Local Shrinkage Rules, Lévy Processes, and Regularized Regression

NICHOLAS G. POLSON
Booth School of Business
University of Chicago

JAMES G. SCOTT
McCombs School of Business
University of Texas at Austin

Original version: June 2010
Revised: October 2010

Abstract

We use Lévy processes to generate joint prior distributions for a location parameter $\beta = (\beta_1, \ldots, \beta_p)$ as p grows large. This leads to the class of local-global shrinkage rules. We extend this framework to large-scale regularized regression for $p > n$ problems, and provide thorough comparisons with current methodologies.

Keywords: Lévy processes; normal scale mixtures; shrinkage; sparsity; PCR; PLS.

1 Introduction

This paper considers two questions relevant to the topic of Bayesian regularized regression, where $(\mathbf{y}|\beta) \sim \mathrm{N}(X\beta, \sigma^2 I)$. First, where do good default priors for the parameter vector β come from? Second, how can these default priors be used most fruitfully in high-dimensional regression problems? We are most interested in so-called "large p" problems, where the number of predictors far exceeds the number of observations, since these problems elude most attempts at a simple solution.

The traditional answer to the first question is that shrinkage priors come from scale mixtures of normals. This class of priors has been used to generate many popular procedures for regularized regression, most notably the lasso (Tibshirani, 1996) and Bayesian variants (Park and Casella, 2008; Hans, 2009).

Sections 2 and 3 of our paper offer a more general answer: that shrinkage priors come from Lévy processes. This approach, which generalizes normal scale-mixture priors to infinite-dimensional, conditionally non-Gaussian settings, and provides an intuitive framework for generating new regularization penalties and shrinkage rules. The use of Lévy processes in high-dimensional Bayesian modeling has been gaining in popularity (e.g. Wolpert and Taqqu, 2005; Wolpert et al., 2010). Our approach differs from this line of work, in that we wish to use the theory of Lévy processes to provide a general framework of penalty functions, shrinkage priors with exchangeable structure, and the relationship between them.

The traditional answer to the second question involves an exchangeable global–local shrinkage hierarchical model for β:

$$\begin{aligned} (\beta_j \mid \tau^2, \lambda_j^2) &\sim \mathrm{N}(0, \tau^2 \lambda_j^2) && (1) \\ \lambda_j^2 &\sim p(\lambda_j^2) && (2) \\ (\tau^2, \sigma^2) &\sim p(\tau^2, \sigma^2)\,. && (3) \end{aligned}$$

Each λ_j^2 is called a local variance component, while τ^2 is the global variance component. Different choices for $p(\lambda_j^2)$ lead to different marginal priors for β_j.

Sections 4 and 5 offer a complementary answer to this question, by placing local shrinkage priors on certain linear combinations of the β_j's, and not on the β_j's themselves. These linear combinations are given by the right-singular vectors of the design matrix. Our approach therefore builds upon the work of Frank and Friedman (1993), Clyde et al. (1996), Denison and George (2000), and West (2003). These authors provide a unified framework for ridge regression (RR), principal-component regression (PCR), partial least-squares (PLS), and the g-prior. We generalize this framework to $p > n$ problems in a way that is intimately related to other recent work on generalized g-priors (Maruyama and George, 2010).

On a wide variety of real and simulated large-p regression problems, our approach is competitive with existing state-of-the-art methods. The fact that it is both simple and easily implemented—essentially, it reduces a regression problem of size p to a normal-means problem of size n —does not seem to preclude good empirical performance.

2 Global–local shrinkage priors

In machine learning and classical statistics, the dominant approach to regularized regression is penalized least-squares, where β is chosen to minimize

$$l(\beta) = \|\mathbf{y} - X\beta\|^2 + \nu \sum_{i=1}^{p} \psi(\beta_i^2) \tag{4}$$

for some regularization penalty ψ. The penalty term ν is usually chosen by cross validation. Under certain choices of ψ—for example, the lasso penalty of Tibshirani (1996)—some of the β_i's may be estimated to be zero. As many previous authors have observed, the sum in (4) can be interpreted as the log posterior density for β under a prior $(\beta_i \mid \nu) \propto \exp\{-\nu\psi(\beta_i^2)\}$. 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 (1–3). This subclass includes widely known forms such as the t and the double-exponential, along with some of the more recent proposals on the following list.

Horseshoe prior, a special case of a normal/inverted-beta class, where $\lambda_i^2 \sim \mathrm{IB}(a,b)$ has an inverted-beta distribution (Carvalho et al., 2010; Polson and Scott, 2009).

Normal/Jeffreys, where $p(\beta_i) \propto |\beta_i|^{-1}$ (Figueiredo, 2003; Bae and Mallick, 2004). It arises from placing Jeffreys' prior upon each local variance: $p(\lambda_i^2) \propto 1/\lambda_i^2$.

Normal/exponential-gamma, where $\lambda_j^2 \sim \mathrm{Ex}(r)$, and where there is a second-level $\mathrm{Ga}(c,1)$ prior for the exponential rate parameter r (Griffin and Brown, 2005). Marginally, this gives $p(\lambda_i^2) \propto \left(1+\lambda_i^2\right)^{-(c-1)}$.

Normal/gamma and normal/inverse-Gaussian, where the local variances receive gamma or inverse-Gaussian mixing densities (Caron and Doucet, 2008; Griffin and Brown, 2010).

Full posterior inference under these priors can be viewed as a Bayesian analogue of penalized-likelihood estimation.

An obvious question is: why would Bayesians consider such an approach to a sparse problem, when these priors do not explicitly allow for the possibility that some of the β_j's are zero? We can think of at least three reasons.

First, suppose that one proceeds in the traditional Bayesian way, by averaging over different submodels in proportion to their posterior probabilities. These model-averaged coefficients will be nonzero with probability 1 under the sampling distribution for $\mathbf{y}$, regardless of β, and hence may be operationally indistinguishable from the posterior mean of a carefully chosen shrinkage prior.

Second, many Bayesians oppose testing point null hypotheses, and would rather shrink than select, on the grounds that point nulls are unrealistic. Sparse shrinkage priors offer a nice compromise. They discount the possibility that $\beta_j = 0$, yet they sift signals from noise more aggressively than a traditional elliptically symmetric prior.

Finally, the pure-shrinkage answer can offer computational gains over Bayesian model averaging. For a normal linear model with conjugate priors, the difference may be small. But for cases where marginal likelihoods of different regression hypotheses cannot be computed in closed form, the difference may be substantial, and the shrinkage approach can be used to approximate the model-averaged solution.

To illustrate this third argument, we simulated data from a probit model with $p = 25$ and $n = 500$:

$$\begin{aligned} y_i &= 1_{z_i > 0} \text{ for } i = 1, \ldots, n \\ z &\sim \mathrm{N}(X\beta, I), \end{aligned}$$

where β contained 20 zeros along with 5 nonzero entries, all equal to $\sqrt{5}$—a so-called "r-spike signal" with $r = 5$ and $||\beta||^2 = p$. The rows of X were simulated from a multivariate normal distribution whose covariance matrix was drawn from an inverse-Wishart distribution, centered at I_p and with $p+2$ degrees of freedom.

