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

Shrink Globally, Act Locally: Sparse Bayesian Regularization and Prediction

Nicholas G. Polson & James G. Scott
University of Chicago, USA *University of Texas, USA*
ngp@chicagobooth.edu james.scott@mccombs.utexas.edu

Summary

We study the classic problem of choosing a prior distribution for a location parameter $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ as p grows large. First, we study the standard "global-local shrinkage" approach, based on scale mixtures of normals. Two theorems are presented which characterize certain desirable properties of shrinkage priors for sparse problems. Next, we review some recent results showing how Lévy processes can be used to generate infinite-dimensional versions of standard normal scale-mixture priors, along with new priors that have yet to be seriously studied in the literature. This approach provides an intuitive framework both for generating new regularization penalties and shrinkage rules, and for performing asymptotic analysis on existing models.

Keywords and Phrases: Lévy processes; shrinkage; sparsity

1. ONE-GROUP ANSWERS TO TWO-GROUP QUESTIONS

Suppose that $(\mathbf{y} \mid \boldsymbol{\beta}) \sim \mathrm{N}(\boldsymbol{\beta}, \sigma^2 I)$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ is believed to be sparse. Many Bayesians, and at least some frequentists, would assume an exchangeable discrete-mixture prior, $\beta_i \sim w \cdot g(\beta_i) + (1 - w) \cdot \delta_0$, and report

$$w(y) = \frac{w \cdot f_1(y)}{w \cdot f_1(y) + (1 - w) \cdot f_0(y)}\,, \tag{1}$$

where $f_0(y) = \mathrm{N}(y \mid 0, \sigma^2)$ and $f_1(y) = \int \mathrm{N}(y \mid \beta, \sigma^2)\ g(\beta)\, d\beta$ are the marginal densities of y under the null and the alternative models, respectively.

Following Efron [2008], we call this the two-groups answer to the two-groups question. Many of this framework's asymptotic properties are well understood, both as the number of means (p) and the number of replicated observations (n) grow [Johnstone and Silverman, 2004, Scott and Berger, 2006, 2010, Muller et al., 2006, Bogdan et al., 2008a,b].

Polson is Professor of Econometrics and Statistics at the Chicago Booth School of Business. Scott is Assistant Professor of Statistics at the University of Texas at Austin.

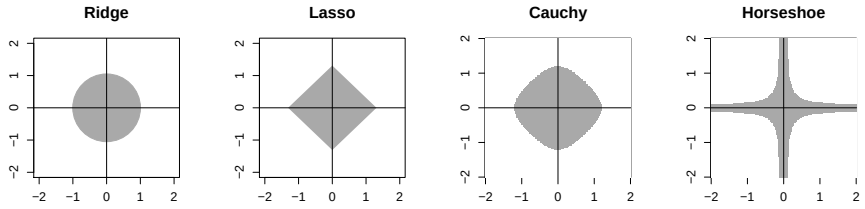

Figure 1: The penalty functions associated with some common priors.

One appealing feature of (1) is that it offers a tentative methodological unification to the multiple-testing problem: Bayesians can interpret $w(y)$ as the posterior probability that y is a signal, while frequentists can interpret $1 - w(y)$ as a local false-discovery rate. Certainly each school of thought calls for nuisance parameters to be handled in different ways. Yet it is comforting that a Bayesian and a frequentist can use essentially the same procedure, and report essentially the same summaries, even if they disagree about their interpretation.

Now consider a sparse regression problem, $(\mathbf{y} \mid \boldsymbol{\beta}) \sim \mathrm{N}(X\boldsymbol{\beta}, \sigma^2 I)$. This is superficially similar to the normal-means problem, yet the tentative unification falls apart. Bayesians are apt to persist in using a two-groups model for the regression parameters. But in machine learning and neoclassical statistics, the dominant approach to sparse regression is penalized least-squares, where $\boldsymbol{\beta}$ is chosen to minimize

$$l(\boldsymbol{\beta}) = \|\mathbf{y} - X\boldsymbol{\beta}\|^2 + \nu \sum_{i=1}^{p} \psi(\beta_i^2) \tag{2}$$

for some regularization penalty ψ (with ν usually chosen by cross validation or marginal maximum likelihood). Under certain choices of ψ, some β_i's may collapse to zero—as in, for example, the lasso penalty of Tibshirani [1996]. Model selection is thereby recast as optimization. For further discussion on this and other similar approaches in machine learning, see [Clarke et al., 2009].

As many previous authors have observed, the sum in (2) can be interpreted as the log posterior density for $\boldsymbol{\beta}$ under a prior $\pi(\beta_i \mid \nu) \propto \exp\{-\nu\psi(\beta_i^2)\}$. Hence the penalized-likelihood solution can be interpreted as a posterior mode (MAP). Within this class of estimators, there has been widespread interest in normal scale-mixture priors, a class that includes widely known forms such as the t and the double-exponential, along with more recent proposals such as the normal/exponential-gamma, the normal/gamma, the improper normal/Jeffreys, and the horseshoe. Figure 1 shows the bivariate penalty functions associated with some common priors.

This might be called the one-group answer to the original two-groups question. Barring the rare case of a true "0–1" loss function, the use of the posterior mode lacks any Bayesian rationale. It is therefore hard to see the potential for true methodological unification in the one-group answer to sparse regression, which seems to dodge the fundamental two-group question of "signal versus noise" altogether.

Nonetheless, the one-group model merits serious attention from Bayesians. For one thing, sparsity can be construed in a weaker sense, where all of the entries in $\boldsymbol{\beta}$ are nonzero, yet most are small compared to a handful of large signals. For example, $\boldsymbol{\beta}$ may be of small ℓ^α norm for some suitably small α, or its entries may decay in absolute value according to some power law [e.g. Johnstone and Silverman, 2004]. This view of sparsity may appeal to Bayesians who oppose testing point null hypotheses, and would rather shrink than select.

Second, not even the staunchest of Bayesians can demand zeros when averaging over models: model-averaged coefficients will be nonzero with probability 1 under the sampling distribution for $\mathbf{y}$, regardless of $\boldsymbol{\beta}$. This simple fact opens the door to the one-group model when the goal is estimation or prediction—albeit only after choosing a one-group model that acts, in some sense, a like a two-groups model.

Finally, the one-group answer can offer substantial computational savings over full-bore model averaging. For a conjugate normal linear model, the difference may be small; for a probit model, where marginal likelihoods of different regression hypotheses cannot be computed in closed form, the difference is substantial, and the one-group model can be used to approximate the model-averaged solution.

The study of oracle properties provides a unifying framework in the classical literature, but no such framework exists for Bayesians. In this paper, we hope to offer a few elements that might form the beginnings of such a framework. First, we review the standard hierarchical-Bayes formulation of global-local shrinkage rules for finite dimension p. Our focus here is on advancing some criteria for evaluating different sparsity priors in terms of their suitability as a default one-group model. We will then discuss the results of some numerical experiments in Section 3.

We then go on to embed the finite-dimensional in a suitable infinite-dimensional generalization by identifying $\boldsymbol{\beta}$ with the increments of a discretely observed Lévy process. This provides a natural setting in which the dimension p grows without bound. In particular, Theorems 3 and 4, along with the associated discussion, establish a mapping from Lévy processes to a wide class of penalty functions.

2. GLOBAL-LOCAL SHRINKAGE RULES

2.1. *The framework*

We will work within the class of global-local scale mixtures of normals:

$$\begin{aligned}(\beta_i \mid \tau^2, \lambda_i^2) &\sim \mathrm{N}(0, \tau^2\lambda_i^2)\\ \lambda_i^2 &\sim \pi(\lambda_i^2)\\ (\tau^2, \sigma^2) &\sim \pi(\tau^2, \sigma^2)\,.\end{aligned}$$

Each λ_i^2 is called a local variance component, while τ^2 is the global variance component (or the regularization parameter ν in the penalized-likelihood formulation).

