

Alternative Global–Local Shrinkage Priors Using Hypergeometric–Beta Mixtures

NICHOLAS G. POLSON

*Booth School of Business
University of Chicago*

JAMES G. SCOTT

*McCombs School of Business
University of Texas at Austin*

August 2009

Abstract

This paper introduces an approach to estimation in possibly sparse data sets using shrinkage priors based upon the class of hypergeometric-beta distributions. These widely applicable priors turn out to be a four-parameter generalization of the beta family, and are pseudo-conjugate: they cannot themselves be expressed in closed form, but they do yield tractable moments and marginal likelihoods when used as priors for the mean of a normal distribution. These priors are useful in situations where standard priors are inappropriate or ill-behaved. Non-Bayesians will find these priors useful for generating easily computable shrinkage estimators that have excellent risk properties. Bayesians will find them useful for generating computationally tractable priors for a variance parameter. We illustrate the use of these priors on a variety of global and local shrinkage problems, and we prove a theorem that characterizes their risk proprieties when used for estimation of a normal mean under a quadratic loss function.

Keywords: multiple testing; normal scale mixtures; shrinkage; sparsity

1 Introduction

This paper considers the classic normal-means problem, where $y_i \sim N(\theta_i, \sigma^2)$, and where the objective is to estimate $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)$. Our goal is to introduce a new flexible class of priors for $\boldsymbol{\theta}$ based upon hypergeometric–beta mixtures of normals. This class, which we will soon define, is of potential interest both to Bayesians and to non-Bayesians. Non-Bayesians will find these priors useful for the construction of shrinkage estimators that are easy to compute and that have excellent risk properties under quadratic loss. Bayesians, on the other hand, will find them useful for the many situations where computationally tractable priors for a variance component are required, but where the usual

conjugate choices are inappropriate or ill-behaved. They are especially useful for modeling sparse, heavy-tailed signals. The family simultaneously generalizes the robust priors of Strawderman (1971) and the horseshoe prior of Carvalho et al. (2008).

We build on a long line of previous work on the normal-means problem, with a focus on two main areas:

Classical shrinkage estimation, where $(\theta_i | \lambda^2) \sim N(0, \lambda^2)$. Here the goal is to construct an estimator or prior for the global shrinkage parameter λ^2 that, in turn, yields a good estimator for θ in terms of quadratic loss (e.g. James and Stein, 1961; Strawderman, 1971; Stein, 1981; Fourdrinier et al., 1998). The fundamental insight of this literature is that shared dependence upon a global parameter λ^2 , which is to be estimated using the data, can yield drastic improvements in risk over traditional estimators. These formal results, however, do not typically extend to sparse configurations of θ . Indeed, it is known that global-shrinkage estimators can perform quite poorly in these situations.

Models for sparse signals, where $(\theta_i | \lambda_i^2) \sim N(0, \lambda_i^2)$, and where the goal is to construct a prior for the local shrinkage parameters λ_i^2 . The fundamental insight of this literature is that mixing over a set of local variances yields a non-normal density that substantially improves upon global shrinkage priors when estimating signals that are sparse, heavy-tailed, or both. Examples can be found in the work of Tibshirani (1996), Denison and George (2000), Tipping (2001), Figueiredo (2003), Park and Casella (2008), Hans (2008), and Carvalho et al. (2008). This literature usually does not, however, consider global shrinkage to be of central concern; any hyperparameters that govern the prior distribution of the λ_i 's are usually treated as nuisance parameters to be handled by, for example, empirical Bayes or cross-validation.

This paper adopts a combined “global–local” view of shrinkage: we use a new family of local-shrinkage priors to model sparse signals, but attempt to do so in a way that pays close attention to the lessons about global shrinkage to be found in the literature on classical estimation under quadratic loss.

Specifically, we assume an exchangeable mixture prior for θ of the form

$$p(\theta_1, \dots, \theta_p | \Gamma) = \prod_{t=1}^p p(\theta_t | \Gamma). \tag{1}$$

where Γ is a vector of common hyperparameters. We assume further that $p(\theta_i | \Gamma)$ is itself a normal scale mixture:

$$p(\theta_i | \lambda_i^2, \Gamma) = \int N(\theta_i | \lambda_i^2) p(\lambda_i^2 | \Gamma) d\lambda_i^2.$$

Our approach is to induce a prior on the local variances λ_i^2 using the transformation $\kappa_i = 1/(1 + \lambda_i^2)$, or equivalently $\lambda_i^2 = (1 - \kappa_i)/\kappa_i$. This transformed variable has an interpretation as the amount of shrinkage toward the origin, since the posterior mean for

θ_i can be expressed as:

$$\mathbb{E}(\theta_i | y_i, \lambda_i^2) = \frac{\lambda_i^2}{1 + \lambda_i^2} y_i + \frac{1}{1 + \lambda_i^2} 0 = (1 - \kappa_i) y_i.$$

Hence our sparse-estimation approach arises from a random convex combination of the data and the prior mean of zero, where the random weight placed on the prior mean has a hypergeometric–beta prior. This family of priors has many advantages for modeling normal variances: it is conjugate with a normal likelihood, it is highly flexible, and it yields moments and marginal densities that are easy to compute.

These random weights κ_i , moreover, share a set of global shrinkage parameters. Indeed, a fully Bayesian view of (1) makes it clear that additional marginalization is required to yield the posterior for $\boldsymbol{\theta}$, given the data $\mathbf{y} = (y_1, \dots, y_p)$:

$$p(\theta_1, \dots, \theta_p | \mathbf{y}) = \int \prod_{t=1}^p p(\theta_t | \mathbf{y}, \Gamma) \, dp(\Gamma | \mathbf{y}),$$

where $p(\Gamma | \mathbf{y})$ is the marginal posterior for Γ under an assumed prior $p(\Gamma)$. Classically, this can be viewed as a form of Rao-Blackwellization, where

$$\mathbb{E}(\theta_i | \mathbf{y}) = \mathbb{E}\{\mathbb{E}(\theta_i | y_i, \Gamma)\}, \tag{2}$$

the outer expectation being taken with respect to the marginal posterior distribution for Γ , given \mathbf{y} .

This paper makes three main contributions to the large and growing literature on the sparse normal-means problem. First, we adopt a pure “global shrinkage” view and assume that $\kappa = 1/(1 + \lambda^2)$ follows a hypergeometric–beta distribution. Our goal here is not to consider sparsity *per se*, but rather to study our approach for handling for the top-level variance component in a hierarchical model for sparse means. After defining the family of hypergeometric–beta scale-mixture priors, we construct a decomposition of the classical risk of the posterior mean that is tractable for all members of the family. This decomposition is based upon Stein’s approach. We prove a theorem that can be used to compute the risk $\mathbb{E}_{\mathbf{y}|\boldsymbol{\theta}}(\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|^2)$ as a function of $\|\boldsymbol{\theta}\|$, and we use this to study the risk gains near the origin that are possible using hypergeometric–beta priors. These potential risk gains are particularly important when the underlying signal may be sparse. We also give conditions for minimaxity in situations where the likelihood is a scale mixture of normals, using a recent result from Fourdrinier et al. (2008).

Second, we extend the horseshoe prior of Carvalho et al. (2008) to a much richer class of local shrinkage rules. The horseshoe prior assumes that $\kappa_i = 1/(1 + \lambda_i^2)$, and that $\kappa_i \sim \text{Be}(1/2, 1/2)$. We generalize to the case where κ_i has a hypergeometric–beta prior $\text{HB}(a, b, \tau, s)$, and describe how each of the four hyperparameters affects the implied density for $p(\theta_i)$. We also show how the advantages of global shrinkage can easily be embedded into a local-shrinkage framework for sparse means.

Finally, we demonstrate how hypergeometric–beta priors can be used to construct flex-

ible, heavy-tailed multiple-testing procedures, generalizing the work of Scott and Berger (2006) and Johnstone and Silverman (2004).

2 Global Shrinkage with Hypergeometric–Beta Priors

2.1 Overview

In the first part of the paper, we develop shrinkage rules of the form $\hat{\boldsymbol{\theta}}(\mathbf{y}) = \{1 - g(Z)\}\mathbf{y}$ for $Z = \|\mathbf{y}\|^2$. This is a form shared by many other similar procedures (e.g. James and Stein, 1961; Strawderman, 1971; Stein, 1981; Fourdrinier et al., 1998). The central issue is how to identify “nice” functions $g(Z)$, and how to understand priors for global variance components in terms of the behavior of the estimators they yield.

The so-called “constraint to rationality”—namely, the requirement that there exists a prior $p(\kappa)$ such that, for all Z , $g(Z) = E(\kappa|Z)$ under the posterior $p(\kappa | Z)$ —rules out a wide class of potential estimators. The function $g(Z)$ cannot, for example, be a polynomial of order two or greater. Indeed, the functional form of a $g(Z)$ that respects admissibility will typically be quite complicated.

It is natural to look in the class of estimators where $g(Z) = p(Z)/q(Z)$, a ratio of power-series expansions. We construct such a $g(Z)$ by assuming that $(\boldsymbol{\theta} | \lambda^2) \sim N(0, \lambda^2 I)$, and then defining $\hat{\boldsymbol{\theta}}(\lambda^2) = E(\boldsymbol{\theta} | \lambda^2, \mathbf{y})$. After removing the dependence upon λ^2 by marginalizing, this leads to

$$\hat{\boldsymbol{\theta}} = E_{\lambda^2|\mathbf{y}}\{\hat{\boldsymbol{\theta}}(\lambda^2)\} = \{1 - E(\kappa | Z)\}\mathbf{y},$$

recalling that $\kappa = 1/(1 + \lambda^2)$. We can therefore identify $g(Z)$ with $E(\kappa | Z)$, the posterior expectation of κ , given Z .

