Bayesian Analysis (2011) **6**, Number 1, pp. 37–42

Comment on Article by Polson and Scott

Chris Hans*

1 Overview

What at first may appear to be "just" a clever bit of calculus turns out to cast a new light on the support vector machine (SVM). I would like to congratulate Nicholas Polson and Steven Scott on an interesting paper that opens a door to many new applications of the SVM. The representation of the SVM pseudo-likelihood as a mean-variance mixture of normals is by no means obvious (to most of us!). Placing the SVM in this framework provides an easy mechanism for developing principled Bayesian models around the core SVM structure. This may well lead to interesting new methods for high-dimensional classification; the spike-and-slab prior extensions in Section 4.2 and the application thereof in Section 5 are a promising start down this path.

A potential criticism of the paper (that you won't hear from me) is: Why use EM or MCMC when convex optimization is so fast? Criticisms along this line, that focus solely on computational efficiency, miss the importance of the work. Anticipating such criticisms, Polson and Scott remark in the introduction that "these algorithms replace the conventional convex optimization algorithm for SVM's, which is fast but unfamiliar to many statisticians, with what is essentially a version of iteratively re-weighted least squares...the latent variable representation brings all of conditional linear model theory to SVM's." While a better understanding of convex optimization would certainly be beneficial for many of us, the point is that casting an estimation procedure in a model-based context instantaneously provides new insight into the approach. The fact that the model-based context in this particular case happens to be conditional linear model theory — perhaps the most widely studied area of statistics — is remarkable. Polson and Scott provide several new insights right away, including the reinterpretation of a support vector in the context of weighted least squares. New insights are sure to follow, not least among them modeling of dependence structures across features and the construction of prior distributions that incorporate context-specific information.

Polson and Scott choose to work with the unnormalized SVM criterion, which corresponds to a pseudo-likelihood and hence generates a pseudo-posterior. They note that this could be avoided by working with $\tilde{L}_i$, a normalized version of the SVM criterion, but that this would break the direct connection to the traditional SVM estimate. The lack of a proper likelihood function seems to hinder formal Bayesian prediction, as this causes the posterior predictive distribution to be not well defined. In the absence of a formal likelihood, and hence Bayes-optimal prediction, the "plug-in" approach of predicting future observations based on the sign of $E(\beta \mid y)^T \mathbf{x}$, where the expectation is taken with respect to the pseudo posterior, may still provide good prediction. Building a fully Bayes model, where the regularization parameters ν and α are learned and av-

*Department of Statistics, The Ohio State University, Columbus, OH mailto:hans@stat.osu.edu

 DOI:10.1214/11-BA601C

eraged over (a smooth process that should be insensitive to small changes in the data), may provide a better estimate of β than in the classical analysis where the amount of penalization must be chosen via a heuristic such as cross validation (a less-smooth process that can depend heavily on local features of the data). Regardless, future study of prediction in this Bayesian model-based version of the SVM may provide even more insight into the structure of the classifier.

In Section 2 below I offer a different application of the same mean-variance mixture results of Andrews and Mallows (1974) that are used to prove Polson and Scott Theorem 1. The proof of their theorem, and the form of the mean-variance mixture of normals, reminded me of the form of a particular regularization prior with which I have worked. The mean-variance mixture I describe below provides what was at first glance an unexpected result and emphasizes the importance of performing Bayesian regularization based on principles rather than convenience.

2 Mean-Variance Mixtures of Normals

Scale-mixtures of normal distributions have recently received special attention in the area of Bayesian regularization. The use of a mean-variance mixture of normals in Polson and Scott Theorem 1, while not employed there for prior regularization, is nonetheless notable as such mixtures have been discussed less frequently during the recent resurgence of this framework. Interestingly, the mean-variance mixture of normal results of Andrews and Mallows (1974) and Polson and Scott Theorem 1 can be applied to a Bayesian formulation of elastic net regression.

The elastic net (Zou and Hastie 2005) is a penalized optimization procedure that estimates the regression coefficients β in a linear model by

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta) + \frac{\lambda_1}{2\sigma^2}|\beta|_1 + \frac{\lambda_2}{2\sigma^2}|\beta|^2,$$

where $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$ are regularization parameters, $|\beta|_1$ and $|\beta|^2$ are the L_1- and L_2-norms of β, respectively, and the entire expression is scaled by $2\sigma^2$ here to make a connection to the normal likelihood. Other scalings are also possible, and perhaps preferable, however scaling the entire quantity by $2\sigma^2$ preserves the direct interpretation of the symbols λ_1 and λ_2 as the elastic net penalty parameters. Notice that if $\lambda_2 = 0$, $\hat{\beta}$ is the usual lasso estimate. The L_2-norm piece of the penalization was added by Zou and Hastie (2005) to address perceived deficiencies with the lasso penalty.

