Contributed discussion on article by Finegold and Drton

Comment by Abdolreza Mohammadi and Ernst C. Wit

Mohammadi, Abdolreza; Wit, Ernst C.

DOI
10.1214/13-BA856D

Publication date
2014

Document Version
Final published version

Published in
Bayesian Analysis

Bayesian Analysis (2014) **9**, Number 3, pp. 561–590

Contributed Discussion on Article by Finegold and Drton

Comment by Guido Consonni[1] and Luca La Rocca[2]

This is a very interesting paper providing both theoretical and computational results for robust structure estimation in decomposable graphical models. Finegold & Drton (F&D hereafter) do a splendid job in motivating and illustrating the various ramifications of this attractive research path. We will comment on prior specification, hoping to add further insights to a paper already rich in content. Notice that model choice results strongly depend on prior specification; see, e.g., O'Hagan and Forster (2004, ch. 7).

Priors on graphs Formula (3) of F&D specifies a product of Bernoulli priors with fixed edge inclusion probability d. As F&D mention in their Discussion, one could place a prior on d. We suggest exploring this avenue in real terms, because recent results suggest that substantial improvements can be obtained by placing, say, a beta prior on d; see for instance Scott and Berger (2010) and Castillo and van der Vaart (2012).

Priors on matrices The Hyper Inverse Wishart (HIW) prior on Ψ, or Σ in the Gaussian case, requires the hyperparameters δ and Φ. F&D choose $\delta = 1$ and $\Phi = cI_p$, referring to Armstrong et al. (2009) for alternative choices of Φ. A related option would be using the Fractional Bayes Factor (FBF) to implement model choice based on objective improper priors: a fraction of the likelihood would be used to make the prior proper, then its complementary fraction would be used for inference (avoiding double use of data); see O'Hagan and Forster (2004, ch. 7).

In the Gaussian case the FBF turns a default improper HIW prior on Σ into a proper HIW prior with Φ proportional to the sample covariance matrix (Carvalho and Scott 2009) and this results in a markedly improved performance with respect to the standard choice $\Phi = cI_p$ (applied to the whole likelihood). The problem is that vague priors assign too much probability to parameter values not supported by the data, and this alters the evidence conveyed by the marginal likelihoods. We note that Consonni and La Rocca (2012) extend the results of Carvalho and Scott (2009) to the larger class of Directed Acyclic Graphs (DAGs).

Implementing the FBF in the setup of F&D should be feasible, because, conditionally on τ, the results for the Gaussian case extend in a straightforward way. Since the

[1]Dipartimento di Scienze Statistiche, Università Cattolica del Sacro Cuore, Milano, Italy, guido.consonni@unicatt.it

[2]Dipartimento di Scienze Fisiche, Informatiche e Matematiche, Università di Modena e Reggio Emilia, Modena, Italy, luca.larocca@unimore.it

 DOI:10.1214/13-BA856D

problem it solves is general, we expect the FBF to give more reliable posterior probabilities also in this setting. We would love to see some experiments in this direction, as well as in the direction of applying the robust approach of F&D to compare general DAG models.

Non-local priors F&D write in their Discussion that for large p the highest posterior probability graph is difficult to find, and not necessarily informative, so that one may like to focus on marginal posterior edge inclusion probabilities. Indeed, we remark that the latter are enough to define the *median probability graph*: the graph containing exactly those edges whose probability is above 50%. Barbieri and Berger (2004) discuss the useful properties of the *median probability model* in the context of linear models.

A way of mitigating the dilution of posterior probability on model space for large p is using non-local parameter priors; see Johnson and Rossell (2012) in the context of linear models. Altomare et al. (2013) use non-local priors, obtained with the FBF, to compare Gaussian DAG models for a given ordering of the variables. Implementing non-local priors, which are characterized by (being continuous and) vanishing on the subspace which characterizes the submodel, in the setup of F&D would be quite expensive, as far as we can see, but also rewarding.

References

Altomare, D., Consonni, G., and La Rocca, L. (2013). "Objective Bayesian search of Gaussian directed acyclic graphical models for ordered variables with non-local priors." *Biometrics*, 69: 478–487.

Armstrong, H., Carter, C. K., Wong, K. F. K., and Kohn, R. (2009). "Bayesian covariance matrix estimation using a mixture of decomposable graphical models." *Statistics and Computing*, 19: 303–316.

Barbieri, M. M. and Berger, J. O. (2004). "Optimal predictive model selection." *The Annals of Statistics*, 32: 870–897.

Carvalho, C. M. and Scott, J. G. (2009). "Objective Bayesian model selection in Gaussian graphical models." *Biometrika*, 96: 497–512.

Castillo, I. and van der Vaart, A. (2012). "Needles and straw in a haystack: posterior concentration for possibly sparse sequences." *The Annals of Statistics*, 40: 2069–2101.

Consonni, G. and La Rocca, L. (2012). "Objective Bayes factors for Gaussian directed acyclic graphical models." *Scandinavian Journal of Statistics*, 39: 743–756.

Johnson, V. E. and Rossell, D. (2012). "Bayesian model selection in high-dimensional settings." *Journal of the American Statistical Association*, 107: 649–660. Corrections: *ibidem* p. 1656.

O'Hagan, A. and Forster, J. J. (2004). *Kendall's Advanced Theory of Statistics, Vol. 2b: Bayesian Inference*. Sevenoaks, UK: Arnold, second edition.

Scott, J. G. and Berger, J. O. (2010). "Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem." *The Annals of Statistics*, 38: 2587–2619.

Comment by Adrian Dobra[1]

I would like to congratulate the authors for their key contribution that makes an essential connection between the literature on Bayesian graphical models and the literature on robust Bayesian inference. The paper introduces an approach for Gaussian graphical models determination in the presence of outliers through flexible generalizations of multivariate t-distributions. First, the alternative t-distribution is defined as a scaled version of a multivariate normal with independent Gamma distributed scaling factors. Second, an adaptive clustering of the scaling factors is induced by a Dirichlet process with a Gamma baseline measure which gives a Dirichlet t-distribution. The papers clearly shows that reducing the effect of the outliers influences the structure of the graphs that are inferred from the data. The sampling methods presented in the paper seem to work well which opens the possibility of applying this methodology to numerous real world problems in which outliers are likely to be occur.

One such real world application is presented in Section 6 and involves the analysis of gene expression data. Typical gene expression datasets have sample sizes in the order of tens or hundreds, but contain expression levels of tens or hundreds of thousands of genes. In such high-dimensional settings in which the number of variables p is several orders of magnitude larger than the sample size n, the curse of dimensionality (Bellman 1961) phenomena create significant challenges for sound statistical inference. As the volume of space spanned by the p variables increases at an exponential rate, the available samples become sparse. Each sample could be regarded an outlier and, for that reason, attempting to reduce the weight of extreme observations might not have the same meaning in high dimensional applications as opposed to low dimensional applications in which the number of samples exceeds the number of variables.

The Bayesian approach to graphical modeling described in this paper assumes that a single graph expresses the conditional independence relationships of all available samples. The multivariate normal distribution constrained by the conditional independence graph is assumed to be independent of the Dirichlet t-distribution which defines clusters of samples. Graphical modeling of gene expression data is relevant because the inferred graph can be linked to biological pathways with known and unknown components. The observed samples are associated with various combinations of experimental conditions that can turn on and off relevant biological pathways. In such cases, learning multiple graphs associated with different groups of samples becomes key since potential changes in the structure of the graphs can be indicative of the dynamics of the underlying biological processes. It would be quite useful if the authors could comment on how multiple experimental conditions can be accommodated in their framework.

