Bayesian Analysis (2009) 4, Number 3, pp. 457–464

Rejoinder

Stefano Monni[1] and Mahlet G. Tadesse[2]

We thank the discussants for their remarks and insights. We will organize our replies by topics, as some questions were raised by more than one discussant.

Nature of the Correlation and Effective Sample Size

Some of the observations made by the discussants are on the nature of the correlation captured by our model. We agree with Professors Chipman, George and McCulloch (hereafter referred to as CGM) that the correlation captured by our method is not the same as that captured by the method of Brown et al. (1998) (BVF). In our model, the error terms ε_j in a component are assumed to be independent. Nevertheless, the outcomes Y_j are correlated because they have the same dependence on the predictor variables $\sum_r X_{ri}\beta_r$ (Breiman and Friedman 1997). We recognize that the totality of the correlation among outcomes may not be captured by assuming independent errors, and that ignoring a potential dependence among the error terms biases the posterior variance of the model parameters (Gelman et al. 1995). However, we believe that this bias is somehow mitigated in that we are not drawing inference on β, but simply identifying associations between X and Y variables. In addition, this assumption allows us to gain in (algorithmic) simplicity and efficacy. If we were to allow for correlation among the error terms and specify $\varepsilon_j \sim N(0, \Sigma)$ as in BVF, it would not be possible to integrate out the regression coefficients, for the prior covariance of the β could not be related to Σ (unlike BVF, where instead $B_{p\times q} \sim \mathcal{N}(B_0, H_{p\times p} \otimes \Sigma_{q\times q})$). Accordingly, updating of the regression coefficients would be required at each MCMC iteration and an appropriate reallocation scheme for these parameters would need to be defined when splitting and merging components, with a consequent complication of the algorithm. Furthermore, by taking the noise terms among the outcome variables to be independent, we are able to circumvent the high-dimensionality problem and convert the situation into one with an effective sample size equal to $N \cdot n_k$ in each component k, where N is the true sample size and n_k the number of outcomes in that component, as noted by CGM and Professor Li.

CGM noted that BVF have to estimate many more regression coefficients than we do when assessing variables in a component. This is true and it is exactly what we are avoiding by exploiting the cluster structure in the data. Outcomes are allocated to the same component because of their identical dependence on the same set of covariates, thus a single m_k-vector β is used rather than an $m_k \times n_k$ matrix of regression coefficients. On a similar note, Li wonders about the possibility of clustering response variables affected by the same predictors with different regression coefficients. In our current formulation,

[1]Department of Public Health, Weill Cornell Medical College, New York, NY, mailto:stm2013@med.cornell.edu

[2]Department of Mathematics, Georgetown University, Washington, DC, mailto:mgt26@georgetown.edu

 DOI:10.1214/09-BA416REJ

this can be solved by post-processing the MCMC output and locating components that contain identical subsets of regressors.

Comparison with other methods

Professor Stern states that the competitors for our model are not only Bayesian variable selection methods but also multivariate exploratory tools. In particular, he invites us to compare our method with canonical correlation analysis (CCA) and partial least squares regression (PLS). A direct comparison with our method is however not quite possible. In CCA, pairs of linear combinations of the original variables X and Y, called canonical variates or coordinates, are constructed that are linearly correlated, with the first pair maximizing the correlation and subsequent pairs maximizing the residual correlation with the additional requirements that each pair be uncorrelated with the previous pairs. Canonical variates are difficult to interpret, in general, and, in our case, cannot be compared with the components of our model. To be more concrete, let us apply CCA, as implemented by González et al. (2008), to the first simulated data set described in the paper (see Table 1 for the underlying model). If we represent the original variables in the space of the first two and three canonical directions in the X space (Figure 1) we can see that, based on their correlations, the Y variables are clustered so as to recover the groups of the underlying model. This is however not all that we want to do with our method, as we need to identify associated sets of X variables. While we were preparing this rejoinder, we discovered that there are some recent papers on CCA, where a sparsity condition is imposed that results in modified canonical variates that have sparse loadings. This might simplify the interpretation of the canonical variates, which however still do not correspond to our components. Nonetheless, we have applied sparse CCA as implemented in Parkhomenko et al. (2007) to our simulated data. The first canonical X-variate contains 110 regressors, far more than those that appear in our simulation. In addition, the associated first canonical Y-variate contains the Ys in S_8 and S_9, which are not correlated to any X. We are presenting these examples only to support our earlier statement that a comparison of CCA with our method is difficult. Similarly, the goal of PLS is different from ours and is closer to that of CCA. We have employed the R package pls (Mevik and Wehrens 2007) to carry out a PLS analysis, and, although it is possible to visually identify some of the clusters for the outcome variables by plotting the Y variables along the first two or three latent components (Figure 2), it is difficult to identify the associated covariates. For these reasons, we see our method as additional, rather than competing, to those already developed.

