Bayesian Analysis (2011) **6**, Number 1, pp. 43–48

Rejoinder

Nicholas G. Polson[*] and Steven L. Scott[†]

We thank all the discussants for their insights and comments on the article. Due to the subject matter specialization, *Bayesian Analysis* has a more homogeneous readership than journals that cater to a more general audience, so it is not surprising to find substantial agreement among the discussants and ourselves. Of course, readers may be disappointed by the lack of blood-sport normally associated with discussion articles. We apologize for this, and promise to write a more provocative article in the future.

1 Mallick et al.

Mallick *et al.* rightly point out that our focus on posterior inference for model parameters is only indirectly related to the classification performance that typically interests SVM users. The simulations provided by Mallick *et al.* are a welcome correction to our omission. The simulations show that the SVM criterion can in fact reduce the misclassification error compared to probit regression. Many Bayesians (including us) approach support vector machines with a wary suspicion that they are simply logistic regression's poor, non-probabilistic cousin. Simulations like this are useful data exercises that should force us to update that viewpoint. We have replicated the simulations in Table 1 with logistic regression in place of probit. The logistic regression and the SVM were both run, using spike-and-slab priors, on the spam data set from Section 5. We used the algorithm from Tüchler (2008) for the logit model.

Prediction is a common theme among the discussants. Lindley (1968) provides the theoretical analysis of prediction-based Bayesian variable selection in the presence of costs, as well as a beautiful discussion of the faults of commonly used classical procedures. The upshot is that, to select variables for a model that predicts best (in an MSE sense), one needs to find the linear combination that best fills in for the linear combination of variables that you leave out. Brown et al. (1998, 1999, 2002) illustrate the advantages of this framework in large scale predictive regression systems. This approach trades-off the cost of variable inclusion with the gain in MSE predictive power. We are not presently in a position to provide the equivalent predictive analysis for SVM's but Hans' proposal of basing prediction on the posterior mean via the linear combination $E(\beta|y)'x_f$ for a future covariate x_f seems sensible. Implementing the Lindley analysis requires some posterior standard errors, which we can directly obtain from our MCMC algorithm.

Another interesting direction for future research is showing the interplay between sparse estimators, variable selection, and prediction in the original Mallows (1973) C_p paper. That paper also contains a very useful discussion of the C_L criteria, corresponding to a linear Bayes ridge rule. Again our representation makes such a discussion

[*] Booth School of Business, Chicago, IL, mailto:ngp@chicagobooth.edu

[†] Google Corporation, mailto:stevescott@google.com

DOI:10.1214/11-BA601REJ

Logit	SVM	Difference	Extra Correct
0.891	0.909	0.017	8
0.909	0.911	0.002	1
0.907	0.915	0.009	4
0.893	0.917	0.024	11
0.915	0.924	0.009	4
0.900	0.913	0.013	6
0.917	0.911	-0.007	-3
0.904	0.917	0.013	6
0.885	0.902	0.017	8
0.913	0.924	0.011	5

Table 1: *Correct classification rates for SVM and logistic regression under spike and slab priors. The last column gives the number of additional successful classifications from the SVM. Each row describes out-of-sample results for a different random 10% cross validation holdout sample.*

applicable to SVM's. Mallows' analysis also illustrates how hyper-parameter selection affects variable inclusion.

Mallick *et al.* suggest three potential generalizations of our method, including multi-category classification, basis expansions of the predictors (the "kernel trick"), and the normalized pseudo-likelihood. The first two are straightforward. The usual trick for handling multiple classes is to produce an ensemble of binary classifiers. We see no obstacle to applying our methods to each member of such an ensemble. It is possible that an alternative data augmentation could more elegantly handle the multi-class problem in a manner akin to what Scott (2011), and Frühwirth-Schnatter and Frühwirth (2007, 2010) have proposed for multinomial logistic regression. Likewise, nothing in our data augmentation strategy assumes linearity among the predictors, so there are no obstacles to using our methods with the "kernel trick," whether that means using actual kernels, trees, splines, neural networks, or other nonparametric regression methods.

