Bayesian Analysis (2011) **6**, Number 1, pp. 31–36

Comment on Article by Polson and Scott

Babak Shahbaba[*], Yaming Yu[†] and David A. van Dyk[‡]

1 Introduction

Polson and Scott's paper presents the enlightening observation that the standard SVM can be embedded into a statistical latent variable model. This is aligned with other recent work, in which the penalty term in the convex optimization for several popular non-Bayesian models has been replaced by a prior distribution in order to develop an alternative Bayesian approach. See, for example, the Bayesian lasso model by Park and Casella (2008) and Hans (2009), and the Bayesian bridge regression model by Armagan (2009). Following the work of Andrews and Mallows (1974) and West (1987), the prior distributions used in these methods are expressed as scale mixtures of normal distributions. For example, in bridge regression (Frank and Friedman 1993), which includes both ridge regression (Hoerl and Kennard 1970) and lasso (Tibshirani 1996) as special cases, the regression parameters are estimated by minimizing the penalized residual sum of squares (using centered data),

$$\hat{\beta} = \arg\min_{\beta} \left[(y - X\beta)^T (y - X\beta) + \lambda \sum_{j=1}^{p} |\beta_j|^{\gamma} \right]$$

where $\beta = (\beta_1, \ldots, \beta_p)$. In the Bayesian framework, the penalty term can be replaced by a prior distribution of the form $P(\beta) \propto \exp(-\lambda|\beta_j|^{\gamma})$. When $0 < \gamma \leq 2$, the penalty can be represented as a scale mixture of normal distributions (West 1987).

The current paper follows a similar approach in its replacement of the regularization term in SVM with a prior distribution. The authors also followed similar steps to specify the likelihood since unlike ridge regression, the likelihood is not readily available for SVM. In particular, they insightfully replace the part of the objective function that depends on the data with $\exp[-2\sum_{i=1}^{n} \max(1-y_i X_i^T \beta, 0)]$ and use results from Andrews and Mallows (1974).

2 A Bayesian SVM model or a Bayesian model with SVM properties

While the authors presented the critical first step of formulating a Bayesian model that encompasses SVM, in general it is not necessary to limit our choice of prior distributions in a Bayesian model by forcing mathematical compatibility with the penalty term in the

[*] Department of Statistics, University of California, Irvine, CA, mailto:babaks@uci.edu
[†] Department of Statistics, University of California, Irvine, CA, mailto:yamingy@ics.uci.edu
[‡] Department of Statistics, University of California, Irvine, CA, mailto:dvd@ics.uci.edu

DOI:10.1214/11-BA601B

corresponding non-Bayesian model. Such penalty terms usually serve a specific purpose in the original form of these models, where the estimates are obtained using convex optimization. In the process of transforming the penalty term to a prior distribution, we may formulate a new procedure with different or even improved properties, or we may arrive at properties similar to the original method, but by different means. For example, the Bayesian lasso model proposed by Park and Casella (2008) replaces the L_1 penalty term with a prior distribution that is a scale mixture of normal distributions, but the resulting model lacks what has made lasso popular: its sparsity. We can of course achieve sparsity using a spike and slab prior (as the authors did in Section 4.2) if that is what we desire, but this strategy completely circumvents lasso's standard L_1 regularization term. Thus we can mimic the desired properties (e.g., sparsity) of popular non-Bayesian models without striving for mathematical equivalence.

The authors use the term $\sum_{i=1}^{n} \max(1 - y_i X^T \beta, 0)$ in the objective function to specify their (pseudo) likelihood function. This term, which the authors refer to as "the awkward SVM optimality criterion", is in fact the constraint function in the original form of SVM, and it is quite reasonable in that context. Assuming that the two classes are perfectly separable, and we scale β such that $|X_i^T \beta|$ (i.e., the distance from the hyperplane) is equal to 1 for all points on the boundary of the slab, the above constraint is imposed so the two classes fall on the correct sides of the separating hyperplane; that is, $y_i X_i^T \beta \geq 1$ for $i = 1, \ldots, n$. In this case, the objective function is in fact $||\beta||^2$, whose minimization is equivalent to the maximization of the margin (width) of the slab. Now the question is: should we start from a constraint function, that makes perfect sense in an optimization problem, and attempt to find a mathematically equivalent model, or would it be more appropriate to focus on what is desirable (e.g., using the kernel trick to create a rich class of nonlinear models) in SVM, in order to define an alternative Bayesian model?

