Bayesian Analysis (2013) **8**, Number 3, pp. 543–548

Comment on Article by Scutari

Hao Wang *

Scutari's paper studies properties of the distribution of graphs $p(\mathcal{G})$. This is an interesting angle because it differs from many works that focus on distributions over parameter spaces for a given graph $p(\Theta \mid \mathcal{G})$. The paper's investigation of $p(\mathcal{G})$ centers around its implied covariance matrix $\mathbf{\Sigma} = \text{Cov}(\mathcal{G})$. The major theoretical results, as I see it, concern eigenvalues of $\mathbf{\Sigma}$ as well as variance and covariance elements of $\mathbf{\Sigma}$ in the *maximum entropy* case for DAGs (i.e., a uniform distribution over all DAGs). While these results are certainly very worth noting by their own intellectual merits, what practical difference they might make is unclear to me. Eigenvalues of $\mathbf{\Sigma}$ might be hard to interpret in terms of their intuitive connections with underlying graph structures. The maximum entropy case is somehow limited as it is rarely the case for posterior graph distributions and is also often less preferred than sparser cases for prior graph distributions. More discussions on the implications of these theoretical results on real data analysis will be very helpful.

The more general point raised by the paper is more interesting to me. It calls attention to deeper investigation on statistical properties of distributions of graphs $p(\mathcal{G})$. In the literature of my own research topic of Gaussian graphical models (Dempster 1972), existing studies usually only focus on a point estimation of $\mathcal{G}$ from $p(\mathcal{G})$ – the mean or the mode of $p(\mathcal{G})$ is often used to represent prior belief or to summarize posterior information. The paper's framework extends this sort of simple summary to the covariance matrix $\mathbf{\Sigma}$ of $p(\mathcal{G})$. It is then tempting to ask what will be gained from these extra efforts. Specific questions include how to construct a prior $p(\mathcal{G})$ with a consideration beyond the implied mean or mode graphs, and how to put $\mathbf{\Sigma}$ into a perspective that better illustrates graph structures than a point estimate alone.

I attempt to explore these questions in this discussion from a more applied point of view than the paper. In addition, I have some doubts about the paper's argument of using variability measures in choosing learning algorithms or hyperparameters. The context of my discussion is Gaussian graphical models under a fully Bayesian treatment (Jones et al. 2005). Generalizations of the following points might be made to other undirected graphs or even DAGs too.

1 Distribution $p(\mathcal{G})$ and its covariance matrix

Similar to the paper, I use the edge set to represent a graph $\mathcal{G}$. Let e_{ij} be the binary edge inclusion indictor variable, that is, $e_{ij} = 1$ if there is an edge between nodes i and j in $\mathcal{G}$, and $e_{ij} = 0$ otherwise. Then the set of $k = p(p-1)/2$ binary variables $\mathcal{E} = \{e_{ij}\}_{1 \leqslant i < j \leqslant p}$ can be used in place of $\mathcal{G}$. The distribution of graphs is $p(\mathcal{E})$ and the implied $k \times k$ covariance matrix is $\mathbf{\Sigma} = \text{Cov}(\mathcal{E})$.

*Department of Statistics, University of South Carolina, Columbia, SC, U.S.A. haowang@sc.edu

© 2013 International Society for Bayesian Analysis DOI:10.1214/13-BA840

I start with the prior distribution of graphs $p(\mathcal{E})$. Perhaps the most commonly used $p(\mathcal{E})$ is the independent and identically distributed Bernoulli priors, namely

$$e_{ij} \overset{\text{iid}}{\sim} \text{BERN}(\beta), \quad 1 \leqslant i < j \leqslant p, \quad 0 < \beta < 1. \tag{1}$$

The iid assumption makes the investigation of prior properties simple. One just needs to focus on the marginal distribution $\text{BERN}(\beta)$. In this sense, the author's results based on $\mathbf{\Sigma}$ are less meaningful because here $\mathbf{\Sigma}$ has a simple diagonal matrix structure $\mathbf{\Sigma} = \beta(1-\beta)\mathbf{I}_k$. The significance of the proposed framework for the prior only comes when these binary variables in $\mathcal{E}$ are dependent. This poses a challenge about hyperparameter specification for such multivariate Bernoulli distributions as the number of hyperparameters explodes quickly as p grows. The covariance matrix of the multivariate Bernoulli alone requires $k(k+1)/2 = \mathcal{O}(p^4)$ parameters. Therefore, moving beyond the independent Bernoulli case (1) requires an easy approach to construct interpretable dependent Bernoulli priors for $\mathcal{E}$. The question is how to develop such approaches.