We don't have much to say regarding the normalized version of the SVM pseudo-likelihood. We share the concern about the SVM criterion being non-probabilistic, but we thought it appropriate to study SVM's as they are actually (and widely) used. Mallick et al. (2005) suggested using a normalized SVM and provide details of Bayesian inference. We haven't worked out the details in our framework, but we agree that it would be interesting direction for future work. As Shahbaba *et al.* point out, the unnormalized version is a natural byproduct of the separating hyperplane construction that is foundational to the SVM procedure. We look forward to the forthcoming work by Mallick *et al.* on the subject, and hope that it maintains or improves the superior classification performance highlighted in their discussion.

2 Shahbaba et al.

Shahbaba *et al.* provide several insightful suggestions. The majorization-minimization (MM) algorithm and the alternative use of Gaussian process certainly deserve further exploration. Gramacy and Polson (2010) have shown how particle methods can be used to implement Gaussian process classification while avoiding the clumsy $0(N^3)$ matrix inversion, thus making direct comparisons with SVM in high dimensions feasible. Our manuscript mentions the possibility of working parameter methods (either in MCMC or EM), and we assume that a faster working-parameter algorithm exists, but we have not explored this option because it is not clear what the working parameter should be. Perhaps there is an artificial identifiability constraint in one of the GIG distributions that could be relaxed.

We were a little confused by Section 2 from Shahbaba *et al.*, which seems to fault us first for blindly using the penalty term from a non-Bayesian model as a prior, and then later for abandoning L_1 normalization in favor of spike-and-slab priors. We would also like to point out that we did not simply "use results of Andrews and Mallows" as suggested near the end of Section 1. Our result is a non trivial extension of Andrews and Mallows' work, which dealt exclusively with scale mixtures (and which was based in turn on Pollard (1946)!).

Another avenue for future research is to use the theoretical frequentist and Bayesian properties of ridge regression estimators in large p small n problems. Our SVM estimator is a Rao-Blackwellized ridge regression estimator of the form $E(\beta|y) = E\left\{E(\beta|\Lambda, \Omega, y)\right\}$ where $E(\beta|\Lambda, \Omega, y)$ is a weighted least squares ridge estimator. By data adaptively estimating the latent variables we obtain a marginalized posterior mean. Theoretical results developed by Ishwaran and Rao (2005) and more recently used in Armagan et al. (2011) for generalized sparse ridge regression estimators appear to apply to this nonlinear context as well.

We agree with the point (also raised by Hans), that one need not stick with priors that are analytically or computationally convenient. However life is a lot easier when you do, and our complete data pseudo-likelihood dramatically increases the number of "convenient" priors available to the analyst.

Finally, we share Shahbaba *et al.*'s concerns about the fact that the SVM criterion does not arise from a probabilistic model. The simulations done by Mallick *et al.* show that improved classifications are possible using the SVM criterion relative to popular probit/logit alternatives. We recognize that classifiers are often used when probabilities of class membership would be more appropriate, but performance should outweigh ideology in cases where classification really is the primary goal.

3 Hans

Hans provides a number of insights into the elastic net procedure using our representation result which we found very intriguing. He presents a cautionary tale showing the unintended consequences of choosing priors based on convenience. One of the central points of the article (as Hans points out) is that our complete data likelihood plays nicely with a very wide set of priors developed for conditionally Gaussian linear models. Thus if one prior fails to induce the desired behavior, another can be easily substituted. In fact, our methods provide a useful computational advantage when used with the orthant normal prior introduced by Hans (2009). In that work, Hans used a Gibbs sampler that drew one component of β at a time, and found that it mixed faster than the sampler that updated the entire β vector using the Andrews and Mallows mixture. Univariate updating can be expensive in large n, large p problems where $p \ll n$, because each update requires a loop over the data. In linear models the loop can be efficiently reformulated in terms of the $p \times p$ cross product matrix. The loop over n observations cannot be avoided in nonlinear models, but data augmentation ensures that there is only one such loop per iteration. There is a cost to be paid in terms of a slower mixing rate, though Shahbaba *et al.* point out that the cost can be mediated by finding the appropriate working variable.

