Bayesian Analysis (2013) **8**, Number 3, pp. 533–538

Comment on Article by Scutari

Adrian Dobra *

This is an interesting and thought-provoking paper which focuses on defining new prior distributions on graphical structures. Priors on graphs represent a key component of any Bayesian approach for graphical models, hence the identification of new prior distributions for graphs is a very important topic. The author proceeds by modeling the possible edges of a graph through appropriate joint probability distributions. This idea receives a good treatment in this writing, but it is certainly not as novel as the author might seem to suggest by not mentioning many other papers who used various priors on graphs which are different from the uniform prior given in equation (2) page 2 of the paper. In fact, the Bayesian literature dedicated to graphical models has a longstanding track of using priors that encourage sparsity in order to increase interpretability and avoid the risk of overfitting. Some of these priors are constructed precisely by treating edges as random variables. In the context of DAGs, a typical prior specification starts with the traditional Bayesian variable selection prior for regressions which is defined by assuming a constant probability of inclusion β of each variable x_i, $i \in V = \{1, 2, \ldots, p\}$, in the regression model. This leads to a prior $\Pr(k) \propto (\beta/(1-\beta))^k$ associated with a regression with k predictors. Independent priors for regressions in the compositional specification of a DAG D,

$$x_i = \sum_{j \in pa(i)} \gamma_{ij} x_j + \epsilon_i, \text{ for each } i \in V,$$

where $pa(i)$ are the parents of vertex i in D, lead to the following prior for D (see, for example, Dobra et al. (2004)):

$$\Pr(D) \propto (\beta/(1-\beta))^{\sum_{i=1}^{p} \#pa(i)}.$$

The DAG D becomes sparser as $\binom{p}{2}\beta$ gets smaller. In the context of Gaussian graphical models, a usual choice is the uniform prior $\Pr(G) \propto 1$. Alternative priors on $\mathcal{G}_p$, the set of graphs with p vertices, have been developed by assuming a constant probability of inclusion $\beta \in (0, 1)$ of each edge. This leads to a prior for a graph $G \in \mathcal{G}_p$ (Dobra et al. 2004; Jones et al. 2005)

$$\Pr(G) \quad \propto \quad (\beta/(1-\beta))^{size(G)}, \tag{1}$$

where $size(G)$ is the number of edges in G. Sparse graphs receive high prior probabilities when $\binom{p}{2}\beta$ is small. By assuming $\beta \sim \mathsf{Beta}(a, b)$, Carvalho and Scott (2009) integrate out β in (1) to obtain the following prior on $\mathcal{G}_p$:

$$\Pr(G) \quad \propto \quad B\left(a + size(G), b + \binom{p}{2} - size(G)\right)/B(a, b), \tag{2}$$

*Department of Statistics, Department of Biobehavioral Nursing and Health Systems, and Center for Statistics and the Social Sciences, University of Washington adobra@uw.edu

 DOI:10.1214/13-BA841

where $B(\cdot, \cdot)$ is the beta function. Armstrong et al. (2009) introduced the size based prior on $\mathcal{G}_p$ which gives equal probability to the size of a graph and equal probability to graphs of each size. The size based prior is obtained by setting $a = b = 1$ in (2). We note that the expected size of a graph under the size based prior is $\binom{p}{2}/2$, which is also the expected size of a graph under the uniform prior on $\mathcal{G}_p$. Due to their bias towards middle size graphs, the size based prior and the uniform prior should be avoided if sparse graphs are desired. Instead, the prior (1) with small expected graph sizes $\binom{p}{2}\beta$ should be preferred.

It would be interesting to see how the aforementioned priors on graphs compare with the new priors proposed by the author in this writing. Besides comparisons focusing on how the mean and variance of the graph size change as a function of the number of vertices, I would find an in-depth discussion of the problem of sampling from these new priors to be extremely relevant. Sampling is very important since, if one cannot sample from a prior in some efficient manner, actually using that prior to explore large graphs becomes quite problematic. For example, direct sampling from the uniform prior on graphs is easy when any graphical structure is allowed. But direct sampling from the uniform prior over the space of decomposable graphs is certainly an open problem which does not have, to the best of my knowledge, a good solution.