Let $\Lambda = \mathrm{diag}(\lambda_1^2, \ldots, \lambda_p^2)$. A natural Bayesian approach is to use the posterior distribution $\pi(\Lambda \mid \tau^2, \sigma^2, \mathbf{y})$ to compute the adaptive ridge estimator

$$\hat{\boldsymbol{\beta}}(\tau^2) = \mathrm{E}_{\Lambda|\tau^2,\sigma^2,\mathbf{y}} \left\{ \left(X'X + \sigma^2\tau^2\Lambda^{-1}\right)^{-1} X'\mathbf{y} \right\}. \tag{3}$$

An alternative is to specify a prior in the space defined by an orthogonal matrix U such that, for $Z = XU$ and $\boldsymbol{\alpha} = U'\boldsymbol{\beta}$, $Z'Z = U'X'XU = D$, the diagonal matrix of eigenvalues of $X'X$. Then set $(\boldsymbol{\alpha} \mid \Lambda, \tau^2, \sigma^2) \sim \mathrm{N}(0, \sigma^2\tau^2 n D^{-1}\Lambda)$. In turn, this

implies that $(\boldsymbol{\beta} \mid \Lambda, \tau^2, \sigma^2) \sim \mathrm{N}(0, \sigma^2\tau^2 nUD^{-1}\Lambda U')$. If $\Lambda = I$, the familiar g-prior is recovered. But if $\lambda_i^2 \sim \pi(\lambda_i^2)$, then the resulting "generalized g-prior" will adaptively shrink the principal components of X using the familiar scale-mixture trick.

Either way, one faces the question: which "sparsity" prior to choose? In approaching the literature on this subject, one encounters a thicket of options, of which the following list comprises only a limited subset:

Student-t, $\beta_i \sim t_\xi$, with an inverse-gamma mixing density. The relevance vector machine of Tipping [2001] involves computing posterior modes to find sparse solutions when ξ, the degrees-of-freedom parameter, goes to 0.

Double-exponential, with an exponential mixing density. See, for example, West [1987], Carlin and Polson [1991], Pericchi and Smith [1992], Tibshirani [1996], Park and Casella [2008], and Hans [2009]

Normal/Jeffreys, where $p(\beta_i) \propto |\beta_i|^{-1}$ [Figueiredo, 2003, Bae and Mallick, 2004]. This improper prior is induced by placing Jeffreys' prior upon each local shrinkage term, $p(\lambda_i^2) \propto 1/\lambda_i^2$.

Strawderman–Berger, which has no analytic form, but can easily be written as a scale-mixture model: $(\beta_i \mid \kappa_i) \sim \mathrm{N}(0, \kappa_i^{-1} - 1)$, with $\kappa_i \sim \mathrm{Be}(1/2, 1)$ [Strawderman, 1971, Berger, 1980]. In addition, Johnstone and Silverman [2004] study this model as a possible choice of g in the two-groups model.

Normal/exponential-gamma, with an exponential mixing density and a second-level $\mathrm{Ga}(c, 1)$ prior for the exponential rate parameter [Griffin and Brown, 2005]. This leads to $p(\lambda_i^2) \propto \left(1 + \lambda_i^2\right)^{-(c-1)}$.

Normal/gamma and normal/inverse-Gaussian, respectively characterized by gamma and inverse-Gaussian mixing densities [Caron and Doucet, 2008, Griffin and Brown, 2010].

Horseshoe prior, a special case of a normal/inverted-beta class, where $\lambda_i^2 \sim \mathrm{IB}(a, b)$ has an inverted-beta (or "beta-prime") distribution. Carvalho, Polson, and Scott [2010] study the case where $a = b = 1/2$, while Polson and Scott [2009] generalize the horseshoe model to a wider class of variance mixtures based on power laws.

All of these priors have been nominated, in one way or another, as suitable default models for sparse vectors. This paper will catalogue still other possibilities for $\pi(\lambda_i^2)$. Navigating this thicket demands a set of criteria to help guide modeling choices.

Our preferred approach is to cajole the one-group model into behaving like a two-groups model, where

$$\beta_i \sim w \cdot g(\beta_i) + (1 - w) \cdot \delta_0 \tag{4}$$

for an unknown, common mixing probability w. Assuming g is appropriately heavy-tailed, the posterior mean for β_i under this model is

$$\mathrm{E}(\beta_i \mid w, y_i) \approx w(y_i) \cdot y_i \,,$$

with $w(y_i)$ as in (1). The posterior means the adapt to the level of sparsity in the data through shared dependence upon the unknown mixing probability w.

This effect can most easily be seen if one imagines testing a small number of signals in the presence of an increasingly large number of noise observations. As the noise comes to predominate, the posterior distribution for w concentrates near 0, making it increasingly more difficult for most of the means to be large. Yet any individual y_i can still escape the pull of w's gravity; as long as g is heavy-tailed enough, the likelihood can still overwhelm the prior probabilities in (1).

The same logic can be applied to the one-group model, where the analogue of w is τ^2, the global variance component:

$$\mathrm{E}(\beta_i \mid \lambda_i^2, \tau, y_i) = \left(1 - \frac{1}{1+\tau^2\lambda_i^2}\right) y_i \,. \tag{5}$$

To squelch noise and shrink all of the means toward zero, τ^2 should be small. Yet in order for large signals to override this effect, λ_i^2 must be allowed to be quite large.

These considerations point to two guidelines for the sparse one-group model:

(i) $\pi(\lambda_i^2)$ should have heavy tails.

(ii) $\pi(\tau^2)$ should have substantial mass near zero.

In this formulation, the sparseness problem is the mirror image of the outlier problem [see, for example, West, 1984]. Strong global shrinkage handles the noise; the local λ_i's act to detect the signals, which are outliers relative to τ^2.

We first focus on $\pi(\lambda_i^2)$. The following two theorems help clarify the role of this prior in controlling the behavior of a global-local shrinkage rule.

2.2. *Tail robustness*

Theorem 1 (Tail equivalence). *Suppose that* $(y \mid \beta) \sim \mathrm{N}(\beta, 1)$*, and that* $\pi(\beta) = \int \mathrm{N}(\beta \mid 0, \lambda^2)\pi(\lambda^2)\, d\lambda^2$*. Suppose further that* $\pi(\lambda^2) \sim (\lambda^2)^{a-1} e^{-\eta\lambda^2} L(\lambda^2)$ *as* $\lambda^2 \to \infty$ *for some slowly varying function* L *such that for every* $t > 0$*,* $L(tx)/L(x) \to 1$ *as* $x \to \infty$*. Let* $b = 1$ *if* $\eta > 0$*, and* 0 *otherwise. Then as* $y \to \infty$*,* $m(y) = \int \mathrm{N}(y \mid \beta, 1)\pi(\beta)\, d\beta)$ *satisfies, up to the score of the slowly varying function,*

$$\frac{d}{dy} \ln m(y) \sim \frac{2^b a - 1}{y} - \sqrt{2\eta} \,.$$

Proof. See the Appendix. □

The result is phrased as $y \to \infty$, but with a reversal of sign would also apply as $y \to -\infty$. Note the interesting discontinuity between $\eta = 0$ and $\eta > 0$.

This theorem is useful for pairing with the well known result that

$$\mathrm{E}(\beta \mid y) = y + \frac{d}{dy} \ln m(y) \,,$$

versions of which appear in Masreliez [1975], Polson [1991], Pericchi and Smith [1992], and Carvalho et al. [2010]. Applying this result together with Theorem 1, we see that

$$\lim_{y\to\infty} \{y - \mathrm{E}(\beta \mid y)\} = \sqrt{2\eta} \,,$$

implying that any variance mixture where $\pi(\lambda^2)$ has exponential (or lighter) tails will always shrink observations back to zero by some nondiminishing amount, no matter how large those observations may be.

This becomes a problem when information is shared across components through a global variance component τ^2. Suppose, for example, we have p normal means and choose a double-exponential prior,

$$\pi(\boldsymbol{\beta}) = \frac{1}{2\tau} \exp\left\{ -\sum_{i=1}^{p} \frac{|\beta_i|}{\tau} \right\} .$$

If most of the β_i's are zero, then τ must be small. But then for any $|y_i|$ that are large, the exponential mixing density for λ_i^2 implies that

$$|y_i - \mathrm{E}(\beta_i \mid y_i, \hat{\tau})| \approx \sqrt{2}/\tau \,,$$

an amount of shrinkage that will grow inappropriately severe as one makes τ small enough to squelch the noise. The goal of shrinking the noise toward zero lies in direct conflict with the equally laudable goal of leaving the large signals unshrunk.