We simulated 100 data sets from this model, and compared four approaches for estimating β using the probit link function: (1) maximum likelihood, using the `glm` function

Table 1: Median and mean sum of squared errors in reconstructing the probit r-spike signal in 100 simulated data sets.

	MLE	Lasso-CT	Lasso-CV	HS
Median SSE	19.0	15.3	12.3	0.7
Mean SSE	68.6	15.4	11.7	1.6

in R; (2) lasso-CT, using the lasso penalty and choosing $\nu = \sqrt{2 \log p}$ as in Candes and Tao (2007); (3) lasso-CV, with ν chosen by generalized cross-validation; and (4) HS, the horseshoe posterior-mean estimator (Carvalho et al., 2010). We measured accuracy in estimating β by squared-error loss.

Bayesian model averaging would be difficult here. The issue is that marginal likelihoods of regression submodels are not available in closed form, even assuming a conditionally conjugate prior for β. Either high-dimensional numerical integration or a Laplace approximation must be used instead. By contrast, a pure-shrinkage model is no harder to fit for binary data than it is for continuous data, using the simple trick of data augmentation.

Table 1 shows the median and mean sum of squared errors realized over the 100 simulations. The pure-shrinkage Bayesian model outperformed the alternatives by a wide margin.

This discussion leaves open the question of how one should choose a prior $\pi(\lambda_j^2)$, or equivalently a penalty function. As the probit r-spike example illustrates, different choices can lead to large differences in performance. There are many options which have primarily been evaluated in a Bayesian framework. Only the lasso/double-exponential approach has been evaluated extensively under both.

A natural question is: how can we translate between the Bayesian and penalized-likelihood formulations? For certain penalty functions, the corresponding scale-mixture representation is known. Likewise, for certain choices of $p(\lambda_j^2)$, the marginal prior $p(\beta_j)$, and thus the corresponding penalty function, is known. In the following section, we use the theory of Lévy processes to establish a series of three (successively more general) characterizations of shrinkage priors and their relationship with penalty functions.

3 Priors from Lévy processes

3.1 Penalty functions and scale mixtures

We begin with two simple definitions.

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

Global linearity A penalty function $\omega(\beta, \nu)$ is globally linear if $\omega(\beta, \nu) = \nu \psi(\beta)$.

Separable penalty functions correspond to exchangeable priors—that is, those with a structure of conditional independence for the β_j's. The penalty function in (4), for example, is both separable and globally linear. These definitions provide the context for a simple theorem that allows us to reinterpret some classic results on normal scale mixtures.

Theorem 1. *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 log moment-generating function of T_s corresponds to a separable, globally linear penalty function*

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

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:

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

Proof. See the Appendix. □

Theorem 1 is useful for at least four reasons.

First, it provides a rich source of new shrinkage rules generated from separable, globally linear penalty functions, since any subordinator 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—in particular, its behavior near zero.

Second, the theorem provides an alternate method for proving that certain distributions—namely, those whose log densities can be identified as the Laplace exponent of a pure-jump, nondecreasing Lévy process—have a scale-mixture representation. For example, it leads to the fact that powered-exponential priors are normal scale mixtures (West, 1987).

Example Let $\log p(\beta_i) = -\nu|\beta_i|^\alpha$. Write this instead as $-\nu(\beta_i^2)^{\alpha/2}$. This is easily recognized as the log moment-generating function, evaluated at β_i^2, of a positive alpha-stable subordinator T_ν with stability index $\alpha/2$.

The Stable$(1/2)$ is equivalent to an inverse-Gaussian distribution. Therefore, the double-exponential prior (and thus the lasso penalty function) can be characterized by an inverse-Gaussian subordinator on a precision scale. This complements the lasso's well-known characterization in terms of an exponential mixing distribution for λ_j^2.

Third, the theorem demonstrates that, for a wide class of priors $p(\nu)$, fully Bayesian marginalization over ν can be done automatically, without the need for cross validation or further estimation procedures. We formulate this in the next theorem.

Theorem 2. *Suppose that*

$$p(\beta) \propto E_\nu \left[\exp\left\{ -\nu \sum_{i=1}^{p} \psi(\beta_i^2) \right\} \right] , \tag{5}$$

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

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

Proof. See the Appendix. □

Upon marginalizing over ν with respect to some prior, the mixture regularization penalty loses its global linearity, and the prior loses its structure of conditional independence. Consider the example of bridge estimation with an alpha-stable prior for the regularization parameter.

Example Let $\log p(\beta_i \mid \nu) = -\nu |\beta_i|$, 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, in particular its peakedness near zero. 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.

Fourth and finally, these two theorems are useful for the further generalizations that they suggest. Many shrinkage priors do not correspond to separable, globally linear penalty functions, and therefore Theorem 1 does not pertain to these priors. Nonetheless, the theorem suggests interesting connections between time-changed Brownian motion, Lévy processes, and shrinkage rules.

3.2 Shrinkage priors as time changes of Brownian motion

A nice fact about subordinators—indeed, all Lévy processes—is that they are infinitely divisible. For example, suppose that we identify the local precisions of p different β_i's with the increments of T_s, a subordinator, observed on a regular grid. The sum of the p local precisions—an easily interpretable aggregate feature of the β sequence—can then be described *a priori* in terms of the behavior of a single random variable T. If we were then to consider $2p$ β_i's instead, but wished to retain the same aggregate features of the (now longer) β sequence, we must merely slice up the increments of the original subordinator on a finer grid.

Self-similarity is a more restrictive, but very appealing, property. It will ensure 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 study a procedure's asymptotic properties.

The relevant aggregate feature of the β sequence that we choose to specify, however, need not be on the precision scale. We now consider two examples that show the generality of this approach, which in many ways is the natural location-vector analogue of the stick-breaking construction for infinite-dimensional probability vectors (c.f. Kingman, 1975). Two inputs are required: (1) a self-similar random variable $z \overset{D}{=} \sum z_i$, with z_i taking values in Ω; and (2) a transformation $g : \Omega \to \mathbb{R}^+$. We then identify the local variancies with the increments of Brownian motion observed at random times: $\lambda_i^2 = g(z_i)$.

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 called subordinated Brownian motion. It is the natural infinite-dimensional generalization of a normal scale mixture.

The normal/gamma distribution is an example of a well-known shrinkage prior 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. We have $\sum_{i=1}^{p} \mathrm{Ga}(a/p, b) \overset{D}{=} \mathrm{Ga}(a, b)$ for all p with g 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.

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 $\lambda_i^2 = c\lambda_i^2 + \varepsilon_c$ in distribution.

We omit the proof of the fact that the inverted-beta distribution is self-decomposable, which is quite difficult; see Example 3.1 in Bondesson (1990). The consequence of this fact, however, is that the horseshoe prior can be represented as subordinated Brownian motion.

Because the proof is not constructive, however, the subordinator itself is unknown. 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). A characteristic function of this form makes it very 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, just as a self-similar representation of the lasso model can be found on the precision scale. Appendix B derives this self-similar characterization.