On the subject of speed (which Hans mentions not mentioning), we did not do a rigorous test to compare the computational speed of our EM algorithms with convex optimization, but we were surprised at how quickly our methods performed relative to the R package we compared them against in Figure 2. One factor in our favor is that our methods do not require cross validation to select ν. However, even when we set the number of cross validation samples to 1 in `penalizedSVM` (i.e. we turned off cross validation) our algorithms were competitive speed-wise. Of course, the performance of an R package often has more to do with characteristics of R than with the potential efficiency of the underlying algorithm, so it is not clear what the speed difference would have been in a competition between two speed-optimized versions of the software.

Hans proposes a sensible prediction rule based on the optimal Bayes "plug-in" estimator $E(\beta|y)'x$. He also suggests to compute $E(\beta|y)$ one should average over the posterior on regularization parameters ν, α and discusses why this is more stable than other choices. One word of caution on selecting hyper-parameters without their own regularization penalty is that the marginal likelihood $p(y|\nu)$ can have its mode at zero precisely when the parameter vector is sparse, see condition 6.2 of Tiao and Tan (1966) and Polson and Scott (2010) for a discussion of the linear p-means problem, the mode of $p(y|\nu)$ is exactly zero. Finally, we thank all the discussants for their contributions.

References

Armagan, A., Dunson, D., and Lee, J. (2011). "Bayesian generalized double Pareto shrinkage." Technical report, Duke University. 45

Brown, B., Fearn, T., and Vannucci, M. (1999). "The choice of variables in multivariate regression: a non-conjugate Bayesian decision theory approach." *Biometrika*, 86(3): 635. 43

Brown, P., Vannucci, M., and Fearn, T. (1998). "Multivariate Bayesian variable selection and prediction." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(3): 627–641. 43

— (2002). "Bayes model averaging with selection of regressors." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3): 519–536. 43

Frühwirth-Schnatter, S. and Frühwirth, R. (2007). "Auxiliary Mixture Sampling with Applications to Logistic Models." *Computational Statistics and Data Analysis*, 51: 3509–3528. 44

— (2010). "Data augmentation and MCMC for binary and multinomial logit models." In Kneib, T. and Tutz, G. (eds.), *Statistical Modelling and Regression Structures – Festschrift in Honour of Ludwig Fahrmeir*, 111–132. Heidelberg: Physica-Verlag. Available online `http://www.ifas.jku.at/e2550/e2756/index_ger.html`, IFAS Research Paper Series 2010-48. 44

Gramacy, R. and Polson, N. G. (2010). "Particle Learning of Gaussian process models for sequential design and optimisation." *working paper*. 45

Hans, C. (2009). "Bayesian lasso regression." *Biometrika*, 96(4): 835–845. 46

Ishwaran, H. and Rao, J. (2005). "Spike and slab variable selection: frequentist and Bayesian strategies." *Annals of statistics*, 730–773. 45

Lindley, D. (1968). "The choice of variables in multiple regression." *Journal of the Royal Statistical Society. Series B (Methodological)*, 30(1): 31–66. 43

Mallick, B. K., Ghosh, D., and Ghosh, M. (2005). "Bayesian classification of tumours by using gene expression data." *Journal of the Royal Statistical Society, Series B, Statistical Methodology*, 67(2): 219–234. 44

Mallows, C. L. (1973). "Some Comments on C_p." *Technometrics*, 15: 661–675. 43

Pollard, H. (1946). "The representation of $e^{-x^{\lambda}}$ as a Laplace integral." *Bull. Amer. Math. Soc.*, 52(10): 908–910. 45

Polson, N. and Scott, J. (2010). "Shrink Globally, Act Locally: Sparse Bayesian Regularization and Prediction." *Bayesian Statistics 9*. 46

Scott, S. L. (2011). "Data Augmentation, Frequentist Estimation, and the Bayesian Analysis of Multinomial Logit Models." *Statistical Papers*, 52(1): 87 – 109. 44

Tiao, G. and Tan, W. (1966). "Bayesian analysis of random-effect models in the analysis of variance." *Biometrika*, 53(3-4): 477. 46

Tüchler, R. (2008). "Bayesian Variable Selection for Logistic Models Using Auxiliary Mixture Sampling." *Journal of Computational and Graphical Statistics*, 17(1): 76–94. 43