, no. 34, pp. 8794–8801, December 2005. [Online]. Available: http://dx.doi.org/10.1200/JCO.2005.01.7632