There are many other examples of priors derived from time-changed Brownian motion. Barndorff-Nielsen and Shephard (2001) study the class of normal/modified-stable processes, where the mixing distribution is based on exponential and power tempering (or tilting) of a positive α-stable subordinator. Another interesting generalisation is the Normal-Lamperti distribution with mixing density

$$p(\lambda_j^2) = \frac{\sin(\pi\alpha)}{\pi} \frac{(\lambda_j^2)^{\alpha-1}}{(\lambda_j^2)^{2\alpha} + 2(\lambda_j^2)^{\alpha} cos(\pi\alpha) + 1}$$

for $\lambda_j^2 > 0$. Depending on the choice of α, this density can share the two distinguishing properties of the inverted-beta class: polynomial tails, and the possibility of diverging near zero. These two behaviors are, however, coupled by a single parameter under the Lamperti, whereas the inverted-beta has separate parameters for tail decay and behavior near the origin. The horseshoe mixing distribution, $\mathrm{IB}(1/2, 1/2)$, is also a special case of this family ($\alpha = 1/2$).

Finally, the normal/exponential-gamma model of Griffin and Brown (2005) also has polynomial tails:

$$p(\lambda_i^2) \propto \left(1 + \lambda_i^2\right)^{-(c-1)} \, ,$$

a special case of the inverted-beta that results from assuming that λ_i^2 is conditionally exponential with a gamma-distributed rate parameter.

Table 2 lists categorizes these shrinkage priors along two further lines: tail-robustness and super-efficiency. Intuitively, tail robustness (TR) refers to whether a prior has sufficiently heavy tails to avoid over-shrinking large signals in the presence of a large volume of noise. Super-efficiency (SE) refers to whether the prior density $p(\beta_j)$ is unbounded at zero. For details, we refer the reader to Polson and Scott (2010).

Table 2: Selected normal variance mixtures based on self-decomposable mixing distributions.

Class	Sub-class	Examples and comments	TR	SE
Generalized z distributions $(\sigma, \alpha, \beta, \delta, \mu)$	z-distributions	Case when $\delta = 1/2$; 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

3.3 The general Lévy-process case

A general formualtion is available. Let $\Delta = p^{-1}$, and let

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

for some arbitrary Lévy process Z_s having Lévy measure $\mu(dx)$. Then upon observing $\mathbf{y} = (y_1, \ldots, y_p)$ with $y_i \sim \mathrm{N}(\beta_i, \sigma^2)$, identify $\mathbf{y}$ with the increments of the interlacing process $X_s = Z_s + \sigma W_s$:

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

The observations are then a superposition of signals (a Lévy process Z_s) and noise (a scaled Wiener process W_s).

The familiar discrete mixture model, whereby β_j is either "in" or "out" of the model, arises as a special case of this Lévy-process framework: namely, when the Lévy measure μ is that of a compound Poisson process. With probability 1, process will have a finite number of jumps on any finite interval. These jumps correspond to the nonzero signals in β; all other increments of the Z process will be zero. The Lévy density of Z_s describes the distribution of the signals, while the unknown jump rate describes their relative abundance.

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, where the Lévy measure is merely sigma-finite. This would mean that the underlying process had an infinite number of very tiny jumps—in other words, that no β_j's are zero, but that most are of insignificant size compared to σ.

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.

In the following section, we will extend this local-global mixture framework to high-dimensional regularized regression problems. This will provide a link between the normal-means problem, ridge regression, principal component regression, partial least squares, Bayesian local shrinkage rules, and the g-prior.

4 Regularized regression when $p < n$

4.1 Connections among RR, PCR, PLS, and the g-prior

We now turn to the question of how these local shrinkage priors can be used most fruitfully in regression problems. Instead of the traditional approach in (1)–(3), we use of local-shrinkage priors in the coordinate system defined by the principal components of $X'X$. This approach will generalize more easily to the $p > n$ case.

Let $X = UDW'$ represent the singular-value decomposition of the design matrix X. If $n > p$, then X is of full column rank, and $D = \mathrm{diag}(d_1, \ldots, d_p)$ is a diagonal matrix of nonzero singular values ordered $d_1 > \cdots > d_p$. Both U and W are orthogonal matrices, of dimensions $n \times p$ and $p \times p$, respectively. Moreover, W is also the matrix of eigenvectors $\{w_j\}$ for the cross-product matrix $S = X'X$, with corresponding eigenvalues d_j^2.

The original regression relationship may be re-expressed in the orthogonalized (or principal-component) space as $y = Z\alpha + \varepsilon$, where $Z = UD$ and $\alpha = W'\beta$. The ordinary least-squares (OLS) estimate for α is $\hat{\alpha} = (Z'Z)^{-1}Z'y = D^{-1}U'y$.

Following Frank and Friedman (1993), the shrinkage structures for many common regularization approaches can be understood by expanding their solutions in the original coordinate system in terms of the eigenvectors $\{w_1, \ldots, w_p\}$ and the OLS coefficients $\hat{\alpha}$:

$$\hat{\beta}^M = \sum_{j=1}^{p} \kappa_j^M \hat{\alpha}_j w_j \, . \tag{6}$$

Here M denotes the method, and the κ_j^M's are method-specific shrinkage weights that scale the OLS solution along each of the directions w_j.

Both ridge regression and principal-components regression use shrinkage weights that do not depend on the response values $\mathbf{y}$. The ridge-regression solution is $\kappa_j^{RR} = d_j^2/(\nu + d_j^2)$ for a fixed regularization parameter ν, while the K-component PCR solution is

$$\kappa_{jK}^{PCR} = \begin{cases} 1, & d_j^2 \geq d_K^2 \\ 0, & d_j^2 < d_K^2 \end{cases} \, .$$

The posterior mean under the g-prior also fits in this shrinkage structure; it corresponds to $\kappa_j^g = g/(1+g)$, thereby shrinking the solution vector along all eigen-directions by a common factor.

The shrinkage weights under partial least squares, on the other hand, depend nonlinearly upon the response values $\mathbf{y}$ through the OLS solution $\hat{\alpha}$. Using the expressions in Frank and Friedman (1993), for the K-component solution we have

$$\kappa_{jK}^{PLS} = \sum_{k=1}^{K} \theta_k d_j^{2k} \, ,$$

where $\theta = \{\theta_1, \ldots, \theta_K\}'$ is equal to $W^{-1}\eta$, with

$$\eta_k = \sum_{j=1}^{p} \hat{\alpha}_j^2 d_j^{2(k+1)} \text{ and } W_{kl} = \sum_{j=1}^{p} \hat{\alpha}_j^2 d_j^{2(k+l+1)} \, .$$

4.2 A Bayesian interpretation

These four procedures differ only in the way that they scale the OLS estimates for the regression parameter in the orthogonal coordinate system defined by W. It is therefore natural to consider them as special cases of an encompassing Bayesian model.

The g-prior estimator is explicitly Bayesian, wherein $\beta \sim \mathrm{N}\{0, \sigma^2 g(X'X)^{-1}\}$ *a priori*, or equivalently $\alpha \sim \mathrm{N}(0, \sigma^2 g D^{-2})$. This prior biases the direction of α along the axes of the principal-component coordinate system.