Fraley cites several papers on Gaussian mixture models and clusterwise regression. We appreciate the added bibliography, but as we were not writing a review paper we thought it best to cite works we felt were directly connected to our own, which we classify as a *variable selection* method in multivariate regression. (The italics are to emphasize the important words omitted by Fraley in her quote of our statement). For completeness, we outline the difference of our work with the four papers Fraley cites in the multivariate regression context. Breiman and Friedman (1997) proposed a method for multivariate

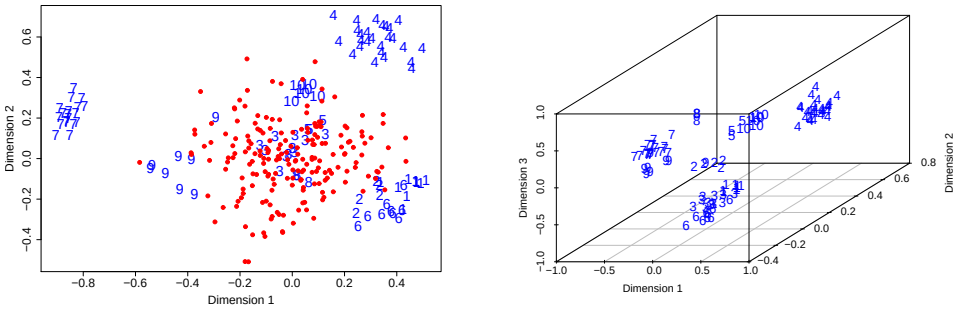

Figure 1: CCA applied to simulated data: Plot of the original variables in the space of the first two and three canonical directions in the X space. The Ys are labeled according to their true cluster membership and the Xs are shown as red dots.

prediction, which can be thought of as a shrunken canonical correlation model, and thus is very different from ours, where selections of sets of the original variables are carried out. In addition, we consider both p and q greater than N, which in 1997 seemed unusual as the authors state in their rejoinder. Turlach et al. (2005) extend the LASSO method to a multivariate response setting, but similarly to Brown et al. (1998) they consider a few outcome variables and select the same set of regressors for all response variables and, accordingly, the flexibility of selecting different sets of predictors for different outcomes is lacking. Brusco et al. (2003) consider multiple outcomes but in a setting where $N \gg p, N \gg q$ and they are concerned with clustering the observations using all available covariates, i.e., they do not perform variable selection. Gupta and Ibrahim (2007) also consider clustering observations with multiple outcomes and perform variable selection in a setting where $N \gg p$ and $N \gg q$ (in fact $q = 2$ both in their simulation and real data examples). The MCMC approach they use to fit their model is also quite different from ours.

Finally, Li states that our model is a minor modification of that of Khalili and Chen (2007). Actually, the goal, model formulation, and model fitting of our method are intrinsically different from theirs. Khalili and Chen (2007) want to identify clusters of homogeneous observations from a univariate outcome and determine predictors related to each component in a setting where $q = 1$ and, it seems from their applications and simulations, $N > p$. Their model fitting is accomplished by introducing a cluster membership indicator z_{ik} for each univariate outcome y_i, defining penalty functions for variable selection, and using an EM algorithm to maximize the penalized log-likelihood.