The theorem makes it clear, moreover, that any prior where $\pi(\lambda_i^2)$ has an exponential tail will force such a tradeoff in sparse problems. This class of priors includes both the normal/gamma and normal/inverse-Gaussian. If $\eta = 0$, on the other hand, then $\pi(\lambda^2)$ has a polynomial tail, and the amount of shrinkage goes to zero for large signals no matter how small the global variance component. Such priors with redescending score functions are said to be *tail robust*.

2.3. *Predictive efficiency*

The next result relates the behavior of $\pi(\lambda^2)$ to the resulting model's efficiency in reconstructing the true sampling distribution $p(y \mid \beta_0)$. It is a direct consequence of Proposition 4 in Barron [1988] and is a restatement of Lemma 1 in Carvalho et al. [2010]; we therefore omit the proof, but refer to Clarke and Barron [1990] for more on the information theory and Bayes asymptotics.

Let β_0 to denote the true value of the parameter, $p_\beta = p(y \mid \beta)$ denote a sampling model with parameter β, and $\mu(A)$ denote the prior or posterior measure of some set A. Also, let $L(p_1, p_2) = \mathrm{E}_{p_1}\{\log(p_1/p_2)\}$ denote the Kullback–Leibler divergence of p_2 from p_1.

Theorem 2 (Kullback–Leibler risk bounds). *Let $A_\epsilon = \{\beta : L(p_{\beta_0}, p_\beta) \leq \epsilon\} \subset \mathbb{R}$ denote the Kullback–Leibler information neighborhood of size ϵ, centered at β_0. Let $\mu_n(d\beta)$ be the posterior distribution under $\pi(\beta)$ after observing data $y_{(n)} = (y_1, \ldots, y_n)$, and let $\hat{p}_n = \int p_\beta \; \mu_n(d\beta)$ be the posterior mean estimator.*

Suppose that the prior $\pi(\beta)$ is information dense at p_{β_0}, in the sense that $\mu(A_\epsilon) > 0$ for all $\epsilon > 0$. Then the following bound for R_n, the Cesàro-average risk of the Bayes estimator $\hat{p}_n$, holds for all $\epsilon > 0$:

$$R_n = \frac{1}{n} \sum_{j=1}^{n} L(p_{\beta_0}, \hat{p}_j) \leq \epsilon - \frac{1}{n} \log \int_{\beta_0 - \sqrt{\epsilon}}^{\beta_0 + \sqrt{\epsilon}} \pi(\beta) \, d\beta \,.$$

The more mass that the prior $\pi(\beta)$ has in a neighborhood near the true value β_0, the better this bound will be. For any prior whose density function is bounded above by $C/2$ in a neighborhood of β_0,

$$\int_{\beta_0-\sqrt{\epsilon}}^{\beta_0+\sqrt{\epsilon}} \pi(\beta)\, d\beta \leq C\sqrt{\epsilon}\,,$$

where $C < 1$ is typical for most priors. On the other hand, if the prior density has a pole at the true value ($\beta_0 = 0$ being the case of special interest in sparse problems), then the risk bound can be improved. Under the horseshoe prior, for example,

$$\int_{-\sqrt{\epsilon}}^{\sqrt{\epsilon}} \pi(\beta)\, d\beta \geq \sqrt{\epsilon}\log\left(1+\frac{4}{\epsilon}\right) + 2\int_{4/\epsilon}^{\infty} \frac{1}{u^{\frac{1}{2}}(1+u)}\, du\,,$$

a bound proven in Carvalho et al. [2010]. This second integral is easily computed and of order $\epsilon^{1/2}$. Therefore, a prior with a pole at zero can more rapidly recover the true sampling density in sparse situations. We use the term *KL super-efficient* to describe such a prior; for example, the normal/gamma can also be KL super-efficient for certain choices of hyperparameters.

2.4. *The global variance components*

We now turn to $\pi(\tau^2, \sigma^2)$, the prior for the global variance components. An excellent reference on hyperpriors for variance components can be found in Gelman [2006]. We highlight the main options here, and discuss their role in sparse inference.

The standard conjugate choice for $\pi(\tau^2)$ is the inverse-gamma prior. This is quite inappropriate for sparse problems, since it artificially forces τ^2 away from zero. It should be used only with some extrinsic (i.e. subjective) justification.

At least three possibilities avoid this poor behavior. Jeffreys' prior is

$$\pi_J(\sigma^2, \tau^2) \propto \sigma^{-2}(\sigma^2+\tau^2)^{-1}\,,$$

which despite being improper still yields a proper posterior. (Placing independent Jeffreys' priors on σ^2 and τ^2 does not.) Scott and Berger [2006], meanwhile, use a "proper Jeffreys" prior that works for model selection, when it is important to ensure that $(\tau^2 \mid \sigma^2)$ is proper:

$$\pi_{PJ}(\sigma^2, \tau^2) \propto \frac{\sigma^2}{(\sigma^2+\tau^2)^2} \cdot \frac{1}{\sigma^2} = (\sigma^2+\tau^2)^{-2}\,.$$

Finally, Gelman [2006] proposes a half-Cauchy prior on the scale: $\tau \sim \mathrm{C}^+(0, \sigma)$. All three priors are scaled by the error variance σ^2, following Jeffreys [1961].

We are persuaded by the main argument leading to the half-Cauchy prior: that $\pi(\tau)$ evaluates to a positive constant at the origin, and therefore does not overwhelm the marginal likelihood of the data at the globally sparse solution $\tau = 0$. Polson and Scott [2009] also provide an alternative justification for this prior based on its classical risk properties near the origin. These facts, coupled with its mild quadratic decay, make the half-Cauchy an appealing default option. There are surely data sets where it can be beaten, but we have not seen examples where it leads to obviously silly behavior.

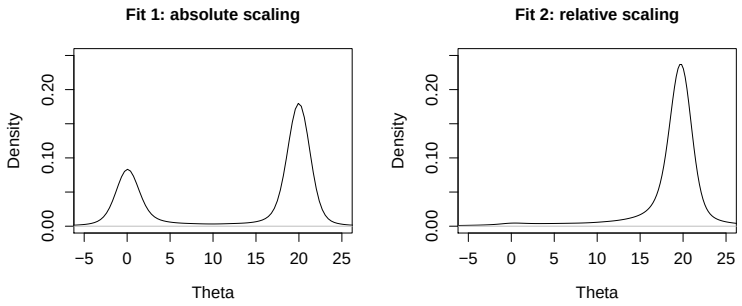

Figure 2: Example 1. Left: the posterior for β when $\tau \sim C^{+}(0,1)$. Right: the posterior when $\tau \sim \mathrm{C}^{+}(0,\sigma)$.

There are many reasons to be leery of empirical-Bayes and cross-validated solutions leading to plug-in estimates for σ^2 and τ^2. For one thing, the marginal maximum-likelihood solution for τ^2 is always in danger of collapsing to the degenerate $\hat{\tau} = 0$ [Tiao and Tan, 1965]. This danger becomes even more acute in sparse problems. Moreover, σ^2 and τ^2 will typically have an unknown, often nonelliptical correlation structure that should ideally be averaged over. Indeed, as the following toy example illustrates, careful handling of uncertainty in the joint distribution for τ and σ can be crucial.

Example 1 Suppose the true model is $\beta = 20$ and $\sigma^2 = 1$. Two observations are available: $y_1 = 19.6$ and $y_2 = 20.4$. Two different versions of the horseshoe prior, where $\lambda_i^2 \sim \mathrm{IB}(1/2, 1/2)$, are entertained. In both cases, σ^2 is unknown and assigned the noninformative prior $1/\sigma^2$. In Model 1, τ is assigned a $\mathrm{C}^{+}(0,1)$ prior; in Model 2, τ is assigned a $\mathrm{C}^{+}(0,\sigma)$ prior, which scales with the unknown error variance.

The posterior distributions for β under Models 1 and 2 are shown in Figure 2. In the first fit using absolute scaling for τ, the posterior is bimodal, with one mode around 20 and the other around 0. This bimodality is absent in the second fit, where τ was allowed to scale relative to σ.

