Bayesian Analysis (2011) **6**, Number 1, pp. 25–30

Comment on Article by Polson and Scott

Bani K. Mallick*, Sounak Chakraborty† and Malay Ghosh‡

We congratulate the authors for a very interesting article. The key contribution of this paper, as we see it and as suggested in the title, is the introduction of latent parameters to carry out Bayesian analysis with support vector machines. The basic identities (4) and (6) are particularly useful in this regard, which enable one to overcome much of the complexities of a non smooth loss resulting in a non smooth likelihood. As an anecdote, from a Bayesian angle, it is extremely convenient to view the loss as the negative of the loglikelihood (for example associating squared error loss with the normal likelihood) and the penalty part with the prior. This is the approach taken in this paper, and also earlier in Mallick *et al.* (2005). Based on a loss function, we can obtain the normalized or the non-normalized (or pseudo) likelihood. The authors considered the non-normalized likelihood and that way obtained the pseudo posterior distribution. This pseudo posterior distribution may not be suitable to make probabilistic inference. Mallick et al. (2005) considered both the normalized and the non-normalized likelihoods and the classification performances were compatible. It will be interesting to see how the proposed method can be adapted for the model with the normalized likelihood.

The introduction of latent parameters facilitates Bayesian variable selection with LASSO (Park and Casella (2008); Bae and Mallick (2007)) and its generalizations such as grouped LASSO, fused LASSO and elastic net (Kyung *et al.*, 2010; Chakraborty and Guo, 2010). Not surprisingly, this helps also in classification problems with a penalty function which is the same as in LASSO. One interesting feature in this paper is the consideration of a general α in (6) rather than the conventional $\alpha = 1$ or 2. Corollary 4 in this paper seems to be an interesting result, especially because of its importance in developing the necessary algorithm.

The major emphasis of this paper seems to be on posterior inference. In many real problems, the emphasis should be on prediction rather than estimation. Particularly, the predictive distribution is useful to compare different classification models. The latent development of this paper should be exploited also in that framework. Specifically, for classification, this will amount to estimating probabilities of misclassification of future observations.

The authors have considered only the linear SVM model. The nonlinear SVM model will require more complex analysis due to the presence of parameters in the $\boldsymbol{X}$ (design)

*Department of Statistics, Texas A&M University, College Station, TX, mailto:bmallick@stat.tamu.edu

†Department of Statistics, University of Missouri, Columbia, MO, mailto:chakrabortys@missouri.edu

‡Department of Statistics, University of Florida, Gainesville, FL, mailto:ghoshm@stat.ufl.edu

 DOI:10.1214/11-BA601A

matrix and hopefully this method can be extended in that situation.

We now provide a comparative analysis of some existing Bayesian methods like Bayesian Probit Regression with a mixture prior for variable selection (BPR) (Lee *et al.*, 2003), Bayesian Additive Regression Trees (BART) (Chipman *et al.*, 2010), and Bayesian Hybrid Huberized Support Vector Machine (BHHSVM) (Chakraborty and Guo, 2010).

Bayesian Probit Regression with mixture prior for variable selection (BPR) (Lee *et al.*, 2003): In BPR we use the Bayesian binary regression model with the probit link function and a linear predictor. Bayesian mixture priors are assigned to the coefficients of the linear predictor to perform the variable selection. BPR is much simpler than BART and BHHSVM. The implementation is easy and faster than the other two competing methods. We adopted the same prior specifications as suggested by Lee *et al.* (2003). In the first column of Table 1 we list the average misclassification error for BPR and the selected covariates.

Bayesian Additive Regression Trees (BART) (Chipman *et al.*, 2010): Here the binary response is connected with the covariates using a probit link function and the linear predictor is modeled using an ensemble of several small trees. Prior distributions are assigned to the tree parameters and the latter are estimated through posterior simulations. Chipman *et al.* (2010) proposed a novel Bayesian back-fitting algorithm to fit the BART model. In this paper we used 200 trees in the BART model and all prior parameters are adopted following the suggestion of Chipman *et al.* (2010). BART model cannot select the covariates and produce a sparse solution. However, it can report the importance of a covariate based on the number of times it is used in a tree decision rule over all 200 trees. In Table 1 second column, we report the accuracy of BART and list covariates according to decreasing order of the variable importance.