Our approach is to define a class of priors for κ indexed by (a, b, τ, s) such that

$$g(Z) = E(\kappa|Z) = \frac{a + p/2}{a + b + p/2} \frac{\Phi_1(b, 1; a + b + p/2 + 1; s + Z/2, 1 - 1/\tau^2)}{\Phi_1(b, 1; a + b + p/2; s + Z/2, 1 - 1/\tau^2)}, \quad (3)$$

where a , b , and τ are positive real numbers; s is any real number; and Φ_1 is the degenerate hypergeometric function of two variables (Gradshteyn and Ryzhik, 1965, 9.261).

This g is a ratio of power series, and can be computed quite rapidly for a given tuple (a, b, τ, s) and a given Z . It leads to a large class of admissible estimators with a wide range of possible behavior. In particular, it includes many estimators that exhibit robustness to large values of Z ; many estimators that offer significant risk reduction near $Z = 0$; and many that do both. This class generalizes the form noted by Maruyama (1999), which contains the positive-part James–Stein estimator as a limiting (improper) case.

The purpose of this section differs from most of this literature, in that we do not seek to give general conditions for minimaxity. Rather, we have two different, more modest, goals—goals that are consistent with our focus on a particular class of priors for variance components, and not on the far more general theory of estimation under quadratic loss.

First, we will derive simple expressions in terms of $\|\boldsymbol{\theta}\|$ for the frequentist risk of the estimators in (3) arising from hypergeometric–beta scale mixtures. These expressions apply

when σ^2 is fixed. In this way, users may easily assess which version of the hypergeometric–beta family is appropriate for use on a given problem, depending upon where potential risk improvements should be focused. The expressions themselves are new as far as we are aware, and apply to more general classes of priors, but they are particularly easy to evaluate for priors in this family.

Second, using these expressions, we will assess the gains near the origin that are possible when using hypergeometric–beta priors instead of other common shrinkage estimators. We will also remark on how some recent results in Fourdrinier et al. (2008) can be used to establish the minimaxity of these estimators when the likelihood is also a scale mixture of normals. Potential gains near the origin are especially relevant when $\boldsymbol{\theta}$ may be sparse, an issue to which we will return in subsequent sections.

2.2 The proposed prior

Our proposed class of normal scale mixtures arises by placing a hypergeometric–beta prior on $\kappa \in [0, 1]$, the amount of shrinkage toward the origin. This prior’s density function is

$$p(\kappa) = C^{-1} \kappa^{a-1} (1 - \kappa)^{b-1} \left\{ \frac{1}{\tau^2} + \left(1 - \frac{1}{\tau^2}\right) \kappa \right\}^{-1} \exp(-s\kappa), \quad (4)$$

where $a, b, \tau > 0$ and $s \in \mathbb{R}$, and where C_1 is a constant of proportionality. We denote the prior by $\kappa \sim \text{HB}(a, b, \tau, s)$.

The normalizing constant,

$$C = \int_0^1 \kappa^{a-1} (1 - \kappa)^{b-1} \left\{ \frac{1}{\tau^2} + \left(1 - \frac{1}{\tau^2}\right) \kappa \right\}^{-1} \exp(-s\kappa) \, d\kappa, \quad (5)$$

can be computed using hypergeometric series. In Appendix A we give details of this computation, which yields

$$C = e^{-s} \text{Be}(a, b) \Phi_1(b, 1, a + b, s, 1 - 1/\tau^2), \quad (6)$$

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 (§9.2 of Gradshteyn and Ryzhik, 1965; Gordy, 1998), making evaluation of (6) quite fast for most allowable choices of the parameters.

The implied density for λ^2 takes the form

$$p(\lambda^2) = C^{-1} (\lambda^2)^{b-1} (\lambda^2 + 1)^{-(a+b)} \exp\left\{-\frac{s}{1 + \lambda^2}\right\} \left\{\tau^2 + \frac{1 - \tau^2}{1 + \lambda^2}\right\}^{-1}. \quad (7)$$

This resembles a modified, tempered version of the inverted-beta distribution, also known as Pearson’s Type VI distribution. Indeed, it reduces to an inverted beta in the special case where $s = 0, \tau = 1$, in which case $a\lambda^2/b$ will follow an $F(2b, 2a)$ density.

The hypergeometric–beta family contains many well-known sub-families of priors for

κ . These include the beta distribution, the generalized beta distribution (McDonald and Xu, 1995), and the Gauss hypergeometric distribution (Armero and Bayarri, 1994). These various sub-families are why we call (4) the hypergeometric–beta prior. The family is itself contained in the class of compound confluent hypergeometric distributions (Gordy, 1998), which has two extra parameters that are not relevant in this context.

2.3 Expressions for moments and marginals

The family in (4) has one major advantage over other similar priors, one to which we have already alluded: there exist easily computable expressions for the posterior mean $E(\boldsymbol{\theta} \mid \mathbf{y})$ and the marginal density $m(\mathbf{y}) = \int N(\mathbf{y} \mid \boldsymbol{\theta}, I) p(\boldsymbol{\theta}) d\boldsymbol{\theta}$. We now derive these expressions.

First, note that the joint distribution for κ and \mathbf{y} is

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

with $a' = a + p/2$, and $s' = s + Z/2\sigma^2$. Hence the posterior for κ is also a hypergeometric–beta distribution, with parameters (a', b, τ, s') .

Next, the moment-generating function of a hypergeometric–beta prior 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 and using the differentiation rules given in Chapter 15 of Abramowitz and Stegun (1964) yields

$$E(\kappa^n \mid \mathbf{y}) = \frac{(a')_n}{(a' + b)_n} \frac{\Phi_1(b, 1, a' + b + n, s', 1 - 1/\tau^2)}{\Phi_1(b, 1, a' + b, s', 1 - 1/\tau^2)}. \quad (8)$$

Since $E(\boldsymbol{\theta} \mid \mathbf{y}) = \{1 - E(\kappa \mid \mathbf{y})\} \mathbf{y}$, (8) provides all that is needed to compute the posterior mean for $\boldsymbol{\theta}$.

Similarly, the marginal density $m(\mathbf{y})$ is a simple expression involving the ratio of prior to posterior normalizing constants:

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

with a' and s' as given.

2.4 Main result

We now provide a result that characterizes the risk of the posterior mean when κ has a hypergeometric–beta prior, and where $g(\mathbf{y}) = g(Z) = E(\kappa \mid Z)$. We assume, without loss of generality, that $\sigma^2 = 1$. Stein (1981) shows that, under this assumption, the

mean-squared error of an estimator can be written as

$$MSE = p + E_{\mathbf{y}} \left(\|g(\mathbf{y})\|^2 + 2 \sum_{i=1}^p \frac{\partial}{\partial y_i} g(\mathbf{y}) \right),$$

where $g(\mathbf{y}) = \nabla \ln m(\mathbf{y})$. In turn this can be written as

$$MSE = p + 4E_{\mathbf{y}} \left(\frac{\nabla^2 \sqrt{m(\mathbf{y})}}{\sqrt{m(\mathbf{y})}} \right).$$

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

Theorem 1. *Suppose that $\boldsymbol{\theta} \sim N_p(0, \lambda^2 I)$, that $\kappa = 1/(1 + \lambda^2)$, and that the prior $p(\kappa)$ is such that $\lim_{\kappa \rightarrow 0,1} \kappa(1 - \kappa)p(\kappa) = 0$. Define*

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

Then the risk of the Bayes rule under $p(\kappa)$ is

$$MSE(\boldsymbol{\theta}) = p + 2E_{Z|\boldsymbol{\theta}} \left\{ Z \frac{m_{p+4}(Z)}{m_p(Z)} - pg(Z) - \frac{Z}{2}g(Z)^2 \right\}, \quad (9)$$

where $g(Z) = E(\kappa | Z)$ and

$$Z \frac{m_{p+4}(Z)}{m_p(Z)} = (p + Z + 4)g(Z) - (p + 2) - E_{\kappa|Z} \left\{ 2\kappa(1 - \kappa) \frac{p'(\kappa)}{p(\kappa)} \right\}. \quad (10)$$

Proof. See Appendix B. □

Theorem 1 is useful because the two important quantities that characterize the risk—the marginal $m_p(Z)$, and the expectation $g(Z) = E(\kappa | Z)$ —are easy to compute under hypergeometric–beta scale mixtures of normals, since

$$(\kappa | Z) \sim \text{HB} \left(a + \frac{p}{2}, b, \tau, s + \frac{Z}{2} \right).$$

Given $\|\boldsymbol{\theta}\|^2$, moreover, Z can be easily simulated:

$$\begin{aligned} Z &= U^2 + V \\ U &\sim N(\|\boldsymbol{\theta}\|^2, 1) \\ V &\sim \chi_{p-1}^2. \end{aligned}$$

The risk of estimators arising from hypergeometric–beta scale mixtures is therefore easy to evaluate as a function of $\|\boldsymbol{\theta}\|^2$ using Monte Carlo sampling.