A situation with only two observations is highly stylized, to be sure, and yet the differences between the two fits are still striking. Note that the issue is not one of failing to condition on σ in the prior for τ; the first fit involved plugging the true value of σ into the prior for τ, which is exactly what an empirical-Bayes analysis aims to accomplish asymptotically. Rather, the issue is one of averaging over uncertainty about σ in estimating the signal-to-noise ratio. Similar phenomena can be observed with other scale mixtures [c.f. Fan and Berger, 1992].

Another fundamental issue is that the act of marginalizing over hyperparameter uncertainty changes the implied regularization penalty. Surprisingly, this difference between Bayesian and plug-in analyses may not disappear even in the limit. Sup-

pose, for example, that $\beta_i = \mu + \tau\eta_i$, where $\eta_i \sim \text{DE}(2)$. Then

$$\pi(\boldsymbol{\beta} \mid \mu, \tau) \propto \tau^{-p} \exp\left(-\frac{1}{\tau}\sum_{i=1}^{p}|\beta_i - \mu|\right),$$

leading to the following joint distribution with regularization penalty ν:

$$p(\boldsymbol{\beta}, \mathbf{y} \mid \mu, \nu) \propto \nu^p \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{p}(y_i - \beta_i)^2 + \nu\sum_{i=1}^{p}|\beta_i - \mu|\right)\right\}.$$

The plug-in solution is to estimate μ and ν by cross-validation or marginal maximum likelihood. Meanwhile, a reasonable fully Bayesian solution, at least in the known-σ^2 case, is to use the noninformative prior $\pi(\mu, \tau) \propto 1/\tau$. This yields a marginal prior distribution for $\boldsymbol{\beta}$ that depends upon the order statistics $\beta_{(j)}$ [Uthoff, 1973]. Specifically, define $v_j(\boldsymbol{\beta}) \equiv v_j = \sum_{i=1}^{p}|\beta_{(i)} - \beta_{(j)}|$. Then

$$\begin{aligned}
\pi(\boldsymbol{\beta}) &= (p-2)!\; 2^{-p+1}\sum_{i=1}^{p} w_j^{-1} \qquad (6)\\
w_j &= \begin{cases} 4v_j^{p-1}\left(j - \frac{p}{2}\right)\left(\frac{p}{2} + 1 - j\right), & j \neq \frac{p}{2}, \frac{p}{2}+1 \\ 4v_j^{p-1}\left[1 + (p-1)\left(\beta_{(p/2+1)} - \beta_{(p/2)}\right)v_j^{-1}\right], & j = \frac{p}{2}, \frac{p}{2}+1 \end{cases}.
\end{aligned}$$

Therefore the non-Bayesian estimates $\boldsymbol{\beta}$ using π_{EB} and the Bayesian using π_{FB}:

$$\pi_{EB}(\boldsymbol{\beta} \mid \mathbf{y}) \propto \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{p}(y_i - \beta_i)^2 + \hat{\nu}^{-1}\sum_{i=1}^{p}|\beta_i - \hat{\mu}|\right)\right\} \qquad (7)$$

$$\pi_{FB}(\boldsymbol{\beta} \mid \mathbf{y}) \propto \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{p}(y_i - \beta_i)^2\right) + \frac{(p-2)!}{2^{p-1}}\log\left(\sum_{i=1}^{p}\frac{1}{w_i(\boldsymbol{\beta})}\right)\right\}. \qquad (8)$$

The former is the traditional double-exponential prior, while the latter prior exhibits a rather complicated dependence upon the order statistics of the β_i's (which do not appear in the plug-in expression). It is by no means certain that the two procedures will reach similar answers asymptotically, since this difference in functional form persists for all p [see, for example, Scott and Berger, 2010].

The double-exponential prior coupled with the noninformative prior on μ and τ is just one example where the marginalization in (6) is analytically tractable. But it serves to convey the essence of the problem, which is quite general. The Bayes and plug-in approaches for estimating τ imply fundamentally different regularization penalties for $\boldsymbol{\beta}$, regardless of whether $\boldsymbol{\beta}$ is estimated by the mean or the mode, and regardless of whether marginal maximum likelihood or cross-validation is used.

Neither prior is wrong *per se*, but the stark difference between (7) and (8) is interesting in its own right, and also calls into question the extent to which the plug-in analysis can approximate the fully Bayesian one. While some practitioners may have different goals for empirical Bayes or cross-validation, such comparison is at least reasonable. Many Bayesians use empirical-Bayes as a computational simplification, and many non-Bayesians appeal to complete-class theorems that rely upon

an empirical-Bayes procedure's asymptotic correspondence with a fully Bayesian procedure. Hence questions about where the two approaches agree, and where they disagree, is of interest both to Bayesians and non-Bayesians.

For all these reasons we prefer the Rao-Blackwellized estimator of $\boldsymbol{\beta}$,

$$\mathrm{E}_{\tau|\mathbf{y}}\{\hat{\boldsymbol{\beta}}(\tau^2)\} = \mathrm{E}_{\tau|\mathbf{y}}\{\mathrm{E}_{\Lambda|\tau,\mathbf{y}}(\boldsymbol{\beta} \mid \mathbf{y}, \tau, \Lambda)\}\,,$$

which Bayes' theorem shows to be equivalent to the posterior mean after τ has simply been marginalized away *a priori*.

One approach for estimating $\nu = 1/\tau$ that arises repeatedly in the classical literature is to set $\hat{\nu} = \sqrt{\log p}$, a choice for which interesting asymptotic results obtain. See, for example, Candes and Tao [2007] and Bickel et al. [2009].

This choice can be interpreted as a form of Bonferroni-like correction. Since

$$\int_{-\nu}^{\nu} \frac{\nu}{2} e^{-\nu|\beta_i|}\, d\beta_i = 1 - e^{-\nu^2}\,,$$

the choice of $\nu = \sqrt{\log p}$ implies that

$$P\left(|\beta_i| < \sqrt{\log p} \text{ for all } i\right) = \left(1 - \frac{1}{p}\right)^p \approx e^{-1}\,.$$

Of course, for this choice, all information flow across the components is lost. We conjecture that the Rao-Blackwellized estimator where $\tau \sim \mathrm{C}^+\{0, \sigma(\log p)^{-1/2}\}$ could allow borrowing of information while still clearing the same asymptotic hurdles.

3. NUMERICAL EXPERIMENTS

We have examined a global-local framework for understanding why certain sparsity priors make better default one-group models than others. We now provide numerical evidence that the gains in performance for a prior motivated by this framework can often be large. Most intriguingly, we show that shrinkage rules that are both tail robust and super-efficient corresponds quite closely to the answers one would get if one pursued a more familiar Bayesian approach using a two-groups model. This "BMA mimicry" can result in a lower computational burden than full Bayesian model averaging.

3.1. *Regularized regression*

In our first example, we test the performance of the one-group model against a highly regarded two-groups model. We simulated 500 data sets from the following sparse model with t-distributed signals, $n = 60$, and $p = 40$:

$$\begin{aligned}(\mathbf{y} \mid \boldsymbol{\beta}) &\sim \mathrm{N}(X\boldsymbol{\beta}, I)\\ (\beta_j \mid w) &\sim w \cdot t_3 + (1-w)\cdot \delta_0\\ w &\sim \mathrm{Be}(1, 4)\,,\end{aligned}$$

reflecting signals that were 80% sparse, on average. The elements of the design matrices were independent standard-normal draws.

We then compared three approaches for estimating $\boldsymbol{\beta}$: (1) Bayesian model averaging under the two-groups model, assuming Zellner-Siow priors for each unique

Table 1: Mean sum of squared errors in estimation and prediction for 500 sparse-regression data sets.

	BMA	HS	Lasso-CV
Prediction SSE	89.2	92.2	128.9
Estimation SSE	0.9	0.8	8.6

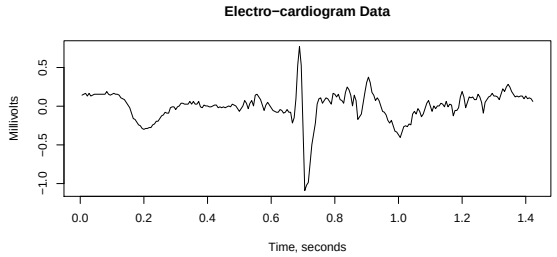

Figure 3: Electro-cardiogram data used as the "true" function f in the wavelet de-noising experiment.

regression model [Zellner and Siow, 1980]; (2) lasso-CV, where ν was chosen using leave-one-out cross-validation; and (3) the horseshoe prior with $\tau \sim \mathrm{C}^+(0, \sigma)$. (Through this section, we use the horseshoe prior, since it is a well-studied example of a prior that is both tail robust and super-efficient.) We measured performance by squared error in estimating $\boldsymbol{\beta}$, and squared error in predicting new values of $\mathbf{y}$ out of sample. To fit the lasso and horseshoe models, we used the R package `monomvn`, described by Gramacy and Pantaleo [2010].