Ridge regression also has a well-known Bayesian interpretation as the posterior mean under the conjugate normal prior $\beta \sim \mathrm{N}(0, \sigma^2 \tau^2 I)$, where the global variance $\tau^2 = 1/\nu$. This prior is agnostic with respect to the orientation of the regression vector, depending only upon its Euclidean norm.

These procedures, along with PCR, are all special cases of a more general prior:

$$(\alpha \mid \sigma^2, \tau^2, \Lambda) \sim \mathrm{N}(0, \sigma^2 \tau^2 \Lambda) \, , \tag{7}$$

where τ^2 is a global variance component and $\Lambda = (\lambda_1^2, \ldots, \lambda_p^2)$ is a diagonal matrix of local variance components. The posterior distribution of α under this prior is conditionally normal, with mean

$$m_j = \kappa_j \hat{\alpha}_j = \left(\frac{\tau^2 \lambda_j^2 d_j^2}{1 + \tau^2 \lambda_j^2 d_j^2} \right) \hat{\alpha}_j \, ,$$

with the α_j's being mutually independent given τ^2, σ^2, and the data.

The classical g-prior therefore corresponds to $\tau^2 = g$ and $\lambda_j \equiv d_j^{-2}$. Ridge regression corresponds to $\lambda_j^2 = 1$. And PCR corresponds to

$$\lambda_j^2 = \begin{cases} \infty, & d_j^2 \geq d_K^2 \\ 0, & d_j^2 < d_K^2 \end{cases}$$

for the K-component solution.

Rather than estimating α under fixed choices of the local variances λ_j^2, the natural fully Bayesian approach is to use the shrinkage weights

$$\kappa_j^{FB} = \mathrm{E}_{(\lambda_j^2, \tau^2 \mid X, \mathbf{y})} \left(\frac{\tau^2 \lambda_j^2 d_j^2}{1 + \tau^2 \lambda_j^2 d_j^2} \right) \, , \tag{8}$$

where the expectation is over the posterior distribution of local and global variance components.

Different choices for the priors $p(\lambda_j^2)$ and $p(\tau^2)$ can center the Bayesian model at different classical regularization approaches, while still allowing the data to dictate otherwise. Choosing $p(\lambda_j^2)$ to concentrate near 1, for example, will center the model near the classical ridge solution. On the other hand, if $\lambda_j^2 \equiv d_j^{-2} \nu_j^2$, then choosing $p(\nu_j^2)$ to concentrate near 1 will center the model near the g-prior. Placing a further prior on τ^2 will replicate the mixtures of g-priors studied by Liang et al. (2008).

Mixing over a further prior $p(\Lambda)$, however, will lead to even more flexible mixtures of g-priors. In particular, the classical g-prior prefers coefficient vectors that line up with the principal components, and further mixing over local variance components helps to robustify the model against this assumption.

Even the PCR solution can be chosen as an approximate centering model by selecting a prior $p(\lambda_j^2)$ such that $p(\kappa_j)$ concentrates simultaneously near 0 and 1. For example, if $\tau^2 = 1$ and λ_j^2 follows an inverted-beta (or "beta-prime") distribution $\mathrm{IB}(1/2, 1/2)$, then κ_j will have a $\mathrm{Be}(1/2, 1/2)$ prior, whose density function is unbounded both at 0 and at 1 as required. Marginally this leads to a horseshoe prior for α_j (Carvalho et al., 2010).

Partial least squares, on the other hand, cannot be interpreted in this framework. To see this, observe that the shrinkage weights are identified with the prior variance components via $\kappa_j = \tau^2\lambda_j^2 d_j^2/(1+\tau^2\lambda_j^2 d_j^2)$. Under PLS, some of the shrinkage weights κ_{jK}^{PLS} may be larger than 1. Such weights cannot arise from a valid (non-negative) configuration of λ_j^2's and τ^2. Therefore, PLS cannot be the optimal solution under any prior expressible as a global-local scale mixture of normals.

4.3 When should the full Bayesian model work better? Some intuition and examples

Ridge regression, PCR, and PLS are all operationally similar. They bias the coefficient vector away from directions in which the predictors have low sampling variance—or equivalently, away from the "least important" principal components of X. This leads to a favorable bias-variance tradeoff in the performance of the resulting estimator. The g-prior and mixtures of g-priors, on the other hand, shrink along all eigen-directions equally, and usually not by very much.

Neither of these approaches need work well. When the underlying regression signal is "eigen-sparse"—that is, when only some of the linear combinations of β_i's given by W are meaningful for predicting $\mathbf{y}$—then one should shrink different components of $\hat{\alpha}$ by different amounts. This makes the g-prior inappropriate.

Yet as many previous authors have noted, there is no logical reason that $\mathbf{y}$ cannot be strongly associated with the low-variance principal components of X. Ridge regression and PCR will both do poorly in these situations: RR will necessarily shrink more along low-variance directions, while PCR must include all the higher-variance directions $(j < K)$ in order to include a lower-variance one (K).

The intuition behind the fully Bayes model of (7) is that the shrinkage weights κ_j should indeed be unequal, but that they can be learned from the data, and need not be monotonic in d_j^2. The fully Bayes shrinkage weights, moreover, will depend not merely on X. They will also depend nonlinearly upon $\mathbf{y}$, and upon each other through their mutual dependence upon τ^2.

Consider three illustrative examples. In all cases we have assumed that $\tau^2 \sim \mathrm{IB}(1/2, 1/2)$ and that $\lambda_j^2 \sim \mathrm{IB}(1/2, 1/2)$, thereby specifying a horseshoe prior for α.

First, we analyzed the data from Fearn (1983), consisting of 24 samples of ground wheat. The response variable is the protein concentration in the wheat, while the predictors (L1–L6) are measurements of the samples' reflection of NIR radiation (R), measured at six different wavelengths between 1680 and 2310 nanometers. The predictors are referred to as "log values", since they are measured on a $\log(1/R)$ scale. The goal is to find a linear combination of log values that predicts protein concentration. Both the response and the predictors were centered and rescaled to have variance 1.

The log values are highly multi-collinear, with the smallest pairwise correlation being 0.925. Despite the fact that ridge regression is intended for just these multi-collinear situations, here it performs quite poorly. As Fearn (1983) explains, this happens because the first principal component places nearly equal weight on all six log values (see Table

Table 3: The six principal component variances and loadings for the wheat protein-concentration data.

	PC1	PC2	PC3	PC4	PC5	PC6
L1	0.411	0.213	0.265	-0.353	0.422	0.642
L2	0.410	0.342	-0.446	-0.079	0.465	-0.542
L3	0.411	0.266	-0.367	-0.209	-0.743	0.173
L4	0.411	-0.028	0.731	-0.127	-0.221	-0.481
L5	0.396	-0.874	-0.242	-0.126	0.067	0.023
L6	0.411	0.05	0.05	0.891	0.013	0.182
Variance	5.868	0.101	0.019	0.012	< 0.001	< 0.001

3). The variation described by this component—essentially the sample average of the log values—is due mainly to differences in particle size. It carries little information about protein content, and yet is prefentially treated as the "most important" predictor by the ridge estimator. Contrasting log values are associated with "less important" principal components, and yet these contrasts—mostly the second, third, and fourth—are far more useful for predicting protein concentration. Notice that, by structural necessity, ridge regression shrinks these components more aggressively than the other methods do. It is also worth noting how much uncertainty there seems to be in the posterior distribution for the later shrinkage factors.

