BAYESIAN STATISTICS 9,
J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid,
D. Heckerman, A. F. M. Smith and M. West (Eds.)

Rejoinder to Bayesian models for sparse regression analysis of high dimensional data

Sylvia Richardson, Leonardo Bottolo & Jeffrey S. Rosenthal
Imperial College London, UK *University of Toronto, Canada*
sylvia.richardson@imperial.ac.uk l.bottolo@imperial.ac.uk jeff@math.toronto.edu

We thank the discussants for their stimulating comments and interesting comparison with alternative models and algorithms. We will briefly consider their points in turn.

Extension of the hierarchically related sparse regression model

We agree that in our set-up, dependence between the responses Y beyond that induced by the hierarchical structure of Ω is not accounted for. The framework of Seemingly Unrelated Regressions (SUR) is indeed more general and applicable for moderate q, but we suspect that it would quickly become unfeasible for the large size of q expected in eQTL types of experiments. In the paper of Banerjee *et al.* (2008), even though the implementation of SUR method for QTL of multiple traits is discussed in general terms, the simulation study only involves $q = 2$. Our own experience of sparse regression with a multiple Gaussian response model also confirms our observation that modelling the covariance between the responses is computationally demanding and can be unstable. In Petretto *et al.* (2010), the ESS algorithm is applied to a multiple response model (gene expression in four tissues) under the restrictive assumption that the selection indicators are the same for all the responses. In the new associations that are illustrated, the correlation between the responses is entirely explained by a pair of markers.

In this work, we were primarily focussed on considering a large number of responses, and we did not pursue the SUR direction. On the other hand, there are different extensions of our set-up that can account for correlation between responses. In the spirit of the recent work on multivariate Sparse Partial Least Squares (SPLS) of Chun and Keleş (2009), a preclustering of the responses could be performed and the likelihood in equation (1) extended to include a random effect common to all the responses in each cluster. This would allow most of the ESS computations and efficiency of parameter integration to go through. A drawback of this approach, shared by the sparse SPLS method, is to rely on a preprocessing clustering step to capture adequately the residual correlation between the responses. Alternatively,

the response groups could be defined through the use of external information, for example pathway information, which could be introduced in the specification of the covariance between the responses. This may link our model to the paper presented in this volume by Stigo and Vannucci (2010). To enhance interpretability of the results, external information can also be included in the prior model of the selection probabilities. As queried by the discussants, it would be straightforward to modify the specification of model (iii) to include the use of external information from biological predictors, $Z_j = (z_{js}, 1 \leq s \leq S)$, for example by generalising model (iii) to $\omega_{kj} = \omega_k \times \rho_j \times \exp(\underline{\varphi}^T Z_j), 0 \leq \omega_{kj} \leq 1$.

Finally, we want to stress that in our paper, the dependency induced by the hierarchical structure of model (ii) or (iii) is helpful for recovery of the true associations as well as for uncovering the dependence structures of the responses that are linked to the same predictors. To partially answer the discussants' comments on this point we compared the outputs of model (i) where there is no hierarchical structure and model (iii) for the **Sim5** set-up. Figure 1 shows typical comparative results, displayed here for one replicate on **Sim5**. The correlation between the Y_k induced by $\underline{\beta}_{-\gamma_k}$, the non-zero elements of $\underline{\beta}_{-k}$, is displayed on the left hand side. Figure 1 shows that the hierarchical column structure of model (iii) encapsulated by ρ_j leads to a clearer recovery of the 10 blocks of responses that were simulated. Figure 2 highlights nicely that in the case of model (iii), within each block, the marginal posterior probabilities of inclusion are highly correlated and homogeneous, whereas there is considerably more heterogeneity within blocks for model (i). Hence in a partial answer to the discussant query, we see that the hierarchical structure of model (iii) is able to capture well the dependence between the Y_k induced by $\underline{\beta}_{-\gamma_k}$.

Alternative priors

We agree that the choice of priors for the regression coefficients has important consequences on the variable selection performance. The generalised g-prior of Gupta and Ibrahim (2007, 2009) has two fixed hyperparameters, one similar to our shrinkage coefficient g and an additional one λ, in the line of ridge regression. How to fix these hyper-parameters and sensitivity to this choice (choice which is not explicited here) is a delicate issue that led us to put a prior on g instead. We note that in Gupta and Ibrahim (2009), comparison between this prior and g-prior is focussed on predictive performance where, indeed, ridge penalisation would be expected to help. Our focus here is on selection of a small number of important predictors and it is not clear to us what benefits the generalised g-prior would have in this context.

