Bayesian Analysis (2012) **7**, Number 2, pp. 1–16

On the Half-Cauchy Prior for a Global Scale Parameter

Nicholas G. Polson[*] and James G. Scott[†]

Abstract. This paper argues that the half-Cauchy distribution should replace the inverse-Gamma distribution as a default prior for a top-level scale parameter in Bayesian hierarchical models, at least for cases where a proper prior is necessary. Our arguments involve a blend of Bayesian and frequentist reasoning, and are intended to complement the case made by Gelman (2006) in support of folded-t priors. First, we generalize the half-Cauchy prior to the wider class of hypergeometric inverted-beta priors. We derive expressions for posterior moments and marginal densities when these priors are used for a top-level normal variance in a Bayesian hierarchical model. We go on to prove a proposition that, together with the results for moments and marginals, allows us to characterize the frequentist risk of the Bayes estimators under all global-shrinkage priors in the class. These results, in turn, allow us to study the frequentist properties of the half-Cauchy prior versus a wide class of alternatives. The half-Cauchy occupies a sensible middle ground within this class: it performs well near the origin, but does not lead to drastic compromises in other parts of the parameter space. This provides an alternative, classical justification for the routine use of this prior. We also consider situations where the underlying mean vector is sparse, where we argue that the usual conjugate choice of an inverse-gamma prior is particularly inappropriate, and can severely distort inference. Finally, we summarize some open issues in the specification of default priors for scale terms in hierarchical models.

Keywords: hierarchical models; normal scale mixtures; shrinkage

1 Introduction

Consider a normal hierarchical model where, for $i = 1, \ldots, p$,

$$\begin{aligned} (y_i \mid \beta_i, \sigma^2) &\sim \mathrm{N}(\beta_i, \sigma^2) \\ (\beta_i \mid \lambda^2, \sigma^2) &\sim \mathrm{N}(0, \lambda^2\sigma^2) \\ p(\lambda^2, \sigma^2) &= p(\lambda^2)\, p(\sigma^2)\,. \end{aligned}$$

This prototype case embodies a general problem in Bayesian inference: how to choose default priors $p(\lambda^2)$ and $p(\sigma^2)$ for top-level variances in a hierarchical model.

The routine use of Jeffreys' prior for the error variance, $p(\sigma^2) \propto \sigma^{-2}$, poses no practical issues. This is not the case for $p(\lambda^2)$, however, as the improper prior $p(\lambda^2) \propto$

[*] Booth School of Business, Chicago, IL, nicholas.polson@chicagobooth.edu

[†] McCombs School of Business, University of Texas at Austin, Austin, TX, james.scott@mccombs.utexas.edu

ba0001

λ^{-2} leads to an improper posterior. This can be seen from the marginal likelihood:

$$p(y \mid \lambda^2) \propto \prod_{i=1}^{p} (1+\lambda^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2} \sum_{i=1}^{p} \frac{y_i^2}{1+\lambda^2} \right) ,$$

where we have taken $\sigma^2 = 1$ for convenience. This is positive at $\lambda^2 = 0$; therefore, whenever the prior $p(\lambda^2)$ fails to be integrable at the origin, so too will the posterior. A number of default choices have been proposed to overcome this issue. A classic reference is Tiao and Tan (1965); a very recent one is Morris and Tang (2011), who use a flat prior $p(\lambda^2) \propto 1$.

We focus on a proposal by Gelman (2006), who studies the class of half-t priors for the scale parameter λ:

$$p(\lambda \mid d) \propto \left(1 + \frac{\lambda^2}{d} \right)^{-(d+1)/2}$$

for a given d. The half-t prior has the appealing property that its density evaluates to a nonzero constant at $\lambda = 0$. This distinguishes it from the usual conjugate choice of an inverse-gamma prior for λ^2, whose density vanishes at $\lambda = 0$. As Gelman (2006) points out, posterior inference under these priors is no more difficult than it is under an inverse-gamma prior, using the simple trick of parameter expansion.

These facts lead to a simple, compelling argument against the use of the inverse-gamma prior for variance terms in models such as that above. Since the marginal likelihood of the data, considered as a function of λ, does not vanish when $\lambda = 0$, neither should the prior density $p(\lambda)$. Otherwise, the posterior distribution for λ will be inappropriately biased away from zero. This bias, moreover, is most severe near the origin, precisely in the region of parameter space where the benefits of shrinkage become most pronounced.

This paper studies the special case of a half-Cauchy prior for λ with three goals in mind. First, we embed it in the wider class of hypergeometric inverted-beta priors for λ^2, and derive expressions for the resulting posterior moments and marginal densities. Second, we derive expressions for the classical risk of Bayes estimators arising from this class of priors. In particular, we prove a result that allows us to characterize the improvements in risk near the origin (that is, when $\boldsymbol{\beta}$ is nearly 0 in Euclidean norm) that are possible using the wider class. Having proven our risk results for all members of this wider class, we then return to the special case of the half-Cauchy. We find that the frequentist risk profile of the resulting Bayes estimator is quite favorable: it is admissible, but otherwise similar to the James–Stein estimator. Therefore Bayesians can be comfortable using the prior on purely frequentist grounds.

Third, we attempt to provide insight about the use of such priors in situations where $\boldsymbol{\beta}$ is expected to be sparse. We find that the arguments of Gelman (2006) in favor of the half-Cauchy are, if anything, amplified in the presence of sparsity, and that the inverse-gamma prior can have an especially distorting effect on inference for sparse signals.

Overall, our results provide a complementary set of arguments to those of Gelman (2006) that support the routine use of the half-Cauchy prior: its excellent frequentist risk properties, and its sensible behavior in the presence of sparsity compared to the usual conjugate alternative. Bringing all these arguments together, we contend that the half-Cauchy prior is a sensible default choice for a top-level variance in Gaussian hierarchical models. We echo the call for it to replace inverse-gamma priors in routine use (e.g. Spiegelhalter et al. 1994, 2003; Park and Casella 2008), particularly given the availability of a simple parameter-expanded Gibbs sampler for posterior computation (Gelman 2006).