Inference, Model Complexity, and Overfitting

CGM question if our assessment of BVF may have been unfair by reporting MAP estimates and not measures of uncertainty. We hope not; we did in fact look at both the

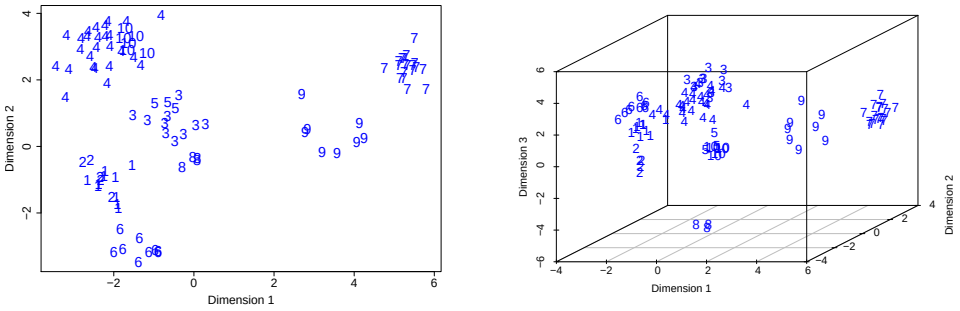

Figure 2: PLS applied to simulated data: Plot of the original Y variables in the space of the first two and three latent components. The points are labeled according to their true cluster membership.

MAP and the marginal/pairwise posterior probabilities for inference. In the simulated data, for components with a relatively large number of response variables, no regressors were selected by BVF using the MAP model or marginal posterior probabilities as low as 0.1.

Fraley makes some comments about quadratic complexity in p and q, which would affect our method. In order to make some clarity in relation to the complexity of the method (and we use the term loosely here), we offer some general observations. There is a difference between complexity of the model and complexity of the inference. The complexity to which Fraley refers (or so we think since the sentence she quotes from our paper is about inference) relates to the way in which the information from the MCMC output is summarized. Such complexity would always arise in any inferential strategy that took into account multiple models. Only if we identify a single model (or a small number of models), will we avoid resorting to pairwise posterior probabilities. When Fraley vaguely suggests possible alternative approaches related to penalized regression methods to remove the complexity of one of our inferential methods, she may have in mind approaches that identify a unique model. Of course, we could identify a single model with our method as well (the MAP configuration) which will do away with this complexity, but as we stated in our paper, we prefer to combine both inferential strategies. A useful suggestion to avoid computing marginal/pairwise posterior probabilities and still consider the most likely models is made by Stern, namely locate different modes of the posterior probability.

Fraley also raises concern about overfitting. She applied two Gaussian mixture models to the gene expression data: one with an unconstrained covariance structure and the second with a covariance matrix proportional to the identity. Based on the results she obtained from the BIC criterion, which selected a small number of components in the former case and a much larger number of components in the latter, she concluded that our model as well would behave similarly, *viz.* it would fit a very large number of com-

ponents. What Fraley does not take into consideration is that in our model regressors do play an important part in the correlation and accordingly her argument would at best apply only to $(0, n)$ components. Rather than belaboring this point, we prefer to give some experimental evidence. Let us consider the simulated data with $q = 1000$ outcome variables and $K = 35$ components presented in the paper. In this case we know the true solution. Fitting a Gaussian mixture model with an unconstrained covariance or assuming covariance $\sigma_k^2 I$ using the R package MCLUST (Fraley and Raftery 2002, 2006) gives BIC plots similar to the ones drawn by Fraley for the real data. As shown in Figure 3, the BIC peaks at three components for the unconstrained model, while the model with covariance matrix proportional to the identity adds large numbers of components and has lower BIC values. Our algorithm, however, recovered a configuration very similar to the simulated one. Even our worst result with non-tuned hyper-parameters, which we reported in the paper, identified the (m, n) components of the true model. We believe that is because the covariance is modulated by a contribution carried along by the regressors. In addition, the prior distribution of a configuration (Equation (5) of the paper) plays a role in this context, as well as in limiting the selection of irrelevant covariates into components (another potential source of overfitting).