I disagree with the author's view that Bayesian inference should be split in two steps: structure learning (in which the edges of the graph are determined) and parameter learning (in which parameters of the underlying joint distribution are estimated given the graph determined at the first step). There are two very good reasons for performing both steps together by sampling from the joint posterior distribution of the model parameters and the graphs. The first reason comes from applications of graphical models for high-dimensional datasets with a small number of observed samples. For such datasets, it is likely that the highest posterior probability graph receives only a small (almost zero) posterior probability. Furthermore, changing a few edges in this graph could lead to graphs with comparable posterior probabilities. When model uncertainty is high, Bayesian model averaging becomes key because it avoids the need to perform inference by making an explicit choice about which edges are present or absent. Sampling from the joint posterior of graphs and model parameters performs Bayesian graph averaging for model parameter estimates and for the edge inclusion probabilities. By conditioning on one or on a relatively small number of graphs, the uncertainty in the model parameter estimates can be severely underestimated for high-dimensional data applications.

The second reason for not separating Bayesian inference in two steps is conceptual and it is precisely related to prior specification for graphs. The structure learning step is performed by integrating out the model parameters and making use of the marginal (integrated) likelihood to obtain the posterior probability of a graph — see equation (1) page 1 of the paper. While taking this approach can be justified, for example, from a computational perspective as it avoids having to sample the model parameters (an expensive and, quite possibly, less than trivial task in some settings), the use of the marginal likelihood effectively disconnects the model parameters from the underlying graph. This aspect is key because the complexity of a model is not necessarily reflected in the number of edges of the graph. Here I will give an example which shows that

removing one edge from the same graph can lead to setting one parameter to zero for one class of graphical models, but it leads to setting a large number of parameters to zero for another class of graphical models.

Consider the graph G from Figure 1 which gives the conditional independence relationships among four random variables A, B, C and D, and the graph G' obtained by removing the edge between B and D from G. First, we assume that the four random

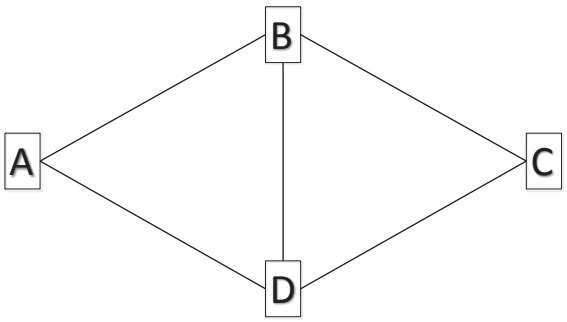

Figure 1: A conditional independence graph with four vertices and five edges.

variables are continuous and we model their joint distribution with a Gaussian graphical model $\mathsf{N}_4\left(0, K_G^{-1}\right)$ defined by the graph G (Whittaker 2009), where the precision matrix is parametrized as

$$K_G = \begin{pmatrix} k_{AA} & k_{AB} & 0 & k_{AD} \\ & k_{BB} & k_{BC} & k_{BD} \\ & & k_{CC} & k_{CD} \\ & & & k_{DD} \end{pmatrix}$$

Note that $k_{AC} = 0$ in K_G because there is no edge between A and C in G. The Gaussian graphical model defined by the graph G' is $\mathsf{N}_4\left(0, K_{G'}^{-1}\right)$, where $K_{G'}$ is obtained from K_G by setting $k_{BD} = 0$. Second, we assume that the random variables A, B, C and D are discrete with I_A, I_B, I_C and I_D categories, respectively. In this case, we model their discrete distribution with a graphical loglinear model defined by the graph G (Fienberg 2007). This loglinear model involves four main effects associated with each random variable, five two-factor interaction terms associated with each edge of G and two three-factor interaction terms associated the two cliques of G, $\{A, B, D\}$ and $\{B, C, D\}$:

