Bayesian Analysis (2009) **4**, Number 3, pp. 439–448

Comment on Article by Monni and Tadesse

Chris Fraley*

The problem of variable selection in a multiresponse setting is a difficult one that deserves attention. With regard to this paper, I see two main areas for discussion: proof of concept in terms of the scope of the examples, and the establishment of context in terms of the relationship to and departure from existing work.

1 Simulated and Real Data Examples

The simulated and 'real data' examples leave a number of fundamental questions unanswered.

Simulated Data Examples. The design of the simulated examples raises two immediate issues:

- In the simulated examples, the outcomes were generated with equal variance throughout — formula (6) on page 10 — whereas in the model underlying the method described on page 3, the variance is allowed to vary among components. How well would the method work in simulations where the variance is allowed to vary?
- Were the outcomes and/or covariates scaled, standardized or normalized in any way? This could affect the resulting analysis and conclusions.

The subsection entitled **Performance in the presence of high collinearity** on page 15 starts with the sentence *"It is reasonable to assume that the presence of highly correlated covariates may complicate the identification of relevant predictors."* Two new covariates, which are linear combinations of existing covariates, are then added to simulated data which had $N = 50$ samples and $p = 200$ covariates, with the conclusion that *"... the algorithm is reasonably resistant to colinearity between predictors."*
It's not clear what's intended here, since the dimensions of the data imply multicolinearlity, and the method is supposed to address cases in which the number of regressors is much larger than the sample size (e.g. page 2).
Questions of interest with regard to multicolinearity would include:

- How would permutations of the covariates and/or outcomes affect the results in the simulated and 'real data' examples?
- Is identifiability an issue?
  Because of the inherent multicolinearity, more than one set of predictors could fit a

*Department of Statistics, University of Washington and Insilicos LLC, Seattle WA, http://www.stat.washington.edu/fraley/

 DOI:10.1214/09-BA416B

given set of responses equally well. The prior weight assigned to each configuration — formula (5) on page 6 — may help avoid this, but not when the sets of predictors and responses in question are equal in size.

The subsection entitled **Performance under deviations from normality**, also on page 15, raises further issues. According to the model formulation, each individual element of an outcome is drawn from a univariate normal distribution (bottom of page 3), with the accompanying description: *"The model we consider is a multivariate Gaussian mixture model . . . "*. These outcomes can be viewed as coming from a multivariate Gaussian mixture, but only under the restriction that each component has a covariance of the form $\sigma_k^2 I$ (no cross-covariances). The 'departure from normality' referred to here is a departure from univariate normality for the individual elements of the outcomes. The t-distributed outcomes generated will form components with fatter tails, while retaining spherical symmetry.

- There are numerous statements to the effect that the proposed method is designed to account for correlation among outcomes (e.g. in the the prior specification on page 6 or the beginning of the concluding section on pages 18 and 21).
  If the outcomes were drawn from a multivariate Gaussian mixture with component covariances Σ_k, where there is nonnegligible cross-covariance, instead of $\sigma_k^2 I$, how well would the method work? Can it be shown to account for the correlation among outcomes in such instances?
- How well would this method work if the mixture components of the simulated data overlap (intersecting clusters), with restricted and/or unrestricted covariance?

Real Data Example. The section on **Real data: CGH and gene expression profiles** summarizes application of the method to publicly available data.

- Were the downloaded outcomes and/or covariates further processed in any way before analysis? This could influence the results and conclusions.

On the download page `http://discover.nci.nih.gov/cellminer/loadDownload.do`, the normalized `Affymetrix HG-U133A` RMA data used by the authors, with a total of $q = 21,225$ possible probes, have since been replaced by the normalized `RNA: Affy U133(A-B)` RMA data, with $q = 35,086$ probes.[1]

To quote from page 9, *"We considered $q = 3291$ probe sets that showed variability across tissue types . . . "*

- How were the $3,291$ probes selected out of the $21,225$ possibilities? This could also influence the results and conclusions.

It might be assumed that the proposed method could model the data with all $21,225$ of the responses, given the premises: *"In this paper we are interested in cases where*