The Laplace prior within a Bayesian shrinkage perspective as proposed by Bae and Mallick (2004) and used in the discussion is an interesting computationally efficient alternative to variable selection. It would be useful to be able to compare, besides the R^2, its ability to recover the true associated variables and associated errors. Recent work on Bayesian sparse signal models by Carvalho *et al.* (2010) propose a different prior, the horseshoe prior, based on Cauchy tails rather than exponential tails for the variances, which is shown to have good theoretical properties and the ability to adapt to different sparsity patterns. Investigating and comparing variable selection with such approaches is an interesting avenue for future research and we thank the discussants for pointing us in that direction.

Alternative algorithms

We welcome the connection made between the adaptive scan and a stochastic approximation of the selection probabilities. We agree that the choice of the decreasing sequence will be important and that more work needs to be done in investigating suitable schemes and their comparative performance. Stochastic approximation is also behind the alternative algorithm to ESS proposed by the discussants. We are intrigued to know more about the implementation of SAMC in this challenging high-dimensional model selection case. How were the partitions defined? What specific MH moves were used to update with the target distribution $f_w(x)$? What importance has the parameter Δ? The model selection examples analysed in the referenced papers are of much smaller size and we are intrigued by the performance of the mSAMC sampler on the simulated data sets. The R^2 reported are extremely high, particularly for Example 2 which has the feature of a contaminated model. With the data simulated in Bottolo and Richardson (2010), we do not reach R^2 higher than 0.81 (ranging between 0.70 and 0.81 in the simulated replicates) when inputting the true variables in the linear model. Hence, we suspect that the data simulated by Mallick *et al.* was not comparable to that used in ESS. Moreover, the distribution of model size reported in Table 1 of the discussion indicates that the models found by SAMC have typically a much smaller size than that of the true model, seemingly contradicting the high R^2 reported in Table 2 for both examples.

We agree with the discussants that the question of how to compare different algorithms is complex and deserves careful consideration. In our article, we were attentive to design a number of scenarios with the dual purpose of (i) evaluating the performance in a range of situations as well as (ii) providing fair ground for comparing to other approaches by including scenarios that were tailored to other approaches. For lack of space, we only reported and compared the algorithms on a limited number of features, emphasizing mostly the hot spot detection performance as this was the main focus of our hierarchical related sparse regression structure, keeping a more comprehensive comparison for follow-up work. In Bottolo and Richardson (2010), we compared ESS with Shotgun Stochastic Search (Hans *et al.*, 2007) with respect to marginal posterior probability of inclusion for the predictors, R^2 of best model visited, average R^2 for the 1,000 top (non-unique) models ranked by their posterior probability and computation time. In a more comprehensive comparison, it would be interesting to report additionally, for example, several distance measures between the simulated and estimated βs, median model size and median 0–1 "test" error (i.e. based on the number of variables which differ between the true ones), following the comparison strategy used by Fan *et al.* (2009) to investigate several methods for feature selection in ultra high dimension. Trying to characterise the complexity of the algorithms to be compared is also an important consideration, and more work needs to be done along this line.

REFERENCES

Bae, K and Mallick, B.K.(1994). Gene selection using a two-level hierarchical Bayesian model.*Bioinformatics* **20**, 3423–3430.

Bottolo, L and Richardson S. (2010). Evolutionary Stochastic Search for Bayesian model exploration. *Bayesian Analysis* **5(3)**, 583–618.

Banerjee, S., Yandell, B. S. and Yi, N. (2008). Bayesian Quantitative Trait Loci mapping for multiple traits. *Genetics* **179**, 2275–2289.

Carvalho, C. M., Polson, N. G. and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* , Advance access.