We have given some examples when $p = 7$ in Figures 1 and 2, which show how the classical risk of the Bayes estimator changes as the parameters a , b , τ , and s change. In

both figures, the risk of the maximum-likelihood estimate and the risk of the James-Stein estimator are plotted for the sake of comparison.

These experiments show that:

- The hypergeometric–beta family provides a large class of Bayes estimators that will perform no worse than the MLE in the tails, i.e. when $\|\boldsymbol{\theta}\|^2$ is large.
- Major improvements over the James–Stein estimator are possible near the origin. 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.
- As other authors have noted (e.g. Berger, 1980), there is a tension between minimaxity and risk improvements near the origin. Particularly if there are grounds for suspecting sparsity in $\boldsymbol{\theta}$, it may be reasonable to give up minimaxity for the sake of these potential gains.

2.5 Remarks

Three further remarks are in order.

First, when $\sigma^2 = 1$, it is difficult to assess the minimaxity of hypergeometric–beta estimators using existing theory. Many priors of this family, despite appearing to yield minimax rules when Monte Carlo simulations are conducted using Theorem 1, do not meet the assumptions of available theorems that characterize minimax estimators (e.g., Theorem 1 of Fourdrinier et al., 1998). One example is the above special case of $a = b = 1/2$, $s = 0$, and $\tau = 1$. By all indications this estimator is minimax when $p \geq 7$ (see Figure 2). Yet the implied prior for $\boldsymbol{\theta}$ does not meet certain boundedness conditions at the origin necessary to decide the matter formally using available tools. The construction of new tools along these lines is, of course, an active area of research. We are investigating possible generalizations to cases where $\pi(\boldsymbol{\theta})$ may be unbounded at the origin, which is the crucial modification required of existing theory.

Second, and perhaps rather surprisingly, it is possible to provide formal a characterization of minimaxity when the sampling variance σ^2 follows a mixing density g (as is the case, for example, for a Student- t likelihood). Here we summarize a set of sufficient conditions for minimaxity, which are given by Theorem 3.1 of Fourdrinier et al. (2008).

1. There exist numbers $K > 0$, $\lambda_0^2 > 0$, and $\alpha < 1$ such that

$$\pi(\lambda^2) \leq K\lambda^{-2\alpha} \tag{11}$$

for $0 < \lambda^2 < \lambda_0^2$.

2. For some $c > 0$ and $\beta < p/2 - 1$,

$$\lim_{\lambda^2 \rightarrow \infty} \frac{\pi(\lambda^2)}{(\lambda^2)^\beta} = c \tag{12}$$

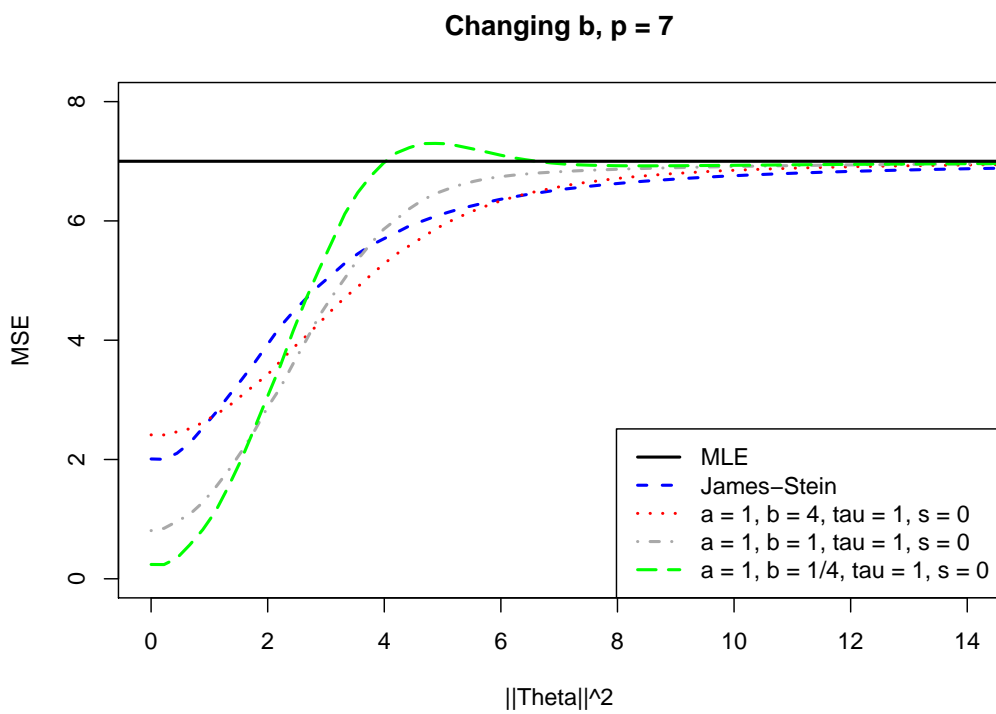
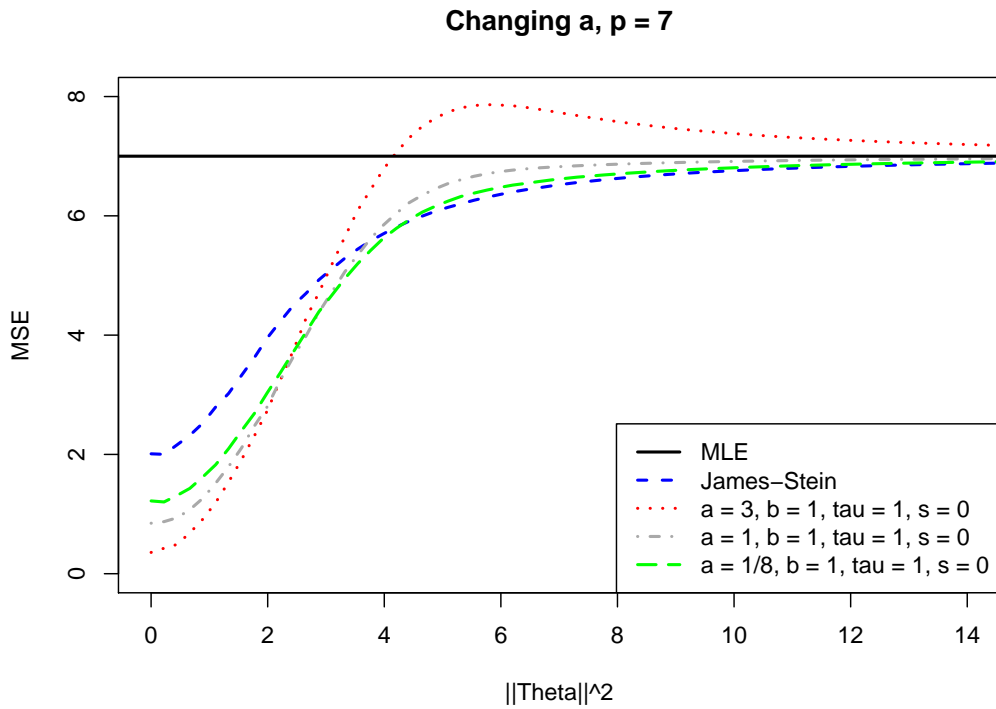


Figure 1: Mean-squared error as a function of $\|\theta\|$: effect of changing a (top) and b (bottom).