2 Inverted-beta priors and their generalizations

Consider the family of inverted-beta priors for λ^2:

$$p(\lambda^2) = \frac{(\lambda^2)^{b-1} \ (1+\lambda^2)^{-(a+b)}}{\mathrm{Be}(a,b)} \, ,$$

where $\mathrm{Be}(a,b)$ denotes the beta function, and where a and b are positive reals. A half-Cauchy prior for λ corresponds to an inverted-beta prior for λ^2 with $a = b = 1/2$. This family also generalizes the robust priors of Strawderman (1971) and Berger (1980); the normal-exponential-gamma prior of Griffin and Brown (2012); and the horseshoe prior of Carvalho et al. (2010). The inverted-beta distribution is also known as the beta-prime or Pearson Type VI distribution. An inverted-beta random variable is equal in distribution to the ratio of two gamma-distributed random variables having shape parameters a and b, respectively, along with a common scale parameter.

The inverted-beta family is itself a special case of a new, wider class of hypergeometric inverted-beta distributions having the following probability density function:

$$p(\lambda^2) = C^{-1} (\lambda^2)^{b-1} \ (\lambda^2+1)^{-(a+b)} \ \exp\left\{ -\frac{s}{1+\lambda^2} \right\} \left\{ \tau^2 + \frac{1-\tau^2}{1+\lambda^2} \right\}^{-1} \, , \qquad (1)$$

for $a > 0$, $b > 0$, $\tau^2 > 0$, and $s \in \mathcal{R}$. This comprises a wide class of priors leading to posterior moments and marginals that can be expressed using confluent hypergeometric functions. In Appendix 5 we give details of these computations, which yield

$$C = e^{-s} \ \mathrm{Be}(a,b) \ \Phi_1(b, 1, a+b, s, 1 - 1/\tau^2) \, , \qquad (2)$$

where Φ_1 is the degenerate hypergeometric function of two variables (Gradshteyn and Ryzhik 1965, 9.261). This function can be calculated accurately and rapidly by transforming it into a convergent series of ${}_2F_1$ functions (S9.2 of Gradshteyn and Ryzhik 1965; Gordy 1998), making evaluation of (2) quite fast for most choices of the parameters.

Both τ and s are global scale parameters, and do not control the behavior of $p(\lambda)$ at 0 or ∞. The parameters a and b are analogous to those of the beta distribution. Smaller values of a encourage heavier tails in $p(\beta)$, with $a = 1/2$, for example, yielding Cauchy-like tails. Smaller values of b encourage $p(\beta)$ to have more mass near the origin, and eventually to become unbounded; $b = 1/2$ yields, for example, $p(\beta) \approx \log(1 + 1/\beta^2)$ near 0.

We now derive expressions for the moments of $p(\boldsymbol{\beta} \mid \mathbf{y}, \sigma^2)$ and the marginal likelihood $p(\mathbf{y} \mid \sigma^2)$ for priors in this family. As a special case, we obtain the posterior mean for $\boldsymbol{\beta}$ under a half-Cauchy prior on λ.

Given λ^2 and σ^2, the posterior distribution of $\boldsymbol{\beta}$ is multivariate normal, with mean m and variance V given by

$$m = \left(1 - \frac{1}{1+\lambda^2}\right)\mathbf{y} \quad , \quad V = \left(1 - \frac{1}{1+\lambda^2}\right)\sigma^2 \, .$$

Define $\kappa = 1/(1+\lambda^2)$. By Fubini's theorem, the posterior mean and variance of $\boldsymbol{\beta}$ are

$$\begin{aligned} E(\boldsymbol{\beta} \mid \mathbf{y}, \sigma^2) &= \{1 - E(\kappa \mid \mathbf{y}, \sigma^2)\}\mathbf{y} & (3) \\ \operatorname{var}(\boldsymbol{\beta} \mid \mathbf{y}, \sigma^2) &= \{1 - E(\kappa \mid \mathbf{y}, \sigma^2)\}\sigma^2 \, , & (4) \end{aligned}$$

now conditioning only on σ^2.

It is most convenient to work with $p(\kappa)$ instead:

$$p(\kappa) \propto \kappa^{a-1} \, (1-\kappa)^{b-1} \, \left\{ \frac{1}{\tau^2} + \left(1 - \frac{1}{\tau^2}\right)\kappa \right\}^{-1} e^{-\kappa s} \, . \qquad (5)$$

The joint density for κ and $\mathbf{y}$ takes the same functional form:

$$p(y_1, \ldots, y_p, \kappa) \propto \kappa^{a'-1} \, (1-\kappa)^{b-1} \, \left\{ \frac{1}{\tau^2} + \left(1 - \frac{1}{\tau^2}\right)\kappa \right\}^{-1} e^{-\kappa s'} \, ,$$

with $a' = a + p/2$, and $s' = s + Z/2\sigma^2$ for $Z = \sum_{i=1}^{p} y_i^2$. Hence the posterior for λ^2 is also a hypergeometric inverted-beta distribution, with parameters (a', b, τ^2, s').