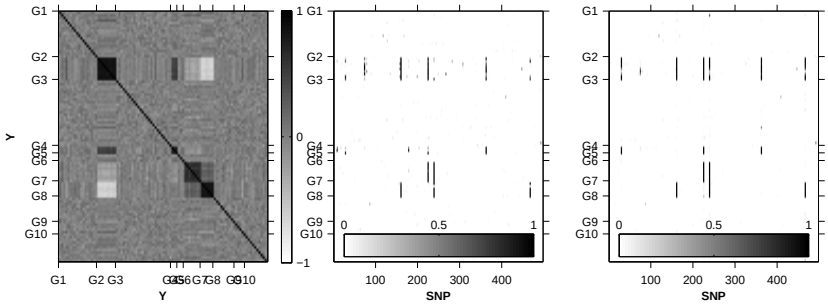

Figure 1: *Pairwise correlation of Y for one replicate of* **Sim5** *(left), marginal posterior probability of inclusion for the independence model (i)(middle) and marginal posterior probability of inclusion for the multiplicative model (iii) (right). The 10 blocks of responses induced by the structure of the simulated βs are indicated on the left G_1 to G_{10}.*

Chun, H. and Keleş, S. (2009). Expression Quantitative Trait Loci mapping with multivariate Sparse Partial Least Square regression. *Genetics* **182**, 79–90.

Fan J., Samworth, R. and Wu, Y. (2009). Ultrahigh dimensional feature selection: Beyond the linear model. *Journal of Machine Learning Research* **10**, 2013–2038.

Gupta, M. and Ibrahim, J.G. (2007). Variable selection in regression mixture modeling for the discovery of gene regulatory networks. *J. Amer. Statist. Assoc.* **102**, 867-880.

Gupta, M. and Ibrahim, J.G. (2009). An information matrix prior for Bayesian analysis in generalised linear models with high-dimensional data, *Statistica Sinica* **19**, 1641–1663.

Hans, C., Dobra, A. and West, M. (2007). Shotgun Stochastic Search for "large p" regression. *J. Amer. Statist. Assoc.* **102**, 507–516.

Petretto, E., Bottolo, L., Langley, S.R., Heinig., M., McDermott-Roe, M.C, Sarwar, R., Pravenec, M., Hübner, N., Aitman, T.J., Cook, S.A. and Richardson, R. (2010). New insights into the genetic control of gene expression using a Bayesian multi-tissue approach. *PLoS Comput. Biol.* **6(4)**, e1000737.

Stingo, F. and Vannucci, M. (2010). Bayesian Models for variable selection that incorporate biological information. In *Bayesian Statistics, Proc. 9th Int. Meeting* (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.), Oxford University Press.

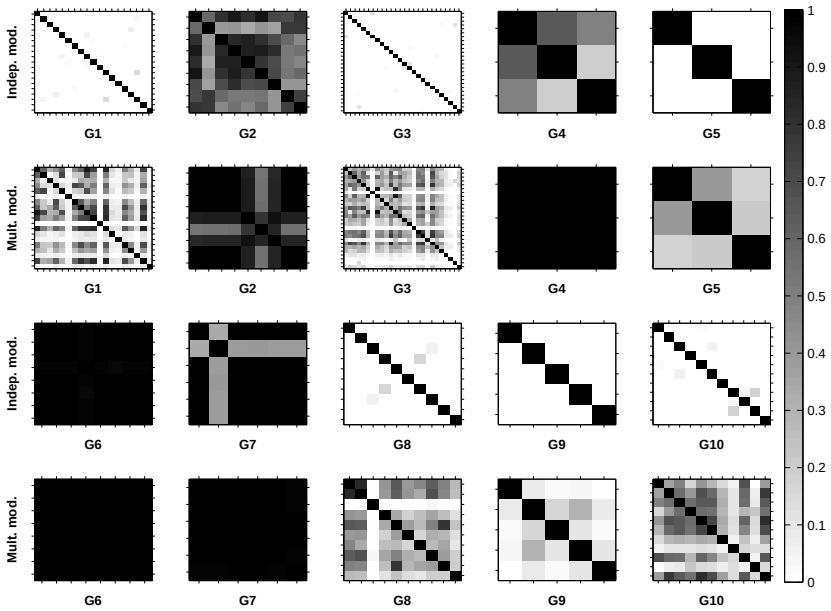

Figure 2: *Empirical correlation between the marginal posterior probabilities of inclusion for the 10 blocks of responses in* **Sim5**. *Comparison of output of the independence model (i) and the multiplicative model (iii).*