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

References

Bellman, R. E. (1961). *Adaptive control processes: a guided tour*. Princeton University Press.

Comment by Jayanta K. Ghosh[1]

Bayesian graphical models for multivariate normal distributions have become very popular. The basic idea is to give a simple graphical structure to the covariance matrix Σ so that a pair of variables Y_j and Y_k are conditionally independent given all the other variables if and only if $\Sigma^{-1}_{jk} = 0$.

To apply Bayesian methods one would need to put a prior on the graphical structure and a prior on the covariance matrix, such that local computations on the graph are possible. A hyper inverse Wishart prior on the covariance matrix and a uniform prior on the decomposable graphs allow such local computations. In particular one gets a closed form for the marginal likelihood. Hence for inference about high dimensional covariance models, this approach is very efficient and hence very popular.

It is interesting that so far the only such models have been for the Gaussian case. Important people in this area are Rajaratnam, Carvalho, and Massam. I am somewhat familiar with the work of Rajaratnam. An earlier paper, Yuan and Huang (2009), makes a major contribution by extending this methodology for a class of t-distributions. The present paper uses a new flexible multivariate t-distribution by modifying a method developed in Finegold et al. (2011). This seems to work better for high dimensions than the usual t-distribution, but requires heavy computation.

In this context, it would also make sense to check whether these small perturbations would affect the Gaussian graphical models in some other way significantly.

"The key new contribution", in the words of the authors, is a further modification that shows how this can be done efficiently by adaptively switching between the classical and an alternative t, using a Dirichlet process clustering. However the new alternative t, being different from the usual t, may disturb the graphical structure. Will this have an effect on the inference?

While I feel quite positive about the paper, it would be nice if the authors can point out some additional applications of their new techniques. Using a Dirichlet process clustering is quite novel and may be applicable elsewhere.

References

Finegold, M., Drton, M., et al. (2011). "Robust graphical modeling of gene networks using classical and alternative t-distributions." *The Annals of Applied Statistics*, 5(2A): 1057–1080.

Yuan, M. and Huang, J. Z. (2009). "Regularized parameter estimation of high dimensional t distribution." *Journal of Statistical Planning and Inference*, 139(7): 2284–2292.
URL http://dx.doi.org/10.1016/j.jspi.2008.10.014

[1]Department of Statistics, Purdue University, ghosh@stat.purdue.edu

Comment by Michele Guindani[1]

The authors provide a lucid argument for the use of flexible t-distributions in the Bayesian estimation of graphical models. The starting point is the realization that some samples may be contaminated, and the contamination might affect only small parts of any given sample. Thus, although the authors focus primarily on the problem of graph recovery, a reader may wonder if these proposals could find application also in the multiple hypotheses testing framework. On the one hand, the improved performance in the estimation of the graph dependence should translate to increased power of the testing procedures (Schwartzman and Lin 2011; Sun et al. 2014). On the other hand, the possibility to downweigh the influence of contaminated observations should increase the accuracy of the procedure. As a matter of fact, recent work has shown the importance of adequately taking into account the heterogeneity of the error variances, either across tests and/or across treatment conditions. Two notable examples are the papers by Shahbaba and Johnson (2013) and Bar et al. (2014). To set the framework, one may start by considering for simplicity the sequence of hypothesis tests,

$$H_0 : \mu_j \in A \quad \text{vs} \quad H_A : \mu_j \notin A^c \quad j = 1, \ldots, p,$$

where $A = (-\varepsilon, +\varepsilon)$ or $A = \{0\}$. We follow the notation in the manuscript, so that p denotes the number of vertices in the set V. In the example presented in Section 6, p is the number of genes in n experiments. Then, equation (18) defines a convenient statistics, which can be used for the purpose of hypothesis testing. Let $t = 1, \ldots, T$ denote the MCMC iterations after burn in. Then, at each iteration, $\mu_j^t = \sum_{i=1}^n \tau_{ij}^t Y_{ij} / \sum_{i=1}^n \tau_{ij}^t$, for both the alternative and the Dirichlet t-distributions. Let δ_j be a binary decision rule such that $\delta_j = 1$ denotes rejection of the null hypothesis, whereas $\delta_j = 0$ denotes the opposite outcome. Under a loss function which is a linear combination of false negative and false positive counts, rules based on thresholding the posterior probabilities of the alternative $v_j = p(\mu_j^t \in A^c)$ are known to be optimal under several criteria (see, e.g., Müller et al. 2007; Wu and Peña 2013; Sun et al. 2014) and can be easily determined on the basis of the MCMC output, since $v_j = E(I(\mu_j \in A)|\text{data}) \approx \sum_{t=1}^T I(\mu_j^t \in A)/T$ and $\delta_j = I(v_j > \tau)$ for a given threshold τ. From equation (18), it's apparent that small values of τ_{ij} downweigh the relevance of sample i in determining μ_j^t, which is the relevant quantity in the hypothesis testing problem. Alternatively, one could consider the sample mean $\bar{\mathbf{Y}}_n$, which conditionally on $\boldsymbol{\tau}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Psi}$, is $\mathcal{N}_p(\boldsymbol{\mu}, \frac{1}{n^2}\text{diag}(\sum_{i=1}^n 1/\sqrt{\boldsymbol{\tau_i}})\, \boldsymbol{\Psi}\, \text{diag}(\sum_{i=1}^n 1/\sqrt{\boldsymbol{\tau_i}}))$. Posterior estimates of $\boldsymbol{\tau}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Psi}$ can be obtained from the MCMC output as described in the manuscript, and can be used to define a single thresholding function for all tests when testing the null hypothesis $H_0 : \mu_j = 0$, for each $j = 1, \ldots, p$, as

$$S(\bar{\boldsymbol{y}}) = \frac{\sum_{j=1}^p \mathcal{N}_p(\bar{\boldsymbol{y}}; \hat{\mu}_j, \frac{1}{n^2}\sum_{i=1}^n \psi_{jj}/\tau_{ij})}{\sum_{j=1}^p \mathcal{N}_p(\bar{\boldsymbol{y}}; 0, \frac{1}{n}\psi_{jj})},$$

where $\hat{\mu}_j = E(\mu_j|\text{data})$, similarly as in Storey (2007) and Storey et al. (2007). See also Bogdan et al. (2008) and Guindani et al. (2009). The null hypothesis is rejected if

[1]University of Texas MD Anderson Cancer Center, mguindani@mdanderson.org

$S(\bar{\boldsymbol{y}}) > \tau$, for some $0 \leq \tau < \infty$. Again, it is apparent the effect of the flexible estimation of the τ_{ij}'s suggested in the manuscript by Finegold and Drton. A small value of τ_{ij}, for any single sample i, does not affect much the posterior mean $\hat{\mu}_j$ and contributes to increase the variance of the Gaussian density in the numerator of $S(\cdot)$. Hence, the overall effect of a single outlier is a decrease in the value of $S(\bar{\boldsymbol{y}})$, which may affect the ranking of the hypotheses (Shahbaba and Johnson 2013). Of course, refinements may be needed to ensure good frequentist properties of such thresholding procedures. However, it seems reasonable to conclude that the flexible extensions of the multivariate t-distribution proposed in the manuscript can find some interesting application also in the multiple comparison framework.