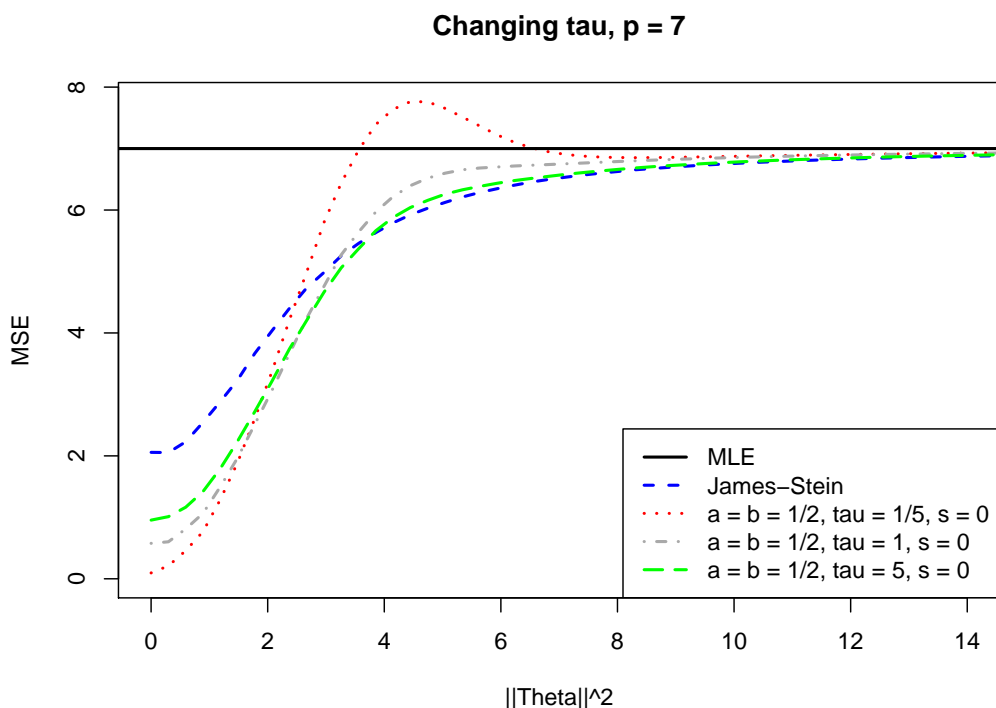
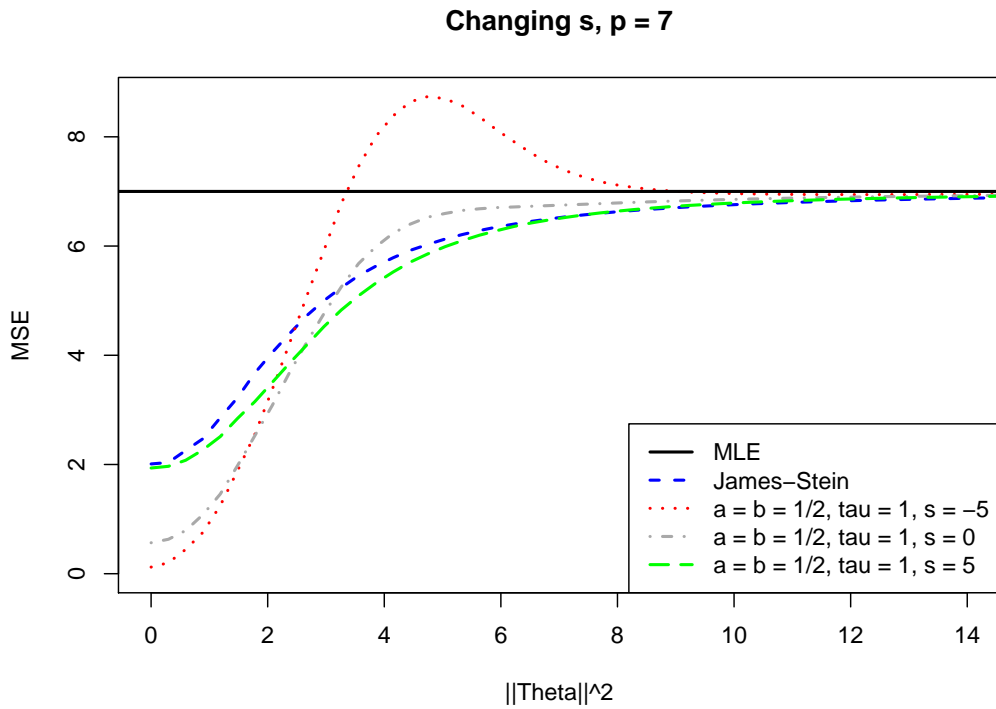


Figure 2: Mean-squared error as a function of $\|\theta\|$: effect of changing s (top) and τ (bottom).

3. For the same β as above,

$$-(p-2) \left\{ \frac{E(\sigma^{-p+2})}{E(\sigma^2) \cdot E(\sigma^{-p})} - \frac{1}{2} \right\} \leq \beta, \quad (13)$$

where all expectations must be assumed finite.

4. The prior $g(\sigma^2)$ must have monotone likelihood ratio as a scale family.

5. The prior $\pi(\lambda^2)$ must also have monotone likelihood ratio as a scale family.

Conditions 3 and 4 will depend entirely upon the choice of g . Meanwhile, conditions 1 and 2 can be checked using the fact that, under a hypergeometric–beta prior, $p(\lambda^2)$ behaves like $(\lambda^2)^{b-1}$ near the origin, and like $(\lambda^2)^{1-a}$ in the upper tail. This gives the allowable range of values for a and b on a given problem. Crucially, condition 2 allows priors that whose density $\pi(\theta)$ unbounded at the origin, which will be the case for any hypergeometric–beta prior in which $b \leq 1/2$. Condition 5 is easily verified using Lemma 3.3 of Fourdrinier et al. (2008) for any hypergeometric–beta scale-mixture with $s > 0$.

Finally, the special case of $a = 1/2$ and $b = 1/2$, leading to a half-Cauchy density for $p(\lambda)$, was recommended by Gelman (2006) as a default prior for variance parameters in Bayesian hierarchical models. This prior tends to a constant at $\lambda = 0$ and yet is quite heavy-tailed, a combination of features that Gelman shows to be quite desirable for a prior on a global variance parameter in high-dimensional inference. The results of this section give a quite different, classical justification for this prior in high-dimensional settings: its excellent mean-squared error risk properties. The fact that two independent lines of reasoning both lead to the same distribution is a strong argument in its favor as a default proper prior for a shared variance component. We note that the hypergeometric–beta class provides a very useful generalization of this prior, in that it allows even greater control over global shrinkage through τ and s .

3 Modeling Sparsity with Hypergeometric–Beta Priors

3.1 Local shrinkage priors

We now extend the hypergeometric–beta prior to a local-shrinkage framework, where

$$\begin{aligned} p(\theta_i | \kappa_i) &\sim N\{0, \sigma^2(1 - \kappa_i)/\kappa_i\} \\ \kappa_i &\sim \text{HB}(a, b, \tau, s). \end{aligned}$$

The connection with global shrinkage can be made explicit by marginalizing over θ_i :

$$(y_i | \lambda_i) \sim N\{0, \sigma^2(1 + \lambda_i^2)\},$$

or equivalently $y_i = \lambda_i \eta_i + \epsilon_i$, where $\eta_i, \epsilon_i \sim N(0, \sigma^2)$ independently. The local shrinkage factors λ_i can therefore also be thought of as location parameters. The mathematics will

be different from the case of normal means, but the intuition that global shrinkage of the λ'_i s can offer substantial risk improvements still remains.

Under our hypergeometric–beta model, the joint distribution for y_i and κ_i takes the form

$$p(y_i, \kappa_i) \propto \kappa_i^{a'-1} (1 - \kappa_i)^{b-1} \left\{ \frac{1}{\tau^2} + \left(1 - \frac{1}{\tau^2}\right) \kappa_i \right\}^{-1} e^{-\kappa_i s'},$$

where now $s' = s + y_i^2/(2\sigma^2)$ and $a' = a + 1/2$. As in the global-shrinkage setting, if κ_i has a hypergeometric–beta prior, it will also have a hypergeometric–beta posterior once normal data have been observed.

Using (8), we get

$$E(\theta_i | y_i) = \left\{ 1 - \frac{a'}{a' + b} \frac{\Phi_1(b, 1, a' + b + 1, s', 1 - 1/\tau^2)}{\Phi_1(b, 1, a' + b, s', 1 - 1/\tau^2)} \right\} y. \quad (14)$$

And by the law of total variance,

$$\begin{aligned} \text{Var}(\theta_i | y_i) &= E\{\text{Var}(\theta_i | y_i, \kappa_i)\} + \text{Var}\{E(\theta_i | y_i, \kappa_i)\} \\ &= \sigma^2 \{1 - E(\kappa_i | y_i)\} + y^2 \text{Var}(\kappa_i | y_i), \end{aligned} \quad (15)$$

will all other posterior moments for θ_i following in turn.

As before, there is also a tractable expression for the marginal likelihood of the data:

$$m(y_i) = C_1^{-1} \int_0^1 \kappa_i^{a'-1} (1 - \kappa_i)^{b-1} \left\{ \frac{1}{\tau^2} + \left(1 - \frac{1}{\tau^2}\right) \kappa_i \right\}^{-1} e^{-\kappa_i s'} d\kappa_i, \quad (16)$$

where again $s' = s + y_i^2/(2\sigma^2)$ and $a' = a + 1/2$. This integral is in the same family as (5), and so by the same series of arguments we obtain

$$m(y_i) = \frac{1}{\sqrt{2\pi}\sigma^2} \exp\left(-\frac{y_i^2}{2\sigma^2}\right) \frac{\text{Be}(a', b)}{\text{Be}(a, b)} \frac{\Phi_1(b, 1, a' + b, s', 1 - 1/\tau^2)}{\Phi_1(b, 1, a + b, s, 1 - 1/\tau^2)}. \quad (17)$$

3.2 Shrinkage profiles

We now turn to the specification of the four hyperparameters, and to the different “local shrinkage profiles” that are accessible through different choices of these parameters.

All normal scale-mixtures have an implied shrinkage profile $p(\kappa_i)$, which describes the amount of shrinkage toward the origin that is expected *a priori*. The prior’s behavior near $\kappa_i = 0$ controls the tail weight of the marginal prior for θ_i , while the behavior near $\kappa_i = 1$ controls the strength of shrinkage near zero.

Table 1 lists four common priors, while Figure 3 plots the implied shrinkage profiles for two of these: the double-exponential and Cauchy priors. Contrast these shrinkage profiles with the wide range of shapes that accessible through the hypergeometric–beta density, some of which are shown in Figure 4.

One important special case of the hypergeometric–beta family is the Strawderman prior (Strawderman, 1971), which corresponds to $a = 1/2$, $b = 1$, $s = 0$, and $\tau = 1$.