[1] author's communication

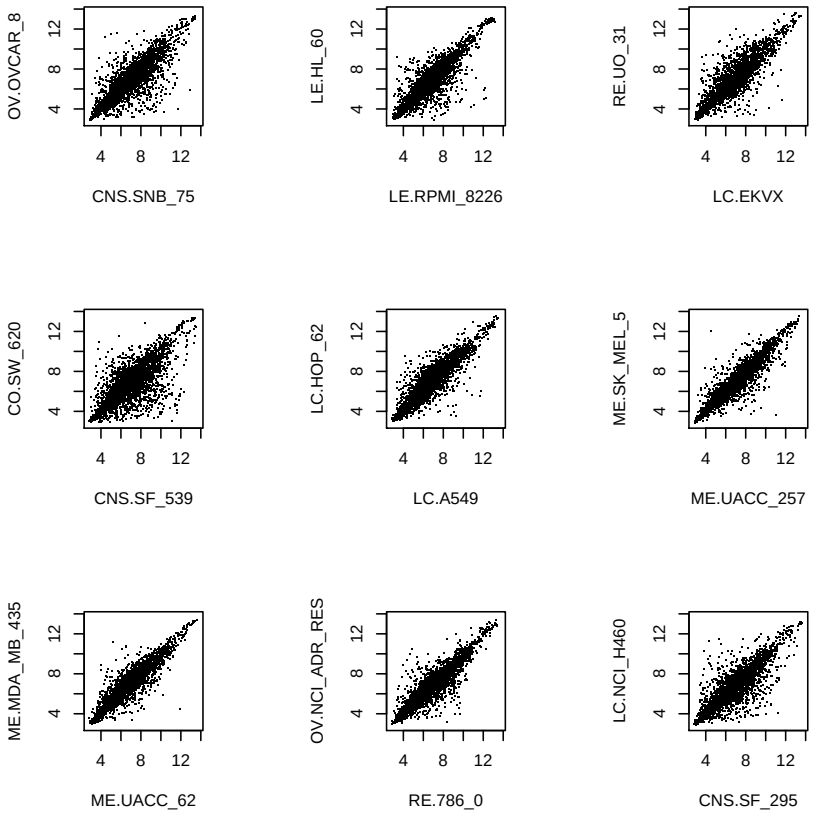

Figure 1: Pairs plots for the `RNA: Affy U133A` RMA probes used by the authors, showing significant correlation among the outcomes.

not only is the number of regressors much larger than the sample size but so too is the number of responses." (introduction, page 2). However, quadratic complexity limits the size of p and q in practice: *"... we have also considered, the $p \times q$ matrix of posterior probabilities of association between a covariate X_i and an outcome Y_j, the $p \times p$ matrix of posterior probabilities that any pair (X_i, X_j) is assigned to the same component (m, n) and the two $q \times q$ matrices of posterior probabilities that any pair Y_i, Y_j is allocated to the same $(0, n)$ or (m, n) component."* (applications section, page 9).

- Given that preselection will be necessary in the proposed method for datasets in which p and/or q are quite large (a situation that often arises in bioinformatics), are there effective strategies or is this choice likely to be highly data dependent?

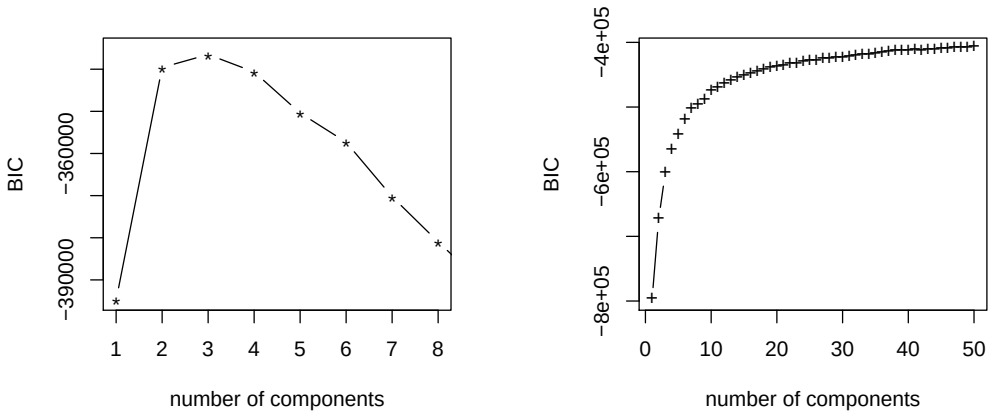

Figure 2: BIC for multivariate Gaussian mixture model fits to the 3291 probe sets used by the authors. LEFT: BIC for the model with unconstrained (ellipsoidal) component covariances. RIGHT: BIC for the model with component covariances constrained to be a multiple of the identity (spherical). Note the difference in scale in both the horizontal and vertical axes of the two plots. The sign is chosen so that higher BIC values are meant to indicate better models. The BIC peaks at three components for the unconstrained model, while BIC values for the constrained model are much lower for the constrained model, which compensates for lack of fit by adding large numbers of components.

Pairs plots for the 3291 Affymetrix probes used by the authors, with coordinates randomly sampled from among the 59 cell lines are displayed in Figure 1, showing considerable correlation among outcomes.

- Do the mixture components produced by the proposed method reflect this correlation structure?

The BIC (Bayesian Information Criterion — Schwarz 1978) is an approximate Bayes factor that is often used for model selection in mixture modeling (e.g. McLachlan and Peel 2000). Figure 2 shows the BIC obtained from fitting Gaussian mixture models to the 3291 outcomes used by the authors.