Bayesian Hybrid Huberized Support Vector Machine (BHHSVM) (Chakraborty and Guo, 2010): The BHHSVM is a Bayesian formulation of the hybrid Huberized support vector machine for binary classification. In BHHSVM the loss function (or negative of the log-likelihood) is as follows (Wang *et al.*, 2008),

$$\phi(yf) = \begin{cases} 0 & \text{for } yf > 1, \\ (1-yf)^2/2\delta & \text{for } 1-\delta < yf \leq 1, \\ 1-yf-\delta/2 & \text{for } yf \leq 1-\delta, \end{cases} \tag{1}$$

where $f(\boldsymbol{x}_i) = \beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta}$ is the linear decision boundary similar to a linear SVM. On the coefficients of the linear classification boundary, we assign the elastic net prior (Zou *et al.*, 2005; Chakraborty and Guo, 2010), which can select variables and group them together simultaneously. The elastic net prior (Chakraborty and Guo, 2010) corresponds to the elastic net penalty, $H(\boldsymbol{\beta}) = (1-\alpha)\sum_{j=1}^{p}|\beta_j| + \alpha\sum_{j=1}^{p}\beta_j^2$ (Zou *et al.*, 2005). The adopted elastic net prior can be derived as a scale mixture of normal and truncated

gamma distribution. The scale mixture formulation of the elastic net prior is as follows,

$$\begin{aligned} \beta_j|\xi_j &\sim \mathrm{N}(\mathbf{0}, \xi_j \mathrm{I}), && (2)\\ u_j = \frac{1}{1-\alpha\xi_j} &\sim \text{Truncated Gamma}(1/2, c^*) I(1 < u_j < \infty), j = 1, \ldots, p\,, && (3)\\ c^* &= \frac{(1-\alpha)^2}{2\alpha}. \end{aligned}$$

In the third column of Table 1 we report the average misclassification error for BHHSVM and the list of the selected covariates.

These three methods are applied to the spam data set described in Polson and Scott (2011). The spam data set has 4601 samples and 57 covariates (excluding the intercept). We randomly split the data into training set (two third samples) and test set (one third samples) fifty times. We record the list of the selected variables from every split. It is natural that for each split the set of covariates selected may be slightly different. The list of the variables reported in Table 1 are the ones that appeared in at least twenty five out of fifty splits. In Table 1 we report the average misclassification error, the standard deviation of the misclassification error, and the list of variables selected by each method.

For all three models we ran multiple MCMC chains to avoid any problem related to multimodality of the posterior distribution and ran the chains until we got satisfactory convergence. The convergence was checked by trace plots of the generated samples and calculating the Gelman-Rubin scale reduction factor (Gelman, 1996) using the *Coda*() package in *R*. From Table 1 we see that BHHSVM and BART work equally well in terms of misclassification error. The BPR results in slightly higher misclassification error but computationally it is faster than BHHSVM and BART. In terms of variable selection BART cannot produce sparse result like BHHSVM or BPR, however covariates are ranked according to the number of times they are used in a tree decision rule over all trees. On an average, BHHSVM selects 33 variables and the BPR selects 31 variables. Following the listed covariates in Table 1 we can see there is a significant overlap of the selected covariates according to these three competing methods and the Bayesian SVM proposed in Polson and Scott (2011). The covariates that are also marked as important in Polson and Scott (2011) are colored red (grey) in Table 1. The above comparisons indicate that at least in the discussed data set the non-linear classifier BART is working better in terms of classification accuracy. However, BART ends up selecting all the covariates due to lack of any in built variable selection technique. On the other hand linear classifiers can perform better with a SVM type likelihood (BHHSVM) rather than Binomial model with probit link function, and variable selection can be easily made by incorporating mixture priors on the coefficients of the linear predictors.

The SVM formulation with data augmentation and pseudo-likelihood introduced by Polson and Scott (2011) is very interesting and certainly opens up new areas of research on Bayesian SVM. Extension of this method to multicategory classification (Chakraborty

et al., 2007) and survival analysis (Maity and Mallick, 2011) problems will be exciting future research directions. Furthermore, apart from the shrinkage penalties discussed in their paper, it would be interesting to study penalties that offer simultaneous shrinkage and grouping as proposed in Kyung *et al.* (2010). These grouping priors have the ability to select genes as a group from genetic pathways rather than picking up individual genes. In a practical situation where several genes are biologically grouped resulting in one outcome or effect, identifying an important pathway is more important than finding a single gene.

References

Bae, K. and Mallick, B. (2004). "Gene selection using a two-level hierarchical Bayesian model." *Bioinformatics*, 20: 3423–3430.

Chakraborty, S. and Guo, R. (2010). "Bayesian Hybrid Huberized SVM and its Applications in High Dimensional Medical Data." *Computational Statistics and Data Analysis*, 55: 1342–1356.