While presenting the standard SVM in a Bayesian form has many advantages, as described by the authors, it is not yet clear how some of the major issues plaguing the standard SVM can be resolved by this Bayesian formulation. For example, one of the main disadvantages of the standard SVM is that its predictions are not probabilistic. Another disadvantage is that the extension to classification problems with multiple classes is not straightforward. Can these issues be addressed by the authors proposed Bayesian formulation?

Indeed, there is an existing class of Bayesian models that is closely related to SVM that answers all the above concerns: Gaussian process (GP) models (Neal 1998). The kernel function of SVM corresponds to the covariance function of a GP (Neal 2004). For nonlinear classification (and regression) models using GP, the prior distributions are quite flexible, the predictions are probabilistic, and the extension to multinomial classification is straightforward.

The main disadvantage of the GP model compared to SVM is its computational cost. In recent years, many researchers have focused on improving the computational time of GP models (e.g., Seeger et al. 2003). By formulating the SVM in terms of latent variables in a model that can be fit with EM or data augmentation, the authors have set the stage for the use of a large and flexible class of sometimes very efficient algorithms. This is the topic of our next section.

3 Computation

With their introduction of latent variables, Polson and Scott illustrate yet another rich class of problems where despite the lack of apparent missing data, the EM and data augmentation (DA) algorithms can be used to derive simple and stable computational schemes. While a similar approach has been proposed for lasso and bridge regression, the idea goes back much further, even before the general formulation of EM in the landmark paper by Dempster et al. (1977). In the context of least absolute deviations (LAD) regression, Schlossmacher (1973) proposed a method which was later identified as an EM algorithm using precisely this idea. In the usual regression context, the LAD problem is to minimize the function

$$D(\theta) = \sum_{i=1}^{n} |y_i - X_i\theta|$$

where $(y_1, \ldots, y_n)$ are the responses and $(X_1^\top, \ldots, X_n^\top)$ is a matrix of covariates. Schlossmacher's approach is an iteratively reweighted least squares algorithm; at each iteration, the weights are inversely proportional to the absolute residuals. There seems to be a common thread that connects LAD, lasso, and now SVM, offering opportunities to borrow ideas between these methods.

Schlossmacher's algorithm brings about another potential point of interest. It is nontrivial to find the latent variable formulation that renders Schlossmacher's algorithm an EM algorithm. However, it is easy to derive Schlossmacher's algorithm using the majorization-minimization (MM) principle (Lange et al. 2000). MM can be regarded as a generalization of EM without missing data. Given a function $l(\theta)$ that is to be maximized, one finds a function $Q(\theta|\tilde{\theta})$ such that $l(\theta) \geq Q(\theta|\tilde{\theta})$ for all θ and $\tilde{\theta}$ and $l(\theta) = Q(\theta|\theta)$. At each iteration, the MM algorithm maximizes the surrogate Q with respect to its first argument. The sequence of $\theta^{(t)}$ monotonically increases $l(\theta)$ because $l(\theta^{(t+1)}) \geq Q(\theta^{(t+1)}|\theta^{(t)}) \geq Q(\theta^{(t)}|\theta^{(t)}) = l(\theta^{(t)})$. Would there be an advantage of using MM in deriving optimization algorithms in the SVM context?

The latent variable formulation has an obvious advantage in that once it is obtained, we have the large arsenal of missing data and imputation methods at our disposal. The authors make use of efficient variants of EM such as the ECME algorithm. There are other algorithms in the EM family that could potentially help. Examples include the AECM algorithm of Meng and van Dyk (1997) and the PXEM algorithm of Liu et al. (1998). Like EM, these algorithms maintain the stable monotonic convergence properties of EM. They cannot always be used of course because they rely on special model structures. If applicable, however, they can result in dramatic improvement in speed while maintaining stability and without sacrificing much of the simplicity, a feature shared by the ECME algorithm. It seems worthwhile to explore the possibility of AECM and PXEM in the SVM context. There are MCMC-counterparts to most EMtype algorithms (van Dyk and Meng 2011) that could be equally attractive for fitting the SVM. Reparameterization (Roberts and Sahu 1997) and parameter expansion (Liu and Wu 1999) are powerful methods that apply to a variety of problems and the partially collapsed Gibbs sampler (van Dyk and Park 2008) offers a stochastic counterpart to ECME. We wonder if such methods or their extensions (Yu and Meng, to appear) could be relevant in the SVM context.