Next, the moment-generating function of (5) is easily shown to be

$$M(t) = e^t \, \frac{\Phi_1(b, 1, a+b, s-t, 1-1/\tau^2)}{\Phi_1(b, 1, a+b, s, 1-1/\tau^2)} \, .$$

See, for example, Gordy (1998). Expanding Φ_1 as a sum of ${}_1F_1$ functions (where ${}_1F_1$ denotes the Kummer confluent hypergeometric function) and using the differentiation rules given in Chapter 15 of Abramowitz and Stegun (1964) yields

$$E(\kappa^n \mid \mathbf{y}, \sigma^2) = \frac{(a')_n}{(a'+b)_n} \frac{\Phi_1(b, 1, a'+b+n, s', 1-1/\tau^2)}{\Phi_1(b, 1, a'+b, s', 1-1/\tau^2)} \, . \qquad (6)$$

Combining the above expression with (3) and (4) yields the conditional posterior mean and variance for $\boldsymbol{\beta}$, given $\mathbf{y}$ and σ^2. Similarly, the marginal density $p(\mathbf{y} \mid \sigma^2)$ involves the ratio of prior to posterior normalizing constants:

$$p(\mathbf{y} \mid \sigma^2) = (2\pi\sigma^2)^{-p/2} \, \exp\left(-\frac{Z}{2\sigma^2}\right) \, \frac{\mathrm{Be}(a', b)}{\mathrm{Be}(a, b)} \, \frac{\Phi_1(b, 1, a'+b, s', 1-1/\tau^2)}{\Phi_1(b, 1, a+b, s, 1-1/\tau^2)} \, .$$

Figure 1 shows a simple example of the posterior mean under the half-Cauchy prior for λ when $p = 10$, calculated for fixed σ using the results of this section.

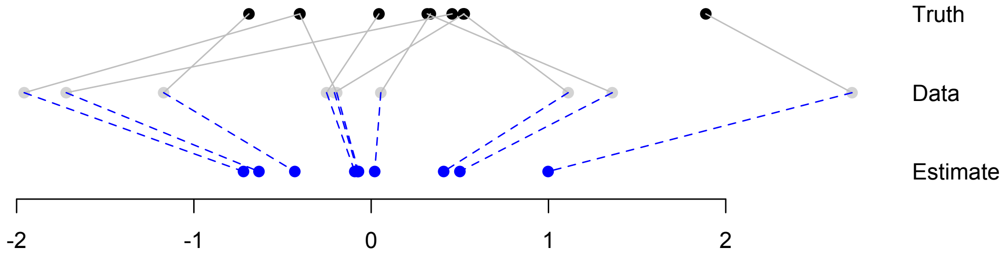

Figure 1: Ten true means drawn from a standard normal distribution; data from these means under standard normal noise; posterior means under the half-Cauchy prior for λ.

3 Classical risk results

These priors are useful in situations where standard priors like the inverse-gamma or Jeffreys' are inappropriate or ill-behaved. Non-Bayesians will find them useful for generating easily computable shrinkage estimators that have known risk properties. Bayesians will find them useful for generating computationally tractable priors for a variance parameter. We argue that these complementary but overlapping goals can both be satisfied for the special case of the half-Cauchy. To show this, we first characterize the risk properties of the Bayes estimators that result from the wider family of priors used for a normal mean under a quadratic loss. Our analysis indicates that noticeable improvements over the James–Stein estimator are possible near the origin. As Figure 2 shows, this can be done in several ways: by choosing a large relative to b, by choosing a and b both less than 1, by choosing s negative, or by choosing $\tau < 1$. Each of these choices involves a compromise somewhere else in the parameter space.

We now derive expressions for the classical risk, as a function of $\|\boldsymbol{\beta}\|$, for the resulting Bayes estimators under hypergeometric inverted-beta priors. Assume without loss of generality that $\sigma^2 = 1$, and let $p(\mathbf{y}) = \int p(\mathbf{y}|\boldsymbol{\beta})p(\boldsymbol{\beta})d\boldsymbol{\beta}$ denote the marginal density of the data. Following Stein (1981), write the the mean-squared error of the posterior mean $\hat{\boldsymbol{\beta}}$ as

$$E(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2) = p + E_{\mathbf{y}|\boldsymbol{\beta}}\left(\|\nabla \log p(\mathbf{y})\|^2 + 2\sum_{i=1}^{p} \frac{\partial}{\partial y_i} g(\mathbf{y})\right) ,$$

where ∇ denotes the gradient operator function and $\|\cdot\|$ is Euclidean norm. In turn this can be written as

$$E(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2) = p + 4E_{\mathbf{y}|\boldsymbol{\beta}}\left(\frac{\nabla^2\sqrt{p(\mathbf{y})}}{\sqrt{p(\mathbf{y})}}\right) .$$

We now state our main result concerning computation of this quantity.

Proposition 3.1. *Suppose that* $\boldsymbol{\beta} \sim N(0, \lambda^2 I)$*, that* $\kappa = 1/(1+\lambda^2)$*, and that the prior* $p(\kappa)$ *is such that* $\lim_{\kappa \to 0,1} \kappa(1-\kappa)p(\kappa) = 0$*. Define*

$$m_p(Z) = \int_0^1 \kappa^{\frac{p}{2}} e^{-\frac{Z}{2}\kappa} p(\kappa) \; d\kappa \tag{7}$$

for $Z = \sum_{i=1}^p y_i^2$*, and let* $g(Z) = E(\kappa \mid Z)$*. Then as a function of* $\boldsymbol{\beta}$*, the quadratic risk of the posterior mean under* $p(\kappa)$ *is*

$$E(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2) = p + 2E_{(Z|\boldsymbol{\beta})}\left[Z\left\{ \frac{m_{p+4}(Z)}{m_p(Z)} \right\} - pg(Z) - \frac{Z}{2}g(Z)^2 \right] , \tag{8}$$

with the expectation taken over the noncentral chi-squared distribution $Z \sim \chi^2(\|\boldsymbol{\beta}\|^2)$*, given* $\boldsymbol{\beta}$*; and with*

$$Z\left\{ \frac{m_{p+4}(Z)}{m_p(Z)} \right\} = (p+Z+4)g(Z) - (p+2) - E_{\kappa|Z}\left\{ 2\kappa(1-\kappa)\frac{p'(\kappa)}{p(\kappa)} \right\} . \tag{9}$$

Proposition 3.1 is useful because it characterizes the risk conditional on Z in terms of known quantities: the integral $m_p(Z)$, and the posterior expectation $g(Z) = E(\kappa \mid Z)$. This reduces the problem to one of evaluating the outer expectation over Z, given $\boldsymbol{\beta}$. One straightforward way to do so is to simulate draws of Z from a non-central chi-squared distribution, allowing the inner terms to be evaluated as a function of Z. An interesting comparison is with George et al. (2006), who consider Kullback–Leibler predictive risk for similar priors.