Zou and Hastie (2005) remark that the estimate $\hat{\beta}$ corresponds to the mode of a Bayesian posterior distribution under a particular prior. When $\lambda_2 = 0$, the prior is $\text{DE}(\lambda_1/(2\sigma^2))$, the double exponential distribution with variance $8\sigma^4/\lambda_1^2$ (Park and Casella 2008; Hans 2009). When λ_2 is not restricted to be zero, Hans (2008) shows that this prior is in the "orthant-normal" family of distributions. Focusing on a single

regression coefficient β, this distribution has density function

$$p(\beta \mid \lambda_1, \lambda_2, \sigma^2) = \begin{cases} \frac{\phi\left(\beta \mid \frac{\lambda_1}{2\lambda_2}, \frac{\sigma^2}{\lambda_2}\right)}{2\Phi\left(\frac{-\lambda_1}{2\sigma\sqrt{\lambda_2}}\right)}, & \beta < 0, \\ \frac{\phi\left(\beta \mid -\frac{\lambda_1}{2\lambda_2}, \frac{\sigma^2}{\lambda_2}\right)}{2\Phi\left(\frac{-\lambda_1}{2\sigma\sqrt{\lambda_2}}\right)}, & \beta \geq 0, \end{cases} \tag{1}$$

where $\phi(\beta \mid \cdot, \cdot)$ is the normal density function and $\Phi(\cdot)$ is the standard normal cumulative distribution function. The only difference between the positive and negative components of (1) is that the location parameters have opposite signs. While this distribution happens to have a scale-mixture-of-normals representation (Hans 2008; Li and Lin 2010; Kyung et al. 2010), here I work with (1) directly to demonstrate an application of mean-variance mixtures in the context of regularization.

The regularization parameters λ_1 and λ_2 play a role that is similar to that of ν in Polson and Scott. As noted in Section 3.3, learning ν by assuming a prior distribution $p(\nu)$ is desirable as it allows the amount of shrinkage to adapt to the data and obviates the need to resort to cross-validation to select a value of the regularization parameter. Perhaps more importantly, estimation of ν using fully Bayes methods incorporates information about ν contained in the pseudo-posterior normalization constant $C_\alpha(\nu)$, which, as Polson and Scott point out, is absent in the classical analysis (see also the discussion of Polson and Scott 2010b).

In the context of the regularized SVM, Polson and Scott choose an inverse gamma prior for ν as it is conditionally conjugate to the exponential power prior that is used for regularization. For elastic net regression, we desire a prior distribution $p(\lambda_1, \lambda_2 \mid \sigma^2)$, which may depend on σ^2, that ideally (i) submits readily to analysis (e.g. providing a closed-form conditional posterior distribution or allowing for simple calculation of a marginal likelihood for model comparison) and (ii) yields, upon marginalization, a prior distribution $p(\beta)$ that has desirable properties (e.g. behavior in the tails and near the origin that allows the prior to handle sparse signals effectively). Ignoring, for the moment, the more important of these two considerations, the term $\Phi(-\lambda_1/(2\sigma\sqrt{\lambda_2}))$ in (1) would appear to preclude any hope of finding a prior to satisfy (i): under "standard" hyperpriors (e.g. placing independent gamma distributions on λ_1 and λ_2, the approach taken in Hans 2008), the $\Phi(\cdot)$ term in the posterior normalizing constant leads to non-standard full conditional distributions and complicates analytical marginalization. We can, however, use results on mean-variance normal mixtures to analytically marginalize λ_2 under a particular choice of hyperprior to yield a surprising (although ultimately disappointing) result.

Theorem 3. *Under the orthant-normal prior for β given by (1) and the (conditional) hyperprior distribution for λ_2 with density function*

$$p(\lambda_2 \mid \lambda_1, \sigma^2) = \frac{\lambda_1^2}{2\sigma^2\lambda_2^2}\Phi\left(-\frac{\lambda_1}{2\sigma\sqrt{\lambda_2}}\right), \quad \lambda_2 > 0, \tag{2}$$

the marginal prior distribution of β (given λ_1 and σ^2) is

$$\begin{aligned} p(\beta \mid \lambda_1, \sigma^2) &= \int_0^\infty p(\beta \mid \lambda_1, \lambda_2, \sigma^2) p(\lambda_2 \mid \lambda_1, \sigma^2) d\lambda_2 \\ &= \frac{\lambda_1}{2\sigma^2} e^{-\lambda_1 |\beta| / \sigma^2}, \quad -\infty < \beta < \infty, \end{aligned}$$

i.e., $\beta \mid \lambda_1, \sigma^2 \sim DE(\lambda_1/\sigma^2)$. In other words, the double exponential distribution can be represented as a mean-variance mixture of orthant-normal distributions.

