Bayesian Analysis (2009) **4**, Number 3, pp. 453–456

Comment on Article by Monni and Tadesse

Hal Stern[*]

Abstract. The article by Monni and Tadesse introduces a model for relating large numbers of predictors and responses. That situation typically occurs when investigators are in an exploratory mode. This discussion argues that in such situations the fairly strong assumptions of Monni and Tadesse (e.g., linear regression models with common coefficients for all variables within a cluster of responses) may be counterproductive. If such models are to be used, it is critical that model fit be assessed before relying on their results.

Keywords: exploratory analysis, model checking

1 Introduction

Monni and Tadesse (MT) deserve credit for the introduction of a novel model that addresses the situation in which scientific researchers want to relate a large number (p) of predictors with a large number (q) of responses. They have also developed a clever Markov chain Monte Carlo algorithm for exploring the posterior distribution of the model parameters (especially over the possible configurations of response and predictor subsets). Their simulations confirm that the model achieves the various advantages they advertise. Specifically the model uses the correlation of the responses to obtain better results than can be obtained from a sequence of q separate multivariate regressions. In addition the model allows for different regressors to be selected for different subsets of the responses which is a considerable benefit relative to trying to find a single set that describe all responses. The model is somewhat sensitive to its assumptions, especially linearity, but this is to be expected. In this discussion several concerns about the approach are raised: each of the concerns is closely related to wondering exactly what scientific question MT are trying to solve and whether their specific parametric form is likely to be helpful for scientistis.

2 What is the question being addressed?

MT remark about their setting (where the number of regressors and responses are both larger than the sample size) that "This important problem has not received the attention it deserves,...". It is surely true that the explosion of available data, especially in the biological and health sciences, has led scientists to contemplate multiple sets of high-dimensional multivariate data more often than ever before. My own work in a brain imaging collaboration has tried to integrate genetic data (100,000+ single nucleotide polymorphisms or SNPs per individual) and functional magnetic resonance imaging (fMRI) brain activity data (200,000+ voxels per individual). But in the face of these

[*] Department of Statistics, University of California, Irvine, CA, mailto:sternh@uci.edu

DOI:10.1214/09-BA416D

large data sets it is important to ask exactly what scientific problem/question we are trying to address before developing a model. The historical recipe for statistical model development involves a scientific issue motivating a mathematical or probability model. Two famous examples are Gossett's early 20th century work on the Student-t distribution motivated by the desire to understand the behavior of averages of small samples at the Guinness brewery and Legendre's early 19th century work on least squares motivated by a desire to fit a curve to empirical observations of an astronomical body's position. The modern high-dimensional data sets don't typically come with very specific scientific questions but rather with a desire to explore the relationships among variables. The goal is to determine if some of the predictor variables are related to some of the response variables. This is an important objective, but note that it is not based on any specific hypotheses. Don't get me wrong, there is definitely nothing wrong with a good exploratory analysis. Such analyses are an outstanding way to generate hypotheses that can be studied in laboratories and/or in subsequent data collections. But if "finding relationships" is the aim, then we must ask if a mixture of regression models with fairly strong assumptions (linear relationships with constant regression coefficients for all the responses in a cluster) is likely to find the relationships of interest.

If we agree that the goal in the MT setting is to identify sets of responses that are related to sets of predictors, then the competition for the MT model is not just Bayesian regression with stochastic search variable selection. The traditional multivariate method of canonical correlations (see, e.g., Anderson (1984)) is another possible analysis method for this situation. Canonical correlation analysis identifies linear combinations of the responses that are highly correlated with linear combinations of the predictors. Admittedly such methods don't work very well with small sample sizes but there are recent efforts to extend these approaches using regularization methods (see, e.g., Gonzalez et al. (2008)). Another method that comes to mind is partial least squares (Mcintosh et al. 1996; Frank and Friedman 1993) which seeks linear combinations of the responses and linear combinations of predictors with high covariances, an approach that has been found useful in brain imaging and in chemometrics, two areas where high dimensional data abound. It would be interesting to see some comparisons of these exploratory methods with the MT model.

3 Assessing model fit

The preceding section questions whether a parametric model of the type proposed by MT is appropriate for exploratory analysis. Assuming we do choose to use their model, shouldn't we be checking the assumptions? MT don't mention model checking at all in their discussion. (They do carry out a range of sensitivity analyses which are extremely valuable.) One might use their MAP estimates to define residuals and use these residuals to see if the data are consistent with linear models and Gaussian errors. The sensitivity analyses of MT in Sections 4.1.4 and 4.1.5 suggest that deviations from normality and linearity can have a big impact on the performance of their approach. It seems likely that nonnormal data (especially outliers) and nonlinearity will be the rule rather than the exception in large data sets, especially in the biological sciences. Residual analyses might suggest modifications that will improve the fit of the model. MT don't appear to have done any model checking in the example of Section 4.2. In fact, complete results are not provided for the real data example. Three clusters of responses and predictors are illustrated but how many others were there and what information do they convey?

4 Computation

MT have clearly put a great deal of effort into developing computational approaches that seek global posterior modes and efficiently sample the space of partitions of variables. Convergence of the MCMC algorithm is a concern given the size of the space on which the algorithm samples but MT are satisfied that they have an effective algorithm. The limited evidence provided suggests that their approach leads to consistent results from different starting values. MT advocate examining both posterior modes (their MAP configuration) and marginal/posterior probabilities that average over a variety of configurations. This is good advice given the form of their model. But it is also true that the goals of an analysis have an impact on how we think about computational issues. Viewing the MT situation as a problem of data exploration should mitigate some of the computational concerns. Are full posterior distributions even necessary in an exploratory analysis? It may well be sufficient to identify lots of different modes and determine what they tell us about the data.

5 Summary

The stochastic partitioning mixture of regression models proposed by MT for associating high dimensional responses and covariates is extremely clever and thought provoking. They have made a real contribution by pointing us in the direction of situations in which the goal is to relate multiple high-dimensional multivarite data sets. This discussion has argued that other (less restrictive) models may prove more useful in the exploratory settings for which MT's model will be used.

References

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*. New York, NY: John Wiley. 454

Frank, I. and Friedman, J. (1993). "A statistical view of some chemometrics regression tools." *Technometrics*, 35(2):109–148. 454

Gonzalez, I., Dejean, S., Martin, P.G.P., and Baccini, A. (2008). "CCA: An R package to extend canonical correlation analysis." *Journal of Statistical Software*, 23(12). 454

Mcintosh, A.R., Bookstein, F.L., Haxby, J.V., and Grady, C.L. (1996). "Spatial pattern analysis of functional brain images using partial least squares." *Neuroimage*, 3:143–157. 454

Acknowledgments

The author would like to thank Editor-in-Chief Brad Carlin for his comments on an earlier draft of this discussion. This work was partially supported by NSF award ATM-0530926 and NIH award RR-0021992.