References

Bar, H. Y., Booth, J. G., and Wells, M. T. (2014). "A Bivariate Model for Simultaneous Testing in Bioinformatics Data." *Journal of the American Statistical Association*, 109(506): 537–547.

Bogdan, M., Ghosh, J., and Tokdar, S. (2008). "A comparison of the Benjamini-Hochberg procedure with some Bayesian rules for multiple testing." In Balakrishnan, N., Peña, E., and Silvapulle, M. (eds.), *Beyond Parametrics in Interdisciplinary Research: Festshcrift in Honor of Professor Pranab K. Sen*, IMS Collections, 211–230. Beachwood, Ohio, USA: Institute of Mathematical Statistics.

Guindani, M., Müller, P., and Zhang, S. (2009). "A Bayesian discovery procedure." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5): 905–925.

Müller, P., Parmigiani, G., and Rice, K. (2007). "FDR and Bayesian Multiple Comparisons Rules." In Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., and West, M. (eds.), *Bayesian Statistics 8*. Oxford, UK: Oxford University Press.

Schwartzman, A. and Lin, X. (2011). "The effect of correlation in false discovery rate estimation." *Biometrika*, 98(1): 199–214.

Shahbaba, B. and Johnson, W. O. (2013). "Bayesian nonparametric variable selection as an exploratory tool for discovering differentially expressed genes." *Statistics in Medicine*, 32(12): 2114–2126.

Storey, J. (2007). "The optimal discovery procedure: a new approach to simultaneous significance testing." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69 (3): 347–368.

Storey, J., Dai, J., and Leek, J. (2007). "The optimal discovery procedure for large-scale significance testing, with applications to comparative microarray experiments." *Biostatistics*, 8: 414–432.

Sun, W., Reich, B. J., Tony Cai, T., Guindani, M., and Schwartzman, A. (2014). "False discovery control in large-scale spatial multiple testing." *Journal of the Royal Statistical Society: Series B (Statistical Methodology), To appear*.

Wu, W. and Peña, E. A. (2013). "Bayes multiple decision functions." *Electronic Journal of Statistics*, 7: 1272–1300.

Comment by Alejandro Jara[1]

The proposal

I congratulate the authors for an interesting paper, and thank them for adding another example to the long list successful applications of Bayesian nonparametric models. Finegold and Drton's paper deals with robust inference for graphical models and proposes a novel class of multivariate t-distributions that arises by (i) replacing the single latent gamma mixing variable with coordinate-specific independent latent variables, in the standard stochastic representation of multivariate t-distributions as a scale mixture of normals, and (ii) by exploiting the discrete nature of Dirichlet processes to clustering the latent mixing variables. Thus, the resulting Bayesian semiparametric model, referred to as 'Dirichlet-t distribution', allows for the clustering of the original Gaussian coordinates according to the 'degree of robustification' needed to adequately fit the data and is an intermediate model having as limiting cases two parametric models: the standard multivariate t-distribution and the multivariate t-distribution arising by assuming coordinate-specific independent gamma latent variables. The first case assumes the same degree of departure from normality for each coordinate of the response vector, and the latter enforces for coordinate-specific potential departures from normality.

The main comment

For some reason, the authors choose to limit their definition of the model to a case that allows for too little borrowing of strength across samples. Since the mixing distribution does not usually change when considering independent samples from a mixture model, one option would have been to treat the τ_{ij}'s as exchangeable across the samples. That is, by considering the hierarchical model given by

$$\boldsymbol{Y}_i \,|\, \boldsymbol{\tau}_i, \boldsymbol{\Psi}, \boldsymbol{\mu} \overset{ind.}{\sim} \mathcal{N}_p\left(\boldsymbol{\mu}, \text{diag}\{1/\sqrt{\boldsymbol{\tau}_i}\} \cdot \boldsymbol{\Psi} \cdot \text{diag}\{1/\sqrt{\boldsymbol{\tau}_i}\}\right), \tag{1}$$

$$\tau_{ij} \,|\, P \overset{i.i.d.}{\sim} P, \tag{2}$$

and

$$P \,|\, \alpha, \nu \sim DP(\alpha, P_0^{\nu}). \tag{3}$$

However, the authors' proposal implies the existence of n sample-specific independent Dirichlet processes and the hierarchical model given by

$$\boldsymbol{Y}_i \,|\, \boldsymbol{\tau}_i, \boldsymbol{\Psi}, \boldsymbol{\mu} \overset{ind.}{\sim} \mathcal{N}_p\left(\boldsymbol{\mu}, \text{diag}\{1/\sqrt{\boldsymbol{\tau}_i}\} \cdot \boldsymbol{\Psi} \cdot \text{diag}\{1/\sqrt{\boldsymbol{\tau}_i}\}\right), \tag{4}$$

$$\tau_{ij} \,|\, P_i \overset{ind.}{\sim} P_i, \tag{5}$$

[1]Department of Statistics, Pontificia Universidad Católica de Chile, atjara@uc.cl

$$P_i \,|\, \alpha, \nu \overset{i.i.d.}{\sim} DP(\alpha, P_0^{\nu}), \tag{6}$$

where the random measures P_i are linked only by the finite dimensional hyper-parameters α and ν. For some applications, the model given by expressions (1) – (3) would enforce too much borrowing by assuming essentially one 'population' P. However, the model given by expressions (4) – (6) allows too little borrowing of strength across samples, which can complicate the inferences on the infinite dimensional parameters P_i when p is not big enough. Therefore, and along the lines of what the authors are trying to do with the latent mixing variables, an intermediate case would be to consider the hierarchical Dirichlet process, originally proposed by Müller et al. (2004). Under this model, the sample-specific random measures P_i would have the following mixture representation

$$P_i = \epsilon H_0 + (1 - \epsilon) H_i$$

where H_0 is a common random measure, shared by all samples, H_i is an idiosyncratic measure which is specific to each sample, and $\epsilon \in [0, 1]$ represents the level of borrowing strength across samples. The model could be completed by assuming $H_i \,|\, \alpha, \nu \overset{i.i.d.}{\sim} DP(\alpha, P_0^{\nu})$ and $H_0 \,|\, \alpha, \nu \sim DP(\alpha, P_0^{\nu})$. This model can be easily implemented, with full conditionals of similar form of the ones described by the authors, and is a more clear way of borrowing strength than the one suggested by the authors in the discussion section, based on a Dirichlet process mixture model for $\boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_n$.

References

Müller, P., Quintana, F. A., and Rosner, G. (2004). "A method for combining inference across related nonparametric Bayesian models." *Journal of the Royal Statistical Society, Series B*, 66: 735–749.

Comment by Juhee Lee[1]

The Dirichlet t-model proposed in the paper by Finegold and Drton is well developed to obtain inferences robust to outliers in a multivariate Gaussian setting. There are several Bayesian approaches to handling outliers, as discussed in Lee and MacEachern (2014). The approach that Finegold and Drton take is to downweight possible outliers through a thick-tailed likelihood, producing posterior inference that is less affected by the outliers. This approach provides reasonable inferences for some inferential targets without requiring elaborate efforts to model the process generating the observations. In particular, this approach works well when the focus is the center of a distribution. However, as shown in Lee and MacEachern (2014), this approach does not work so well for other inferences such as a predictive distribution. The reason is that use of a thick-tailed likelihood is not designed to pick up asymmetry of the distribution, nor need it correctly capture the spread of the distribution. An alternative approach pursues density estimation of both outlying and non-outlying data and handles outliers through use of a robust inference function.