Our interest is in the special case $a = b = 1/2$, $\tau = 1$, and $s = 0$, corresponding to a half-Cauchy prior for the global scale λ. Figure 3 shows the classical risk of the Bayes estimator under this prior for $p = 7$ and $p = 15$. The risk of the James-Stein estimator is shown for comparison. These pictures look similar for other values of p, and show overall that the half-Cauchy prior for λ leads to a Bayes estimator that is competitive with the James–Stein estimator, while retaining admissibility and a fully Bayesian interpretation.

A natural question is: of all the hypergeometric inverted-beta priors, why choose the half-Cauchy? There is no iron-clad reason to do so, of course, and we can imagine many situations where subjective information would support a different choice. But in examining many other members of the class, we have observed that the half-Cauchy seems to occupy a sensible middle ground in terms of frequentist risk. To study this, we are able to appeal to the theory of the previous section. See, for example, Figure 2, which compares several members of the class for the case $p = 7$. Observe that large gains over James–Stein near the origin are possible, but only at the expense of minimaxity. The half-Cauchy, meanwhile, still improves upon the James–Stein estimator near the origin, but does not sacrifice good risk performance in other parts of the parameter space. From a purely classical perspective, it looks like a sensible default choice, appropriate for repeated general use.

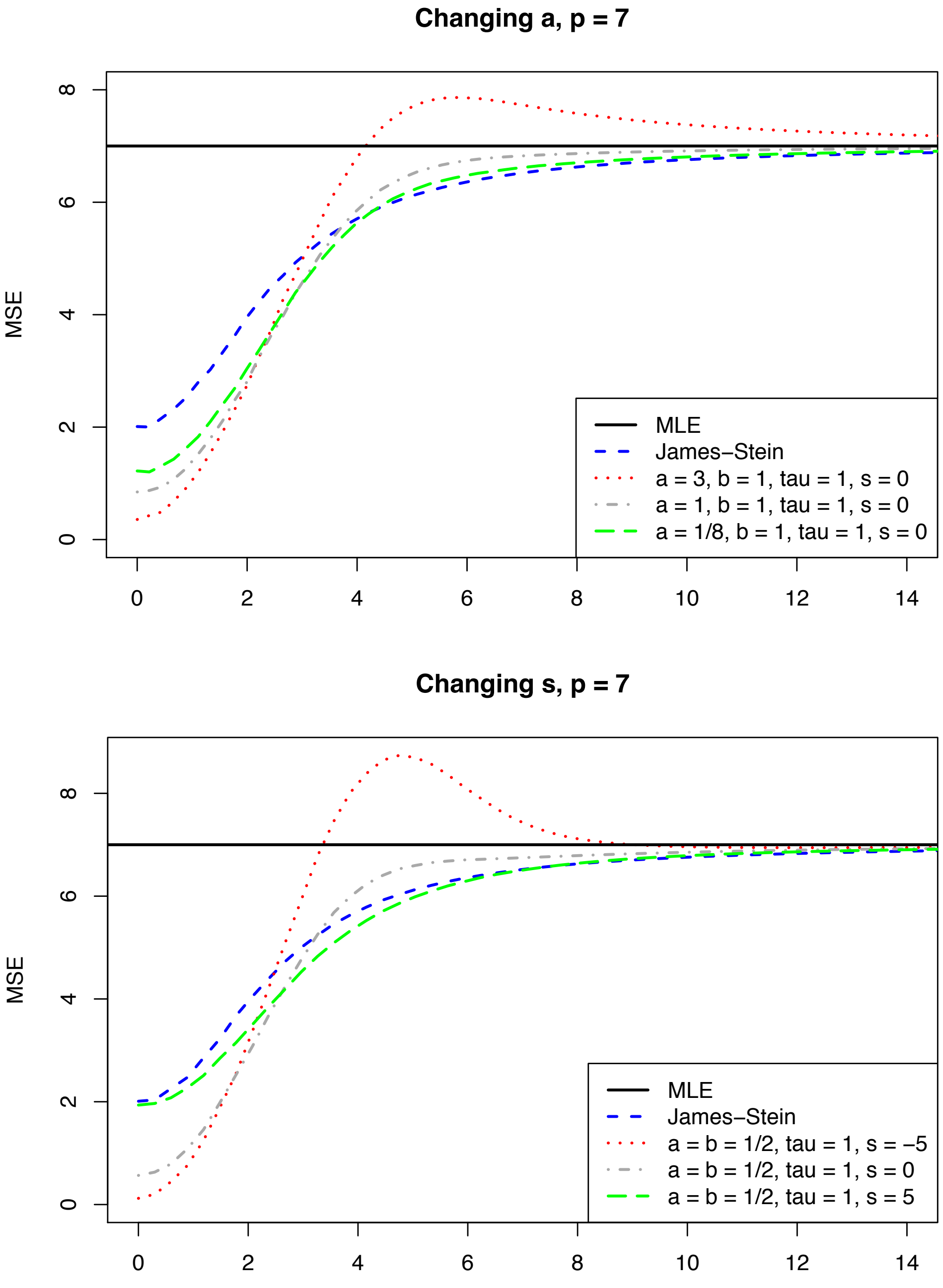

Figure 2: Mean-squared error as a function of $\|\beta\|^2$ for $p = 7$ and various cases of the hypergeometric inverted-beta hyperparameters.