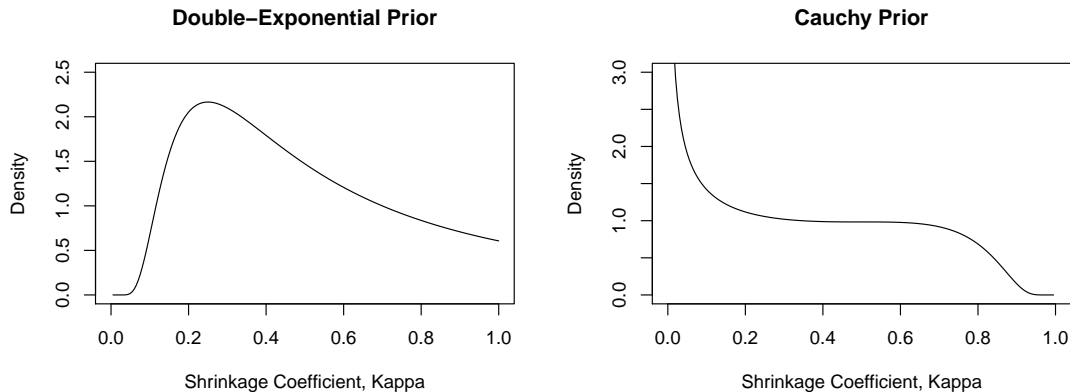


Figure 3: Implied shrinkage profiles for double-exponential and Cauchy priors.

Table 1: Priors for λ_i and κ_i associated with some common local shrinkage rules. Densities are given up to constant terms.

Prior for θ_i	Prior for λ_i	Prior for κ_i
Double-exponential	$\lambda_i \exp\{\lambda_i^2/2\}$	$\kappa_i^{-2} e^{-\frac{1}{2\kappa_i}}$
Cauchy	$\lambda_i^{-2} \exp(-1/2\lambda_i^2)$	$\kappa_i^{-\frac{1}{2}} (1 - \kappa_i)^{-\frac{3}{2}} e^{-\frac{\kappa_i}{2(1-\kappa_i)}}$
Strawderman-Berger	$\lambda_i (1 + \lambda_i^2)^{-3/2}$	$\kappa_i^{-\frac{1}{2}}$
Horseshoe	$(1 + \lambda_i^2)^{-1}$	$\kappa_i^{-1/2} (1 - \kappa_i)^{-1/2}$

Another special case is the half-Cauchy prior on the scale factor λ , studied by Gelman (2006) and Carvalho et al. (2008). This corresponds to $a = b = 1/2$, $s = 0$, and $\tau = 1$. Yet a third special case is the uniform-shrinkage prior, where $a = b = 1$, $s = 0$, and $\tau = 1$. All of these can be seen in the upper-left pane of Figure 4.

Clearly (4) can lead to many standard-looking shapes that are similar to other normal scale mixtures. Yet it can also produce a wide variety of other densities that are inaccessible through other standard families. We now describe the role of each hyperparameter, recalling that more probability near $\kappa = 1$ means more aggressive shrinkage.

First, τ is a global scaling factor, with larger values leading to larger marginal variance in θ . To see this, suppose that all components of θ have a common variance component in addition to their idiosyncratic ones: $(y_i | \theta_i) \sim N(\theta_i, \sigma^2)$ and $\theta_i \sim N(0, \sigma^2 \tau^2 \lambda_i^2)$. The form involving τ in (4) arises from the special case of assuming a half-Cauchy prior for each λ_i , as in the horseshoe prior of Carvalho et al. (2008). The generalization of the scaled half-Cauchy prior to arbitrary a , b , and s then arises quite naturally on the κ

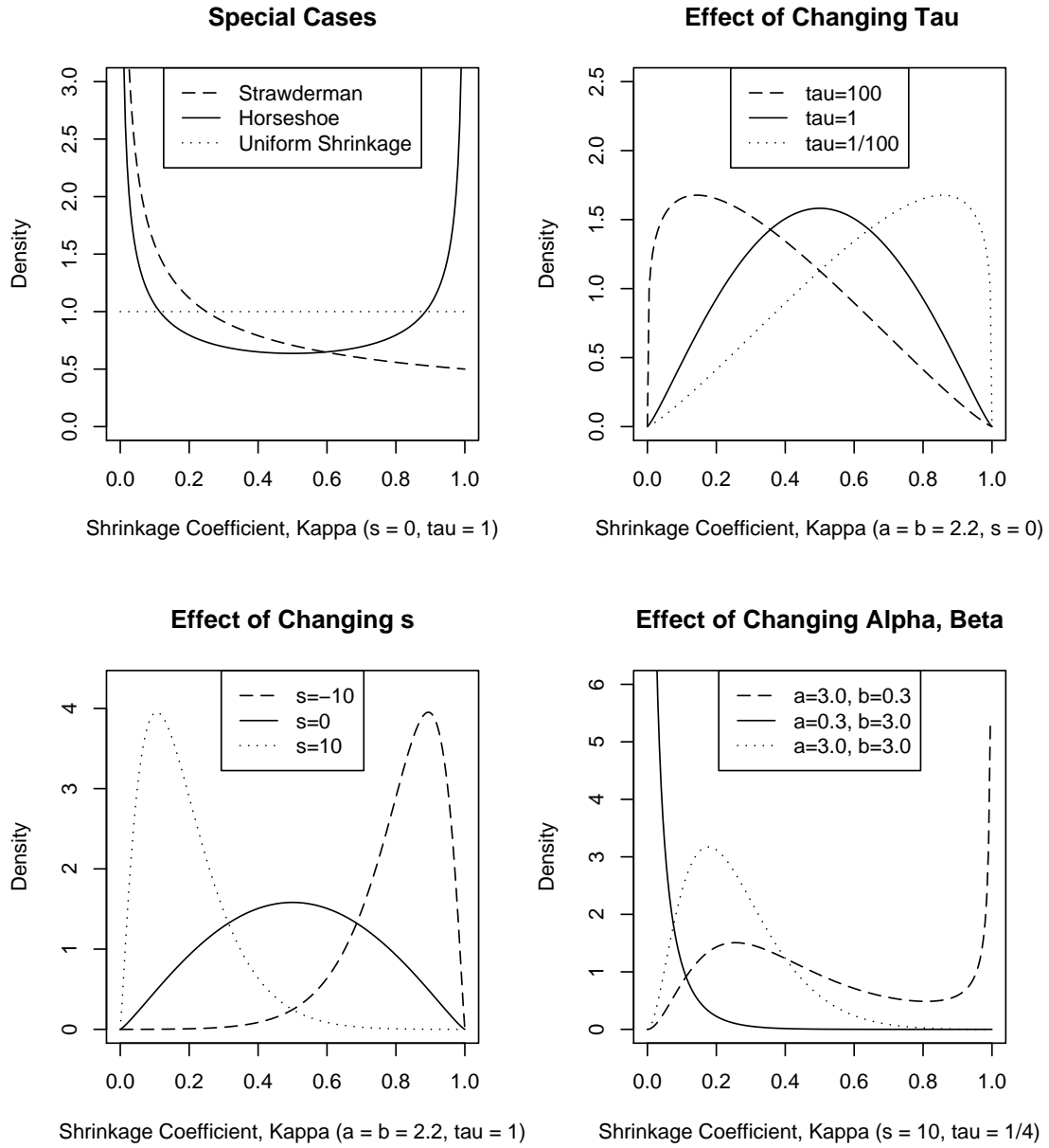


Figure 4: Effect of changing the four parameters (a, b, s, τ) on the density for the shrinkage coefficient κ .

scale. Shifting τ up and down causes the shrinkage profile to be shifted left and right, respectively, controlling the overall aggressiveness of shrinkage.

The parameters a and b are analogous to those of beta distribution, to which (4) reduces when $\tau = 1$ and $s = 0$. Smaller values of a encourage heavier tails in $\pi(\theta)$, with $a = 1/2$, for example, yielding Cauchy-like tails. Smaller values of b encourage $p(\theta)$ to have more mass near the origin, and eventually to become unbounded; $b = 1/2$ yields, for example, $p(\theta) \approx \log(1 + 1/\theta^2)$ near 0.

Finally, s is a second global scaling factor, though with a different effect than τ on the shape of the density. This parameter might, for example, be usefully construed s as a vehicle for regressing shrinkage coefficients upon external covariates $\{x_{ij}\}_{j=1}^m$, since it is straightforward to let $s_i = \sum x_{ij}\beta_j$ for some vector β of regression coefficients. Values of s smaller than zero encourage κ to be smaller, meaning that a positive regression coefficient β_j is associated with increased shrinkage.

The scale parameters τ and s do not control the behavior of $\pi(\lambda)$ at 0 and ∞ . Specifically, $\pi(\lambda)$ behaves like λ_i^{2b-1} near the origin, and like $\lambda_i^{-(2a+1)}$ in the upper tail. Since $\pi(\theta)$ has the same polynomial rate of decay as $\pi(\lambda)$, a can be chosen to reflect the desired tail weight of $\pi(\theta)$.

3.3 The effect of adding global shrinkage

The hypergeometric–beta prior allows a combination of global and local shrinkage that can be both flexible and robust. Figure 5 shows how a very small value of τ , encouraging strong global shrinkage, can be reinforced by a small observation ($y = 1.0$), and yet be almost completely overruled by a large observation ($y = 4.0$). Meanwhile, the marked bimodality for an intermediate observation such as $y = 2.5$ reflects uncertainty about whether such an observation corresponds to signal or noise, with the posterior mean for θ averaging over both possibilities.

This example demonstrates that global shrinkage through τ can be very effective at squelching noise in high-dimensional problems. It is crucial, however, that τ be estimated from the data, and that the prior for κ_i grow sufficiently fast near 0 in order to allow κ_i to escape the strong “gravitational pull” of a small τ when y_i is large (as in this example when $y_i/\sigma = 4$). We recommend setting $a = 1/2$ in sparse problems involving a normal likelihood; see Carvalho et al. (2008) for further discussion. In situations with heavier-tailed sampling models, it may be appropriate to choose a smaller value of a .