The example of the graphical model suggests a style of hybrid model which I have been developing in other contexts. Here, the model lies between a model that is purely flexible for all coordinates and one that is only flexible for some bad coordinates. In particular, one can decompose the model into two parts to describe the generative process. One component precisely models the distribution of the good coordinates (called the "head" of the model). The head of the model incorporates informed prior knowledge and may have a sharp distribution. It will often define the primary characteristics about which inference is to be made. As an example, for the Gaussian graphical model setting of the paper, the good coordinates in an observation share a single τ_i drawn from a gamma distribution with a large shape parameter, or even have $\tau_i = 1$ to reflect normality. The other component (called the "tail") accommodates departures from the head. The tail is based on vaguer information. It picks up model misfit, whether this be systematic departures of modest size from the head, or individual or small pockets of cases that are difficult to describe such as outlying coordinates. Use of a carefully tailored nonparametric Bayesian component for the tail is natural, because such a component is flexible and has full support. The Dirichlet t-distribution developed in the paper can be a good choice for the tail in the discussed problem. Alternative versions provide more control over the relative sizes of clusters of outlying observations. The final piece of the model determines the split between the head and the tail, with a simple split following a beta distribution. Under this comprehensive modeling framework, outlying coordinates within an individual observation are hopefully assigned to the tail, reducing the impact of the outliers for inference for the head. This model can be adapted to account for dependence across the vectors $\mathbf{Y}$. Therefore, local deficiencies of the model will not drive inference while the computation burden is still controlled.

[1]Applied Mathematics and Statistics, University of California Santa Cruz,juheelee@soe.ucsc.edu

References

Lee, J. and MacEachern, S. N. (2014). "Inference Functions in High Dimensional Bayesian Inference." *Statistics and Its Interface (To Appear)*.

Comment by Steven N. MacEachern[1]

Finegold and Drton provide an interesting class of models whose purpose is to provide a more robust fit of a Bayesian model. Their focus is the normal-theory graphical model, but, as they indicate, their technique can be applied much more broadly. The driving forces behind their model are the use of a thick-tailed likelihood to discount modest outliers and essentially drop extreme outliers and the exploitation of the clustering properties of the Dirichlet process to capture pockets of outliers. The details of the model place it squarely in the realm of nonparametric Bayesian methods in spite of the very parametric nature of the model. The resulting model is an excellent example of problem-driven model development.

Nonparametric Bayesian methods, in particular those based on the Dirichlet process, are well-developed and have been applied to a wide variety of problems since the advent of Markov chain Monte Carlo (MCMC) in the 1980s. MCMC computation in nonparametric Bayesian models began with the dissertation work of Escobar (1988, 1994) who developed a basic algorithm which has come to be known as a Gibbs sampler. Performance of the basic algorithm has been improved with strategies that facilitate mixing and that can aid in estimation. The perspectives and techniques in the nonparametric Bayesian literature can be borrowed to enhance computation and to suggest routes for robust graphical modelling. These strategies have proven useful in MCMC problems well beyond nonparametric Bayes.

Marginalization, or "integrating out parameters" typically improves both convergence and mixing of MCMC algorithms. Escobar's basic algorithm for mixture of Dirichlet process models marginalizes the infinite dimensional distribution function P, instead working with Blackwell and MacQueen (1973)'s Polya urn scheme which underlies (21) and (23) of the paper. MacEachern (1994) goes further, extending the basic algorithm to the hierarchical model, marginalizing the cluster locations, here the η_k, and providing a simple theoretical result on marginalization. See Liu et al. (1994) for more satisfying results on marginalization. The η_k can be generated as needed for further inference or other steps in the algorithm. Additional marginalization is possible, in particular of the mass parameter of the Dirichlet process (α in the paper, though it is perhaps better to follow Ferguson (1973)'s notation and reserve α for the base measure of the Dirichlet process). This marginalization can be performed with a pre-integration, done once, before the iterates of the MCMC are performed (MacEachern 1998), eliminating the need for steps (v) and (vi) in Algorithm 4 and adjusting the draws in step (ii) slightly.

Reparameterization of the model at various stages of an MCMC iterate can greatly enhance mixing. Bush and MacEachern (1996) provide one of the earliest examples of this technique, adding a step to the basic algorithm where the cluster locations (the η_k) are generated (step (iii) in Algorithm 4). As Finegold and Drton have found, this remixing step is essential to obtain full mixing of the Markov chain in a run of reasonable length in most contexts. Mixing is further enhanced with the split-merge techniques of Jain and Neal (2000, 2007) and Dahl (2003). These perspectives and others from

[1]Department of Statistics, The Ohio State University, snm@stat.osu.edu

the nonparametric Bayesian literature provide useful views on how to tune up MCMC algorithms. They also apply more generally, having applications to particle filtering and to a variety of approximation methods.

As a final comment, in their penultimate paragraph the authors mention the possibility of hooking the collection of n problems together. The dependent Dirichlet process (MacEachern 1999) and dependent nonparametric processes provide a framework which has proven to be successful for the development of such models. The gene expression data in Section 6 show evidence of "clustering by rows and columns," suggesting links to the work of Lee et al. (2013).

References

Blackwell, D. and MacQueen, J. B. (1973). "Ferguson distributions via Polya urn schemes." *The Annals of Statistics*, 1(2): 353–355.

Bush, C. A. and MacEachern, S. N. (1996). "A semi-parametric Bayesian model for randomized block designs." *Biometrika*, 83: 275–286.

Dahl, D. B. (2003). "An improved merge-split sampler for conjugate Dirichlet process mixture models." Technical Report 1086, Department of Statistics, University of Wisconsin.

Escobar, M. D. (1988). "Estimating the means of several normal populations by nonparametric estimation of the distribution of the means." Ph.D. thesis, Yale University.

— (1994). "Estimating normal means with a Dirichlet process prior." *Journal of the American Statistical Association*, 89: 268–277.

Ferguson, T. S. (1973). "A Bayesian analysis of some nonparametric problems." *The Annals of Statistics*, 1(2): 209–230.

Jain, S. and Neal, R. (2000). "A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model." *Journal of Computational and Graphical Statistics*, 13: 158–182.

— (2007). "Splitting and merging components of a nonconjugate Dirichlet process mixture model." *Bayesian Analysis*, 2: 445–472.

Lee, J., Müller, P., Zhu, Y., and Ji, Y. (2013). "A nonparametric Bayesian model for local clustering with application to proteomics." *Journal of the American Statistical Association*, 108(503): 775–788.

Liu, J. S., Wong, W. H., and Kong, A. (1994). "Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes." *Biometrika*, 81(1): 27–40.

MacEachern, S. (1994). "Estimating normal means with a conjugate style Dirichlet process prior." *Communications in Statistics B*, 23: 727–741.

— (1998). "Computational methods for mixture of Dirichlet process models." In Dey, D., Müller, P., and Sinha, D. (eds.), *Practical Nonparametric and Semiparametric Bayesian Statistics*, 23–44. Springer.

— (1999). "Dependent nonparametric processes." In *ASA Proceedings of the Section on Bayesian Statistical Science*, 50–55.

Comment by Abdolreza Mohammadi[1] and Ernst C. Wit[2]

We congratulate the authors with this thought-provoking paper, which constitutes a valuable contribution in graphical model inference. They propose a robust Bayesian inference method based on the Dirichlet t-distribution, that has clear benefits over Gaussian graphical models. Their method can deal with continuous data that contain outliers for some of their measurements and it is therefore more suitable for large real data sets of variable quality. Here we would like to contribute to the discussion by suggesting an extension to non-decomposable graphs and to suggest a comparison with Copula Gaussian graphical models.