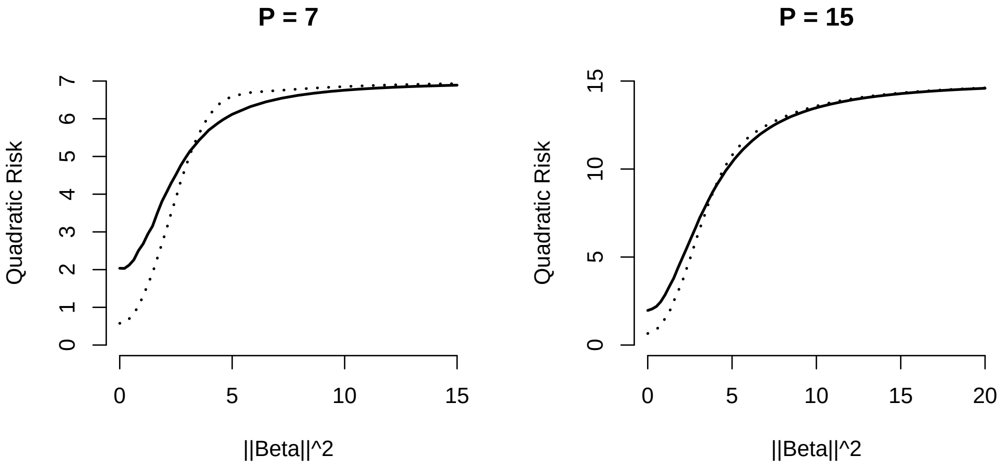

Figure 3: Mean-squared error as a function of $\|\boldsymbol{\beta}\|^2$ for $p = 7$ and $p = 15$. Solid line: James–Stein estimator. Dotted line: Bayes estimator under a half-Cauchy prior for λ.

4 Global scale parameters in local-shrinkage models

A now-canonical modification of the basic hierarchical model from the introduction involves the use of local shrinkage parameters:

$$\begin{aligned}
(y_i \mid \beta_i, \sigma^2) &\sim \mathrm{N}(\beta_i, \sigma^2) \\
(\beta_i \mid \lambda^2, u_i^2, \sigma^2) &\sim \mathrm{N}(0, \lambda^2 \sigma^2 u_i^2) \\
u_i^2 &\sim f(u_i^2) \\
\lambda^2 &\sim g(\lambda^2)\,.
\end{aligned}$$

Mixing over u_i leads to a non-Gaussian marginal for β_i, given the global parameter λ^2. For example, choosing an exponential prior for each u_i^2 results in a Laplace prior, used in the Bayesian lasso model (Park and Casella 2008; Hans 2009). This class of models provides a Bayesian alternative to penalized-likelihood estimation. When the underlying vector of means is sparse, these global-local shrinkage models can lead to large improvements in both estimation and prediction compared with pure global shrinkage rules. There is a large literature on the choice of $p(u_i^2)$, with Polson and Scott (2011) providing a recent review.

As many authors have documented, strong global shrinkage combined with heavy-tailed local shrinkage is why these sparse Bayes estimators work so well at sifting signals from noise. Intuitively, the idea is that λ acts as a global parameter that adapts to the underlying sparsity of the signal. When few signals are present, it is quite common for the marginal likelihood of $\mathbf{y}$ as a function of λ to concentrate near 0, and for the signals

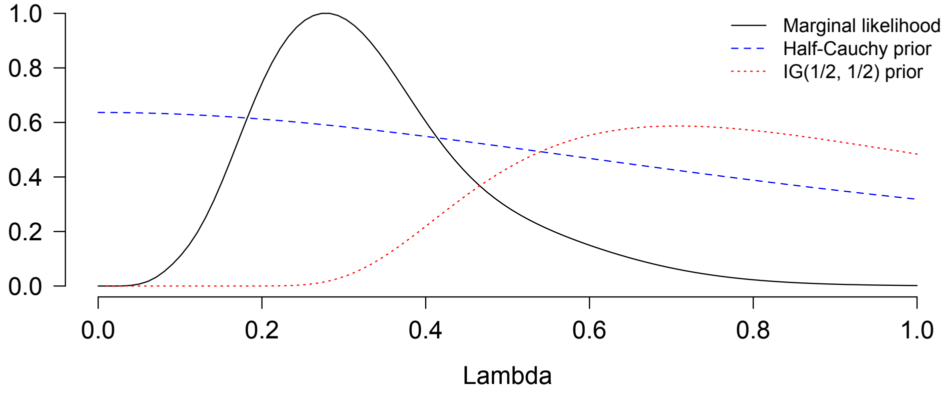

Figure 4: The black line shows the marginal likelihood of the data as a function of λ under a horseshoe prior for each β_i. (The likelihood has been renormalized to obtain a maximum of 1.) The other lines show two priors for λ: the half-Cauchy (dashed line), and that induced by an inverse-Gamma prior on λ^2 (dotted).

to be flagged via very large values of the local shrinkage parameters u_i^2. Indeed, in some cases the marginal maximum-likelihood solution can be the degenerate $\hat{\lambda} = 0$ (Tiao and Tan 1965).

The classical risk results of the previous section no longer apply to a model with these extra local-shrinkage parameters, since the marginal distribution of $\boldsymbol{\beta}$, given λ, is not multivariate normal. Nonetheless, the case of sparsity serves only to amplify the purely Bayesian argument in favor of the half-Cauchy prior for a global scale parameter—namely, the argument that $p(\lambda \mid \mathbf{y})$ should not be artificially pulled away from zero by an inverse-gamma prior.

Figure 4 vividly demonstrates this point. We simulated data from a sparse model where $\boldsymbol{\beta}$ contained the entries $(5, 4, 3, 2, 1)$ along with 45 zeroes, and where $y_{ij} \sim \mathrm{N}(0, 1)$ for $j = 1, 2, 3$. We then used Markov Chain Monte Carlo to compute the marginal likelihood of the data as a function of λ, assuming that each β_i has a horseshoe prior (Carvalho et al. 2010). This can be approximated by assuming a flat prior for λ (here truncated above at 10), and then computing the conditional likelihood $p(\mathbf{y} \mid \lambda, \sigma, u_1^2, \ldots, u_p^2)$ over a discrete grid of λ values at each step of the Markov chain. The marginal likelihood function can then be approximated as the pointwise average of the conditional likelihood over the samples from the joint posterior.