Second, we analyzed data on the softening temperature (y) of $n = 99$ ash samples originating from different biological sources. The predictor matrix comprises $p = 16$ observed mass concentrations for the ash samples' constituent molecules. The measurements are highly multi-collinear, with the eigenvalues of the correlation matrix for X spanning 10 orders of magnitude. The data are available in the R package `chemometrics`, and have been centered and scaled.

Finally, we analyzed synthetic data where X corresponds to a factor model. That is, each row x_i' satisfies

$$x_i = Bf_i + \xi_i \, ,$$

where the loadings matrix B is $p \times k$, $f_i \sim \mathrm{N}(0, I)$ is $k \times 1$, $\xi_i \sim \mathrm{N}(0, \psi I)$ is $p \times 1$, and $k < p$. The predictors that arise from this structure will exhibit multi-collinearity, and when ψ is small compared to the entries in B, this multi-collinearity will be very pronounced. In a factor model, moreover, it need not be the case that y will be associated most strongly with the high-variance principal components of X.

We generated data where $p = 20$, $n = 100$, $k = 5$, and $\psi = 0.1$, with all the entries of B set to 1. The resulting coefficient vector, least-squares estimate, and eigenvalues D are excerpted in Table 4. Principal component 12 is clearly the outlier: it is a strong predictor of y, and yet its corresponding variance is two orders of magnitude smaller than the largest variance.

Figure 1 compares the shrinkage structures of RR, PCR, PLS, and the Bayesian model for all three of these data sets. The components are ordered left to right along the x axis from highest variance (1) to lowest variance (p), while the shrinkage coefficients κ (Equation 6)

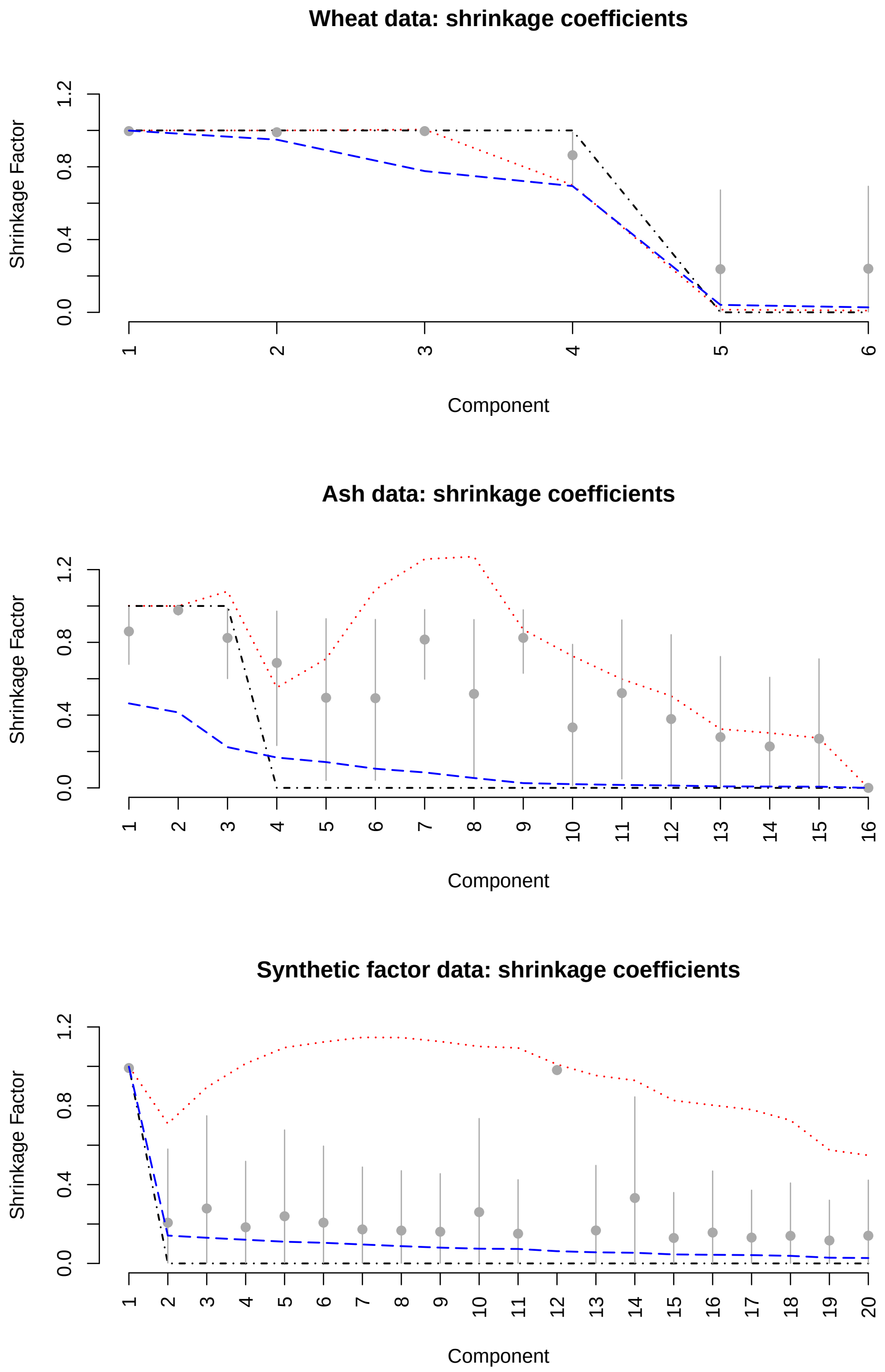

Figure 1: Comparison on three data sets in terms of how much the four methods shrink each principal component. Grey dots (grey lines): posterior means (75% credible intervals) under the fully Bayesian model. Blue dashes: ridge regression. Red dots: partial least squares. Black dots and dashes: principal-component regression.

Table 4: Subset of the true orthgonalized coefficient vector, least-squares estimate $\hat{\alpha}$, and eigenvalues for Example 3, where X is a five-factor model.

Comp.	α	$\hat{\alpha}$	D
1	-0.10	-0.11	91.83
2	-0.02	-0.50	1.41
⋮			
11	0.42	1.36	0.98
12	12.10	12.16	0.91
13	0.04	0.13	0.85
⋮			
19	0.39	-1.35	0.60
20	0.00	-1.87	0.58

are along the y axis. The tuning parameters for the non-Bayesian methods were chosen by cross-validation.

In all three cases, there appears to be a tendency for both PCR and ridge regression to over-shrink coefficients corresponding to low-variance eigen-directions. On the ash data set, components 7 and 9 seem to be important, while for the factor model, component 12 is known to be the most important. Yet all are shrunk nearly to zero by RR and PCR. For the sake of variance reduction, too much bias is introduced.

