Bayesian Analysis (2014) **9**, Number 3, pp. 591–596

Rejoinder

Michael Finegold [*] and Mathias Drton [†]

In a nutshell, our paper treats the problem of how to extract information from partially corrupted observations in multivariate statistical problems—specifically, we focused on the problem of graphical model selection. Our approach considers different versions of multivariate distributions with t-marginals that are obtained by using, to varying extent, distinct Gamma-divisors for the different coordinates of a random vector. As described in the contributions to this discussion, there are a lot of ways one could modify or generalize the models we used, and related ideas have appeared or might be useful in contexts other than graphical modeling. There is also a vast literature on the general themes our paper touches upon, and the discussants have provided many additional references giving a far more comprehensive description of the existing literature than our own paper.

In this rejoinder we try to summarize and comment on the ideas we see emerge from the discussion.

Directions of multivariate tails. Figure 2 in our paper contrasts different versions of t-distributions in a bivariate setting. The figure shows a 'spherical' case, that is, the dispersion matrix $\mathbf{\Psi}$ is the identity. **Anthony O'Hagan**'s Figure 1 shows a case where $\mathbf{\Psi}$ exhibits strong correlation. His figure illustrates nicely that the alternative t-distribution has heavy tails along the coordinate directions, which also underlies the bounds on correlations we mention in our Section 4.2. O'Hagan's Figure 2 shows an example of a different type of t-distribution with heavy tails following principal component directions. We primarily thought of applications in which the latent dependence pattern captured by $\mathbf{\Psi}$ and the pattern of outliers are not tied together, the former being of say biological nature and the latter a matter of experimental technology. In this case focusing on the coordinate directions seems natural to us, but principal component directions or another coordinate system could be of interest in other applications.

Inference on the degrees of freedom. Our numerical work was based on the default choice of $\nu = 3$ degrees of freedom. **Abel Rodriguez** raised the point of inference on the degrees of freedom. Sticking with the precise setup of our paper, we expect that the message regarding the relative merits of the different t-distributions remains the same under inference on ν. However, one would certainly be able to decrease the gap that is visible in the ROC curve for normal data in Figure 3 of our paper. The work of Besag and Higdon (1999) is one example of Bayesian inference on the degrees of freedom. The two used a finite set of values for ν and suggested marginalization of the Gamma-divisors for a block Gibbs step, which poses no problems in their setup of independence ($\mathbf{\Psi}$ diagonal, in our setting). For non-diagonal $\mathbf{\Psi}$, their blocking strategy might prove useful for what we called the classical t-distribution but it seems more

[*] Department of Statistics, Carnegie Mellon University, Pittsburgh, U.S.A. mfinegol@andrew.cmu.edu

[†] Department of Statistics, University of Washington, Seattle, U.S.A. md5@uw.edu

DOI:10.1214/14-BA856

difficult to implement for the other versions of t-distributions.

Different degrees of freedom and related suggestions. Several discussants suggested refined models allowing for possibly different degrees of freedom in different coordinates of an observation or more direct ways to model 'good' components of a multivariate observation as (more or less) Gaussian while using heavier-tailed distributions only for 'bad' components; see the comments of **Franccois Caron & Luke Bornn**, **Anthony O'Hagan**, **Juhee Lee** and **Abel Rodriguez**. We agree that this an interesting avenue to explore in future work. Modeling the degrees of freedom in these ways would offer opportunities for gain in statistical efficiency when a considerable portion of the data is roughly Gaussian. It would also alleviate further the conflict between modeling of multivariate outliers and shrinkage of correlations that becomes more pronounced with smaller degrees of freedom. In contrast to our t-distribution models, which place a particular measurement model on top of the entire Gaussian model, refined models might also be more appealing to those that would like see a stronger reference to an underlying Gaussian model—we interpret the question of **Jayanta Ghosh** as going in that direction. However, in either case any conditional independence interpretation of the edge pattern in graphical modeling would have to make reference to the latent Gaussian vector.

If we write ν_{ij} for the degrees of freedom for divisor τ_{ij} then the approach most closely in line with our Dirichlet t setup would let the Dirichlet process clustering pertain to the pairs $(\tau_{i1}, \nu_{i1}), \ldots, (\tau_{ip}, \nu_{ip})$, as mentioned by Rodriguez. This could also be viewed as yet another version of more general scale mixing, as discussed below. The second approach of Caron & Bornn, which to us looks like one possible implementation of the ideas outlined by Lee, seems just as promising. A comparison would get at the question raised at the end of **Luis Pericchi**'s comment.