Extension to non-decomposable graphs

Restricting themselves to decomposable graphs allows a closed form of the marginal likelihood, which results in an explicit form for the acceptance ratio (see paper eq. 6). However, the space of decomposable graphs is much smaller than the full graph space. For example, the percentages of graphs that are decomposable for $p = 3, 4, 5, 6, 7, 8$ variables are $1, 0.95, 0.80, 0.55, 0.29, 0.12$, respectively (Armstrong 2005, p.149). It shows that decomposability is a serious restriction, even for a small number of variables. Moreover, this restriction has several other computational consequences. Firstly, in each sweep of the main MCMC algorithm (Algorithm 1) one should check by using e.g. the Max-Cardinality algorithm whether the proposed graph is decomposable or not, which is relatively computationally expensive. Secondly, as in high-dimensions most graphs are non-decomposable, moves will have a high rejection rate and result in slow convergence. As the authors mention in the conclusion, one extension of their work should be to the non-decomposable graphs by using e.g. the double reversible jump (Lenkoski and Dobra 2011; Lenkoski 2013) or birth-death MCMC approach (Mohammadi and Wit 2014a), which we proposed recently. In general for high-dimensional graphs, reversible jump algorithms still suffer from high rejection rates. Using the Dirichlet t-distribution in combination with the birth-death MCMC algorithm (Mohammadi and Wit 2014a) would be very promising.

Comparison with copula Gaussian graphical models

Gaussian graphical models are very sensitive to outliers. The current paper proposes an excellent way to deal with this sensitivity. A potentially different way is to use copula Gaussian graphical models (Dobra and Lenkoski 2011; Mohammadi et al. 2014). This method embeds a graph selection procedure inside a semiparametric Gaussian copula, which can deal robustly with many marginal distributions of the data. For copula estimation, the marginal likelihood is only a function of the association parameters. Indeed, graph determination based on the t-distribution is a spacial case of copula Gaussian graphical models (CGGMs) by assuming that the marginal distributions have

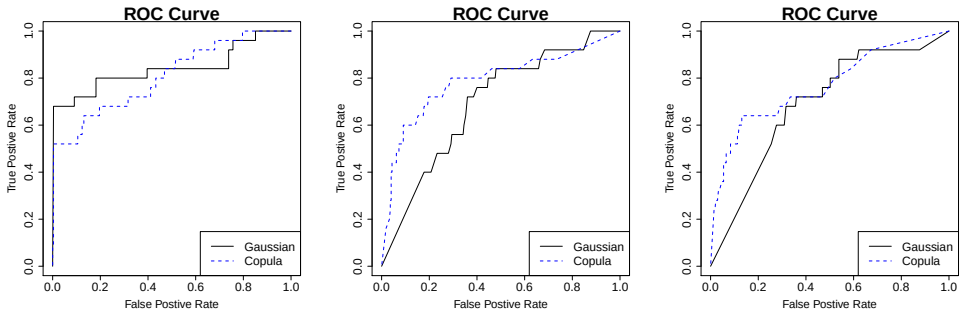
[1]Dept. of Statistics, University of Groningen, Groningen, Netherlands, a.mohammadi@rug.nl
[2]Dept. of Statistics, University of Groningen, Groningen, Netherlands, e.c.wit@rug.nl

a t-distribution. It has the advantage that the full likelihood has an explicit form, however it also means that it can deal only with continuous t-distributed data.

Simulation example: AR1 with $p = 25$

To compare the performance of the t-distribution method proposed in the paper with the CGGMs, we run our simulation based on the same scenario as in subsection 5.1 of the paper. Regrettably the code for the t-distribution method is unavailable to do any direct comparison, so here we focus on a comparison with a standard Gaussian graphical model (GGM). We run two algorithms. One is the birth-death MCMC algorithm for decomposable and non-decomposable GGMs (Mohammadi and Wit 2014a). The other is the birth-death MCMC algorithm for CGGMs (Mohammadi et al. 2014).

The simulation results are summarized in Figure 1, which shows that CGGMs indeed perform better when the data is generated from a model with outliers, such as the t-distribution and alternative t-distribution. For normal data, the CGGM method is only marginally worse than the GGM method. The processing time to fit CGGMs was around $1.2N$ where N is the processing time for the GGMs. The results are based on the R-package BDgraph (Mohammadi and Wit 2014b).



(a) Normal,n=25,p=25 (b) Classical,n=25,p=25 (c) Alternative,n=25,p=25

Figure 1: ROC curves present the performances of the two methods for data generated from a $N_{25}(0, K^{-1})$ distribution, a $t_{25,3}(0, K^{-1})$ distribution and a $t^*_{25,3}(0, K^{-1})$ distribution.

References

Armstrong, H. (2005). "Bayesian Estimation of Decomposable GGMs." *PhD thesis*, The University of New South Wales.

Dobra, A. and Lenkoski, A. (2011). "Copula Gaussian graphical models and their application to modeling functional disability data." *The Annals of Applied Statistics*, 5(2A): 969–993.

Lenkoski, A. (2013). "A direct sampler for G-Wishart variates." *Stat*, 2(1): 119–128.

Lenkoski, A. and Dobra, A. (2011). "Computational aspects related to inference in Gaussian graphical models with the G-Wishart prior." *Journal of Computational and Graphical Statistics*, 20(1): 140–157.

Mohammadi, A., Abegaz, F., and Wit, E. C. (2014). "Efficient Bayesian inference for Copula Gaussian graphical models." *Proceedings of the 29th International Workshop on Statistical Modelling*, 1: 225–230.

Mohammadi, A. and Wit, E. C. (2014a). "Bayesian structure learning in sparse Gaussian graphical models." *Bayesian Analysis*, accepted for publication.

— (2014b). *BDgraph: Graph estimation based on birth-death MCMC*. R package version 2.10.
URL http://CRAN.R-project.org/package=BDgraph

Comment by Anthony O'Hagan[1]

Whilst I imagine that the material in this paper relating to graphical models is its principal focus, I am particularly interested in Finegold and Drton's novel heavy-tailed multivariate distributions. The authors note that, "There is substantial literature on robustness in Bayesian inference," but then cite two very old papers. It seems that they have not actually researched that "substantial literature" fully. A good starting point would be the recent review by O'Hagan and Pericchi (2012).

What Finegold and Drton call their alternative t distribution is closely related to the case of independent t distributions which arises in the special case when $\mathbf{\Psi}$ is diagonal. This case is illustrated in their Figure 2, and has been used by various authors for robust Bayesian modelling. As the authors note, this kind of formulation allows a wider range of robust responses than the simple multivariate t distribution. This is an important point, because a major theme of the work reviewed in O'Hagan and Pericchi (2012) is that different heavy-tailed modelling formulations lead to different robust responses in the posterior distribution, and hence that care must be taken when introducing heavy-tailed models to choose a formulation with the desired posterior behaviour. This is the modelling philosophy first propounded in O'Hagan (1988), and is clearly seen in the authors' reasons for proposing first the alternative t distribution and then the Dirichlet t distribution. However, it merits being emphasised here because there is a common misconception amongst Bayesians that 'robustifying' a model by introducing any convenient form of t distribution will automatically produce desirable posterior behaviour.