Partial least-squares, on the other hand, can identify important low-variance components. Yet it does so by including many other unimportant low-variance components. For the sake of bias reduction, too much variance is introduced.

The fully Bayesian model seems to blend the best of both these techniques. It can successfully pick out important coefficients corresponding to low-variance eigen-directions. Yet at the same time, it can squelch the other unimportant components. Intuitively, this combination should make for a favoriable bias–variance tradeoff in larger problems.

5 Regression when $p > n$

5.1 Generalization to large-p cases

Suppose now that the design matrix X is of rank $r < p$ and has singular-value decomposition $X = UDW'$ with $D = \text{diag}(d_1, \ldots, d_r)$, again ordered from largest (d_1) to smallest (d_r). The approach of the previous section works just as before, with no essential modification:

$$\begin{aligned}
(\hat{\alpha} \mid \alpha, \sigma^2) &\sim \text{N}(0, \sigma^2 D^{-2}) \\
(\alpha \mid \sigma^2, \tau^2, \Lambda) &\sim \text{N}(0, \sigma^2 \tau^2 \Lambda) \\
\lambda_j^2 &\sim p(\lambda_j^2) \\
(\sigma^2, \tau^2) &\sim p(\sigma^2, \tau^2),
\end{aligned}$$

where $\alpha = W'\beta$ and $\hat{\alpha}$ is the corresponding OLS estimate. Instead of a p-dimensional vector to estimate, we now have an r-dimensional one. Moreover, because we have orthogonalized the coefficients, the elements of α are conditionally independent in the posterior distribution, given τ^2 and σ^2. We are faced with a simple normal-means problem, with the only complication being that the singular values d_j enter the likelihood.

This approach is also related to the work of Maruyama and George (2010), who propose a modification of the standard g-prior (Zellner, 1986) for use in Bayesian variable selection when $p > n$. Suppose that

$$p(\beta) = \prod_{j=1}^{r} p_j(w_j'\beta \mid g, \sigma^2) \, .$$

Each $p_j(w_j'\beta \mid g, \sigma^2)$ is a normal density,

$$\mathrm{N}\left(w_j'\beta \mid 0, \frac{\sigma^2}{d_j^2} f_j(1+g) - \frac{\sigma^2}{d_j^2} \right) \, , \tag{9}$$

where w_j is the jth right-singular vector of X, and where $f_j > 1$ is necessary to ensure positive definiteness.

The seemingly strange form of (9) harks back to Strawderman (1971). Structurally, it essentially the same prior as above considered above, with a slight modification made for the sake of ensuring that the marginal distribution $p(\mathbf{y})$ is analytically convenient (see Section 4.7.10 of Berger, 1985). Maruyama and George recommend mixing over a prior for g while fixing $f_j = d_j^2/d_r^2$ in (9). This approximately corresponds to a similar fixed choice for the λ_j^2's in (7).

Under this prior, there exists closed-form expression for the Bayes factor between any two submodels of the full p-variable model. This allows one to perform full Bayesian model selection even when $p > n$.

Our proposal is an alternative generalization appropriate for pure shrinkage solutions, one that incorporates additional mixing over local variances λ_j^2. If we treat W as the canonical pseudo-inverse that maps back to the original coordinate system, then the implied prior for $\beta = W\alpha$ is a singular normal distribution:

$$(\beta \mid \Lambda, \tau^2, \sigma^2) \sim \mathrm{N}(0, \sigma^2\tau^2 W \Lambda W') \, .$$

To see the connection with the g-prior more explicitly, suppose that $\lambda_j^2 = d_j^{-2}$ and that $n > p$, such that X is of full column rank. It is easily verified that $WD^{-2}W' = (X'X)^{-1}$, leading to the original g-prior with $g \equiv \tau^2$. Other authors have considered the same generalization, but with simple conjugate priors for λ_j^2—for example, Clyde et al. (1996), Denison and George (2000), and West (2003). Our approach differs in our emphasis placed upon the choice of prior for λ_j^2, for which the developments earlier in the paper are clearly relevant.

Under this model, the (conditional) posterior mean estimator for α_j is, just as before, given by

$$\left(\frac{\tau^2 \lambda_j^2 d_j^2}{1 + \tau^2 \lambda_j^2 d_j^2} \right) \hat{\alpha}_j \, ,$$

a generalized Bayesian version of the classic ridge estimator.

5.2 Assessing out-of-sample predictive performance

In the following simulation studies, we investigate the performance of the Bayesian model proposed above. We use the horseshoe prior, whereby τ and each λ_j receive independent half-Cauchy priors. We now sketch a brief rationale for this choice. Intuitively, the vectors $\{w_j\}$ can be thought of as contrasts. A nice "default" Bayesian model would express the prior belief that certain contrasts of the β sequence will be strong predictors of $\mathbf{y}$, and that some will be weak predictors. The horseshoe prior does just this: it will shrink most α_j's very strongly, as the posterior mass for τ tends to concentrate near zero. Yet it will leave unshrunk those α_j's corresponding to contrasts that predict $\mathbf{y}$ well—even, it is to be hoped, those that correspond to a low-variance principal components—since the heavy tails of the half-Cauchy prior will allow certain λ_j's to be quite large.

As test cases, we used the following 7 data sets, all of which had more predictors than observations. Only 1 of the 7 data sets is simulated; the other 6 are from chemometrics or genomics. All are available upon request from the authors, and the 6 real data sets are available from the R packages `pls`, `chemometrics`, and `mixOmics`.

factor: the only simulated data set considered. Both X and y were generated jointly from a standard Bayesian factor model, with y loading most heavily on the lowest-variance factors.

nutrimouse: observations of 40 mice where hepatic fatty-acid concentrations are regressed upon the expression of 120 potentially relevant genes measured in liver cells.

cereal: chemometric observations of 15 cereal molecules where starch content is regressed upon NIR spectra at 145 different wavelengths.

yarn: samples of 28 polyethylene terephthalate (PET) yarns, where the density of the yarn sample is regressed upon measurements of NIR spectra at 268 wavelenths.

gasoline: octane numbers of 60 gasoline samples along with NIR spectra at 401 wavelengths.

multidrug: the X matrix comprises observations of the activity of 853 drugs on 60 different human cell lines, expressed as the concentration at which each drug leads to a 50% inhibition of growth for each cell line. The y variable is the measured expression of ABC3A (an ATP-binding cassette transporter) in each cell line.

liver: the X matrix contains the expression scores for 3116 genes in 64 rat subjects. The y variable is the cholesterol concetration in the liver.

We compare the Bayesian model to the three basic techniques (partial least squares, ridge regression, and principal-components regression), along with a new technique called sparse partial least squares (Chun and Keles, 2010) aimed at simultaneous dimension reduction and variable selection. This final method is implemented in the R package `spls`.

Table 5: Average out-of-sample predictive error (SSE) on 50 different train/test splits for 7 data sets where $p > n$. Bayes: the local-shrinkage model with horseshoe priors. PLS: partial least squares. PCR: principal-components regression. RR: ridge regression. SPLS: sparse partial least squares. The smallest entry in each row is in boldface.