The proof follows directly from the results of Andrews and Mallows (1974) and Polson and Scott Theorem 1. Under the change of variables $\lambda = \lambda_2^{-1}$, the integrand in the marginalization is

$$p(\beta \mid \lambda_1, \lambda, \sigma^2) p(\lambda \mid \lambda_1, \sigma^2) = \begin{cases} \frac{\lambda_1^2}{4\sigma^2} \phi(\beta \mid (\lambda_1/2)\lambda, \sigma^2 \lambda), & \beta < 0 \\ \\ \frac{\lambda_1^2}{4\sigma^2} \phi(\beta \mid -(\lambda_1/2)\lambda, \sigma^2 \lambda), & \beta \geq 0, \end{cases}$$

where the awkward term involving $\Phi(\cdot)$ is no longer present due to the careful choice of the hyperprior. The double exponential density function appears after applying the identity $\phi(b \mid au, cu) = \phi(-b \mid -au, cu)$ to the negative piece and then using the result $\int_0^\infty \phi(b \mid -au, cu) du = a^{-1} \exp\{-2 \max(ab/c, 0)\}$ for $a > 0$ and $c > 0$.

While the particular choice of hyperprior (2) satisfies consideration (i) above — marginalization of the prior is easy and yields a well-known prior for β — the choice fails consideration (ii). The Gaussian ($\lambda_2 \neq 0$) component in the penalty (prior) was included specifically to allow for more flexible penalization than would be provided by the DE($\lambda_1/(2\sigma^2)$) prior alone. Oddly, including λ_2 in the prior and then marginalizing it via mixing distribution (2) results in a prior with identical behavior to the prior obtained with the trivial mixing distribution of a point mass at $\lambda_2 = 0$ (although the variances of the two distributions differ by a factor of 4).

The point of this particular application of mean-variance mixtures to regularized regression was not to suggest a constructive approach for choosing regularization hyperpriors. The approach clearly failed here. Rather, we learn that suggestive, obvious or default choices that are made on the basis of analytic or computational ease may have undesirable properties. Careful study and further development of the existing normal mixture literature, as exemplified by Polson and Scott Theorem 1, is essential to the future development of Bayesian regularization methods, particularly in high dimensional problems. Fortunately such work is currently underway. Recent work by Polson and Scott (2010a,b) describes new mechanisms for generating regularization priors that include normal scale-mixtures, while other recent papers (e.g., Carvalho et al. 2010; Griffin and Brown 2010) have examined particular classes of normal mixtures and developed criteria for evaluating the properties of these priors. While there is surely more work to come, the future for Bayesian regularization is bright.

References

Andrews, D. F. and Mallows, C. L. (1974). "Scale Mixtures of Normal Distributions." *Journal of the Royal Statistical Society - Series B*, 36: 99–102. 38, 40

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). "The horseshoe estimator for sparse signals." *Biometrika*, 97: 465–480. 40

Griffin, J. E. and Brown, P. J. (2010). "Inference with normal-gamma prior distributions in regression problems." *Bayesian Analysis*, 5: 171–188. 40

Hans, C. (2008). "Regression Modeling with the Elastic Net Prior." Technical Report 817, Department of Statistics, The Ohio State University, Columbus, Ohio, 43215. 38, 39

— (2009). "Bayesian lasso regression." *Biometrika*, 96: 835–845. 38

Kyung, M., Gill, J., Ghosh, M., and Casella, G. (2010). "Penalized Regression, Standard Errors, and Bayesian Lassos." *Bayesian Analysis*, 5: 369 – 412. 39

Li, Q. and Lin, N. (2010). "The Bayesian Elastic Net." *Bayesian Analysis*, 5: 151–170. 39

Park, T. and Casella, G. (2008). "The Bayesian lasso." *Journal of the American Statistical Association*, 103: 681–686. 38

Polson, N. G. and Scott, J. G. (2010a). "Local shrinkage rules, Lévy processes, and regularized regression." *arXiv:1010.3390v1 [stat.ME]*. 40

— (2010b). "Shrink Globally, Act Locally: Sparse Bayesian Regularization and Prediction." In Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., and West, M. (eds.), *Bayesian Statistics 9*. Oxford, U.K.: Oxford University Press. 39, 40

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. 38

Acknowledgments

The author acknowledges support from National Science Foundation grant DMS-1007682.