By allowing a general $\mathbf{\Psi}$ matrix in their alternative t distribution, the authors have introduced a new class of multivariate heavy-tailed distributions. Figure 2 below shows 100,000 draws from an alternative t distribution with $\psi = \begin{pmatrix} 11 & 9 \\ 9 & 11 \end{pmatrix}$, the strong correlation of $\frac{9}{11}$ implied here being evident in the preponderance of points in the first and third quadrants. Nevertheless, the fact that the heavy tails are operating only along the x and y axes means that the actual correlation in this distribution is much lower. In contrast, consider Figure 3, which shows a bivariate distribution that is also characterised by $\psi = \begin{pmatrix} 11 & 9 \\ 9 & 11 \end{pmatrix}$, but now the heavy tails apply along the principal axes $x = y$ and $x = -y$. The correlation in this distribution is $\frac{9}{11}$. The point is that in multivariate heavy-tailed distributions we can have different tail weights in all directions. The multivariate t has the same tail weight in all directions, whereas the alternative t has heavy tails along the axes. Figure 3 illustrates just one of many other possibilities, while others are discussed in O'Hagan and Le (1994).

The authors introduce the very interesting Dirichlet t distribution to allow for the possibility of clustering. The tails of such a distribution will exhibit complex patterns of thickness. Another example of heavy-tailed modelling to achieve clustering is given in O'Hagan (1988). Another strong theme in this literature is the importance of the

[1]University of Sheffield, a.ohagan@sheffield.ac.uk

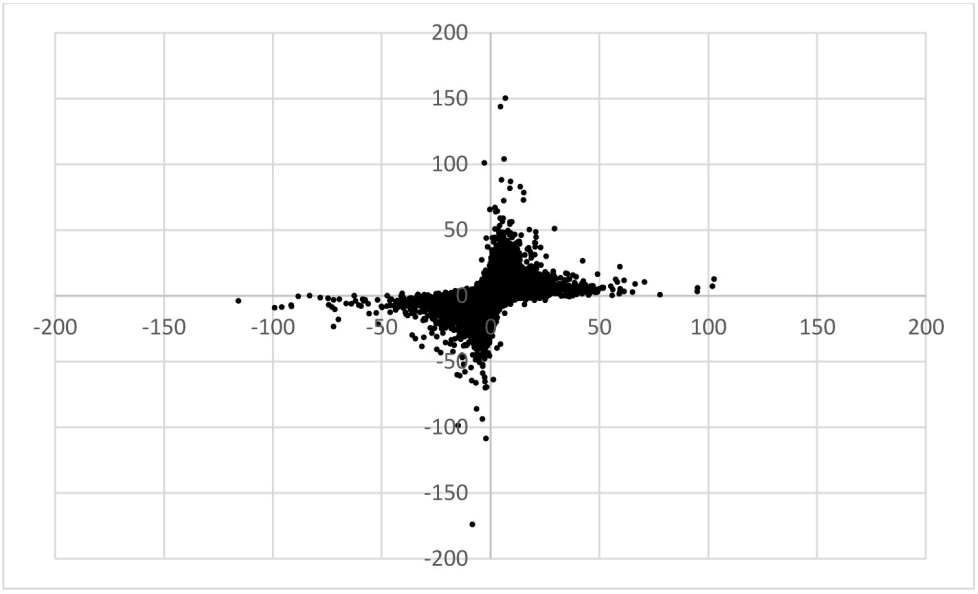


Figure 2: Draws from an alternative t distribution.

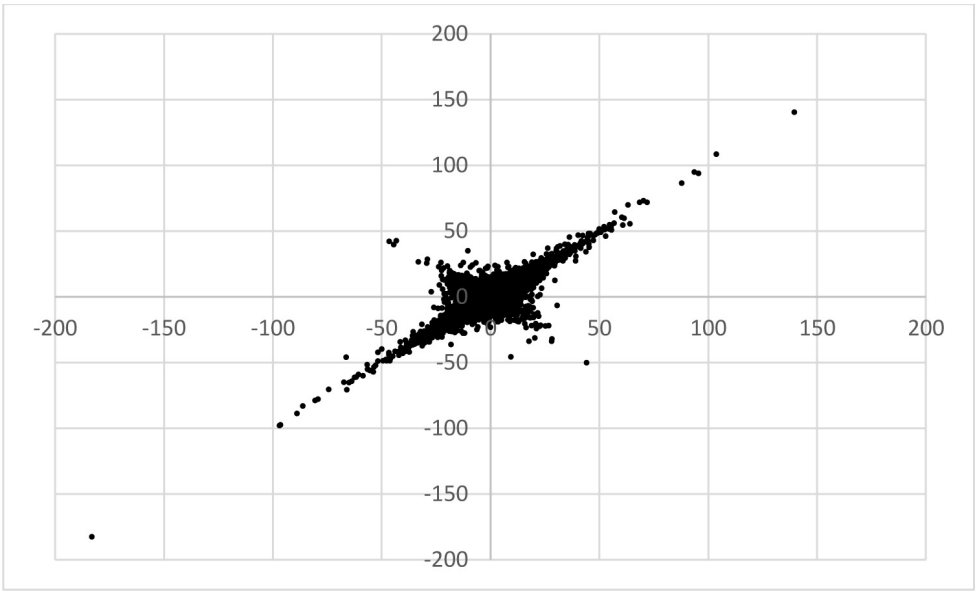


Figure 3: Draws from a bivariate distribution.

degrees of freedom in t modelling components. Figures 2 and 3 employ 3 degrees of freedom, which is the authors' default choice, but changing the relative tail thicknesses of different components can radically change the posterior behaviour; see O'Hagan and Pericchi (2012).

References

O'Hagan, A. (1988). Modelling with heavy tails. *Bayesian Statistics 3*, J. M. Bernardo *et al.* (Eds.), 345–359. Oxford University Press.

O'Hagan, A. and Le, H. (1994). Conflicting information and a class of bivariate heavy-tailed distributions. In *Aspects of Uncertainty: a Tribute to D. V. Lindley*, A. F. M. Smith and P. R. Freeman (eds.), 311–327. Wiley: Chichester.

O'Hagan, A. and Pericchi, L. (2012). Bayesian heavy-tailed models and conflict resolution: a review. *Brazilian Journal of Probability and Statistics* **26**, 372–401.

Comment by Stefano Peluso[1]

We congratulate the authors for the insightful generalization of the classical multivariate t-distribution to the Dirichlet t-distribution that allows to model graphs that account for outliers, still keeping a reasonably low computational burden. In this comment we focus on a possible further generalization aiming at incorporating skewness in the analysis. In more details, if we define $\Omega := \text{diag}(1/\sqrt{\tau_i})\Psi\text{diag}(1/\sqrt{\tau_i})$, then $Y_i|\tau_i, \Psi, \mu \sim N_p(\mu, \Omega)$ in the model can be replaced by

$$\begin{cases} Y_i|\tau_i, \Psi, \mu, \Delta, \omega \sim N_p(\mu + \Delta\omega, \Omega) \\ \omega \sim N_p(0, I_p)1_{\{\omega>0\}} \end{cases}, \tag{7}$$

where Δ is a $p \times p$ skewness matrix and ω is a p-dimensional Gaussian latent variable truncated to the region $\{\omega_i > 0,\ i = 1, \ldots, p\}$. Usually $\Delta = \text{diag}(\delta_1, \ldots, \delta_p)$ or, if the same skewness for all the data is assumed, $\Delta = \delta I_p$. Note that for $\delta_i > 0$ or $\delta_i < 0$ a positively or negatively skewed distribition is obtained, whilst for $\delta_i = 0$ for all i, the original model in the paper is obtained.