I found one simple construction of multivariate dependent Bernoulli distributions already exists in the literature. The hierarchical prior (e.g, Scott and Carvalho 2008, Wang 2010) belongs to the multivariate Bernoulli distribution framework. It modifies the independent prior (1) by treating β as unknown and placing another hierarchy in the form of a beta distribution on it:

$$e_{ij} \overset{\text{iid}}{\sim} \text{BERN}(\beta), 1 \leqslant i < j \leqslant p, \quad \beta \sim \text{BETA}(a, b), a, b > 0. \tag{2}$$

Clearly, marginalizing β in (2) gives the following first and second moments of e_{ij}:

$$\begin{aligned} \text{E}(e_{ij}) &= \frac{a}{a+b}, \quad \text{VAR}(e_{ij}) = \frac{ab}{(a+b)^2}, \\ \text{COV}(e_{ij}, e_{kl}) &= \frac{ab}{(a+b)^2(a+b+1)}, \quad \text{CORR}(e_{ij}, e_{kl}) = \frac{1}{(a+b+1)}. \end{aligned} \tag{3}$$

Thus, e_{ij}'s are no longer independent but have constant values of variances and covariances. Focus on a simpler case with $a = b$. Then the expected value is $\text{E}(e_{ij}) = 1/2$, the variance is $\text{VAR}(e_{ij}) = 1/4$, and the covariance is $\text{COV}(e_{ij}, e_{kl}) = 1/(8a+4)$. Note that the expected value and the variance are equal to those of (1) with $\beta = 1/2$, which is also the *maximum entropy* case. The only difference in $\text{E}(\mathcal{E})$ and $\text{COV}(\mathcal{E})$ between (1) with $a = b$ and (2) with $\beta = 1/2$ lies on the off-diagonal covariance elements, making it an interesting example to illustrate the role of $\mathbf{\Sigma}$.

To characterize $\mathbf{\Sigma}$, the author studies its eigenvalues in Lemma 2.1 and Example 3.1. The exact values of these eigenvalues are hard to interpret in terms of what aspects of graph structures they measure. Motivated by the popularity of the overall sparsity level as a good summary of graph structures in the literature, I suggest to consider the random variable of the total number of edges:

$$m = \sum_{0 \leqslant i < j \leqslant p} e_{ij} \in \{0, 1, \ldots, k\}. \tag{4}$$

Then the covariance matrix $\mathbf{\Sigma}$ determines the variance of m according to the relation $\text{VAR}(m) = 1_k' \mathbf{\Sigma} 1_k$. Comparing $1_k' \mathbf{\Sigma} 1_k$ provides an assessment of the variability of the overall sparsity level. In Prior (1) with $\beta = 1/2$, $\text{VAR}(m) = 1_k' \mathbf{\Sigma} 1_k = k/4$; in Prior (2) with $a = b$,

$$\text{VAR}(m) = \frac{k}{4} + \frac{k(k-1)}{8a+4}.$$

A smaller a implies higher variability of the total number of edges m, although the expected value is the same $\text{E}(m) = k/2$ for all a.

The following example demonstrates the above point that $\text{E}(\mathcal{E})$ may not illustrate interesting patterns of $p(\mathcal{E})$ but $\text{COV}(\mathcal{E})$ may characterize $p(\mathcal{E})$.

Example Consider a $p = 6$ (so $k = 15$) case. Under Prior (2) with $a = b$, let $a = 0.01, 1$, and $+\infty$. Note that the case of $a = +\infty$ corresponds to Prior (1) with $\beta = 1/2$. A summary of variances and correlations of $\mathbf{\Sigma}$ is given in Panel (a) of Table 1. The variances are the same $\text{VAR}(e_{ij}) = 1/4$ for all a. The constant correlations are $\text{CORR}(e_{ij}, e_{kl}) = 0.98, 0.33$, and 0 for $a = 0.01, 1, +\infty$. The eigenvalues are shown in Panel (a) of Table 2. Because of the constant correlation and variance structure, the smallest to the second largest eigenvalues are the same for each a.

Now consider $\text{VAR}(m)$. It is equal to 55.22, 21.25, and 3.75 for $a = 0.01, 1, +\infty$. This indicates that $a = 0.01$ and $a = 1$ cases place substantially higher uncertainties about the overall sparsity level than the *maximum entropy* case. Panel (a) of Figure 1 displays the exact probability distributions of m. The $a = +\infty$ case has a bell shape due to the central limit theorem applied to the sum of independent e_{ij}'s. The $a = 1$ case corresponds to a unform distribution on the edge inclusion probability β. It places equal probabilities on every possible outcome in $\{0, 1, \ldots, 15\}$ for m. When $a = 0.01$, higher probability masses are on the two ends of the range of all possible outcomes. The outcomes of $m = 0$ and $m = 15$ both have probabilities close to 0.5 while all other outcomes have probabilities close to zero. Note that the expectation of $\mathcal{E}$ is the same across a. The difference in distributions of m is driven by higher-moments of $\mathcal{E}$. Another intuitive connection can be drawn between m and $\mathbf{\Sigma}$. A higher correlation between e_{ij}'s indicates that e_{ij}'s tend to be 1 or 0 together, which explains the higher probability masses on the two ends for $a = 0.01$.