			Average out-of-sample error				
Data set	n	p	Bayes	PLS	PCR	RR	SPLS
factor	50	100	**45.8**	66.9	69.2	358	97.6
nutrimouse	40	120	**394**	428	467	**394**	462
cereal	15	145	45.2	46.9	46.3	**42.2**	46.5
yarn	28	268	**2.63**	6.89	20.2	4.18	53.8
gasoline	60	401	0.82	0.87	0.93	**0.72**	1.04
multidrug	60	853	**139**	152	173	143	160
liver	64	3116	**1340**	1457	1475	1407	1470

To test these five methods, we split each of the seven data sets into training and test samples, with 75% of the observations used for training. We then fit each model using the training data, with tuning parameters for the non-Bayesian methods chosen by ten-fold cross validation on the training data alone. We then compared out-of-sample predictive performance on the holdout data, measured by sum of squared prediction errors (SSE). In each case the y variable was centered, and the X variables were centered and scaled.

All of our results in Table 5 represent the average SSE incurred over 50 different train/test splits. There are several interesting things to notice here. For one thing, the Bayes method seems to be the overall winner. It was the outright best on 4 data sets, tied for best on 1 data set, and second-best on the other two data sets. Surprisingly, the next-best method seems to be a venerable classic: ridge regression. The newest method, sparse partial least squares, was either worst or second-worst on all 7 data sets.

The two cases where the Bayesian method offered the biggest improvements—the factor data and the yarn data—are also instructive. In these cases, the y variable was most strongly associated with smaller-variance contrasts w_j, or in other words, those contrasts associated with smaller singular values d_j. Much as we saw in the previous section, classic methods like ridge regression and PCR perform poorly when this is the case, whereas the Bayesian model is quite robust.

In other cases (notably the cereal, gasoline, and nutrimouse data sets), the signal-to-noise ratio seems to be either so favorable, or so poor, that all the methods do almost equally well. This suggests that the extra variance induced by mixing over local λ_j^2's does not pose difficulty for the Bayesian model.

6 Final Remarks

The study of oracle properties provides a unifying framework in the classical literature for the study of regularized regression, but no such framework exists for Bayesians. In this paper, we have offered a few elements that might form the beginnings of such a framework. By identifying β (or α) with the increments of a discretely observed Lévy process, we have embedded the finite-dimensional problem in a suitable infinite-dimensional generalization. This provides a natural setting in which the dimension p grows without bound. In particular, Theorems 1 and 2 establish mappings among Lévy processes, penalty functions, priors, and scale mixtures of normals. This offers a convenient way of using heavy-tailed, infinitely divisible probability distributions, giving Bayesian statisticians a much larger toolbox for building shrinkage models like the kind explored in Section 5.

References

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.

O. E. Barndorff-Nielsen and N. Shephard. Normal modified stable processes. Technical Report 2001-W6, Nuffield College, University of Oxford, 2001.

J. O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, 2nd edition, 1985.

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.

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*, 97(2):465–80, 2010.

H. Chun and S. Keles. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society, Series B (Methodology)*, 72:3–25, 2010.

M. Clyde, H. Desimone, and G. Parmigiani. Prediction via orthogonalized model mixing. *Journal of the American Statistical Association*, 91(435):1197–208, September 1996.

D. Denison and E. George. Bayesian prediction using adaptive ridge estimators. Technical report, Imperial College, London, 2000.

T. Fearn. A misuse of ridge regression in the calibration of a near infrared reflectance instrument. *Journal of the Royal Statistical Society, Series C*, 32(1):73–9, 1983.

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

R. A. Fisher. The mathematical distributions used in the common tests of significance. *Econometrica*, 3(4):353–65, 1935.

I. Frank and J. H. Friedman. A statistical view of some chemometrics regression tools (with discussion). *Technometrics*, 35(2):109–135, 1993.

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.

B. Grigelionis. Processes of Meixner type. *Lithuanian Mathematical Journal*, 39(1):33–41, 1999.

B. Grigelionis. Generalized z-distributions and related stochastic processes. *Lithuanian Mathematical Journal*, 41(3):239–51, 2001.

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

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

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.

D. Madan and M. Yor. CGMY and Meixner subordinators are absolutely continuous with respect to one sided stable subordinators. Technical Report arXiv:math/0601173, ArXiv Mathematics e-prints, 2006.

Y. Maruyama and E. I. George. $g$bf: A fully Bayes factor with a generalized g-prior. Technical report, University of Tokyo, arXiv:0801.4410v2, 2010.

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

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. Shrink globally, act locally: sparse Bayesian regularization and prediction. In *Proceedings of the 9th Valencia World Meeting on Bayesian Statistics*, volume to appear. Oxford Univeristy Press, 2010.

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

M. West. Bayesian factor regression models in the "large p, small n" paradigm. In J. M. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, and M. West, editors, *Bayesian Statistics 7*, pages 723–732. Oxford University Press, 2003.

R. L. Wolpert and M. Taqqu. Fractional Ornstein-Uhlenbeck Lévy processes and the Telecom process: Upstairs and downstairs. *Signal Processing*, 85(8):1523–1545, Aug. 2005.

R. L. Wolpert, M. A. Clyde, and C. Tu. Stochastic expansions using continuous dictionaries: Lévy Adaptive Regression Kernels. Technical Report 2006-08, Duke University Department of Statistical Science, 2010.

A. Zellner. On assessing prior distributions and Bayesian regression analysis with *g*-prior distributions. In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, pages 233–243. Elsevier, 1986.

A Proofs of main results

Proof of Theorem 1

To identify the Laplace exponent of the subordinator as a separable, globally linear penalty function, we simply evaluate the logarithm of the moment-generating function of T_s at $t = \beta_i^2$ and time $s = \nu$. To see this, first note that, since T_s is a subordinator, it is completely characterized by the Lévy representation of its moment-generating function,

$$M_T(t) = \mathrm{E}\{\exp(-tT_s)\} = \exp\{-s\psi(t)\}\,, \tag{10}$$

where

$$\psi(t) = \int_0^\infty \{1 - \exp(tx)\}\; \mu(\mathrm{d}x)\,.$$

Now let $\mathbf{T} = \big(T_s^{(1)}, \ldots, T_s^{(p)}\big)$ be a random variable whose components $T^{(i)}$ are independent, identically distributed subordinators, each satisfying the conditions of the theorem and each observed at time $s = \nu$. Then

$$\begin{aligned} M_{\mathbf{T_s}}(\mathbf{s}) &= \mathrm{E}\{\exp(-\mathbf{t}'\mathbf{T_s})\} \\ &= \mathrm{E}\Big[\prod_{i=1}^{p} \exp\big\{-t_i T_s^{(i)}\big\}\Big] \\ &= \prod_{i=1}^{p} \exp\{-\nu\psi(t_i)\} \\ &= \exp\Big(-\nu \sum_{i=1}^{p} t_i\Big) \end{aligned}$$

Evaluating the logarithm of this moment-generating function at $\mathbf{T} = (\beta_1^2, \ldots, \beta_p^2)$ is thus equivalent to evaluating the log m.g.f. of the one-dimensional subordinator T_s at $t = \sum_{i=1}^{p} \beta_i^2$. This gives

$$\log \mathrm{E}\Big[\exp\big\{-T_\nu\Big(\sum_{i=1}^{p} \beta_i^2\Big)\Big] = -\nu \sum_{i=1}^{p} \psi(\beta_i^2)\,.$$

Next, we recognize the normal scale-mixture representation by writing the expectation in (10), evaluated at $t = \beta_i^2$, as

$$\begin{aligned} \mathrm{E}\{\exp(-tT_s\} &= \int_0^\infty \exp\{-\beta_i^2 T_s\}\, g(T_s)\mathrm{d}T_s \\ &= \int_0^\infty \sqrt{T_s}\exp\big\{-\beta_i^2 T_s\big\}\,\big\{T_s^{-1/2} g(T_s)\big\}\mathrm{d}T_s\,, \end{aligned}$$

By the integrability condition, $T_s^{-1/2} g(T_s)$ is proportional to some prior density, and thus the above expression is clearly proportional to a normal scale mixture.