When $1 - 1/\tau^2$ is very close to 1 (or when $1 - \tau^2$ is very close to 1 for $\tau < 1$), the Φ_1 functions in (8) and (17) may become slow to evaluate due to the slow convergence of the series representations given in the appendix. In our experience, the issue becomes practically significant in a serial computing environment only when τ^2 is larger than 1000 or smaller than $1/1000$. These frontiers are far more expansive in a parallel computing environment, and will no doubt expand further as computers grow faster. But for now, a different approach may be required in situations where the global shrinkage parameter must be either very large or very small compared to the sample variance of \mathbf{y} .

Two such options are available, either of which will avoid any convergence problems associated with extreme arguments of the Φ_1 function. First, global shrinkage can take

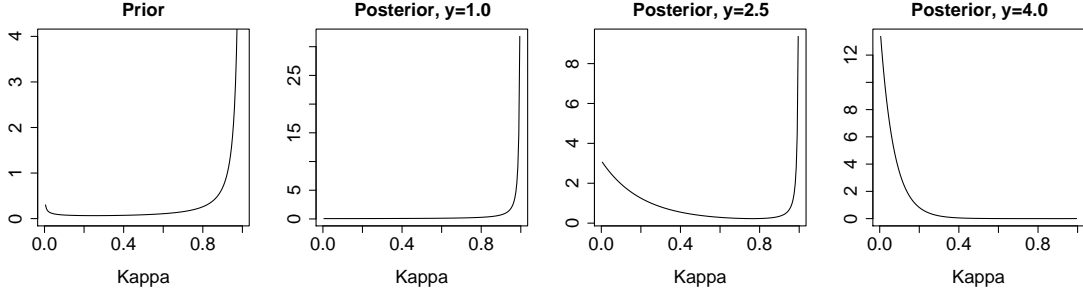


Figure 5: The left pane shows the prior for κ when $\tau = 1/15$, $s = 0$, and $a = b = 1/2$, reflecting a prior bias for strong shrinkage. The next three panes show the different posteriors for κ upon observing a single data point: $y = 1.0$, $y = 2.5$, or $y = 4.0$, respectively.

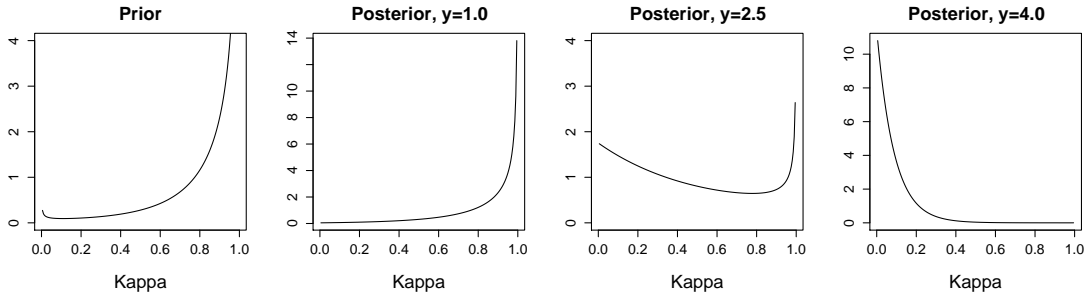


Figure 6: The left pane shows the prior for κ when $\tau = 1$, $a = b = 1/2$, and $s = -4$. The next three panes show the different posteriors for κ upon observing a single data point: $y = 1.0$, $y = 2.5$, or $y = 4.0$, respectively.

place through s rather than τ (with τ being set equal to 1). Then $\kappa_i \sim \text{HB}(a, b, \tau = 1, s)$, and so

$$(\kappa_i | y_i) \sim \text{HB}(a + 1/2, b, \tau = 1, s + y_i^2/2\sigma^2).$$

Figure 6 shows that global shrinkage through s can produce results quite similar to global shrinkage through τ .

Alternatively, one can set adopt a subtly different form of global shrinkage:

$$\begin{aligned} (\theta_i | \kappa_i) &\sim \text{N}\left(0, \frac{\eta^2 + \sigma^2}{2\kappa_i} - \sigma^2\right) \\ \kappa_i &\sim \text{HB}(a, b, 1, s), \end{aligned}$$

with the first step following Section 4.7 of Berger (1985) and requiring that the global shrinkage parameter satisfies $\eta > \sigma$.

In either case, the posterior moments of θ_i and the marginal for y_i involve only a single

${}_1F_1$ evaluation, since

$$\int_0^1 \kappa^{a-1} (1-\kappa)^{b-1} e^{-s\kappa} d\kappa = \text{Be}(a, b) {}_1F_1(a, a+b, -s) = \text{Be}(a, b) e^s {}_1F_1(b, a+b, s).$$

Note that, in the second case, the posterior mean is:

$$\mathbb{E}(\theta_i | y_i, \kappa_i) = \left\{ 1 - \left(\frac{2\sigma^2}{\sigma^2 + \tau^2} \right) \kappa_i \right\} y_i.$$

We prefer the elegance of the original approach involving the full hypergeometric–beta family, since both τ and the λ_i 's can be interpreted as variance components. But we offer these alternatives for handling difficulties that may be encountered in evaluating the Φ_1 function for particular data sets. In practice, we have only encountered these difficulties in extremely high-dimensional problems.

3.4 Sampling from the hypergeometric–beta distribution

The previous expressions show that posterior means under hypergeometric–beta priors are easy to compute for fixed values of the hyperparameters. Averaging over uncertainty with respect to these hyperparameters, however, requires Markov-chain Monte Carlo to compute posterior means. This, in turn, requires a method to generate random draws from the hypergeometric–beta distribution.

Luckily this is straightforward using rejection sampling under a beta proposal. Some algebra yields the required bound on the density function:

$$p(\kappa | a, b, s, \tau) \leq M \cdot \left\{ \frac{\kappa^{a-1} (1-\kappa)^{b-1}}{\text{Be}(a, b)} \right\} \quad (18)$$

$$M = \frac{\tau^2 \cdot \max(1, e^{-s})}{e^{-s} \cdot \Phi_1(b, 1, b, s, 1 - 1/\tau^2) \cdot \min(1, \tau^2)} < \infty. \quad (19)$$

The part of (18) inside braces is the density of a beta random variable. This suggests the following sampler:

1. Draw a beta random variable, $\kappa \sim \text{Be}(a, b)$.
2. Draw $u \sim \text{Unif}(0, 1)$.
3. Accept κ if

$$u \leq \frac{\min(1, \tau^2)}{\max(1, e^{-s})} \cdot \frac{e^{-s\kappa}}{1 + (\tau^2 - 1)\kappa},$$

and otherwise return to Step 1.

In practice, this sampler is efficient when a and b are both less than 1, in the sense that the bound M will be moderate. Better is to sample $\kappa \sim \text{Be}\{\min(a, 1), \min(b, 1)\}$ in Step 1 of the algorithm, with the obvious modification of the acceptance probability in Step 2. This will prevent M from being too large in most common situations.

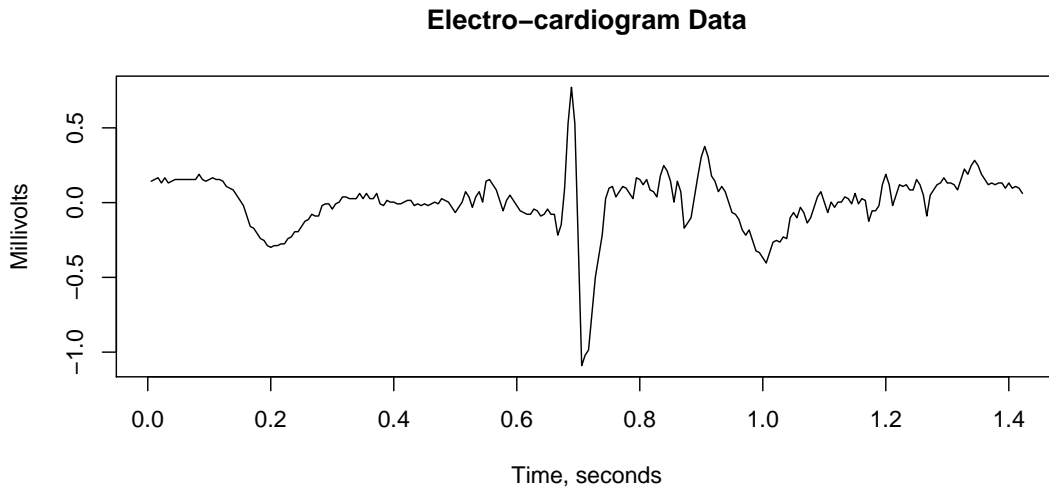


Figure 7: Electro-cardiogram data used as the “true” function f in the wavelet de-noising experiment.

Table 2: Results for the wavelet-denoising experiment under three different noise levels and two different loss functions. The table entries are the average loss across 100 simulated data sets.