Borrowing strength across the sample. Figure 6 in our paper shows small divisors for four genes across a group of 11 experiments. For this and similar examples, it could thus be of interest to explore shared structure in the Dirichlet clustering across the sample. **Franccois Caron & Luke Bornn**, **Steve MacEachern**, **Alejandro Jara** and **Abel Rodriguez** commented and provided references to different work that would be relevant for such extensions, and we agree that constructions involving hierarchical Dirichlet processes would be natural. In fact, the Ph.D. thesis of the first author (Finegold 2010) discusses this in more detail than the possibly misleading comments in the conclusion of our paper. The thesis also spells out the full conditionals needed for a Gibbs sampler, following the setup of Teh et al. (2006). It would be nice to see a full exploration of the idea in a future paper.

Beyond t-distributions. While our focus was solely on t-distributions for modeling of heavy tails/downweighting of outliers, our work could be repeated for many other similar constructions, and the discussants suggest interesting examples. The figures in **Babak Shahbaba**'s comment illustrate some popular distributional choices in scale mixtures of normals, and **Stefano Peluso** shows how skewness could be generated. Skewness is mentioned by **Franccois Caron & Luke Bornn** as well. Moreover, the two point out nice properties of the family of Generalized Inverse Gaussian distributions.

Again we look forward to development and applications of these ideas. In particular, the Generalized Inverse Gaussian distributions could be of interest to the problems involving scale parameters mentioned by **Luis Pericchi**.

Moving away from t-distributions in a rather different way, **Abdolreza Mohammadi & Ernst Wit** consider copula Gaussian graphical models and give pointers to recent literature on Bayesian inference. This is an area that has indeed seen much activity in the last few years (see also Liu et al. 2009, 2012; Xue and Zou 2012; Harris and Drton 2013). We agree that these semiparametric models would likely do quite well in handling the problems we discuss in our paper. However, Gaussian copula models do not overlap with our t-distribution models. In fact, one could consider a copula construction based on our t-distributions instead of the Gaussian; compare, for instance, Han and Liu (2014) and also the comment of Caron & Bornn.

Refinements in inference for graphical models. In our treatment of graphical models, we stuck to a computationally convenient setting, using rather simple priors and restricting to decomposable graphs. While this was sufficient to make our main point about the properties of the different t-distributions, one would expect serious applications to call for refinements. **Guido Consonni & Luca La Rocca** outline a number of possible improvements in prior choice that would indeed be worth pursuing. In particular, in higher-dimensional settings with sparsity, the prior distribution on graphs they suggest can drastically improve model selection. This was also shown recently for Bayesian information criteria for graphical models by Foygel and Drton (2010, 2014) and Gao et al. (2012) who build on the work of Bogdan et al. (2004), Chen and Chen (2008) and Chen and Chen (2012). The last paragraph of the comment of **Franccois Caron & Luke Bornn** points out another interesting possibility for priors over graphs.

The restriction to decomposable graphs was made merely for computational convenience as the associated Hyper Inverse Wishart distributions have a normalizing constant in simple closed form. Traversing the space of decomposable graphs is not as involved as the comments of **Abdolreza Mohammadi & Ernst Wit** suggest, in particular, one does not need to invoke the max-cardinality algorithm at each step. However, we share the view of Mohammadi & Wit that it would be preferable to consider all graphs in applications. In fact, the 'statistical cost' of using only decomposable models has been studied recently by Fitch and Jones (2012). The main difficulty in working with non-decomposable graphs is the fact that the normalizing constant for the Hyper Inverse Wishart distribution no longer has the same simple form as in the decomposable case. To address the issue, different approximation methods have been considered in the literature. All references that come to our mind in this regard are mentioned in the recent manuscript of Uhler et al. (2014) who provide some new mathematical insight in the normalizing constants.

Throughout our paper we treat the case of independent and identically distributed observations. However, this can be somewhat of a stretch for applications such as gene expression, where different observations might be obtained under rather different experimental conditions as emphasized in the comment of **Adrian Dobra**. Different approaches might be useful to address this issue depending on how much is known about how experimental interventions affect the system under observation. With little prior knowledge, building mixture models could be appealing. If more detailed information on the effects of interventions is available, then we would find it more appropriate to work with directed graphical models at the latent Gaussian level and leverage their causal interpretation; see Pearl (2009); Spirtes et al. (2000) for an introduction to this area.