To study posterior distributions, I use the scenario from Wang and Li (2012). Consider a $p \times n$ sample matrix Y consisting of n iid samples from a p-dimensional multivariate normal $\text{N}(0, \Omega^{-1})$. Then $(S = YY', n)$ are sufficient statistics of Ω. So I do not have to generate Y directly; instead I let $S = nA^{-1}$ where $n = 18$ and

$$A = \begin{pmatrix} 1 & 0.5 & 0 & 0 & 0 & 0.4 \\ & 1 & 0.5 & 0 & 0 & 0 \\ & & 1 & 0.5 & 0 & 0 \\ & & & 1 & 0.5 & 0 \\ & & & & 1 & 0.5 \\ & & & & & 1 \end{pmatrix}.$$

This choice of (S, n) represents 18 samples of Y from $\text{N}(0, A^{-1})$. I use the G-Wishart prior (Dawid and Lauritzen 1993; Atay-Kayis and Massam 2005) $\text{W}_{\mathbf{G}}(3, I_6)$ on the

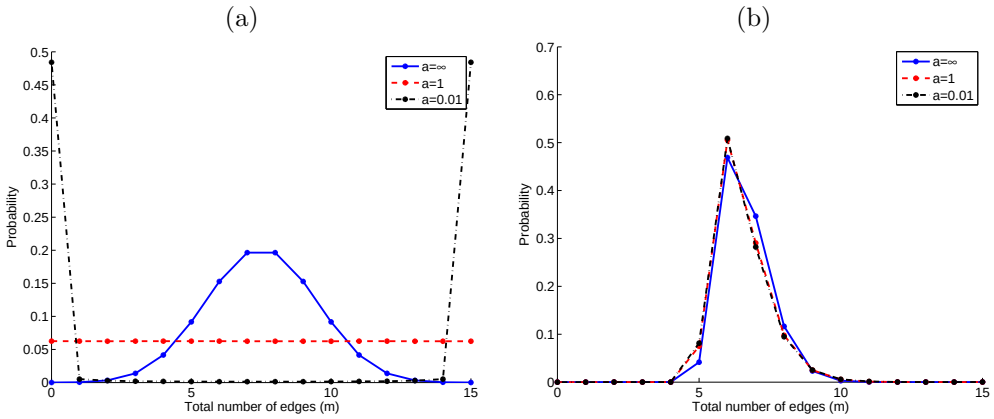

Figure 1: Prior (a) and posterior (b) distributions of the total number of edges m defined in (4) for a $p = 6$ example under Prior (2) with $a = b \in \{0.01, 1, +\infty\}$.

precision matrix Ω. The marginal likelihood $p(Y \mid \mathcal{E})$ can be computed for all possible $\mathcal{E}$ by integrating Ω out. Given these $p(Y \mid \mathcal{E})$ and $p(\mathcal{E})$ of (2), the exact values of $\mathbf{\Sigma}$ can be computed. Summary statistics of the variance and correlation elements in $\mathbf{\Sigma}$ are provided in Panel (b) of Table 1. These correlations are generally close to zero for all a values, although the prior correlations can be as high as 0.98 when $a = 0.01$. Eigenvalues of $\mathbf{\Sigma}$ are in Panel (b) of Table 2, which seem to be hard to interpret in terms of what aspects of graph structures they respond to.

Table 1: Summary of variance elements and correlation elements in $\mathbf{\Sigma}$ in the $p = 6$ example

	Variance					Correlation				
	min	lower quartile	median	upper quartile	max	min	lower quartile	median	upper quartile	max
Panel (a): Prior $\mathbf{\Sigma}$										
$a = 0.01$	0.25	0.25	0.25	0.25	0.25	0.98	0.98	0.98	0.98	0.98
$a = 1$	0.25	0.25	0.25	0.25	0.25	0.33	0.33	0.33	0.33	0.33
$a = \infty$	0.25	0.25	0.25	0.25	0.25	0	0	0	0	0
Panel (b): Posterior $\mathbf{\Sigma}$										
$a = 0.01$	0.018	0.032	0.072	0.087	0.145	-0.236	-0.015	0.000	0.023	0.046
$a = 1$	0.018	0.032	0.072	0.088	0.143	-0.237	-0.015	0.000	0.016	0.040
$a = \infty$	0.017	0.029	0.079	0.095	0.127	-0.240	-0.019	-0.007	-0.001	0.025

The variances $\text{VAR}(m)$ obtained from $\mathbf{\Sigma}$ are 0.88, 0.84, and 0.73 for $a = 0.01, 1$, and $+\infty$, respectively. Combined with the mean value $\text{E}(m) \approx 6.5$ for all three values of a, these variances give a good estimate of the possible range of the overall sparsity level.