This marginal likelihood has been renormalized to obtain a maximum of 1 and then plotted alongside two alternatives: a half-Cauchy prior for λ, and the prior induced by assuming that $\lambda^2 \sim \mathrm{IG}(1/2, 1/2)$. Under the inverse-gamma prior, there will clearly be an inappropriate biasing of $p(\lambda)$ away from zero, which will negatively affect the ability of the model to handle sparsity. For data sets with even sparse signals, the distorting effect of a supposedly non-informative inverse-gamma prior will be even more pronounced, as the marginal likelihood will favor values of λ very near zero (along with a small handful of very large u_i^2 terms).

5 Discussion

On strictly Bayesian grounds, the half-Cauchy is a sensible default prior for scale parameters in hierarchical models: it tends to a constant at $\lambda = 0$; it is quite heavy-tailed; and it leads to simple conditionally conjugate MCMC routines, even in complex settings. All these desirable features are summarized by Gelman (2006). Our results give a quite different, classical justification for this prior in high-dimensional settings: its excellent quadratic risk properties. The fact that two independent lines of reasoning both lead to the same prior is a strong argument in its favor as a default proper prior for a shared variance component. We also recommend scaling the β_i's by σ, as reflected in the hierachical model from the introduction. This is the approach taken by Jeffreys (1961, Section 5.2), and we cannot improve upon his arguments.

In addition, our hypergeometric inverted-beta class provides a useful generalization of the half-Cauchy prior, in that it allows for greater control over global shrinkage through τ and s. It leads to a large family of estimators with a wide range of possible behavior, and generalizes the form noted by Maruyama (1999), which contains the positive-part James–Stein estimator as a limiting, improper case. Further study of this class may yield interesting frequentist results, quite apart from the Bayesian implications considered here. The expressions for marginal likelihoods also have connections with recent work on generalized g-priors (Maruyama and George 2010; Polson and Scott 2012). Finally, all estimators arise from proper priors on λ^2, and will therefore be admissible.

There are still many open issues in default Bayes analysis for hierarchical models that are not addressed by our results. One issue is whether to mix further over the scale in the half-Cauchy prior, $\lambda \sim \mathrm{C}^+(0, \tau)$. One possibility here is simply to let $\tau \sim \mathrm{C}^+(0, 1)$. We then get the following "double" half-Cauchy prior for λ:

$$p(\lambda) = \frac{2}{\pi^2} \int_0^\infty \frac{1}{1+\tau^2} \frac{1}{\tau(1+\frac{\lambda^2}{\tau^2})} d\tau = \frac{\ln|\lambda|}{\lambda^2 - 1} \, .$$

Admittedly, it is difficult to know where to stop in this "turtles all the way down" approach to mixing over hyperparameters. (Why not, for example, mix still further over a scale parameter for τ?) Even so, this prior has a number of appealing properties. It is proper, and therefore leads to a proper posterior; it is similar in overall shape to Jeffreys' prior; and it is unbounded at the origin, and will therefore not down-weight the marginal likelihood as much as the half-Cauchy for near-sparse configurations of $\boldsymbol{\beta}$. The implied prior on the shrinkage weight κ for the double half-Cauchy is

$$p(\kappa) \propto \frac{\ln\left(\frac{1-\kappa}{\kappa}\right)}{1-2\kappa} \frac{1}{\sqrt{\kappa(1-\kappa)}} \, .$$

This is like the horseshoe prior on the shrinkage weight (Carvalho et al. 2010), but with an extra factor that comes from the fact that one is letting the scale itself be random with a $\mathrm{C}^+(0,1)$ prior.

We can also transform to the real line by letting $\psi = \log \lambda^2$. For the half-Cauchy prior $p(\lambda) \propto 1/(1+\lambda^2)$ this transformation leads to

$$p(\psi) \propto \frac{e^{\frac{\psi}{2}}}{1+e^{\psi}} = \left(e^{\frac{\psi}{2}} + e^{-\frac{\psi}{2}}\right)^{-1} = \mathrm{sech}\left(\frac{\psi}{2}\right) .$$

This is the hyperbolic secant distribution, which may provide fertile ground for further generalizations or arguments involving sensible choices for a default prior.

A more difficult issue concerns the prior scaling for λ in the presence of unbalanced designs—that is, when $y_{ij} \sim \mathrm{N}(\beta_i, \sigma^2)$ for $j = 1, \ldots, n_i$, and the n_i's are not necessarily equal. In this case most formal non-informative priors for λ, such as the reference priors for particular parameter orderings, involve complicated functions of the n_i's (Yang and Berger 1997). These expressions emerge from the reference-prior formalism that, in turn, embodies a particular operational definition of "non-informative."

We have focused on default priors that occupy a middle ground between formal non-informative analysis and pure subjective Bayes. This is clearly an important situation for the many practicing Bayesians who do not wish to use noninformative priors, whether for practical, mathematical, or philosophical reasons. An example of a situation in which formal noninformative priors for λ should not be used on mathematical grounds is when $\boldsymbol{\beta}$ is expected to be sparse; see Scott and Berger (2006) for a discussion of this issue in the context of multiple-testing. It is by no means obvious how, or even whether, the n_i's should play a role in scaling λ within this (admittedly ill-defined) paradigm of "default" Bayes.

Finally, another open issue is the specification of default priors for scale parameters in non-Gaussian models. For example, in logistic regression, the likelihood is highly sensitive to large values of the underyling linear predictor. It is therefore not clear whether something so heavy-tailed as the half-Cauchy is an appropriate prior for the global scale term for logistic regression coefficients. All of these issues merit further research.

