Bayesian Analysis (2013) **8**, Number 3, pp. 549–552

Rejoinder

Marco Scutari *

I would like to thank Hao Wang, Adrian Dobra, Christine Peterson and Francesco Stingo for the insightful comments and critiques they contributed to the discussion. The material contained in the paper originated in large part as the theoretical core of my Ph.D. thesis (Scutari 2011), and served the purpose of improving my understanding of the workings of prior and posterior distributions on graph structures as much as that of exploring novel applications. As a result, and as the discussants have observed, the paper provides a useful starting point for further developments while not focusing on specific applications such as prior specification or the analysis of real-world data.

The discussants' remarks highlight key strengths and limitations in the material, and suggest several useful directions for future research. In the following, I will concentrate on four topics that were touched on in all discussions: the development of new priors, sampling random graphs from non-uniform distributions, the applications and interpretation of the covariance matrix and the entropy of $\mathrm{P}(\mathcal{G}(\mathcal{E}))$ and $\mathrm{P}(\mathcal{G}(\mathcal{E}) \mid \mathcal{D})$, and the role of structure learning in graphical modelling.

1 Developing new prior distributions

In the paper much attention is devoted to the uniform prior and the *maximum entropy* case. As remarked by Dobra, other choices are available in the literature that are more flexible and tailored to real-world data. Additional examples inspired by genetics and systems biology are presented, for instance, in Imoto et al. (2003), Werhli and Husmeier (2007) and Mukherjee and Speed (2008). The reason for investigating the uniform prior is two-fold. First of all, it is a limit case in terms of entropy and therefore it is useful as a term of comparison along with *maximum entropy* distributions. Furthermore, the uniform prior is a *de facto* standard for $\mathrm{P}(\mathcal{G}(\mathcal{E}))$ in computer science and artificial intelligence literature on Bayesian networks, to the point that sometimes its use is not even mentioned explicitly but is implied by the fact that imaginary sample size is the only hyperparameter.

Developing new priors using the second order moments of $\mathrm{P}(\mathcal{G}(\mathcal{E}))$ (i.e. arc and edge correlations) in addition to first order moments (i.e. arc and edge probabilities) presents significant challenges due to the number of parameters involved. As the discussants pointed out, achieving sparsity and addressing the need for multiplicity adjustment while keeping hyperparameter specification simple is a difficult task. In my thesis, I addressed a related problem, the regularisation of the covariance matrix of $\mathrm{P}(\mathcal{G}(\mathcal{E}) \mid \mathcal{D})$ with the shrinkage estimators from Ledoit and Wolf (2003) and Schäfer and Strimmer (2005). Such estimators have a Bayesian interpretation and can be used to achieve sparsity by shrinking $\mathrm{diag}(\Sigma)$ and (in turn) edge and arc probabilities towards zero

*Genetics Institute, University College London, United Kingdom, m.scutari@ucl.ac.uk

 DOI:10.1214/13-BA845

when the data does not support the corresponding dependence relationships. Clearly, other approaches for sparse covariance matrix estimation (e.g. Cai et al. 2011) may be used to the same effect.

2 Random graph sampling

The ability to sample efficiently from $\mathrm{P}(\mathcal{G}(\mathcal{E}))$ and $\mathrm{P}(\mathcal{G}(\mathcal{E}) \,|\, \mathcal{D})$ is crucial to investigate the graphical properties of both Markov and Bayesian networks. In the case of Bayesian networks, several Markov chain Monte Carlo approaches have been developed for uniform sampling with and without constraints (Melançon et al. 2000; Ide and Cozman 2002; Ide et al. 2004) and for posterior sampling (Friedman and Koller 2003). The former can be adapted to include arbitrary structural constraints and unequal arc probabilities by controlling the transition probabilities of arc inclusion, reversal and removal. Unfortunately, mixing becomes increasingly slow as the number of nodes increases (because the dimension of the space of the graphs increases super-exponentially) and as the sampling distribution moves away from *maximum entropy* (because the probability mass is unevenly distributed, with sharp peaks and valleys).

As far as Markov networks are concerned, Dobra notes that sampling from the uniform prior on the space of decomposable models is an open problem. It would be interesting to investigate whether that could be solved by sampling from the space of the Bayesian networks with some specific $\mathrm{P}(\mathcal{G}(\mathcal{E}))$ and moralising the resulting DAGs.

3 Practical applications and interpretation

Peterson, Stingo and Wang observe (rightly) that practical interpretation and applications of the variability measures and the second order moments are not thoroughly investigated in the paper. I hope that the geometrical representation of $\mathrm{P}(\mathcal{G}(\mathcal{E}))$ and $\mathrm{P}(\mathcal{G}(\mathcal{E}) \,|\, \mathcal{D})$ on the space $\mathcal{L}$ of the eigenvalues of Σ will prove valuable in that respect. Furthermore, I agree with the discussants that variability measures are not completely intuitive to use as summary statistics, even though they can provide additional information in parameter tuning and model selection when used in combination with other criteria. Rather than considering the overall approach to be without merit, I believe that better summary statistics could be derived from Σ using *minimum* and *maximum entropy* as references.