Two possible mixture models are considered, one with the covariance of each component constrained to be of the form $\sigma_k^2 I$, and the other without this restriction. The sign is chosen so that higher BIC values are meant to indicate better models.

The BIC peaks at three components for the unconstrained model, while BIC values are much lower for the constrained model, which tries to compensate for lack of fit by adding large numbers of components. The mean/median uncertainty of the 3 component unconstrained model fit is `4.62e-3/2.56e-8`, indicating well-separated clusters (e.g.

Fraley and Raftery 2002). This suggests treating the groups of outcomes separately as a plausible approach. The behavior of the constrained model raises concerns about overfitting in the proposed method, as does the statement on page 6 discussing the penalty term on the size of components included in the prior: *"We try to favor smaller components because larger ones tend to fit the noise."*

To quote page 17: *"Some components captured known associations and grouped probe sets of genes which are involved in similar biological processes."* It's interesting that some of the components reflect known biological relationships, but it's not really possible to assess usefulness without analyzing the signal-to-noise ratio.

- What proportion of the output components can be considered noise? Are there any that don't seem to make biological sense or any well-known relationships that are missed or fragmented? If so can these anomalies be explained?
- Would the associations mentioned as being noteworthy also be obtained if each component was regressed separately?

2 Relationship to Existing Work

There appear to be quite a few missing links to the literature.

- From **Section 2.2 Prior specification** (page 5) through **Section 3 Model specification and posterior inference** (page 9), there is one citation (Geyer 1991 in **Section 3.2 A tempering extension**), although the conclusion on page 21 states: *"We have, however, preferred to use standard priors . . . "*. Who's standard?
- Use of the U133 RNA data currently requires citation of Shankavaram et al. (2007) — see `http://discover.nci.nih.gov/cellminer/citing.do`.
- The discussion of the 'real data' example also lacks citations, which could be helpful in evaluating statements like: *"These genes are involved in hematopoietic development and lymphocytic proliferation, known to be implicated in a subset of human T-cell leukemia."* (page 18).
- There are also ties to the Bayesian variable selection, clusterwise regression, and model-based clustering literatures.

  The method draws heavily from Bayesian variable selection, for which a couple of older references are cited (George and McCulloch 1997, Brown et al. 1998), although there is an extensive recent literature (surveyed, for example, in O'Hara and Sillanpää 2009).

  There is also a fairly extensive literature on clusterwise regression, which was introduced by Späth (1979) and later formulated as a mixture model by DeSarbo and Cron (1988). Bayesian approaches are proposed in the Jiang and

Tanner (1999a) paper cited by the authors, as well as in Lenk and DeSarbo (2000). A recent paper by Kahlili and Chen (2007) addresses variable selection in these models. Situations involving multiple dependent variables are addressed in Wedel and Steenkamp (1991), Brusco et al. (2003) and Gupta and Ibrahim (2007). There are known identifiability issues, discussed for example in Jiang and Tanner (1999b) and Hennig (2000).

The authors refer a number of times to *"standard model-based clustering"* (e.g. pages 4 and 15). This seems intended to mean clustering based on Gaussian mixture models, for which references include McLachlan and Basford (1988), McLachlan and Peel (2000), Fraley and Raftery (2002) and — in the context of genomics — Yeung et al. (2001). However, the proposed model constrains the covariance of each component to be a multiple of the identity (as discussed above), which would not usually be considered standard for multivariate data in mixture modeling. Moreover, many clustering methods are based on mathematical and statistical models, so that the term 'model-based clustering' does not necessarily imply mixture modeling (see e.g. Banerjee and Rosenfeld 1993).

- Of the *"few existing methods for multivariate regression"* (page 21), one alternative (Brown et al. 1998) is mentioned and considered by the authors. There are a number of other relevant references, including Breiman and Friedman (1997), Brusco et al. (2003), Turlach et al. (2005), and Gupta and Ibrahim (2007). The latter also proposes a simulation-based variable selection method that uses regression mixture modeling, with application to genomics data.

3 Concluding Remarks

In conclusion, I don't think it's clear that this method can capture situations in which there is significant crosscorrelation among components. Approaches along the modeling lines described in this paper can perhaps offer an improvement over fitting each response separately, but overfitting remains a critical concern. The authors' rationale against clustering the outcomes in advance and then fitting each cluster separately is: *"... this two stage procedure would treat the clusters as known, and would ignore uncertainty in estimating the cluster memberships when searching for relevant predictors, and thus inevitably introduce some bias into the analysis (Bryant and Williamson 1978)."* (page 1–2). When clustering based on mixture models, an assessment of uncertainty is available in terms of the conditional probabilty of each sample being in each component (e.g. Fraley and Raftery 2002), which could be exploited in a Bayesian setting.