$$\begin{aligned} \log m_{ijkl}^{ABCD} &= u + u_i^A + u_j^B + u_k^C + u_l^D + \\ &\quad u_{ij}^{AB} + u_{il}^{AD} + u_{jl}^{BD} + u_{jk}^{BC} + u_{kl}^{CD} + \\ &\quad u_{ijl}^{ABD} + u_{jkl}^{BCD}, \end{aligned} \tag{3}$$

where $1 \leqslant i \leqslant I_A$, $1 \leqslant j \leqslant I_B$, $1 \leqslant k \leqslant I_C$ and $1 \leqslant l \leqslant I_D$. The loglinear model (3) is made identifiable by imposing the usual ANOVA-like constraints that the sum over any index i, j, k or l of a u-term is zero, which implies that the number of fitted parameters of this model is

$$1 + (I_A - 1) + (I_B - 1) + (I_C - 1) + (I_D - 1) +$$
$$(I_A - 1)(I_B - 1) + (I_A - 1)(I_D - 1) + (I_B - 1)(I_D - 1) + (I_B - 1)(I_C - 1) + (I_C - 1)(I_D - 1) +$$
$$(I_A - 1)(I_B - 1)(I_D - 1) + (I_B - 1)(I_C - 1)(I_D - 1).$$

The graphical loglinear model defined by the graph G' involves four main effects and four two-factor interaction terms associated with each edge of G':

$$\begin{aligned} \log m_{ijkl}^{ABCD} &= u + u_i^A + u_j^B + u_k^C + u_l^D + \\ &\quad u_{ij}^{AB} + u_{il}^{AD} + u_{jk}^{BC} + u_{kl}^{CD}. \end{aligned} \tag{4}$$

Remark that the four edges of G' also happen to be the cliques of G'. Therefore the difference in the number of fitted parameters of the graphical loglinear models (3) and (4) is

$$(I_B - 1)(I_D - 1) + (I_A - 1)(I_B - 1)(I_D - 1) + (I_B - 1)(I_C - 1)(I_D - 1).$$

For example, if the random variables A, B, C and D have three categories each, the difference in the number of fitted parameters is 20. Therefore, for graphical loglinear models, removing one edge in the independence graph can lead to a significant change in the number of fitted parameters. On the other hand, for Gaussian graphical models, removing one edge in the independence graph corresponds with setting only one parameter to zero.

Specifying priors for graphs should be directly linked with the number of parameters each edge corresponds to. For Gaussian graphical models, an edge always corresponds to exactly one parameter, but for graphical loglinear models, the same edge could correspond to one or more parameters depending on the graph it belongs to. To see this, consider removing the edge between A and B in the graph G and in the graph G' in the categorical data case. By splitting Bayesian inference in two steps, the author oversimplifies the problem of prior specification for graphs by effectively assuming that edges have similar roles in terms of model complexity. While such an assumption is certainly valid in some cases (e.g., Gaussian graphical models), it is questionable in other cases (e.g., graphical loglinear models). Therefore specifying priors for graphs based on multivariate discrete distributions for edges — the key idea of this paper — might not be an ideal solution for certain classes of graphical models.

References

Armstrong, H., Carter, C. K., Wong, K. F., and Kohn, R. (2009). "Bayesian Covariance Matrix Estimation Using a Mixture of Decomposable Graphical Models." *Statistics and Computing*, 19: 303–316. 534

Carvalho, C. M. and Scott, J. G. (2009). "Objective Bayesian Model Selection in Gaussian Graphical Models." *Biometrika*, 96: 1–16. 533

Dobra, A., Hans, C., Jones, B., Nevins, J. R., Yao, G., and West, M. (2004). "Sparse Graphical Models for Exploring Gene Expression Data." *Journal of Multivariate Analysis*, 90: 196–212. 533

Fienberg, S. E. (2007). *The Analysis of Cross-Classified Categorical Data*. Springer, second edition. 535

Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C., and West, M. (2005). "Experiments in Stochastic Computation for High-dimensional Graphical Models." *Statistical Science*, 20: 388–400. 533

Whittaker, J. (2009). *Graphical Models in Applied Multivariate Statistics*. Wiley. 535