Rikhtehgaran and Kazemi (2013) show that the hierarchical representation in (7) is equivalent to the multivariate Skew-Normal distribution of Azzalini and Dalla Valle (1996). From a computational point of view, ω can be sampled simulating from a multivariate normal until the generated number is positive or, more efficiently, using the distribution function inversion method suggested in Gelfand et al. (1992); see also Robert (1995) and Solgi and Mira (2014) for related sampling algorithms. Then, if τ_i is a sample from a Dirichlet process of Ferguson (1973), the resulting model is a Skew-Dirichlet t-model, generalizing, to the Bayesian nonparametric framework, the multivariate skew t-distribution introduced in Azzalini and Capitanio (2003).

The state space in the Gibbs sampler is then extended from (G, Θ, z, η) - as in the discussed paper - to $(G, \Theta, z, \eta, \Delta, \omega)$. Following the approach in the paper, for computational reasons μ is approximated and not sampled, but in the extended model with skewness its approximation is now $\mu = \frac{\sum_{i=1}^n \tau_i Y_i}{\sum_{i=1}^n \tau_i} - \Delta\omega$. The full conditionals of G, Θ, z and η remain almost unchanged, with the only difference being that μ is replaced by $\mu + \Delta\omega$. For sampling $\Delta = \text{diag}(\delta_1, \ldots, \delta_p)$ in the Gibbs iterations, it can be shown that

$$\omega|G, \Theta, z, \eta, \Delta \sim N_p(\mu_\omega, \Sigma_\omega)1_{\{\omega>0\}},$$

where $\Sigma_\omega = (\Delta\Omega^{-1}\Delta + I_p)^{-1}$ and $\mu_\omega = \Sigma_\omega^{-1}(\Delta\Omega^{-1}(Y_i - \mu))$. Finally, for a prior $(\delta_1, \ldots, \delta_p) \sim N_p(0, \sigma_\Delta^2 I_p)$, a posteriori $\Delta|G, \Theta, z, \eta, \omega \sim N_p(\mu_\Delta, \Sigma_\Delta)$ is obtained, where $\Sigma_\Delta = (\Omega^{-1}\omega\omega' + I_p/\sigma_\Delta^2)^{-1}$ and $\mu_\Delta = (\Sigma_\Delta^{-1}(\Omega^{-1}(Y_i - \mu)\omega'))_{ii}$, for $i = 1, \ldots, p$.

To illustrate the usefulness of the skewness generalization, we extend Figure 2 in the paper, by simulating 100.000 draws from the Skew-Dirichlet t-distribution, for various

[1]Universita della Svizzera italiana, stefano.peluso@usi.ch

skewness matrices. The results in Figure 4 show that the distribution is still able to account for joint ouliers in the four corners, but it can also exhibit an asymmetric behavior. The relevance of the skewed model on real data is under current investigation and will be thoroughly developed in a future work.

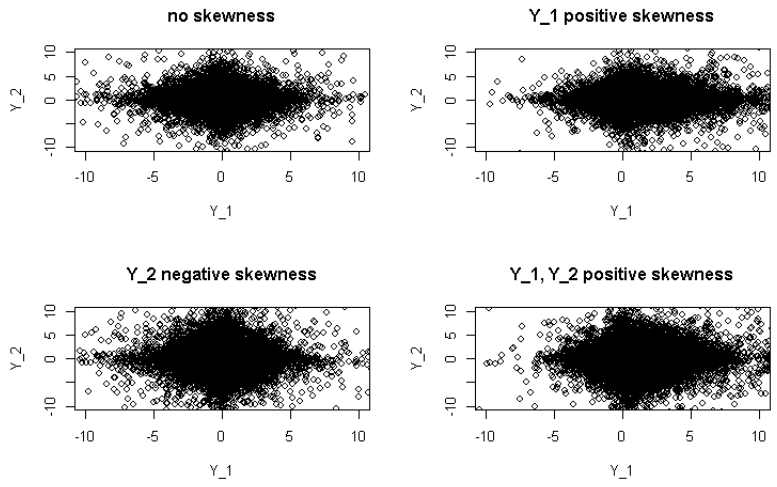


Figure 4: Top left plot: 100.000 draws from the Dirichlet $t_{2,3}^{\alpha=1}(0, I_2)$ (no skewness), corresponding to the Skew-Dirichlet t-distribution with $\delta_1 = 0$ and $\delta_2 = 0$. In the other plots: draws from the Skew-Dirichlet t-distribution under distinct skewness scenarios: $\delta_1 = 1$, $\delta_2 = 0$ (Y_1 positive skewness), $\delta_1 = 0$, $\delta_2 = -1$ (Y_2 negative skewness) and $\delta_1 = 1$, $\delta_2 = 1$ (Y_1, Y_2 positive skewness).

References

Azzalini, A. and Capitanio, A. (2003). "Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t distribution." *Journal of the Royal Statistical Society, Series B*, 65: 367–389.

Azzalini, A. and Dalla Valle, A. (1996). "The multivariate skew-normal distribution." *Biometrika*, 83: 715–726.

Ferguson, T. (1973). "A Bayesian analysis of some nonparametric problems." *Annals of Statistics*, 1: 209–230.

Gelfand, A., Smith, A., and Lee, T. (1992). "Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling." *Journal of the American Statistical Association*, 87: 523–532.

Rikhtehgaran, R. and Kazemi, I. (2013). "Semi-parametric Bayesian estimation of

mixed-effects models using the multivariate skew-normal distribution." *Computational Statistics*, 28: 2007–2027.

Robert, C. (1995). "Simulation of truncated normal variables." *Statistics and Computing*, 5: 121–125.

Solgi, R. and Mira, A. (2014). "Does the HAR model select the right frequencies to predict volatility?" Working paper.

Comment by Luis R. Pericchi[1]

Bayesian Robustness or Bayesian Conflict Resolution (BCR) is one of the best routes to appreciate Bayesian Statistics: in a structured way you arrive from assumptions to conclusions, in a "what if" approach, as opposed to ad-hoc non Bayesian methods. In addition Bayes gives you probabilities and uncertainty measures. This powerful paper is an excellent example, on which very creative alternative versions of multivariate t-Distributions are employed and the objective sought fulfilled. However, BCR is much more than replacing Normal by t-distributions (Pericchi et al. (1993), O'Hagan and Pericchi (2012)). In the latter reference is exposed a theme under-represented in BCR, alternative heavy tailed distributions for scale parameters. For scales, in addition to tail considerations, behavior at the origin is crucial, and Gammas, Inverted-Gammas and their multivariate generalizations may also need replacement as suggested by Gelman (2006), Fuquene, Perez, and Pericchi (2014) and by Perez, Pericchi, and Ramirez (2014) presented at this ISBA conference. But I have to recognize that what thrilled me most was the interpolation achieved between models through the Dirichlet Process. Similar effects may have been obtained treating the problem as a Bayesian Model Averaging problem. The final question is, what are the relationships of these two methods of interpolations between models? Which are the relative merits? This seems to be a substantial and exciting research program.

References

Fuquene, J., Perez, M. E., and Pericchi, L. R. (2014). "An alternative to the Inverted Gamma for the variances to modelling outliers and structural breaks in dynamic models." *Brazilian Journal of Probability and Statistics*, 28(2): 288–299.