Proof of Theorem 2

Since T_s is a subordinator, its moment-generating function is

$$M_s(t) = \mathrm{E}\{\exp(-tT_s)\} = \exp\{-s\chi(t)\}\,,$$

with $\chi(t)$ given above. To compute (5), simply evaluate this moment-generating function for T_1 at $t = \sum_{i=1}^{p} \psi(\beta_i^2)$.

B Derivation of self-similar representation for the horseshoe prior

Suppose $\lambda_i^2 \sim \mathrm{IB}(a,b)$. Then

$$\lambda_i^2 \overset{D}{=} \frac{\kappa_i}{1-\kappa_i}\,,$$

where $\kappa_i \sim \mathrm{Be}(a,b)$. Following Fisher (1935), if $z_i = \log\{\kappa_i/(1-\kappa_i)\}$, then

$$p(z_i) = \frac{1}{\beta(a,b)} \frac{(e^{z_i})^a}{(1+e^{z_i})^{a+b}} \,,$$

where $\beta(a,b)$ is the Beta function. More generally we may assume that $z_i \sim \mathrm{Z}(a,b,\mu,\sigma)$, a z-distribution with density

$$p(z_i) = \frac{2\pi}{\sigma\beta(a,b)} \frac{[\exp\{2p(z_i-\mu)/\sigma\}]^a}{[1+\exp\{2p(z_i-\mu)/\sigma\}]^{a+b}}$$

and characteristic function

$$\phi(t) = \frac{\beta\left(a+\frac{i\sigma t}{2\pi}, b-\frac{i\sigma t}{2\pi}\right)}{\beta(a,b)} \exp(i\mu t)$$

for $a > 0$, $b > 0$, $\sigma > 0$, $\mu \in \mathbb{R}$.

The z distribution can then be recognized as the special case of Grigelionis's class of generalized-z (GZ) distributions, which have characteristic function

$$\phi(t) = \left\{ \frac{\beta\left(a+\frac{i\sigma t}{2\pi}, b-\frac{i\sigma t}{2\pi}\right)}{\beta(a,b)} \right\}^{2\delta} \exp(i\mu t)$$

for $\delta > 0$ (Grigelionis, 2001). This distribution has parameters (a,b,μ,σ,δ) and can also be characterized by its Lévy triple $\{A, 0, \mu(x)dx\}$, where

$$A = \frac{\sigma\delta}{\pi} \int_0^{2\pi/\sigma} \frac{e^{-bx}-e^{-ax}}{1-e^{-x}} \mathrm{d}x + \mu \,, \tag{11}$$

and

$$\mu(x) = \begin{cases} \frac{2\delta \exp\left\{\frac{2\pi b x}{\sigma}\right\}}{x\left\{1-\exp\left(\frac{2\pi x}{\sigma}\right)\right\}} \,, & \text{if } x > 0 \\ \\ \frac{2\delta \exp\left\{\frac{2\pi a x}{\sigma}\right\}}{|x|\left\{1-\exp\left(\frac{2\pi x}{\sigma}\right)\right\}} \,, & \text{if } x < 0 \,. \end{cases}$$

The characteristic function of a generalized-z distribution makes its self-similarity plain: if $z_i \overset{iid}{\sim} \mathrm{GZ}(a,b,\mu/p,\sigma,1/2p)$, then

$$\sum_{i=1}^{p} z_i \overset{D}{=} z \,,$$

where $z \sim \mathrm{Z}(a,b,\mu,\sigma)$. We thus have a self-similar representation, on the log-variance scale, of the normal/inverted-beta class.

This result is of limited use except in special cases where the density of the generalized-z increments is known, which will not hold in general. Luckily the horseshoe prior, where $a = b = 1/2$, corresponds to just such a special case—as do all symmetric cases where $\kappa \sim \mathrm{Be}(a, 1-a)$ and $\lambda_i^2 = \kappa/(1-\kappa)$.

To see this, let $z \sim \mathrm{Z}(a, 1-a, \mu, \sigma)$ for $a \in (0,1)$. Then

$$\phi(t) = \frac{\beta\left(a + \frac{i\sigma t}{2\pi}, 1 - a - \frac{i\sigma t}{2\pi}\right)}{\beta(a, 1-a)} \exp(i\mu t)\,.$$

After some standard manipulations, this reduces to

$$\phi(t) = \frac{\cos(c/2)}{\cosh\left(\frac{\sigma t - ic}{2}\right)} \exp(i\mu t)\,,$$

where $c = \pi(2a-1)$. This is recognizable as the characteristic function of a Meixner process, $z \sim \mathrm{Meix}(\sigma, c, 1/2, \mu)$ (Grigelionis, 1999). The density and Lévy measure of a Meixner random variable are

$$p(z) = \frac{2\cos(c/2)}{\sigma\pi} \exp\left\{\frac{c(z-\mu}{\sigma}\right\} \left|\Gamma\left(\frac{1}{2} + \frac{i(z-\mu)}{\sigma}\right)\right|^2 \tag{12}$$

$$\mu(dx) = \frac{\exp(cx/\sigma)}{2x\sinh(\pi x/\sigma)} \mathrm{d}x\,. \tag{13}$$

For the horseshoe prior, $a = 1-a$ and therefore $c = 0$.

A Meixner process is self-similar: if $z_i \sim \mathrm{Meix}\{a, c, 1/(2p), \mu/p\}$, then

$$\sum_{i=1}^{p} z_i \overset{D}{=} z \sim \mathrm{Meix}(a, c, 1/2, \mu)\,.$$

When $a = 1$ and $\mu = 0$, then the random variable $T \overset{D}{=} e^z$ will have an $\mathrm{IB}(a, 1-a)$ distribution, as required. Therefore, the most intuitive way of passing to a limit under the horseshoe prior is to continue dividing the random variable T, on the log variance scale, into arbitrarily many self-similar increments.

Interestingly, both the z-distribution and the Meixner can themselves be represented as mixtures of normals. The mixing distribution for the z is an infinite convolution of exponentials, a potentially interesting generalization of the lasso model (Barndorff-Nielsen et al., 1982). For the mixing distribution of the Meixner, see Madan and Yor (2006).