BAYESIAN STATISTICS 9,
J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid,
D. Heckerman, A. F. M. Smith and M. West (Eds.)
© Oxford University Press, 2010

REPLY TO THE DISCUSSION

We thank all of the discussants for their valuable insights and elaborations. In particular, we thank Prof. Clarke and Dr. Severinski for their conjectured extension to Theorem 3, the product of many personal discussions both in Austin and in Spain (and probably many more hours of work in Miami). The conjecture seems quite likely to be true, and strikes us as a nice way of understanding adaptive penalty functions and infinite-dimensional versions of the corresponding shrinkage priors.

Rather than respond to each of the six discussions in turn, we have grouped the comments into three rough categories.

Evaluating local shrinkage rules using extrinsic criteria

Clarke and Severinski take a predictivist view in studying the performance of various local shrinkage rules. We agree that the oracle property is of questionable relevance for either a predictivist or a Bayesian. Bridge estimators, for example, satisfy the oracle property, and yet fail many other criteria with more direct Bayesian underpinnings. Moreover, in situations with poor signal-to-noise ratio, the goal of recovering the model may directly conflict with the goal of predicting well. In our experience, Bayesian model averaging (under reasonable priors) tends to be biased downward in estimating model size—often quite severely when the data is thin. Yet it does well at prediction. We can easily envision situations where one could predict better by zeroing out coefficients that one knew with certainty to be nonzero *a priori*, and yet could not be estimated with much precision given the data.

Robert and Arbel, on the other hand, take a classical decision-theoretic view in wondering whether sparsity and minimaxity can be formally connected. We agree with their reservations about MAP estimators and global shrinkage rules; this is why we recommend that they never be used as default tools for sparse problems. MAP estimators have been a special target of scrutiny ever since Bayesians first scratched their heads in confusion at the popularity of the lasso, and we cannot improve upon the list of problems cited by Robert, Arbel, and Hans.

Global shrinkage rules pose issues that are more subtle. To be sure, many common global shrinkage estimators are inadmissible. The James–Stein estimator, for example, is based on the model $y_i = \beta_i + \epsilon_i$ and $\beta_i \sim \mathcal{N}(0, \tau^2)$, with all $\lambda_i \equiv 1$. This dominates the plain MLE but loses admissibility because a plug-in estimate $\hat{\tau}$ of the global shrinkage parameter is used. As Tiao and Tan (1965) point out, the mode of $p(\tau^2 \mid y)$ is zero exactly when the James–Stein shrinkage weight turns negative (condition 6.2). The positive-part James–Stein estimator, though it improves upon the original version, still fails the necessary smoothness constraints.

But the problem runs deeper than mere inadmissibility. The so-called "r-spike signal" shows why sparsity cannot be handled well using rules that apply the same shrinkage weight to all components of $\boldsymbol{\beta}$. Let the true p-dimensional parameter be $\boldsymbol{\beta}_r = (\sqrt{p/r}, \ldots, \sqrt{p/r}, 0, \ldots, 0)$, with r nonzero components. Regardless of $\boldsymbol{\beta}$,

$$\frac{p\|\boldsymbol{\beta}\|^2}{p + \|\boldsymbol{\beta}\|^2} \leq R\left(\hat{\boldsymbol{\beta}}^{JS}, \boldsymbol{\beta}\right) \leq 2 + \frac{p\|\boldsymbol{\beta}\|^2}{p + \|\boldsymbol{\beta}\|^2}\,,$$

meaning that $R(\hat{\boldsymbol{\beta}}^{JS}, \boldsymbol{\beta}_r) \geq (p/2)$ for the r-spike signal. Simple thresholding can beat this, with risk $\sqrt{\log p}$. Different components of $\boldsymbol{\beta}$, in other words, must be shrunk by different amounts, which is exactly what global-local shrinkage rules are designed to do.

Yet as Ed George pointed out during oral discussion of our paper at the Valencia meeting, this componentwise differential shrinkage is quite likely to compromise minimaxity. For these reasons, we see the connection between minimaxity and sparsity as allegorical, rather than formal. The use of heavy-tailed priors for constructing global shrinkage estimators has a long history, and we thank the discussants for their many additional references to this literature. As our results show, many of the priors explored years ago can enjoy fruitful second careers when used for local variance components in sparse problems, even if the resulting Bayes estimators are not provably minimax, or even minimax at all.

It is also possible, of course, to proceed in the other direction. In our work on local shrinkage rules, we discovered a class of scale-mixture priors that we call the hypergeometric inverted-beta class. This class generalizes the work of Maruyama [2004], in addition to the beta-prime distribution described by Pericchi. Many members of this class have excellent classical risk properties when used as a prior for a global variance component [Polson and Scott, 2010].

Bayesian robustness and the problem of hyperparameters

We wholeheartedly agree with Griffin and Brown's observation that a global shrinkage parameter is also needed, in addition to the p local shrinkage parameters—hence the title of our paper.