Procedure	$\sigma = 0.1$		$\sigma = 0.2$		$\sigma = 0.4$	
	ℓ_W^2	ℓ_T^2	ℓ_W^2	ℓ_T^2	ℓ_W^2	ℓ_T^2
DWT	20.4	20.5	81.9	82.0	328.0	328.2
JS	13.6	13.7	36.3	36.4	87.1	87.3
HB	9.3	9.3	26.7	26.8	72.4	72.6

In rare cases where many draws from a single density are required, it may be most efficient to minimize M over (a, b) using a numerical optimization routine. But in our experience, this additional front-end investment does not usually offer a noteworthy payoff.

3.5 Example: wavelet de-noising

Figure 7 represents an electro-cardiogram of approximately one beat of a normal human heart rhythm. The data set contains 256 millivolt readings sampled at 180 Hz, and is available from the R package `wavelets`. The readings have been re-scaled to have a mean of zero, and their standard deviation is approximately 0.2.

We took these data points to represent the “true” function f sampled at equi-spaced intervals, and simulated noisy realizations of f by setting $y_i = f_i + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$ for $i = 1 \dots, 256$. We constructed 100 fake data sets each for three different noise levels:

$\sigma = 0.1$, $\sigma = 0.2$, and $\sigma = 0.4$. Most of the quite standard details concerning Bayes and empirical-Bayes inference in the wavelet domain are omitted here, including how empirical wavelet coefficients should be scaled. For a detailed discussion, see Clyde and George (2000), whose framework we follow.

Specifically, let d_{jk} represent the k th coefficient of the discrete wavelet transform (DWT) at resolution level j , appropriately re-scaled as per Clyde and George (2000). We assume that these coefficients are observed with error according to $d_{jk} = \beta_{jk} + \nu_{jk}$, place a hypergeometric–beta scale-mixture prior on β_{jk} , and estimate β_{jk} by the posterior mean. The DWT of the ECG data are assumed to represent the true β_{jk} ’s, while the DWT of the noisy realizations y are treated as raw data.

Our goal is to assess the performance of hypergeometric–beta priors in the wavelet domain against two benchmarks: the discrete wavelet transform, and the thresholding procedure for normal means described by Johnstone and Silverman (2004), which has been proven to have a strong set of asymptotic optimality properties. We measure the performance of an estimator $\hat{\beta}$ by ℓ^2 loss in both the wavelet domain and the time domain: $\ell_W^2(\hat{\beta}) = \sum_j \sum_k (\hat{\beta}_{jk} - \beta_{jk})^2$, and $\ell_T^2(\hat{\beta}) = \sum_i (\hat{f}_i - f_i)^2$, where \hat{f} is the inverse wavelet transform of the estimated coefficients $\hat{\beta}$.

Table 2 shows the results for overall best performer in the hypergeometric–beta family as it compares to the DWT and the Johnstone/Silverman procedure. This corresponded to $a = b = 1/2$ and $s = 0$, with τ unknown and given a positive-Cauchy hyper-prior. This “default” specification of hyperparameters gave performance that uniformly beat the Johnstone/Silverman procedure, which is the recognized gold standard in the literature on modeling sparse wavelet coefficients.

These results, while preliminary and fairly limited in scope, nonetheless show that our proposed family of priors can generate shrinkage rules with the potential to yield substantial risk improvements over other common estimators.

4 Multiple hypothesis testing with heavy-tailed priors

Hypergeometric–beta scale mixtures of normals are an especially useful class of priors for building discrete mixture models for θ_i , due to the existence of closed form moments and marginals under the hypothesis that θ_i is nonzero:

$$(\theta_i \mid \kappa_i) \sim w \cdot N(0, \kappa^{-1} - 1) + (1 - w) \cdot \delta_0 \quad (20)$$

$$\kappa_i \sim \text{HB}(a, b, \tau, s), \quad (21)$$

where δ_0 indicates a degenerate distribution at 0. The posterior mean under this model is a natural estimator for $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)$, since it averages over uncertainty about whether each component is zero or nonzero. Estimators of this form are explored in Scott and Berger (2006) and Bogdan et al. (2008).

Tables 3 and 4 shows the results of an extensive simulation study where $p = 1000$. We benchmarked a variety of hypergeometric–beta scale mixtures against the procedure from Johnstone and Silverman (2004), where θ_i is estimated by the posterior median under a

mixture of a point mass and a double-exponential (Laplace) prior.

Table 3 summarizes an experiment in which the nonzero means were randomly drawn from a heavy-tailed t distribution with 3 degrees of freedom and scale parameter c . We investigated 12 configurations of different sparsity patterns (40, 100, 500, and 900 nonzero means) and different scales ($c = 1, 2, 3$).

Table 4 then recapitulates the study reported in Table 1 of Johnstone and Silverman (2004). This experiment also involved 12 configurations of different sparsity patterns (5, 50, and 500 nonzero means) and different scales (all nonzero means equal to 3, 4, 5, or 7).

The table entries are the average sum of squared errors in estimating θ over 1000 independent data sets. For all hypergeometric–beta priors, we set $\tau = 1$, while w and s were estimated by marginal maximum likelihood. We also include hard-thresholded versions of the hypergeometric–beta estimators, where θ_i is estimated to be zero if it has less than 50% probability of being nonzero *a posteriori*. These thresholded versions are denoted by a (T) in the tables.

In Experiment 1 (random coefficients), the “conditional uniform shrinkage” prior, where $\kappa_i \sim \text{HB}(1, 1, 1, s)$, outperforms all alternatives when s is estimated by maximum likelihood. The advantages of global shrinkage through s come through quite clearly here. We also note that thresholded versions of the estimators rarely equal, and never improve upon, the performance of the straight estimators. In Experiment 2 (fixed coefficients), Johnstone and Silverman’s procedure performs best overall, though its advantage does not hold for all cases. Thresholding seems to yield slight improvements in some of the configurations.

We attribute these differences to the relative tail weight of the two priors. The double-exponential prior has tails that are heavier than the Gaussian likelihood, but not as heavy as those of the hypergeometric–beta priors we studied. This difference in tail weight becomes much more significant in the experiment with random coefficients, since draws from a t_3 density produce some very large signals—much larger than signals of size 7 in the “fixed coefficients” study. In Experiment 2, however, the heavier-tailed priors are wasting some of their mass in areas of the parameter space far from the origin. Since these areas are pre-destined to be unimportant by the particular choices of fixed signals, it is no surprise that a lighter-tailed prior such as the double-exponential will yield superior results.

These experiments speak to the fact that hypergeometric–beta priors provide a broad class of densities for the alternative hypothesis in Bayesian model selection and hypothesis testing. Default versions of the prior seem to perform at least as well as existing gold-standard techniques.

We note in passing that similar expressions involving hypergeometric–functions appear in a study of mixtures of g -priors for Bayesian variable selection by Liang et al. (2008). We conjecture that hypergeometric–beta priors may offer a useful way of constructing new g -like priors for regression, where $g/(1 + g)$ follows a hypergeometric–beta distribution.

Table 3: Simulation study, random coefficients. Bold entries are the best in the column. Thresholded estimators are denoted by a (T).

Number nonzero Signal-to-noise ratio	40			100			500			900		
	1	2	3	1	2	3	1	2	3	1	2	3
$a = 1, b = 1$	62	88	91	126	184	194	404	604	657	619	837	912
$a = 1/2, b = 1$	63	89	91	128	188	197	409	611	669	642	860	929
$a = 1/2, b = 1/2$	63	89	91	127	188	197	415	611	668	658	867	930
$a = 1/4, b = 1/2$	64	90	93	130	192	203	435	638	676	709	930	983
$a = 1, b = 1$ (T)	64	96	100	133	198	211	404	605	682	619	837	912
$a = 1/2, b = 1$ (T)	65	95	98	132	193	207	409	611	669	642	860	929
$a = 1/2, b = 1/2$ (T)	64	94	98	129	190	206	415	611	668	658	867	930
$a = 1/4, b = 1/2$ (T)	65	93	96	130	193	204	435	638	676	709	930	983
Laplace (median)	65	98	100	143	210	208	530	672	676	832	843	927

Table 4: Simulation study, fixed coefficients. Bold entries are the best in the column. Thresholded estimators are denoted by a (T).

Number nonzero Value	5				50				500			
	3	4	5	7	3	4	5	7	3	4	5	7
$a = 1, b = 1$	35	30	21	10	210	171	119	80	882	909	925	941
$a = 1/2, b = 1$	37	33	21	10	220	182	121	79	931	901	801	684
$a = 1/2, b = 1/2$	37	33	21	10	218	182	121	79	930	886	805	668
$a = 1/4, b = 1/2$	38	36	23	10	228	198	126	76	1032	895	808	658
$a = 1, b = 1$ (T)	37	32	19	9	209	164	104	68	882	908	925	940
$a = 1/2, b = 1$ (T)	38	34	20	8	220	174	108	68	931	902	801	656
$a = 1/2, b = 1/2$ (T)	38	34	21	8	218	177	108	68	930	887	805	637
$a = 1/4, b = 1/2$ (T)	39	37	23	8	228	198	118	67	1032	895	808	631
Laplace (median)	36	31	18	9	212	155	101	73	855	873	782	657

5 Final Discussion

We conclude with three final remarks about hypergeometric–beta scale mixtures.

First, it is known that only certain functions of the data y can be admissible estimators of θ . If one specifies an estimator by some functional $\hat{\theta} = f(y)$ for all y , then in general $\hat{\theta}$ will not be admissible, and will not correspond to any distribution $\pi(\theta)$ that might have described one’s prior uncertainty. (This class of inadmissible estimators includes, for example, fractional powers of y and polynomials in y of order two or greater.) The estimator in (3) is interesting because it represents a broad class of nonlinear functions that satisfy this “constraint to rationality,” in that they arise from proper priors.