References

Abramowitz, M. and Stegun, I. A. (eds.) (1964). *Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables*, volume 55 of *Applied Mathematics Series*. Washington, DC: National Bureau of Standards. Reprinted in paperback by Dover (1974); on-line at http://www.math.sfu.ca/$\sim$cbm/aands/. 4, 13

Berger, J. O. (1980). "A robust generalized Bayes estimator and confidence region for a multivariate normal mean." *The Annals of Statistics*, 8(4): 716–761. 3

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). "The horseshoe estimator for sparse signals." *Biometrika*, 97(2): 465–80. 3, 9, 11

Fourdrinier, D., Strawderman, W., and Wells, M. T. (1998). "On the construction of Bayes minimax estimators." *The Annals of Statistics*, 26(2): 660–71. 14

Gelman, A. (2006). "Prior distributions for variance parameters in hierarchical models." *Bayesian Analysis*, 1(3): 515–33. 1, 2, 3, 10

George, E. I., Liang, F., and Xu, X. (2006). "Improved minimax predictive densities under Kullback-Leibler loss." *The Annals of Statistics*, 34(1): 78–91. 6

Gordy, M. B. (1998). "A generalization of generalized beta distributions." Finance and Economics Discussion Series 1998-18, Board of Governors of the Federal Reserve System (U.S.). 3, 4, 13, 14

Gradshteyn, I. and Ryzhik, I. (1965). *Table of Integrals, Series, and Products*. Academic Press. 3, 13

Griffin, J. and Brown, P. (2012). "Alternative prior distributions for variable selection with very many more variables than observations." *Australian and New Zealand Journal of Statistics*. (to appear). 3

Hans, C. M. (2009). "Bayesian Lasso Regression." *Biometrika*, 96(4): 835–45. 8

Jeffreys, H. (1961). *Theory of Probability*. Oxford University Press, 3rd edition. 10

Maruyama, Y. (1999). "Improving on the James–Stein estimator." *Statistics and Decisions*, 14: 137–40. 10

Maruyama, Y. and George, E. I. (2010). "gBF: A Fully Bayes Factor with a Generalized g-prior." Technical report, University of Tokyo, arXiv:0801.4410v2. 10

Morris, C. and Tang, R. (2011). "Estimating Random Effects via Adjustment for Density Maximization." *Statistical Science*, 26(2): 271–87. 2

Park, T. and Casella, G. (2008). "The Bayesian Lasso." *Journal of the American Statistical Association*, 103(482): 681–6. 3, 8

Polson, N. G. and Scott, J. G. (2011). "Shrink globally, act locally: sparse Bayesian regularization and prediction." In Bernardo, J., Bayarri, M., Berger, J. O., Dawid, A., Heckerman, D., Smith, A., and West, M. (eds.), *Proceedings of the 9th Valencia World Meeting on Bayesian Statistics*. Oxford University Press. 8

— (2012). "Local shrinkage rules, Lévy processes, and regularized regression." *Journal of the Royal Statistical Society (Series B)*. (to appear). 10

Scott, J. G. and Berger, J. O. (2006). "An exploration of aspects of Bayesian multiple testing." *Journal of Statistical Planning and Inference*, 136(7): 2144–2162. 11

Spiegelhalter, D. J., Thomas, A., Best, N. G., Gilks, W. R., and Lunn, D. (1994, 2003). *BUGS: Bayesian inference using Gibbs sampling*. MRC Biostatistics Unit, Cambridge, England. 3

Stein, C. (1981). "Estimation of the mean of a multivariate normal distribution." *The Annals of Statistics*, 9: 1135–51. 5

Strawderman, W. (1971). "Proper Bayes minimax estimators of the multivariate normal mean." *The Annals of Statistics*, 42: 385–8. 3

Tiao, G. C. and Tan, W. (1965). "Bayesian analysis of random-effect models in the analysis of variance. I. Posterior distribution of variance components." *Biometrika*, 51: 37–53. 2, 9

Yang, R. and Berger, J. O. (1997). "A Catalog of Noninformative Priors." Technical Report 42, Duke University Department of Statistical Science. 11

Appendix 1

In this appendix we present the details for computings moments and marginals of the hypergeometric-beta distribution. The normalizing constant in (2) is

$$C = \int_0^1 \kappa^{\alpha-1} \, (1-\kappa)^{\beta-1} \, \left\{ \frac{1}{\tau^2} + \left(1 - \frac{1}{\tau^2}\right) \kappa \right\}^{-1} \exp(-s\kappa) \, \mathrm{d}\kappa \, . \tag{10}$$

Let $\eta = 1 - \kappa$. Using the identity that $e^x = \sum_{m=0}^{\infty} x^m/m!$, we obtain

$$C = e^{-s} \sum_{m=0}^{\infty} \left[\frac{s^m}{m!} \int_0^1 \eta^{\beta+m-1} (1-\eta)^{\alpha-1} \{1 - (1 - 1/\tau^2)\eta\}^{-1} \, \mathrm{d}\eta \right] .$$

Using properties of the hypergeometric function ${}_2F_1$ (Abramowitz and Stegun 1964, S15.1.1 and S15.3.1), this becomes, after some straightforward algebra,

$$C = e^{-s} \, \mathrm{Be}(\alpha, \beta) \, \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} \frac{(\beta)_{m+n}}{(\alpha+\beta)_{m+n} \, m! \, n!} \, s^m \, (1 - 1/\tau^2)^m \, , \tag{11}$$

where $(a)_n$ is the rising factorial. Appendix C of Gordy (1998) proves that, for all $\alpha > 0$, $\beta > 0$, and $1/\tau^2 > 0$, the nested series in (11) converges to a positive real number, yielding

$$C = e^{-s} \, \mathrm{Be}(\alpha, \beta) \, \Phi_1(\beta, 1, \alpha+\beta, s, 1 - 1/\tau^2) \, , \tag{12}$$

where Φ_1 is the degenerate hypergeometric function of two variables (Gradshteyn and Ryzhik 1965, 9.261).