A two-stage approach is suggested by Fraley to achieve the same goal as our method: in stage 1, a cluster analysis procedure would be used to group correlated outcomes, and in stage 2 a variable selection procedure would be applied to each component. This is similar to the comparison we have illustrated in the first simulated example, where we applied the multivariate variable selection approach of Brown et al. (1998) at the second stage. As we discussed there, even if the outcome variables could be clustered correctly, there would still remain the problem of performing variable selection in a multivariate response setting, because clusters of outcomes will have cardinality greater than one.

Li advocates the use of a penalized likelihood approach implemented with an EM algorithm and asks about the potential advantages of our method over this approach. An obvious advantage is that our MCMC sampler explores a much larger space of possible configurations compared to the suggested EM algorithm and is less prone to being trapped in a local mode. Another advantage is that our method provides a posterior model probability for all visited configurations. This allows us to perform model averaging for inference, as well as identify several models that have high posterior probabilities. The procedure suggested by Li, however, presumes the existence of one best model, which is not realistic when exploring high-dimensional data sets. This said, we recognize that it would still be interesting to perform a thorough comparison of the two algorithms.

Analysis of array CGH and gene expression data

Li asked if it would be preferable to estimate the copy numbers, and then link them to the gene expression data. We opted against this approach since it would require choosing among the many available methods for copy number detection and this estimation would introduce another level of uncertainty/error in the data. Li and Fraley asked if the associations between array CGH and transcript profiles identified by our algorithm

could also be identified by univariate analysis. Some are. But the univariate analysis tends to select many CGH clones as being associated to each gene expression, and as our simulations showed us, we think many are false positives.

Fraley questioned the independence of the samples in the data sets. It is important to note that when thousands of variables are considered and very few of them are significantly different between samples, the correlation across samples using all variables will appear very high. Or, put it differently, a substantial correlation of the genes can induce the appearance of a correlation among the samples (Efron 2008).

Stern remarked that we showed limited results. Because of space constraints and since the focus of the paper (and the journal) is on the statistical method, we just selected a few associations to illustrate some of the results. Several of the associations, for example those presented in the paper, made biological or technical sense (*e.g.* probes representing the same gene being assigned to the same component) and we hope that others may be good candidates for further investigation. Naturally there will be some that will have no biological meaning.

We agree with Stern that it would be valuable to check the assumptions prior to fitting the model, but performing an extensive check for all variables and trying to identify the optimal transformation in this high-dimensional setting is practically infeasible. Here, we have relied on the fact that gene expression levels on the log-scale often follow a fairly symmetric distribution and are commonly assumed to be normally distributed. The assumptions of normality and linear association are indeed strong and may not be satisfied in most applications. We are now investigating non-parametric models that would weaken these assumptions.

To conclude this rejoinder, we thank the discussants for their contributions, and we hope more methods will be developed to deal with variable selection in multivariate regression and more generally to address the problem of combining different high-dimensional data sets.

References

Breiman, L. and Friedman, J. (1997). "Predicting multivariate responses in multiple linear regression (with discussion)." *Journal of the Royal Statistical Society, Ser. B*, 59: 3–54. 457, 458

Brown, P., Vannucci, M., and Fearn, T. (1998). "Multivariate Bayesian variable selection and prediction." *Journal of the Royal Statistical Society, Ser. B*, 60: 627–641. 457, 459, 461

Brusco, M., Cradit, J., and Tashchian, A. (2003). "Multicriterion clusterwise regression for joint segmentation settings: An application to customer value." *Journal of Marketing Research*, 40: 225–234. 459

Efron, B. (2008). "Are a set of microarrays independent of each other?" Technical Report 244, Stanford University, Department of Statistics. 462

Fraley, C. and Raftery, A. E. (2002). "Model-based clustering, discriminant analysis and density estimation." *Journal of the American Statistical Association*, 97: 611–631. 461

— (2006). "MCLUST version 3 for R: Normal mixture modeling and model-based clustering." Technical Report 504, University of Washington, Department of Statistics. 461

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (1995). *Bayesian Data Analysis*. London, UK: Chapman & Hall. 457