As these results show, both BMA and the horseshoe prior systematically outperformed the lasso, without either one enjoying a noticeable advantage.

3.2. *Wavelet de-noising*

Our second data set (Figure 3) contains 256 electro-cardiogram millivolt readings of one beat of a normal human heart rhythm sampled at 180 Hz, and is available in 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 equispaced intervals, and simulated noisy realizations of f by setting $y_i = f_i + \epsilon_i$, $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$ for $i = 1 \ldots, 256$. 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.

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. DWT: discrete wavelet transform. JS: Johnstone/Silverman. HS: horseshoe prior

	$\sigma = 0.1$		$\sigma = 0.2$		$\sigma = 0.4$	
Procedure	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
HS	9.3	9.3	26.7	26.8	72.4	72.6

Specifically, let d_{jk} represent the kth 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.

We assessed the performance of the horseshoe one-group model against two benchmarks: the discrete wavelet transform, and the two-groups model for normal means described by Johnstone and Silverman [2004]. We measure the performance of an estimator by quadratic 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 $\hat{\beta}$.

As Table 2 shows, the horseshoe prior consistently beat the Johnstone/Silverman procedure, which is the recognized gold standard in the literature on modeling sparse wavelet coefficients. This echoes the results of Scott [2009], who finds the same pattern to hold when the horseshoe prior and the Johnstone/Silverman method are both used to fit a sparse needlet basis to spherical data.

4. PRIORS FROM LÉVY PROCESSES

4.1. *Penalty functions and scale mixtures*

We have phrased the problem of sparse inference in the one-group model as one of estimating a vector of variances: $\pi(\lambda_1^2, \ldots, \lambda_p^2, \tau^2 \mid \mathbf{y})$. The analogy with a stochastic volatility model is instructive, and permits further generalization.

We begin with two simple criteria for characterizing penalty functions—that is, functions $\omega(\boldsymbol{\beta}, \nu)$ such that the minimum of

$$l(\boldsymbol{\beta}) = \|\mathbf{y} - X\boldsymbol{\beta}\|^2 + \omega(\boldsymbol{\beta}, \nu)$$

defines an ω-penalized least-squares estimator for a global penalty parameter $\nu > 0$.

Definition 1 (Separability). *A penalty function $\omega(\boldsymbol{\beta}, \nu)$ is separable if $\omega(\boldsymbol{\beta}, \nu) = \sum_{i=1}^p \psi(\beta_i^2, \nu)$.*

Definition 2 (Global linearity). *A penalty function $\omega(\boldsymbol{\beta}, \nu)$ is globally linear if $\omega(\boldsymbol{\beta}, \nu) = \nu\psi(\boldsymbol{\beta})$.*

Separable penalty functions naturally correspond to exchangeable priors. A penalty function like (2) is both separable and globally linear. These definitions provide the context for a simple theorem from Polson and Scott [2010] that allows us to reinterpret some classic results on normal scale mixtures.

Theorem 3 (Subordinators and penalty functions). *Let T_s, $s \in [0, \nu]$, be a subordinator—that is, a nondecreasing, pure-jump Lévy process—with Lévy measure $\mu(dx)$. Then the cumulant-generating function of T_s corresponds to a separable, globally linear penalty function*

$$\omega(\boldsymbol{\beta}, \nu) = \nu \sum_{i=1}^{p} \psi(\beta_i^2)\,,$$

via the Laplace exponent of the subordinator T_s,

$$\psi(t) = \int_0^\infty \{1 - \exp(tx)\}\mu(dx)\,,$$

Suppose in addition that $\int_0^\infty T_s^{-1/2}\; g(T_s)\, dT_s < \infty$, where $g(T_s)$ is the marginal density of the subordinator at time s. Then the ω-penalized least-squares solution is the posterior mode under an exchangeable normal scale-mixture prior whose mixing measure is expressible in terms of the density of the subordinator:

$$p(\beta_i) \propto \exp\{-\psi(\beta_i^2)\} = \int_0^\infty \mathrm{N}(\beta_i \mid 0, T_\nu^{-1})\; \{T_\nu^{-1/2} g(T_\nu)\}\, dT_\nu\,.$$

Proof. See Polson and Scott [2010]. □

Theorem 3 is useful for several reasons. First, it provides a potentially rich source of new shrinkage rules generated from separable, globally linear penalty functions, since any pure-jump Lévy process with Lévy measure concentrated on $\mathbb{R}^+$ corresponds to such a rule. The behavior of such a shrinkage rule, moreover, can be interpreted in terms of properties of the underlying Lévy measure.

Second, it provides an elegant method for proving that certain distributions—namely, those whose log densities can be identified as the Laplace exponent of some known subordinator—are normal scale mixtures. This naturally leads to the standard generalized-ridge-regression interpretation of most penalty functions. The theorem, for example, suggests a single-line proof of the widely known result that powered-exponential priors are normal scale mixtures [West, 1987].

Example 2 (Powered-exponential priors). Suppose $\log p(\beta_i) = -\nu|\beta_i|^\alpha$. Equivalently, this is $-\nu(\beta_i^2)^{\alpha/2}$, which is easily recognized as the cumulant generating function, evaluated at β_i^2, of a stable subordinator T_ν with index $\alpha/2$.

The Stable(1/2) is equivalent to an inverse-Gaussian distribution, meaning that the lasso can be characterized by an inverse-Gaussian subordinator on a precision scale.

Third, the theorem shows how, for a wide class of priors $\pi(\nu)$, marginalizing over ν can be done via a simple argument appealing to moment-generating functions. This leaves no further hyperparameters to be estimated, as shown by the following theorem, proven in Polson and Scott [2010].

Theorem 4 (Rao-Blackwellized penalty functions). *Suppose*

$$\pi(\boldsymbol{\beta}) \propto \mathrm{E}_{\nu}\left[\exp\left\{-\nu\sum_{i=1}^{p}\psi(\beta_i^2)\right\}\right], \tag{9}$$

where the expectation is with respect to $\pi(\nu)$ defined by the equivalence $\nu \stackrel{D}{=} T_1$, given a subordinator T_s with Lévy measure $\mu(dx)$. Then

$$\begin{aligned}\log\pi(\boldsymbol{\beta}) &= -\chi\left\{\sum_{i=1}^{p}\psi(\beta_i^2)\right\}\\ \chi(t) &= \int_0^{\infty}\{1-\exp(tx)\}\mu(dx)\,,\end{aligned}$$

a composition of the global and local Laplace exponents.

Recall that $\nu = 1/\tau$ in the conditionally normal representation for $\pi(\boldsymbol{\beta})$. Notice that, when the data are allowed to inform the choice of ν in a principled Bayesian way, the mixture regularization penalty loses its global linearity, and the prior loses its structure of conditional independence.

An example helps to demonstrate the theorem's utility.

Example 3 (α-stable mixing). Suppose $\log p(\beta_i \mid \nu) = -\nu|\beta_i|$, where ν is assumed equal in distribution to a standard α-stable subordinator, $0 < \alpha < 1$, observed at time $s = 1$. Then $\psi(\cdot)$ is the square-root function, and $\chi(t) = |t|^{\alpha}$. Therefore the mixture penalty function is

$$\chi\left\{\sum_{i=1}^{p}\psi(\beta_i^2)\right\} = \left(\sum_{i=1}^{p}|\beta_i|\right)^{\alpha}.$$

As before, we see how global mixing changes the functional form of the prior; for example, as $\alpha \to 0$, the density becomes more peaked around zero. A strange situation of idempotence results from the limiting case as $\alpha \to 1$: the limit of this mixture penalty is the same as the original penalty with no global parameter.

One can also attempt to run Theorem 4 in the opposite direction, by recognizing the underlying combination of global and local priors corresponding to a penalty function that takes the compositional form $\chi\left\{\sum_{i=1}^{p}\psi(\beta_i^2)\right\}$.