The Φ_1 function can be written as a double hypergeometric series,

$$\Phi_1(\alpha, \beta; \gamma; x, y) = \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} \frac{(\alpha)_{m+n} (\beta)_n}{(\gamma)_{m+n} m! n!} \, y^n \, x^m \, . \tag{13}$$

We use three different representations of $\Phi_1(\alpha, \beta, \gamma, x, y)$ for handling different combinations of arguments, all from Gordy (1998). When $0 \leq y < 1$ and $x \geq 0$,

$$\Phi_1(\alpha, \beta, \gamma, x, y) = \sum_{n=0}^{\infty} \frac{(\alpha)_n}{(\gamma)_n} \frac{x^n}{n!} \; {}_2F_1(\beta, \alpha + n; \gamma + n; y) \,. \tag{14}$$

When $0 \leq y < 1$ and $x < 0$,

$$\Phi_1(\alpha, \beta, \gamma, x, y) = e^x \; \sum_{n=0}^{\infty} \frac{(\gamma - \alpha)_n}{(\gamma)_n} \frac{(-x)^n}{n!} \; {}_2F_1(\beta, \alpha; \gamma + n; y) \,. \tag{15}$$

Finally, when $y < 0$,

$$\Phi_1(\alpha, \beta, \gamma, x, y) = e^x \; (1 - y)^{-\beta} \; \Phi_1(\tilde{\alpha}, \beta, \gamma, -x, \tilde{y}) \,, \tag{16}$$

where $\tilde{\alpha} = \gamma - \alpha$ and $\tilde{y} = y/(y - 1)$. Then either (14) or (15) may be used to evaluate the righthand side of (16), depending on the sign of x.

Appendix 2

To prove Proposition 3.1, we begin with Stein's decomposition of risk. Following Equation (10) of Fourdrinier et al. (1998), we have

$$\|\nabla p(\mathbf{y})\| = \|\mathbf{y}\| \int_0^1 \kappa^{\frac{p}{2}+1} p(\kappa) e^{-\frac{Z}{2}\kappa} \, \mathrm{d}\kappa \,.$$

The score can be written as

$$\frac{\|\nabla p(\mathbf{y})\|}{p(\mathbf{y})} = \|\mathbf{y}\| \frac{m_{p+2}(\|\mathbf{y}\|)}{m_p(\|\mathbf{y}\|)} = \|\mathbf{y}\| \; E(\kappa \mid Z) \,,$$

with $m_p(\cdot)$ defined as in Equation (7). We may also write the Laplacian term in Stein's decomposition as

$$\Delta p(\mathbf{y}) = \int_0^1 (Z\kappa - p) \, \kappa^{\frac{p}{2}+1} p(\kappa) e^{-\frac{Z}{2}\kappa} \, \mathrm{d}\kappa \,.$$

Combining these terms, we have,

$$\begin{aligned} \frac{\Delta p(\mathbf{y})}{p(\mathbf{y})} &= \frac{\int_0^1 (Z\kappa - p) \, \kappa^{\frac{p}{2}+1} p(\kappa) e^{-\frac{Z}{2}\kappa} \, \mathrm{d}\kappa}{\int_0^1 \kappa^{\frac{p}{2}} p(\kappa) e^{-\frac{Z}{2}\kappa} \, \mathrm{d}\kappa} \\ &= Z \frac{m_{p+4}(Z)}{m_p(Z)} - p \frac{m_{p+2}(Z)}{m_p(Z)} \,. \end{aligned}$$

The risk term $\Delta \sqrt{p(\mathbf{y})} / \sqrt{p(\mathbf{y})}$ is then computed using the identity

$$\frac{\nabla^2 \sqrt{p(\mathbf{y})}}{\sqrt{p(\mathbf{y})}} = \frac{1}{2} \left[\frac{\Delta p(\mathbf{y})}{p(\mathbf{y})} - \frac{1}{2} \left\{ \frac{\|\nabla p(\mathbf{y})\|}{p(\mathbf{y})} \right\}^2 \right] ,$$

which reduces to

$$\frac{1}{2}\left\{Z\frac{m_{p+4}(Z)}{m_p(Z)} - pg(Z) - \frac{Z}{2}g(Z)^2\right\}$$

for $g(Z) = E(\kappa \mid Z)$.

Secondly, note that

$$Z\{m_{p+2}(Z) - m_{p+4}(Z)\} = 2\int_0^1 \kappa^{\frac{p}{2}+1}(1-\kappa)p(\kappa)d\left(-e^{-\frac{Z}{2}\kappa}\right) . \tag{17}$$

Therefore,

$$\begin{aligned} Z\left\{\frac{m_{p+2}(Z) - m_{p+4}(Z)}{m_p(Z)}\right\} &= \int_0^1 \left\{(p+2)(1-\kappa) - 2\kappa + 2\kappa(1-\kappa)\frac{p'(\kappa)}{p(\kappa)}\right\} h(\kappa)\ \mathrm{d}\kappa \\ h(\kappa) &= \frac{\kappa^{\frac{p}{2}} e^{-\frac{Z}{2}\kappa} p(\kappa)}{m_p(Z)} . \end{aligned}$$

Then under the assumption that $\lim_{\kappa \to 0,1} \kappa(1-\kappa)p(\kappa) = 0$, integration by parts gives (9). Hence

$$E(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2) = p + 2E_{Z|\boldsymbol{\beta}}\left[(Z+4)g(Z) - (p+2) - \frac{Z}{2}g(Z)^2 - E_{\kappa|Z}\left\{2\kappa(1-\kappa)\frac{p'(\kappa)}{p(\kappa)}\right\}\right] .$$

Acknowledgments

The authors would like to thank an anonymous referee, the editor, the associate editor, and the production editor, whose many useful suggestions have helped us to improve the manuscript.