Second, we note that the use of heavy-tailed priors for constructing robust shrinkage estimators has a long history, with prominent examples to be found in Strawderman (1971) and Berger (1980). Jeffreys, meanwhile, observed as early as 1939 that heavy-tailed priors play an important role in Bayesian hypothesis testing (see Jeffreys, 1961, a later edition). His arguments have been recapitulated in the context of linear models by Zellner and Siow (1980) and, more recently, Liang et al. (2008). The practical issue in both problems is roughly the same: that heavy-tailed priors lead to a desirably mild rate of tail decay in the marginal likelihood $m(\mathbf{y})$, but that very few known priors are both heavy-tailed and

analytically tractable. Any prior that possesses both properties, as our proposed family has to potential to do with certain hyperparameter choices, is therefore of great potential interest to Bayesians and non-Bayesians alike.

Third, it is known that a wide class of stochastic processes can be generated by subordinating Brownian motion to a random clock. This construction is the continuous-time analogue of local scale mixtures of normals; prominent examples include the variance–gamma process and the normal–inverse-Gaussian process. The modified, tempered inverted-beta density in (7) yields many possible distributions for the random clock, and can therefore generate a wide range of stochastic processes. This as-yet-unexplored area represents a interesting set of possibilities for future research.

A Details of hypergeometric–beta integrals

Theorem 2. *The hypergeometric–beta density is proper for all $a, b, \tau > 0$ and $s \in \mathbb{R}$.*

Proof. The normalizing constant in (4) is

$$C = \int_0^1 \kappa^{\alpha-1} (1-\kappa)^{\beta-1} \left\{ \frac{1}{\tau^2} + \left(1 - \frac{1}{\tau^2}\right) \kappa \right\}^{-1} \exp(-s\kappa) \, d\kappa. \quad (22)$$

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

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

Using properties of the hypergeometric function ${}_2F_1$ (Abramowitz and Stegun, 1964, §15.1.1 and §15.3.1), this becomes, after some straightforward algebra,

$$C = e^{-s} \operatorname{Be}(\alpha, \beta) \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} \frac{(\beta)_{m+n}}{(\alpha + \beta)_{m+n} m! n!} s^m (1 - 1/\tau^2)^m, \quad (23)$$

where $\operatorname{Be}(\cdot, \cdot)$ is the beta function and $(a)_n$ 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 (23) converges to a positive real number, yielding

$$C = e^{-s} \operatorname{Be}(\alpha, \beta) \Phi_1(\beta, 1, \alpha + \beta, s, 1 - 1/\tau^2), \quad (24)$$

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

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

$$\Phi_1(\alpha, \beta; \gamma; x, y) = \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} \frac{(\alpha)_{m+n} (\beta)_n}{(\gamma)_{m+n} m! n!} y^n x^m, \quad (25)$$

where $(c)_n$ is the rising factorial. 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 $x \geq 0$,

$$\Phi_1(\alpha, \beta, \gamma, x, y) = \sum_{n=0}^{\infty} \frac{(\alpha)_n x^n}{(\gamma)_n n!} {}_2F_1(\beta, \alpha + n; \gamma + n; y). \quad (26)$$

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

$$\Phi_1(\alpha, \beta, \gamma, x, y) = e^x \sum_{n=0}^{\infty} \frac{(\gamma - \alpha)_n (-x)^n}{(\gamma)_n n!} {}_2F_1(\beta, \alpha; \gamma + n; y). \quad (27)$$

Finally, when $y < 0$,

$$\Phi_1(\alpha, \beta, \gamma, x, y) = e^x (1 - y)^{-\beta} \Phi_1(\tilde{\alpha}, \beta, \gamma, -x, \tilde{y}), \quad (28)$$

where $\tilde{\alpha} = \gamma - \alpha$ and $\tilde{y} = y/(y - 1)$. Then either (26) or (27) may be used to evaluate the righthand side of (28), depending on the sign of x .

Alternative representations for Φ_1 involving ${}_1F_1$ functions are also available. In our experience, however, these take longer to converge than those given above.

B Proof of Theorem 1

Proof. Begin with Stein's decomposition of risk. Following Equation (10) of Fourdrinier et al. (1998), we have

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

The score can be written as

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

And the Laplacian term is $\Delta m(\mathbf{y}) = \int_0^1 (Z\kappa - p) \kappa^{\frac{p}{2}+1} p(\kappa) e^{-\frac{Z}{2}\kappa} d\kappa$. Combining these terms, we have,

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

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

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

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 $g(Z) = E(\kappa | Z)$.

Secondly, note that

$$Z\{m_{p+2}(Z) - m_{p+4}(Z)\} = 2 \int_0^1 \kappa^{\frac{p}{2}+1} (1-\kappa) p(\kappa) d\left(-e^{-\frac{Z}{2}\kappa}\right) \quad (29)$$

Therefore,

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

Then under the assumption that $\lim_{\kappa \rightarrow 0,1} \kappa(1-\kappa)p(\kappa) = 0$, integration by parts gives (10). Therefore we get

$$MSE = p + 2\mathbb{E}_{Z|\theta} \left[(Z+4)g(Z) - (p+2) - \frac{Z}{2}g(Z)^2 - E_{\kappa|Z} \left\{ 2\kappa(1-\kappa) \frac{p'(\kappa)}{p(\kappa)} \right\} \right].$$

□

References

- M. Abramowitz and I. A. Stegun, editors. *Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables*, volume 55 of *Applied Mathematics Series*. National Bureau of Standards, Washington, DC, 1964. Reprinted in paperback by Dover (1974); on-line at <http://www.math.sfu.ca/~cbm/aands/>.
- C. Armero and M. Bayarri. Prior assessments for predictions in queues. *The Statistician*, 43:139–53, 1994.
- J. O. Berger. A robust generalized Bayes estimator and confidence region for a multivariate normal mean. *The Annals of Statistics*, 8(4):716–761, 1980.
- J. O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, 2nd edition, 1985.
- M. Bogdan, J. K. Ghosh, and S. T. Tokdar. A comparison of the Benjamini-Hochberg procedure with some Bayesian rules for multiple testing. In *Beyond Parametrics in*

- Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen*, volume 1, pages 211–30. Institute of Mathematical Statistics, 2008.
- C. M. Carvalho, N. G. Polson, and J. G. Scott. The horseshoe estimator for sparse signals. Discussion Paper 2008-31, Duke University Department of Statistical Science, 2008.
- M. Clyde and E. I. George. Flexible empirical Bayes estimation for wavelets. *Journal of the Royal Statistical Society, Series B (Methodology)*, 62(4):681–98, 2000.
- D. Denison and E. George. Bayesian prediction using adaptive ridge estimators. Technical report, Imperial College, London, 2000.
- M. Figueiredo. Adaptive sparseness for supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1150–9, 2003.
- D. Fourdrinier, W. Strawderman, and M. T. Wells. On the construction of Bayes minimax estimators. *The Annals of Statistics*, 26(2):660–71, 1998.
- D. Fourdrinier, O. Kortbi, and W. Strawderman. Bayes minimax estimators of the mean of a scale mixture of multivariate normal distributions. *Journal of Multivariate Analysis*, 99:74–93, 2008.
- A. Gelman. Prior distributions for variance parameters in hierarchical models. *Bayesian Anal.*, 1(3):515–33, 2006.
- M. B. Gordy. A generalization of generalized beta distributions. Technical report, Board of Governors of the Federal Reserve System, Finance and Economics Discussion Series, 1998.
- I. Gradshteyn and I. Ryzhik. *Table of Integrals, Series, and Products*. Academic Press, 1965.
- C. M. Hans. Bayesian lasso regression. Technical report, Ohio State University, 2008.
- W. James and C. Stein. Estimation with quadratic loss. In *Proc. Fourth Berkeley Symp. Math. Statist. Prob.*, volume 1, pages 361–79, 1961.
- H. Jeffreys. *Theory of Probability*. Oxford University Press, 3rd edition, 1961.
- I. Johnstone and B. W. Silverman. Needles and straw in haystacks: Empirical-Bayes estimates of possibly sparse sequences. *The Annals of Statistics*, 32(4):1594–1649, 2004.
- F. Liang, R. Paulo, G. Molina, M. Clyde, and J. Berger. Mixtures of g -priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103:410–23, 2008.
- Y. Maruyama. Improving on the James–Stein estimator. *Statistics and Decisions*, 14: 137–40, 1999.
- J. B. McDonald and Y. J. Xu. A generalization of the beta distribution with applications. *Journal of Econometrics*, 66:133–52, 1995.

- T. Park and G. Casella. The Bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–6, 2008.
- J. G. Scott and J. O. Berger. An exploration of aspects of Bayesian multiple testing. *Journal of Statistical Planning and Inference*, 136(7):2144–2162, 2006.
- C. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9:1135–51, 1981.
- W. Strawderman. Proper Bayes minimax estimators of the multivariate normal mean. *The Annals of Statistics*, 42:385–8, 1971.
- R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B.*, 58(1):267–88, 1996.
- M. Tipping. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–44, 2001.
- A. Zellner and A. Siow. Posterior odds ratios for selected regression hypotheses. In *Bayesian Statistics: Proceedings of the First International Meeting held in Valencia*, pages 585–603, 1980.