4 Structure and parameter learning

In his discussion, Dobra argues that the split between structure and parameter learning has a potentially negative impact on modelling and inference, depending on the parametric assumptions and the shape of $\mathrm{P}(\mathcal{G}(\mathcal{E}) \,|\, \mathcal{D})$. While this is true to a certain extent, common operating practices in some settings emphasise structure learning over parameter learning or skip parameter learning altogether.

An example of such a workflow is the analysis in Sachs et al. (2005), which reconstructs a causal protein signalling network with a very high accuracy. Structure learning was performed using a mixture of observational and interventional discretised data. The structure of the final Bayesian network was the result of model averaging over a set of *maximum a posteriori* structures learned with simulated annealing initialised from the *maximum entropy* distribution. In other words, both sampling and model averaging were applied only to the network structure. Subsequent investigations on the signalling pathways were then performed with additional, targeted experiments, not with parameter estimation and Bayesian network inference algorithms. This choice can be justified by the fact that the authors' main interest was in the presence or the absence of particular causal relationships, and because the data were discretised at the beginning of the analysis, thus reducing the quality of quantitative inference results.

Another situation in which sampling from structure and parameter space at the same time is problematic is when structure learning is not explicitly implemented within the framework of Bayesian statistics. This is the case for many state-of-the-art learning algorithms (e.g. the semi-interleaved Hiton-PC from Aliferis et al. 2010), which perform structure learning with frequentist or information theoretic criteria and parameter learning with Bayesian posterior estimates.

In conclusion, while the approach to learning outlined by Dobra is certainly preferable on theoretical grounds, I feel that investigating structure learning as a self-contained part of graphical modelling is worthwhile and may be informative about the behaviour of current practices in applied data analysis.

References

Aliferis, C. F., Statnikov, A., Tsamardinos, I., Mani, S., and Xenofon, X. D. (2010). "Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification Part I: Algorithms and Empirical Evaluation." *Journal of Machine Learning Research*, 11: 171–234. 551

Cai, T., Liu, W., and Luo, X. (2011). "A Constrained l_1 Minimization Approach to Sparse Precision Matrix Estimation." *Journal of the American Statistical Association*, 106(494): 594–607. 550

Friedman, N. and Koller, D. (2003). "Being Bayesian about Bayesian Network Structure: A Bayesian Approach to Structure Discovery in Bayesian Networks." *Machine Learning*, 50(1–2): 95–126. 550

Ide, J. S. and Cozman, F. G. (2002). "Random Generation of Bayesian Networks." In *Proceedings of the 16th Brazilian Symposium on Artificial Intelligence*, 366–375. Springer-Verlag. 550

Ide, J. S., Cozman, F. G., and Ramos, F. T. (2004). "Generating Random Bayesian Networks with Constraints on Induced Width." In *Proceedings of the 16th European Conference on Artificial Intelligence*, 323–327. IOS Press. 550

Imoto, S., Higuchi, T., Goto, T., Tashiro, K., Kuhara, S., and Miyano, S. (2003). "Combining Microarrays and Biological Knowledge for Estimating Gene Networks via Bayesian Networks." In *Proceedings of the IEEE Computer Society Bioinformatics Conference*, 104–113. 549

Ledoit, O. and Wolf, M. (2003). "Improved Estimation of the Covariance Matrix of Stock Returns with an Application to Portfolio Selection." *Journal of Empirical Finance*, 10: 603–621. 549

Melançon, G., Dutour, I., and Bousquet-Mélou, M. (2000). "Random Generation of DAGs for Graph Drawing." Technical Report INS-R0005, Centre for Mathematics and Computer Sciences, Amsterdam. 550

Mukherjee, S. and Speed, T. P. (2008). "Network inference using informative priors." *Proceedings of the National Academy of Sciences*, 105(38): 14313–14318. 549

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. (2005). "Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data." *Science*, 308(5721): 523–529. 551

Schäfer, J. and Strimmer, K. (2005). "A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics." *Statistical Applications in Genetics and Molecular Biology*, 4(32): 1175–1189. 549

Scutari, M. (2011). "Structure Variability in Bayesian Networks." Ph.D. thesis, Department of Statistical Sciences, University of Padova. 549

Werhli, A. V. and Husmeier, D. (2007). "Reconstructing Gene Regulatory Networks with Bayesian Networks by Combining Expression Data with Multiple Sources of Prior Knowledge." *Statistical Applications in Genetics and Molecular Biology*, 6(1): 1–45. 549