4.2. *Shrinkage priors as time changes of Brownian motion*

Finally and most importantly, these two theorems are useful as allegory. Many shrinkage priors do not correspond to separable, globally linear penalty functions, and these priors therefore cannot easily be characterized along the lines of Theorem 3 using a subordinator on the precision scale. Nonetheless, the theorem suggests interesting connections between time-changed Brownian motion and shrinkage rules. These connections merit deeper exploration.

A key fact about subordinators is that they are infinitely divisible. Suppose that, as above, we identify the local precisions of p different β_i's with the increments of T, a subordinator, observed on a regular grid. The sum of the p local precisions—an easily interpretable aggregate feature of the $\boldsymbol{\beta}$ sequence—can then be described *a priori* in terms of the behavior of a single random variable T.

Now suppose we want to consider $2p$ β_i's instead, while retaining the same aggregate features of the $\boldsymbol{\beta}$ sequence (now twice as long). This changes requires only that we observe the increments of the original subordinator on a finer grid. Such a scenario is less far-fetched than it sounds; in genomic studies, for example, there is only so much physiological variation to explain, but many successively finer scales of analysis on which to explain it.

From an analytical (and aesthetic) standpoint, the nicest subordinators are the self-similar ones. Self-similar processes have the same distributional form no matter the scale: inverse-Gaussian processes, for example, have inverse-Gaussian increments, no matter how finely one slices them.

The appeal of self-similarity is that we may specify some aggregate feature of the $\boldsymbol{\beta}$ sequence; keep this feature (or its prior) fixed as p grows; and allow the priors for each $\boldsymbol{\beta}_i$ to, in some sense, take care of themselves without our having to worry about their functional form. Put another way: self-similarity ensures that, as p grows and we divide the subordinator into arbitrarily fine increments, the probabilistic structure of the local precisions remains the same—a useful fact if one wishes to contemplate, for example, certain asymptotic features of the double-exponential model.

Formally, let W_t be a standard Wiener process, and define a Lévy process $Z_s = W_{T_s}$, where T_s is a subordinator that defines a random, irregular time scale. The process Z_s is known as subordinated Brownian motion. Its increments will be normal-variance mixtures, with local variances given by the corresponding increments of the subordinator T_s.

The normal/gamma is an example of a prior that divides naturally in this way. If $T_s \sim \mathrm{Ga}(as, b)$ is a gamma subordinator, then its increments follow a gamma distribution at all scales, and one gets normal-gamma β_i's from the increments of W_{T_s} no matter how finely we slice T_s. Slightly abusing notation, we have

$$\sum_{i=1}^{p} \mathrm{Ga}(a/p, b) \stackrel{D}{=} \mathrm{Ga}(a, b)$$

for all p. Here g is the identity mapping from $\mathbb{R}^+$ to $\mathbb{R}^+$.

The normal/inverse-Gaussian distribution has the same property of closure under summation [see, e.g. Barndorff-Nielsen, 1997] and will therefore also be self-similar on the variance scale. Both the normal/inverse-Gaussian and the normal/gamma are examples of self-decomposable mixtures from the class of generalized hyperbolic (GH) distributions [Barndorff-Nielsen, 1978]. The mixing distribution of a GH distribution is characterized by three parameters ($a \in \mathbb{R}, b \geq 0, c \geq 0$):

$$p(\lambda_i^2) = \frac{(c/b)^{a/2}}{2K_a\left(\sqrt{bc}\right)} \; (\lambda_i^2)^{a-1} \; \exp\left\{-\frac{1}{2}\left(b/\lambda_i^2 + c\lambda_i^2\right)\right\} ,$$

where $K_a(\cdot)$ is a modified Bessel function. The resulting mixtures have semi-heavy tails, and so will not yield redescending score functions.

Table 3: A phylogeny of selected normal variance mixtures based on self-decomposable mixing distributions. TR: indicates whether the prior can be tail-robust for certain choices of hyperparameters. SE: indicates whether the prior can be KL super-efficient for certain choices of hyperparameters.

Class	Sub-class	Examples and comments	TR	SE
Generalized z-distributions $(\sigma, \alpha, \beta, \delta, \mu)$	z-distributions	Corresponds to $\delta = 1/2$; well known examples include the $\log F$ and logistic distributions.	N	N
	Meixner	Used in mathematical finance; can be represented as normal variance mixtures.	N	N
Variance mixtures based on power laws	Normal/inverted-beta	Mixing distribution can be represented as an exponentiated z random variable. Examples include the horseshoe prior and Strawderman prior.	Y	Y
	Normal/Lamperti	Mixing distribution can be represented as a ratio of positive stable random variables.	Y	Y
	Normal/ Exponential-Gamma	Special case of the normal/inverted-beta. Similar to the normal/Pareto, which is also known as a Type–II modulated normal distribution.	Y	N
Generalized hyperbolic distributions (a, b, c)	Normal/inverse-Gaussian	Infinite-variation process; corresponds to $a = -1/2$.	N	N
	Normal/gamma	Also known as the variance-gamma process, widely used in finance; corresponds to $b = 0$, $a = c > 0$; related to the Dirichlet process via the gamma subordinator.	N	Y
Variance mixtures based on stable processes	Normal/positive-stable	Related to the Pitman-Yor process via mixtures of alpha-stable subordinators.	Y	Y
	Normal/tempered stable	Widely used in mathematical finance as the CGMY model.	N	N

The horseshoe prior of Carvalho et al. [2010] provides an example that does not submit so readily to either of these approaches. In the usual hierarchical representation of this prior, one specifies a standard half-Cauchy distribution for the local scales: $\lambda_i \sim \mathrm{C}^+(0,1)$. This corresponds to

$$p(\lambda_i^2) \propto (\lambda_i^2)^{-1/2}(1+\lambda_i^2)^{-1}\,,$$

an inverted-beta distribution denoted $\mathrm{IB}(1/2,1/2)$.

This generalizes to the wider class of normal/inverted-beta mixtures [Polson and Scott, 2009], where $\lambda_i^2 \sim \mathrm{IB}(a,b)$. These mixtures satisfy the weaker property of being self-decomposable: if $\lambda_i^2 \sim \mathrm{IB}(a,b)$, then for every $0 < c < 1$, there exists a random variable ϵ_c independent of λ_i^2 such that $\lambda_i^2 = c\lambda_i^2 + \epsilon_c$ in distribution.

Self-decomposability follows from the fact that the inverted-beta distribution is in Thorin's class of generalized gamma convolutions, which are to the gamma distribution what Lévy processes are to the Poisson. If $p(z)$ is a generalized gamma convolution (hereafter GGC), then its moment-generating function can be represented as

$$M(t) = \exp\left\{at + \int_0^\infty \log\left(\frac{1}{1-s/x}\right)\gamma(dx)\right\}\,,$$

where $a = \sup_{[0,\infty)}\{z : p(z) = 0\}$. The measure $\gamma(dx)$ is known as the Thorin measure, and must satisfy some basic integrability conditions similar to those required of a Lévy measure.

Since the gamma distribution is also a Poisson mixture, the Thorin measure is related to the Lévy measure by the Laplace transform

$$\mu(dx) = \frac{dx}{x}\int \exp(-zx)\gamma(dz)\,.$$

We recognize this as the Lévy measure of a Cauchy process, up to the tempering function $h(x) = \int \exp(-zx)\gamma(dz)$. Hence the Thorin measure controls the degree of tempering in a straightforward way.

All GGCs are continuous and unimodal, and all generate self-decomposable normal-variance mixtures with known (though possibly quite complicated) Lévy representations. The density function of a GGC can be represented as

$$p(x) = Cx^{K-1}h(x)\,,$$

where K is the total Thorin measure, and $h(x)$ is completely monotone; up to some further regularity conditions on h, the converse is also true. The class of normal/GGC mixtures seems to contain virtually all commonly used shrinkage priors, but is much more general.

We omit the proof of the fact that the inverted-beta distribution is a GGC, which is surprisingly involved; see Example 3.1 in Bondesson [1990]. The upshot of this result, however, is that the horseshoe prior can be represented as subordinated Brownian motion: the Lévy measure of the inverted-beta is concentrated on $\mathbb{R}^+$, and the corresponding independent-increments process therefore increases only by positive jumps.