References

Andrews, D. F. and Mallows, C. L. (1974). "Scale Mixtures of Normal Distributions." *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(1): 99–102. 31

Armagan, A. (2009). "Variational Bridge Regression." In van Dyk, D. and Welling, M. (eds.), *The 12th International Conference on Artificial Intelligence and Statistics*, 5:17–24. Clearwater Beach, Florida USA: JMLR W&CP. 31

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1): 1–38. 33

Frank, I. E. and Friedman, J. H. (1993). "A Statistical View of Some Chemometrics Regression Tools." *Technometrics*, 35(2): 109–135. 31

Hans, C. (2009). "Bayesian lasso regression." *Biometrika*, 96(4): 835–845. 31

Hoerl, A. E. and Kennard, R. W. (1970). "Ridge regression: Application to nonorthogonal problems." *Technometrics*, 12(1): 55–67. 31

Lange, K., Hunter, D., and Yang, I. (2000). "Optimization transfer using surrogate objective functions (with discussion)." *Journal of Computational and Graphical Statistics*, 9: 1–59. 33

Liu, C., Rubin, D. B., and Wu, Y. N. (1998). "Parameter expansion to accelerate EM – the PX-EM algorithm." *Biometrika*, 85(4): 755–770. 33

Liu, J. S. and Wu, Y. N. (1999). "Parameter expansion for data augmentation." *Journal of the American Statistical Association*, 94: 1264–1274. 34

Meng, X. L. and van Dyk, D. (1997). "The EM Algorithm–An Old Folk-Song Sung to a Fast New Tune." *Journal of the Royal Statistical Society. Series B (Methodological)*, 59(3): 511–567. 33

Neal, R. M. (1998). "Regression and classification using Gaussian process priors." *Bayesian Statistics*, 6: 471–501. 32

— (2004). "Tutorial on Bayesian Methods for Machine Learning."
URL `http://www.cs.toronto.edu/~radford/ftp/bayes-tut.pdf` 32

Park, T. and Casella, G. (2008). "The Bayesian Lasso." *Journal of the American Statistical Association*, 103(482): 681–686. 31, 32

Roberts, G. O. and Sahu, S. K. (1997). "Updating Schemes, Correlation Structure, Blocking and Parameterisation for the Gibbs Sampler." *Journal of the Royal Statistical Society, Series B*, 59: 291–317. 34

Schlossmacher, E. J. (1973). "An iterative technique for absolute deviations curve fitting." *Journal of American Statistical Association*, 68(344): 857–859. 33

Seeger, M., Williams, C. K. I., and Lawrence, N. (2003). "Fast forward selection to speed up sparse Gaussian process regression." In Bishop, C. and Frey, B. J. (eds.), *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*. Key West, Florida: Society for Artificial Intelligence and Statistics. 32

Tibshirani, R. (1996). "Regression Shrinkage and Selection Via the Lasso." *Journal of the Royal Statistical Society, Series B*, 58(1): 267–288. 31

van Dyk, D. and Park, T. (2008). Partially collapsed Gibbs samplers: Theory and methods. *Journal of the American Statistical Association*, 103:790–796. 34

van Dyk, D. A. and Meng, X.-L. (2011). Cross-fertilizing strategies for better EM mountain climbing and DA field exploration: A graphical guide book. *Statistical Science*, in press. 34

West, M. (1987). "On scale mixtures of normal distributions." *Biometrika*, 74(3): 646–648. 31

Yu, Y. and Meng, X. L. (to appear). "To Center or Not to Center, That is Not the Question: An Ancillarity-Sufficiency Interweaving Strategy (ASIS) for Boosting MCMC Efficiency." *Journal of Computational and Graphical Statistics*.