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