Even still, this proof is not constructive, and is of no use whatsoever for actually computing the distribution of the increments. The difficulty becomes plain upon

inspecting the characteristic function of an inverted-beta distribution:

$$\phi(t) = \frac{\Gamma(a+b)}{\Gamma(b)}\, U(a, 1-b, -it)\,,$$

where $U(x, y, x)$ is a confluent hypergeometric function (Kummer function of the second kind). We are not aware of any applicable results for powers of Kummer functions, making it difficult to compute the distribution of sums of inverted-beta random variables.

Representing the horseshoe prior in terms of the increments of a self-similar Lévy process would therefore seem out of reach. But only, it turns out, on the variance scale. If instead we move to a log-variance scale, a self-similar representation can indeed be found, thereby clarifying how the asymptotics of normal/inverted-beta class can be understood intuitively. This self-similar representation is based on the theory of z-distributions.

Table 3 shows the stochastic-process version of many common priors. For details, we refer the reader to Polson and Scott [2010].

5. WHY LÉVY PROCESSES?

5.1. *Some further motivation*

These models all are special cases of the following general form. Let $\Delta = p^{-1}$, and suppose that

$$\beta_i \stackrel{D}{=} Z_{j\Delta} - Z_{(j-1)\Delta}$$

for some arbitrary Lévy process Z_s having Lévy measure $\mu(dx)$. Then upon observing $\mathbf{y} = (y_1, \ldots, y_p)$ with $y_i \sim \mathrm{N}(\beta_i, \sigma^2)$, as in the normal-means problem, we may identify $\mathbf{y}$ with the increments of an interlacing process:

$$y_j \stackrel{d}{=} X_{i\Delta} - X_{(i-1)\Delta}\,,$$

where $X_s = Z_s + \sigma W_s$, a superposition of signals (a Lévy process Z_s) and noise (a scaled Wiener process W_s).

Even though the use of Lévy processes as prior distributions has a well established tradition in Bayesian statistics [e.g. Wolpert et al., 2003], our framework may at first seem overly complex. But we find that it illuminates several aspects of the normal-means problem, and believe it to be worth pursuing.

All of our reasons for thinking so can be subsumed under one basic principle: that in the absence of strong prior information, inferences within the one-group framework should correspond to actual Bayesian models, using reasonable default priors and loss functions. This principle seems almost banal, yet it has serious consequences for the relevance of an estimator's oracle properties. Berger and Pericchi [2001] express this view eloquently:

> One of the primary reasons that we . . . are Bayesians is that we believe that the best *discriminator between procedures* is study of the prior distribution giving rise to the procedures. Insights obtained from studying overall properties of procedures (e.g. consistency) are enormously crude in comparison (at least in parametric problems, where such properties follow automatically once one has established correspondence of the procedure with a real Bayesian procedure). Moreover, we believe that one of the best ways of studying any biases in a procedure is by examining the corresponding prior for biases.

To which we would add only that a procedure's implied loss function can be illuminating, as well.

Theorems 3 and 4 provide the machinery for reverse-engineering the global-local Bayesian models implied by certain penalty functions. The important question is not "How does this penalty function behave?" Rather, it is "What are we assuming about $\boldsymbol{\beta}$ in using this penalty function?"

To illustrate the point, observe that the familiar two-groups model arises as a special case of the general Lévy-process framework: namely, when the Lévy measure μ is that of a compound Poisson process with jump density g and unknown jump rate r. With probability 1, process will have a finite number of jumps on any finite interval. These jumps correspond to the nonzero signals in $\boldsymbol{\beta}$; all other increments of the Z process will be zero.

The discrete-mixture prior is an example of a finite-activity process where the total Lévy measure is finite, but one could also use an infinite-activity process, corresponding to μ being merely sigma-finite. Intuitively, this would correspond to a situation in which the underlying process had an infinite number of small jumps—a natural asymptotic description of a "weakly sparse" vector.

The one-group model and the two-groups model can therefore be subsumed into this single framework, which seems very appealing. Indeed, by the Lévy-Khinchine theorem, any model that preserves the conditional-independence property of the β_i's will fall into this framework, since any stationary càdlàg process with independent increments is completely characterized by its Lévy measure.

By casting the finite-dimensional problem in terms of the marginal distributions of a suitable infinite-dimensional problem, the Lévy process view provides an intuitive framework for asymptotic calculations. Such analysis can be done under one, or both, of two assumptions: that we observe the process longer, or that we observe it on an ever finer grid. Each scenario corresponds quite naturally to a different assumption about how the data's signal-to-noise ratio behaves asymptotically.

From a Bayesian perspective, asymptotic analysis is useful less as a validation step and more as a tool for illuminating what we may, in principle, discover about the underlying "signal" process Z_s on the basis of observing X_s.

For example, it is impossible to recover the entire Lévy measure μ of a discretely observed process that has both a diffusion and a jump component, even as the discretization becomes arbitrarily fine [Aït-Sahalia and Jacod, 2009]. This corresponds to the claim that it is impossible to learn all distributional features of the underlying $\boldsymbol{\beta}$ sequence, even with a huge amount of data.

It is, however, possible to learn certain vague features of the prior, such as its behavior near zero or its tail weight, in the same way that it is possible to learn a higher-level variance component. These are knowable unknowns. Other features, however, are unlearnable in principle, and hence must truly be set in stone by a prior.

Asymptotic investigations, therefore, can help us know where to stop in the "turtles all the way down" approach to hyperparameter specification: first mix over the first-level hyperparameters, then over the second-level, then over the third, and so forth. These are important considerations; if there is one thing our study has clarified, it is the lack of consensus in the literature about what default prior to use for such a basic statistical problem.

We have phrased the problem as one of recovering the $\boldsymbol{\beta}$ sequence. But it is also possible to phrase the problem strictly in terms of claims about observables. Here, the claim would be that, given some Lévy measure, the data look like the increments

of the corresponding stationary, independent-increments process with Lévy triple $\{A, B, \mu(dx)\}$. One can describe the Lévy measure of this process without ever appealing to the notion of a parameter; any subsequent interpretation of the non-Brownian jumps of this process as "signals" is purely optional.

There are also intimate connections between this view of shrinkage and nonparametric Bayesian analysis, in which the goal is to construct distributions over the weights in a countably infinite mixture model. These connections, explored by Kingman [1975] in the context of gamma subordinators, raise the possibility that existing work on regularization can lead to novel priors for sparse infinite mixtures using the normalized jumps of an appropriate subordinator, generalizing the venerable Dirichlet process in practically fruitful ways.

5.2. *Characterizing the signal process*

One natural way of understanding the sparsity of an infinite $\boldsymbol{\beta}$ sequence is through its Blumenthal–Getoor (or sparsity) index, defined as

$$\alpha = \inf \left\{ \delta \geq 0 : \int_{|x| \leq 1} x^{\delta} \mu(\, dx) < \infty \right\} ,$$

where $\mu(dx)$ is the Lévy measure giving rise to increments β_i. This is equal to the index of stability for an alpha-stable process, and provides a straightforward notion of sparsity, since it measures the activity of the small jumps in the process. For a compound Poisson process, $\alpha = 0$. Estimating this index is equivalent to performing model selection for the prior $\pi(\beta_i)$.

To understand the classical approach for estimating the sparsity index, it helps first to imagine a "noiseless" version of the normal-means problem, where $\sigma^2 = 0$. Suppose there are two possible models. Under Model 1, the signals arise from the increments of a tempered stable process Z_s having Lévy measure

$$\mu(dx) = D \exp(-b|x|) \frac{1}{|x|^{1+\alpha}} .$$

Under this model, the log arrival-rate of jumps is linear in jump size and the log of jump size:

$$\log \mu(dx) = -b|x| - (1 + \alpha) \log |x| + \log D .$$

Under Model 2, the signals are from a compound Poisson process with Gaussian jumps. Then the log arrival rate is linear in size and the square of size:

$$\log \mu(dx) = -b|x| - c|x|^2 + K .$$

Hence the model-choice problem for the Lévy measure—that is, the problem of choosing between two possible priors for the signals—boils down to a choice of which linear model best describes the log arrival rate of jumps.