Chakraborty, S., Mallick, B., Ghosh, D., Ghosh, M., and Dougherty, E. (2007). "Hierarchical Bayesian vector machines for gene expression-based glioma classification." *Sankhya*, 69: 514–547.

Chipman, H., George, E., and McCulloch, R. (2010). "BART: Bayesian Additive Regression Trees." *Annals of Applied Statistics*, 4: 266–298.

Gelman, A. (1996). "Inference and monitoring convergence." In W. Gilks, S. R. and Spiegelhalter, D. J. (eds.), *Markov Chain Monte Carlo in Practice*, 131–140. Chapman & Hall.

Kyung, M., Gilly, J., Ghosh, M., and Casella, G. (2010). "Penalized Regression, Standard Errors, and Bayesian Lassos." *Bayesian Analysis*, 5: 369–412.

Lee, K., Naijun, S., Dougherty, E., Vannucci, M., and Mallick, B. (2003). "Gene Selection: A Bayesian Variable Selection Approach." *Bioinformatics*, 19: 90–97.

Maity, A. and B, M. (2011). "Proportional Hazards Regression Using Bayesian Kernel Machines." In Dey, Ghosh, and Mallick (eds.), *Bayesian Modeling in Bioinformatics*. Chapman & Hall.

Mallick, B., Ghosh, D., and Ghosh, M. (2005). "Bayesian classification of tumors using gene expression data." *JRSSB*, 67: 219–234.

Park, T. and Casella, G. (2008). "The Bayesian Lasso." *JASA*, 103: 681–686.

Wang, L., Zhu, J., and Zou, H. (2008). "Hybrid Huberized support vector machines for microarray classification and gene selection." *Bioinformatics*, 24: 412–419.

Zou, H. and Hastie, T. (2005). "Regularization and variable selection via the elastic net." *JRSSB*, 67: 301–320.

Table 1: Spam Data: Classification Accuracy and Variables Selected.

	BPR	BART	BHHSVM
Misclassification error	8.996	6.1694	6.6375
SD	0.7577	0.6867	0.5147
		Variables Selected	
	word_freq_remove	word_freq_remove	word_freq_address
	char_freq_!	word_freq_000	word_freq_our
	char_freq_$	word_freq_hp	word_freq_remove
	word_freq_hp	word_freq_george	word_freq_will
	word_freq_george	word_freq_free	word_freq_free
	word_freq_free	char_freq_$	word_freq_000
	word_freq_meeting	word_freq_edu	word_freq_hp
	word_freq_000	word_freq_our	word_freq_george
	word_freq_edu	char_freq_!	word_freq_lab
	word_freq_our	word_freq_internet	word_freq_data
	word_freq_1999	word_freq_meeting	word_freq_cs
	capital_run_length_total	word_freq_money	word_freq_meeting
	word_freq_internet	word_freq_your	word_freq_project
	word_freq_money	word_freq_business	word_freq_re
	word_freq_re	word_freq_re	word_freq_edu
	word_freq_your	word_freq_email	word_freq_conference
	word_freq_business	word_freq_over	char_freq_;
	word_freq_will	word_freq_credit	char_freq_(
	word_freq_over	word_freq_data	char_freq_$
	capital_run_length_longest	word_freq_hpl	char_freq_!
	word_freq_hpl	word_freq_you	word_freq_business
	word_freq_project	word_freq_font	word_freq_hpl
	word_freq_font	word_freq_will	word_freq_85
	word_freq_credit	word_freq_project	word_freq_3d
	word_freq_mail	word_freq_address	word_freq_money
	word_freq_report	capital_run_length_average	word_freq_credit
	word_freq_order	word_freq_3d	capital_run_length_longest
	word_freq_data	capital_run_length_longest	capital_run_length_total
	word_freq_lab	word_freq_mail	word_freq_original
	word_freq_85	word_freq_all	word_freq_your
	word_freq_conference	word_freq_1999	word_freq_pm
		capital_run_length_total	char_freq_#
		char_freq_;	word_freq_650
		word_freq_make	
		word_freq_order	
		word_freq_receive	
		word_freq_people	
		word_freq_report	
		word_freq_addresses	
		word_freq_650	
		word_freq_lab	
		word_freq_labs	
		word_freq_telnet	
		word_freq_857	
		word_freq_415	
		word_freq_85	
		word_freq_technology	
		word_freq_parts	
		word_freq_pm	
		word_freq_direct	
		word_freq_cs	
		word_freq_original	
		word_freq_table	
		word_freq_conference	
		char_freq_(	
		char_freq_[	
		char_freq_#	

Overlap with Polson and Scott (2011) are marked by red (grey).