To see the role played by our global parameter τ in adapting to the overall sparsity level, let $\kappa_i = 1/(1+\tau^2\lambda_i^2)$, and let $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_p)$. For the horseshoe prior, $p(\lambda_i) \propto 1/(1+\lambda_i^2)$, and thus

$$p(\kappa_i \mid \tau) \propto \kappa_i^{-1/2}(1-\kappa_i)^{-1/2} \; \frac{1}{1+(\tau^2-1)\kappa_i} \, .$$

This in turn leads to

$$p(\mathbf{y}, \boldsymbol{\kappa}, \tau^2) \propto p(\tau^2) \; \tau^p \; \prod_{i=1}^{p} \frac{e^{-\kappa_i y_i^2/2}}{\sqrt{1-\kappa_i}} \; \prod_{i=1}^{p} \frac{1}{\tau^2\kappa_i + 1 - \kappa_i} \, .$$

The global shrinkage parameter τ is thus estimated by the sparsity of the whole vector: if p is large, the conditional posterior distribution for τ^2, given $\boldsymbol{\kappa}$, is well approximated by substituting $\bar{\kappa} = p^{-1}\sum_{i=1}^{p} \kappa_i$ for each κ_i. Up to the contribution of the prior and a constant term, this gives

$$\begin{aligned} p(\tau^2 \mid \boldsymbol{\kappa}) \quad &\approx \quad (\tau^2)^{-p/2} \; \left(1 + \frac{1-\bar{\kappa}}{\tau^2\bar{\kappa}}\right)^{-p} \\ &\approx \quad (\tau^2)^{-p/2} \; \exp\left\{ -\frac{1}{\tau^2} \; \frac{p(1-\bar{\kappa})}{\bar{\kappa}} \right\} \, , \end{aligned}$$

or approximately a $\mathrm{Ga}(\frac{p+2}{2}, \frac{p-p\bar{\kappa}}{\bar{\kappa}})$ distribution for $1/\tau^2$. When $\boldsymbol{\beta}$ is sparse, then $\bar{\kappa}$ will be close to 1, and τ^2 will be very small with high probability.

There is also the matter of choosing a prior for τ. Griffin and Brown characterize our approach as involving a "default, vague" choice of prior. While we agree that our goal is to find a sensible default, we disagree that our recommended priors are vague.

Our approach in Section 2.4 mirrors that of Jeffreys: to scale the prior for the $\beta_i's$ using the error variance σ^2, which provides the only obvious source of information about the natural scale of the problem. A vague prior when the observations are measured in miles will be a tight prior if the units are changed to millimeters. Our recommendations for the conditional prior $p(\tau^2 \mid \sigma^2)$ will implicitly rescale the prior for β_i using the observables y_i, and will elegantly solve the problem.

Rather, it will solve the problem provided that one uses a prior with sufficiently heavy tails not to be too dogmatic! Heavy-tailed priors need not be vague, but as Pericchi notes, they are an elegant hedge against model and prior misspecification. We echo his hope that the importance of robustness when choosing priors will become even more widely recognized, since models are only getting more complicated.

We would also call attention to a point raised by Hans about the advantages of estimating τ by fully Bayes methods (see the references in his discussion). Typically the normalizing constant $C(\tau)$ in the marginal posterior distribution for $\boldsymbol{\beta}$ contains extra information about τ. This information is often ignored by empirical-Bayes methods, but is naturally incorporated into the fully Bayes answer.

Extensions to wider classes and harder problems

Clarke/Severini and Griffin/Brown raise issues with the proposed methodology when $p > n$, and we agree that there are many wide-open issues associated with this difficult problem. We are particularly intrigued by Griffin and Brown's comments about the potential undesirability of heavy tails in $p > n$ problems; this point merits serious investigation.

Of course, in our formulation of the sparse normal-means problem, $n = 1$ regardless of p; this is a simple example of a $p > n$ problem where heavy tails are crucial. We suspect the issue has more to due with the complexity of the design matrix, an important point raised by Hans. This is borne out by our (admittedly incomplete) reading of the classical literature on penalty functions and the oracle property, where simplifying assumptions about the asymptotic behavior of the design matrix must inevitably be made.

Clyde and Wolpert propose priors directly for the β_i's based on a stochastic integral of an indicator function with respect to a random measure. These LARK models offer considerable promise for dealing with some of the issues raised by the discussants concerning cases where prior independence is "undesirable if not unbelievable." In particular, we hope to see further investigations of models that use generating functions with marks, periodicities, covariates, or other elaborations. With these tools in hand, a whole new approach to Bayesian nonparametrics—that is, one not based upon countable discrete mixtures—may be possible.

REFERENCES

Y. Maruyama. Stein's idea and minimax admissible estimation of a multivariate normal mean. *Journal of Multivariate Analysis*, 88(2):320–34, 2004.

N. G. Polson and J. G. Scott. On the half-cauchy prior for a global scale parameter. Technical report, University of Texas at Austin, 2010.