González, I., Déjean, S., Martin, P., and Baccini, A. (2008). "CCA: An R package to extend canonical correlation analysis." *Journal of Statistical Software*, 23(2). 458

Gupta, M. and Ibrahim, J. (2007). "Variable selection in regression mixture modeling for the discovery of gene regulatory networks." *Journal of the American Statistical Association*, 102: 867–880. 459

Khalili, A. and Chen, J. (2007). "Variable selection in finite mixture of regression models." *Journal of the American Statistical Association*, 102: 1025–1037. 459

Mevik, B.-H. and Wehrens, R. (2007). "The pls package: Principal component and partial least squares regression in R." *Journal of Statistical Software*, 18(2). 458

Parkhomenko, E., Trichler, D., and Beyene, J. (2007). "Genome-wide sparse canonical correlation of gene expression with genotypes." *BMC Proceedings*, 1(Suppl 1:): S119. 458

Turlach, B., Venables, W., and Wright, S. (2005). "Simultaneous variable selection." *Technometrics*, 47: 349–363. 459

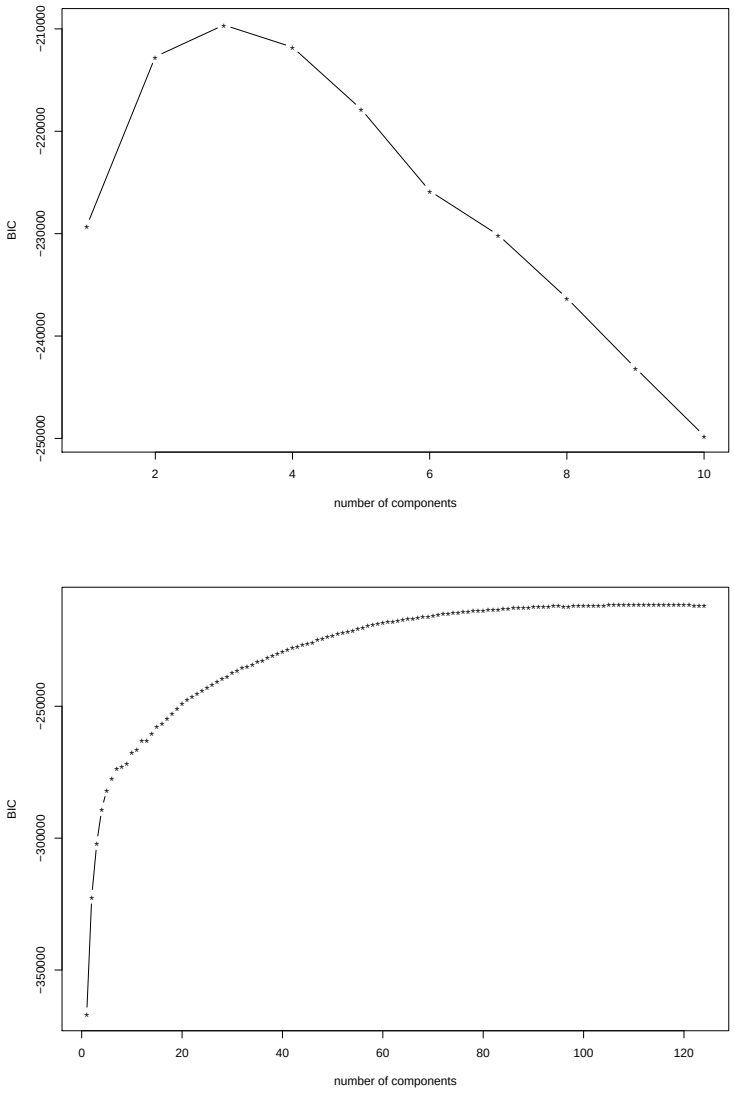

Figure 3: BIC for multivariate Gaussian mixture model fits to the simulated data with $q = 1000$ outcomes and $K = 35$ components included in paper. TOP: BIC for the model with unconstrained covariance. BOTTOM: BIC for the model with covariance taken to be a multiple of the identity (spherical).