A crude non-Bayesian approach is to bin up the jumps into disjoint intervals defined by their size; compute arrival rates by counting how many jumps fall into each bin; and regress log arrival rate on jump size, plus either log jump size or the square of jump size. If one linear model fits the arrival rates better than the other, the corresponding Lévy measure and sparsity index are supported. This approach, and similar ones, are well studied in mathematical finance, where the

need to account for jumps in the movement of asset prices has long been recognized [e.g. Eraker et al., 2002].

Remarkably, such an approach for recovering the sparsity index still works even in the presence of a Brownian component. This runs contrary to all intuition: an infinite number of arbitrarily small jumps would seem impossible to separate from Gaussian noise, which itself can be thought of as an aggregation of tiny, independent effects. Nonetheless, disentanglement is possible. For example, Aït-Sahalia and Jacod [2009] define the power-variation estimator of a process X_s, $s \in [0, T]$, as

$$\alpha = \inf_{\delta} \left\{ \lim_{\Delta_p \to 0} \sum_{i=1}^{T/\Delta_p} |X_{i\Delta_p} - X_{(i-1)\Delta_p}|^{\delta} \right\},$$

and are able to estimate this quantity consistently as $\Delta_p \to 0$. This recovers the sparsity index of the underlying jump process.

Such estimators are typically quite inefficient, and make use of asymptotic arguments that are likely anathema to most Bayesians. They do, however, point the way to one essential fact: that there is information in the data, however poor, about the sparsity of the signal process. The asymptotic assumptions, moreover, are quite similar to the assumptions made by, for example, Bogdan et al. [2008a] in their characterization of the limiting performance of the Bayesian two-groups model. The door would seem open for a formal Bayesian treatment of the problem.

REFERENCES

Y. Aït-Sahalia and J. Jacod. Estimating the degree of activity of jumps in high frequency data. *The Annals of Statistics*, 37:2202–44, 2009.

K. Bae and B. Mallick. Gene selection using a two-level hierarchical Bayesian model. *Bioinformatics*, 20(18):3423–30, 2004.

O. Barndorff-Nielsen. Hyperbolic distributions and distributions on hyperbolae. *Scandinavian Journal of Statistics*, 5(151–7), 1978.

O. Barndorff-Nielsen. Normal inverse Gaussian distributions and stochastic volatility modeling. *Scandinavian Journal of Statistics*, 24:1–13, 1997.

O. Barndorff-Nielsen, J. Kent, and M. Sorensen. Normal variance-mean mixtures and z distributions. *International Statistical Review*, 50:145–59, 1982.

A. R. Barron. The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions. Technical Report 7, University of Illinois at Urbana–Champaign, 1988.

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 and L. Pericchi. Objective Bayesian methods for model selection: introduction and comparison. In *Model Selection*, volume 38 of *Institute of Mathematical Statistics Lecture Notes – Monograph Series*, pages 135–207. Beachwood, 2001.

P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of lasso and Dantzig selector. *Annals of Statistics*, 37:1705–32, 2009.

M. Bogdan, A. Chakrabarti, and J. K. Ghosh. Optimal rules for multiple testing and sparse multiple regression. Technical Report I-18/08/P-003, Wroclaw University of Technology, 2008a.

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, 2008b.

L. Bondesson. Generalized gamma convolutions and complete monotonicity. *Probability Theory and Related Fields*, 85:181–94, 1990.

E. Candes and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–51, 2007.

B. P. Carlin and N. G. Polson. Inference for nonconjugate Bayesian models using the gibbs sampler. *The Canadian Journal of Statistics*, 19(4):399–405, 1991.

F. Caron and A. Doucet. Sparse Bayesian nonparametric regression. In *ICML '08: Proceedings of the 25th international conference on Machine learning*, pages 88–95. ACM, 2008.

C. M. Carvalho, N. G. Polson, and J. G. Scott. The horseshoe estimator for sparse signals. *Biometrika*, to appear, 2010.

B. Clarke and A. R. Barron. Information-theoretic asymptotics of Bayes methods. *IEEE Trans. Inf. Theory*, 36:453–71, 1990.

B. Clarke, E. Fokoue, and H. H. Zhang. *Principles and Theory for Data Mining and Machine Learning*. Springer, 2009.

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.

B. Efron. Microarrays, empirical Bayes and the two-groups model (with discussion). *Statistical Science*, 1(23):1–22, 2008.

B. Eraker, M. Johannes, and N. Polson. The impact of jumps in volatility and returns. *Journal of Finance*, 58:1269–1300, 2002.

T. Fan and J. O. Berger. Behaviour of the posterior distribution and inferences for a normal mean with t prior distributions. *Stat. Decisions*, 10:99–120, 1992.

M. Figueiredo. Adaptive sparseness for supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1150–9, 2003.

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

R. Gramacy and E. Pantaleo. Shrinkage regression for multivariate inference with missing data, and an application to portfolio balancing. *Bayesian Analysis*, 5(2), 2010.

J. Griffin and P. Brown. Alternative prior distributions for variable selection with very many more variables than observations. Technical report, University of Warwick, 2005.

J. Griffin and P. Brown. Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, 5(1):171–88, 2010.

C. M. Hans. Bayesian lasso regression. *Biometrika*, 96(4):835–45, 2009.

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.

J. Kingman. Random discrete distributions. *Journal of the Royal Statistical Society (Series B)*, 37(1):1–22, 1975.

C. Masreliez. Approximate non-Gaussian filtering with linear state and observation relations. *IEEE. Trans. Autom. Control*, 1975.

P. Muller, G. Parmigiani, and K. Rice. FDR and Bayesian multiple comparisons rules. In *Proceedings of the 8th Valencia World Meeting on Bayesian Statistics*. Oxford University Press, 2006.

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

L. R. Pericchi and A. Smith. Exact and approximate posterior moments for a normal location parameter. *Journal of the Royal Statistical Society (Series B)*, 54(3):793–804, 1992.

N. G. Polson. A representation of the posterior mean for a location model. *Biometrika*, 78:426–30, 1991.

N. G. Polson and J. G. Scott. Alternative global–local shrinkage rules using hypergeometric–beta mixtures. Technical Report 14, Duke University Department of Statistical Science, 2009.

N. G. Polson and J. G. Scott. Local shrinkage rules, Lévy processes, and regularized regression. Technical report, University of Texas at Austin, 2010.

J. G. Scott. Flexible learning on the sphere via adaptive needlet shrinkage and selection. Technical Report 09, Duke University Department of Statistical Science, 2009.

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.

J. G. Scott and J. O. Berger. Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, 2010. to appear.

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

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

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.

V. Uthoff. The most powerful scale and location invariant test of the normal versus the double exponential. *The Annals of Statistics*, 1(1):170–4, 1973.

M. West. Outlier models and prior distributions in Bayesian linear regression. *Journal of the Royal Statistical Society (Series B)*, 46(3):431–9, 1984.

M. West. On scale mixtures of normal distributions. *Biometrika*, 74(3):646–8, 1987.

R. Wolpert, K. Ickstadt, and M. Hansen. A nonparametric Bayesian approach to inverse problems. In *Bayesian Statistics 7*. Oxford University Press, 2003.

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.

APPENDIX

Proof of Theorem 1

Write the likelihood as

$$(y \mid z) \sim \mathrm{N}(0, z) \;\text{ where } z = 1 + \lambda^2 \sim \pi(z)$$

where $z = 1 + \lambda^2$ with induced prior $\pi(z)$. If $\pi(\lambda^2)$ satisfies the tail condition of the theorem, then so will $\pi(z)$:

$$\pi(z) \sim z^{a-1} e^{-\eta z} L(z) \;\text{ as }\; z \to \infty$$

Then the marginal likelihood of the observation y is a scale mixture of normals,

$$m(y) = \int_1^\infty \frac{1}{\sqrt{2\pi z}} e^{-\frac{y^2}{2z}} \pi(z)\, dz\,.$$

The rest of the proof follows Theorem 6.1 of Barndorff-Nielsen et al. [1982], which shows that

$$m(y) \sim \begin{cases} |y|^{2a-1} L(y^2) & \text{if } \eta = 0 \\ |y|^{a-1} e^{-\sqrt{2\eta}|y|} L(|y|) & \text{if } \eta > 0 \end{cases}$$

as $y \to \infty$. The form of the score function then follows immediately.