Table 2: Eigenvalues of $\mathbf{\Sigma}$ in the $p = 6$ example

Panel (a): Prior $\mathbf{\Sigma}$

$\boldsymbol{\lambda}_{a=0.01}$ =	0.167	0.167	0.167	0.167	0.167	0.167	0.167	0.167	0.167	0.167	0.167	0.167	0.167	0.167	1.417
$\boldsymbol{\lambda}_{a=1}$ =	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	3.681
$\boldsymbol{\lambda}_{a=\infty}$ =	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250

Panel (b): Posterior $\mathbf{\Sigma}$

$\boldsymbol{\lambda}_{a=0.01}$ =	0.016	0.019	0.020	0.030	0.030	0.066	0.069	0.072	0.078	0.080	0.083	0.087	0.093	0.101	0.161
$\boldsymbol{\lambda}_{a=1}$ =	0.016	0.019	0.020	0.029	0.030	0.067	0.070	0.073	0.079	0.080	0.084	0.088	0.094	0.098	0.159
$\boldsymbol{\lambda}_{a=\infty}$ =	0.016	0.018	0.019	0.027	0.027	0.073	0.076	0.077	0.083	0.087	0.091	0.096	0.098	0.102	0.153

When compared with the prior variances of m, these numbers indicate that a substantial amount of information about the sparsity level is learned from the data. Panel (b) of Figure 1 shows the exact posterior distributions of m. They appear to be similar across a – all curves have the same mode $m = 6$ and center around this mode alike.

2 Measures of variability as metrics for comparing algorithms or choosing tuning parameters

In Section 4, the author proposes three measures of variability based on $\mathbf{\Sigma}$ and further recommends that they can be used for comparing different structure learning algorithms or choosing tuning parameters, which usually requires metrics that are based on a "golden standard". I have some doubts about this claim.

The claim seems to ignore the fact that both variance and bias are important in measuring the performance of structure learning and to solely focus on the variance part. An algorithm or a hyperparameter with a smaller variability measure does not necessarily mean it is better because it might generate graphs systematically far away from the true structure. Consider the previously used $p = 6$ node example. For an extreme case, let the prior be almost like a point mass on the empty graph. This can be achieved by letting $a = 0.0001$ and $b = 10000$ so $\beta \approx 0$. The posterior will be tightly concentrated around the empty graph, like a *minimum entropy* case; its covariance matrix $\mathbf{\Sigma}$ is close to $\mathbf{0}$; and its variability measures such as VAR_T, VAR_G and VAR_F will be extremely small. Then, according to the paper, the choice of $a = 0.0001$ and $b = 10000$ should be favored against many other reasonable values of a and b, but apparently this is a bad choice of tuning parameters. In fact, any $\mathbf{\Sigma}$-based measures of variability including VAR_T, VAR_G and VAR_F seem to only contain information about the variance of $p(\mathcal{E})$ and ignore the bias. They may not be used alone for comparing algorithms or choosing tuning parameters. How to effectively use them for the purpose of comparing algorithms or choosing tuning parameters is an open and very interesting question.

3 Conclusion

The paper is interesting in its unique focus on distributions of graphs instead of distributions of parameters. Using Gaussian graphical models as illustrating examples, I have discussed three challenges related to the paper's new framework: how to construct dependent multivariate Bernoulli distributions, how to better characterize distributions of graphs $p(\mathcal{E})$ using its second moment rather than the first moment alone, and how to effectively use the variability measures for comparing algorithms and choosing tuning parameters.

References

Atay-Kayis, A. and Massam, H. (2005). "The marginal likelihood for decomposable and non-decomposable graphical Gaussian models." *Biometrika*, 92: 317–335. 545

Dawid, A. P. and Lauritzen, S. L. (1993). "Hyper-Markov laws in the statistical analysis of decomposable graphical models." *Annals of Statistics*, 21: 1272–1317. 545

Dempster, A. (1972). "Covariance selection." *Biometrics*, 28: 157–175. 543

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. 543

Scott, J. G. and Carvalho, C. M. (2008). "Feature-Inclusion Stochastic Search for Gaussian Graphical Models." *Journal of Computational and Graphical Statistics*, 17(4): 790–808. 544

Wang, H. (2010). "Sparse seemingly unrelated regression modelling: Applications in finance and econometrics." *Computational Statistics & Data Analysis*, 54(11): 2866–2877. 544

Wang, H. and Li, S. Z. (2012). "Efficient Gaussian graphical model determination under G-Wishart prior distributions." *Electronic Journal of Statistics*, 6: 168–198. 545