For outcomes effectively modeled by a mixture of well-separated components, it seems unlikely that the proposed method would have an advantage over the two-stage alternative. Moreover, approaches related to elastic net (Zou and Hastie 2005) and lasso (Tibshirani 1996) can avoid quadratic complexity in p and q (e.g. bottom of page 9) — not to mention simulation overhead — by either using an active set approach (Turlach et al. 2005 — see also Osborne et al. 2000 and Efron et al. 2004) or coordinate descent

(e.g. Friedman et al. 2009). New methods may need to be developed to handle cases in which clusters of responses overlap.

References

Banerjee, S. and Rosenfeld, A. (1993). "Model-based cluster analysis." *Pattern Recognition*, 26: 963–974.

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

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

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

Bryant, P. and Williamson, A. J. (1978). "Asymptotic behavior of classification maximum likelihood estimates." *Biometrika*, 65: 273–278.

Bussey, K. J., Chin, K., Lababidi, S., Reimers, M., Rheinhold, W. C., Luo, W., Gwadry, F., Ajay, Kouros-Mehr, H., Fridlyand, J., Jain, A., Collins, C., Nishizuka, S., Tonon, G., Roschke, A., Gehlhaus, K., Kirsch, I., Scudiero, D. A., Gray, J. W., and Weinstein, J. N. (2006). "Integrating data on DNA copy number with gene expression levels and drug sensitivities in the NCI-60 cell line panel." *Molecular Cancer Therapeutics*, 6(4): 853–867.

DeSarbo, W. S. and Cron, W. L. (1988). "A maximum likelihood methodology for clusterwise linear regression." *Journal of Classification*, 5: 249–282.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). "Least angle regression (with discussion)." *Annals of Statistics*, 32: 407–499.

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

Friedman, J., Hastie, T., and Tibshirani, R. (2008, revised in 2009). "Regularization paths for generalized linear models via coordinate descent." manuscript, Department of Statistics, Stanford University.

George, E. and McCulloch, R. (1997). "Approaches for Bayesian variable selection." *Statistica Sinica*, 7: 339–373.

Geyer, C. (1991). "Markov chain Monte Carlo maximum likelihood." In Keramigas, E. (ed.), *Computing Science and Statistics*, 156–163. Interface Foundation.

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

Hennig, C. (2000). "Identifiability of Models for Clusterwise Linear Regression." *Journal of Classification*, 17: 273–296.

Jiang, W. and Tanner, M. A. (1999a). "Hierarchical mixtures-of-experts for exponential family regression models: approximation and likelihood estimation." *Annals of Statistics*, 27: 987–1011.

— (1999b). "On the identifiability of mixtures-of-experts." *Neural Networks*, 12: 1253–1258.

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

Lenk, P. J. and DeSarbo, W. S. (2000). "Bayesian inference for finite mixtures of generalized linear models with random effects." *Psychometrika*, 65: 93–119.

McLachlan, G. J. and Basford, K. E. (1988). *Mixture Models : Inference and Applications to Clustering*. Marcel Dekker.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. Wiley.

O'Hara, R. B. and Sillanpää (2009). "A review of Bayesian variable selection: what, how, and which?" *Bayesian Analysis*, 4: 85–118.

Osborne, M. R., Presnell, B., and Turlach, B. A. (2000). "On the LASSO and its dual." *Journal of Computational and Graphical Statistics*, 9: 319–337.

Schwarz, G. (1978). "Estimating the dimension of a model." *Annals of Statistics*, 6: 461–464.

Shankavaram, U. T., Rheinhold, W. C., Nishizuka, S., Major, S., Morita, D., Chary, K. K., Reimers, M. A., Sherf, U., A, K., Dolginow, D., Coosman, J., Kaldjian, E. P., Scudiero, D. A., Petricoin, E., Liotta, L., Lee, J. K., and Weinstein, J. N. (2007). "Transcript and protein expression profiles of the NCI-60 cancer cell panel: an integromic microarray study." *Molecular Cancer Therapeutics*, 6(3): 820–832.

Späth, H. (1979). "Clusterwise linear regression." *Computing*, 22: 367–373.

Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society, Series B*, 58: 267–288.

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

Wedel, M. and Steenkamp, J. (1991). "A clusterwise regression method for simultaneous fuzzy market structuring and benefit segmentation." *Journal of Marketing Research*, 28: 384–396.

Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E., and Ruzzo, W. L. (2001). "Model-based clustering and data transformation for gene expression data." *Bioinformatics*, 17: 977–987.

Zou, H. and Hastie, T. (2005). "Regularization and variable selection via the elastic net." *Journal of the Royal Statistical Society, Series B*, 67: 301–320.

Acknowledgments

We thank Dr. Mahlet Tadesse for providing the Affymetrix data used for the 'real data' example in the original paper, as well as for answering a number of questions about the Cellminer data.