Other areas of application. As **Jayanta Ghosh** comments, our work could prove useful in contexts other than graphical modeling. We agree that the application to graphical modeling is in no way specific to the robustness issues we aimed to address. Any other multivariate statistical problem could have been the base problem. Being more specific, **Michele Guindani** discusses possible applications in multiple testing, **Babak Shahbaba** mentions closely related work in the context of genome-wide association studies, and **Pablo Verde** suggests meta-analysis as a further interesting field of application.

Computation. As commented on by some of the discussants, our work involves a considerable amount of Markov chain Monte Carlo computation. **Steve MacEachern** provides a nice review of the sampling strategies that have been used to handle Dirichlet process-based models. Based on our numerical work, our methods can definitely handle problems with a few hundred variables and samples, but this is of course somewhat reliant on the data. For larger scale problems, it will be necessary to take computational shortcuts. For instance, we might opt for simpler clustering schemes for the divisors τ_{ij} instead of our fully Bayesian treatment of the Dirichlet t model, and we might consider subsampling methods to cope with large sample size. Scaling up Bayesian methods is an area of great recent activity, and we expect others to be able to propose less naive approaches to speed up computations.

To conclude, we are very grateful to all of the discussants for their comments, with special thanks going to Fran**c**cois Caron and Babak Shahbaba for taking on the role of oral discussants at the 2014 ISBA meeting. The discussion offers a wide range of ideas beyond what is mentioned in our paper, and we look forward to seeing some of the suggested ideas in full development.

References

Besag, J. and Higdon, D. (1999). "Bayesian analysis of agricultural field experiments." *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 61(4): 691–746. With discussion and a reply by the authors. 591

Bogdan, M., Ghosh, J. K., and Doerge, R. W. (2004). "Modifying the Schwarz Bayesian Information Criterion to Locate Multiple Interacting Quantitative Trait Loci." *Genetics*, 167: 989–999. 593

Chen, J. and Chen, Z. (2008). "Extended Bayesian information criteria for model selection with large model spaces." *Biometrika*, 95(3): 759–771. 593

— (2012). “Extended BIC for small-n-large-P sparse GLM.” *Statistica Sinica*, 22(2): 555–574. 593

Finegold, M. A. (2010). *Robust network inference with multivariate t-distributions*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)–The University of Chicago. 592

Fitch, A. M. and Jones, B. (2012). “The cost of using decomposable Gaussian graphical models for computational convenience.” *Computational Statistics & Data Analysis*, 56(8): 2430–2441. 593

Foygel, R. and Drton, M. (2010). “Extended Bayesian information criteria for Gaussian graphical models.” *Advances in Neural Information Processing Systems*, 23: 2020–2028. 593

— (2014). “High-dimensional Ising model selection with Bayesian information criteria.” ArXiv:1403.3374. 593

Gao, X., Pu, D. Q., Wu, Y., and Xu, H. (2012). “Tuning parameter selection for penalized likelihood estimation of Gaussian graphical model.” *Statistica Sinica*, 22(3): 1123–1146. 593

Han, F. and Liu, H. (2014). “Scale-Invariant Sparse PCA on High-Dimensional Meta-Elliptical Data.” *Journal of the American Statistical Association*, 109(505): 275–287. 593

Harris, N. and Drton, M. (2013). “PC algorithm for nonparanormal graphical models.” *Journal of Machine Learning Research*, 14: 3365–3383. 593

Liu, H., Han, F., Yuan, M., Lafferty, J., and Wasserman, L. (2012). “High-dimensional semiparametric Gaussian copula graphical models.” *The Annals of Statistics*, 40(4): 2293–2326. 593

Liu, H., Lafferty, J., and Wasserman, L. (2009). “The nonparanormal: semiparametric estimation of high dimensional undirected graphs.” *Journal of Machine Learning Research*, 10: 2295–2328. 593

Pearl, J. (2009). *Causality*. Cambridge University Press, Cambridge, second edition. Models, reasoning, and inference. 594

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, prediction, and search*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, second edition. With additional material by David Heckerman, Christopher Meek, Gregory F. Cooper and Thomas Richardson, A Bradford Book. 594

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). “Hierarchical Dirichlet processes.” *Journal of the American Statistical Association*, 101(476): 1566–1581. 592

Uhler, C., Lenkoski, A., and Richards, D. (2014). “Exact formulas for the normalizing constants of Wishart distributions for graphical models.” ArXiv preprint arXiv:1406.4901. 593

Xue, L. and Zou, H. (2012). "Regularized rank-based estimation of high-dimensional nonparanormal graphical models." *The Annals of Statistics*, 40(5): 2541–2571. 593