Gelman, A. (2006). "Prior distributions for variance parameters in hierarchical models(Comment on Article by Browne and Draper)." *Bayesian Analysis*, 1(3): 515–534.

O'Hagan, A. and Pericchi, L. R. (2012). "Bayesian heavy-tailed models and conflict resolution: a review." *Brazilian Journal of Probability and Statistics*, 26(372-401): 372–401.

Perez, M. E., Pericchi, L. R., and Ramirez, I. (2014). *The Scaled Beta2 Distribution as a robust prior for scales, and an Explicit Horseshoe Prior for locations*. Mexico: Presented at the ISBA World Meeting.

Pericchi, L. R., Sanso, B., and Smith, A. F. M. (1993). "Posterior Cumulant relationships in Bayesian inference involving the Exponential Family." *Journal of the American Statistical Association*, 88: 1419–1426.

[1]University of Puerto Rico, Rio Piedras Campus, Department of Mathematics, luis.pericchi@upr.edu

Comment by Abel Rodríguez[1]

I would like to start by congratulating the authors for a very interesting paper. Gaussian graphical models have become an important tool in applied fields such as genomics and finance. However, inferences derived from them are sensitive to the presence of outliers. The authors address this issue by considering graphical models based on heavy-tailed distributions that can be written as scale mixtures of Gaussians. In particular, the main contribution is a new sparse multivariate t distribution in which the over-dispersion parameters are assigned a nonparametric prior based on a Dirichlet process. Their specification allows for differential shrinkage along each dimension while attempting to preserve parsimony by reducing the number of distinct parameters being estimated. Furthermore, the fact that the model is conditionally Gaussian allows the authors to leverage well-known Markov chain Metropolis Hastings algorithms for posterior inference.

One important concern associated with the kind of t-variate graphical models presented in the paper relates to the interpretation of the underlying graph. Indeed, in the t models discussed by the authors, zeros in the precision matrix *do not imply* that the corresponding variables are conditionally independent. Of course, conditional on the over-dispersion parameters $\{\tau_{i,j}\}$ (the "divisors", in the language of the paper) we do have normality, and therefore conditional independence. However, since divisors are different for different observations and, potentially, different dimensions, the interpretation of the model is very awkward. Note that a similar problem arises in the context of countable mixtures (see, for example, Rodriguez et al. 2011). However, in that case observations within each cluster follow the same Gaussian distribution, so interpretations that condition on the group structure are straightforward. I believe that it would be useful if the authors can comment on the interpretation of the underlying graph (particularly in the context of their application), and the differences between using their models for testing hypotheses about structural relationships in the data and for prediction (in which case the issues I just raised are moot).

A number of natural extensions of the models presented here come to mind. Probably the most obvious arises by treating the number of degrees of freedom ν as unknown. This is computationally straightforward and would allow the data to automatically inform the model about the weight of the tails. Moreover, we could allow the behavior of the tails of each marginal distribution to be different by considering a different value ν_j for the distribution of each sequence $\tau_{1,j}, \ldots, \tau_{n,j}$ (in the case of the alternative t distribution considered in Section 3.3) or mixing on both ν and τ (in the case of the Dirichlet-t models discussed in Section 4). In a different direction, we could further borrow strength across observations by modeling P_0 non-parametrically by letting $P_0 \sim \mathsf{DP}(\beta, \Gamma(\nu/2, \nu/2))$ instead of letting $P_0 = \Gamma(\nu/2, \nu/2)$ (leading to a hierarchical Dirichlet process prior on $\{\tau_{i,j}\}$, Teh et al. 2006). I am somewhat surprised that the authors did not consider using a different (unknown) number of degrees of freedom for each marginal distribution, which I believe could have an impact on the results at least as big as the introduction

[1]Department of Applied Mathematics and Statistics, University of California, Santa Cruz, abel@soe.ucsc.edu

of a nonparametric prior on the over-dispension coefficients.

One final point that the authors could expand upon refers to the scalability of the computational algorithms discussed in the paper. Although the authors consider higher dimensional examples as part of their simulation study, the real data is relatively low dimensional ($p = 8$) for today's standard. What are the execution times (per posterior equivalent sample size) of their algorithms, and how do they scale as the number of variables and sparsity of the graph increase?

References

Rodriguez, A., Lenkoski, A., Dobra, A., et al. (2011). "Sparse covariance estimation in heterogeneous samples." *Electronic Journal of Statistics*, 5: 981–1014.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). "Sharing Clusters among Related Groups: Hierarchical Dirichlet Processes." *Journal of the American Statistical Association*, **101**: 1566–1581.

Comment by Pablo E. Verde[1]

My congratulations to the authors for this interesting paper. I found the extension of the classical t-distribution by using a scale mixture of normal distributions per coordinate quite useful in practice and the Dirichlet t-distribution an elegant approach. I would like to make the following practical comments:

Statistical inference of the parameter α in the Dirichlet t-distribution looks challenging. The authors Michael Finegold and Mathias Drton applied two strategies: one by fixing α to different values and another one by applying a Gamma prior distribution with parameters equal to 1, which gives a prior $E(\alpha) = 1$. In applications, I would recommend to make a prior to posterior analysis of this parameter in order to understand if we could learn something about α from the data at hand. The same strategy should be applied to the degrees of freedom parameter ν.

In my work in multi-parameters meta-analysis (Verde 2010; Verde and Sykosch 2011) I found that the single component scale mixture is useful enough for outliers' identification and for down-weighting pieces of evidence with unusual results. However, the introduction of the Dirichlet t-distribution opens an interesting possibility in the detection of conflict of evidence in meta-analysis and in the detection of structural outliers in Bayesian hierarchical modeling.

The conflict assessment is the deconstructionist side of meta-analysis, where each piece of evidence is put aside from the full model and compared to the rest of the evidence. One possibility for this type of analysis is to embed a meta-analysis model in a more general model where the non-conflict situation is a particular case. For example in Verde et al. (2014), we applied a scale mixture of multivariate normal distributions in a meta-analysis combining randomized and non-randomized evidence and we made conflict diagnostics by direct interpretation of the scale weights. Another alternative is presented by Presanis et al. (2013), where the authors described how to generalize the conflict p-value proposed by Marshall and Spiegelhalter (2007) to complex evidence modeling. In summary, by using a Dirichlet t-distribution conflict of evidence can be generalized and performed for each parameter in a multi-parameter meta-analysis.

References

Marshall, E. C. and Spiegelhalter, D. J. (2007). "Identifying outliers in Bayesian hierarchical models: a simulation-based approach." *Bayesian Analysis*, 2: 409–444.

Presanis, A. M., Ohlssen, D., Spiegelhalter, D., and Angelis, D. D. (2013). "Conflict diagnostic in directed acyclic graphs, with applications in Bayesian evidence synthesis." *Statistical Science*, 28: 376–397.

Verde, P. E. (2010). "Meta-analysis of diagnostic test data: a bivariate Bayesian modeling approach." *Statistics in Medicine*, 30(29): 3088–3102.

[1]Coordination Center for Clinical Trials, University of Duesseldorf, Germany, pabloemilio.verde@hhu.de

Verde, P. E., Ohmann, C., Icks, A., and Morbach, S. (2014). "Bayesian evidence synthesis and combining randomized and nonrandomized results: a case study in diabetes." *Statistics in Medicine*, (under review).

Verde, P. E. and Sykosch, A. (2011). "bamdit: Bayesian meta-analysis of diagnostic test data." *CRAN: R package version 1.1-1*.