The horseshoe estimator for sparse signals
Author(s): Carvalho, Polson and Scott, Biometrika, Vol. 97, No. 2 (JUNE 2010), pp. 465-480

1. Introduction

1·1. *The proposed estimator*

Suppose we observe a p-dimensional vector $y \mid \theta \sim N(\theta, \sigma^2 I)$. If θ is believed to be sparse, we propose using the following model for estimation and prediction:

$$\theta_i \mid \lambda_i \sim N(0, \lambda_i^2), \quad \lambda_i \mid \tau \sim \mathrm{C}^+(0, \tau), \quad \tau \mid \sigma \sim \mathrm{C}^+(0, \sigma),$$

where $\mathrm{C}^+(0, a)$ is a standard half-Cauchy distribution on the positive reals with scale parameter a. Crucially, each θ_i is mixed over its own λ_i, and each λ_i has a half-Cauchy prior with common scale τ. Additionally, we assume Jeffreys' prior for the variance, $p(\sigma^2) \propto 1/\sigma^2$. The prior for τ also follows the treatment of Jeffreys, in that it is scaled by σ, the standard deviation of the error model (Jeffreys, 1961, Ch. 5).

We estimate θ using the posterior mean under this model, which we call the horseshoe prior. This name arises from the observation that, for fixed values $\sigma^2 = \tau^2 = 1$,

$$E(\theta_i \mid y) = \int_0^1 (1 - \kappa_i) y_i \, p(\kappa_i \mid y) \, \mathrm{d}\kappa_i = \{1 - E(\kappa_i \mid y)\} y_i,$$

where $\kappa_i = 1/(1 + \lambda_i^2)$, and where $E(\kappa_i \mid y)$ can be interpreted as the amount of shrinkage towards zero, a posteriori. The half-Cauchy prior on λ_i implies a horseshoe-shaped $\mathrm{Be}(1/2, 1/2)$ prior for the shrinkage coefficient κ_i. The left side of the horseshoe, $\kappa_i \approx 0$, yields virtually no shrinkage, and describes signals. The right side of the horseshoe, $\kappa_i \approx 1$, yields near-total shrinkage and describes noise.

Table 1. *Priors for λ_i and κ_i associated with some common local shrinkage rules. For the normal-exponential-gamma prior, it is assumed that $d = 1$. Densities are given up to constants.*

Prior for θ_i	Density for λ_i	Density for κ_i
Double-exponential	$\lambda_i \exp(-\lambda_i^2/2)$	$\kappa_i^{-2} \exp\{-1/(2\kappa_i)\}$
Cauchy	$\lambda_i^{-2} \exp\{1/(2\lambda_i^2)\}$	$\kappa_i^{-1/2}(1-\kappa_i)^{-3/2} \exp[-\kappa_i/\{2/(1-\kappa_i)\}]$
Strawderman–Berger	$\lambda_i\,(1+\lambda_i^2)^{-3/2}$	$\kappa_i^{-1/2}$
Normal-exponential-gamma	$\lambda_i\,(1+\lambda_i^2)^{-(c+1)}$	κ_i^{c-1}
Normal-Jeffreys	λ_i^{-1}	$\kappa_i^{-1}(1-\kappa_i)^{-1}$
Horseshoe	$(1+\lambda_i^2)^{-1}$	$\kappa_i^{-1/2}(1-\kappa_i)^{-1/2}$

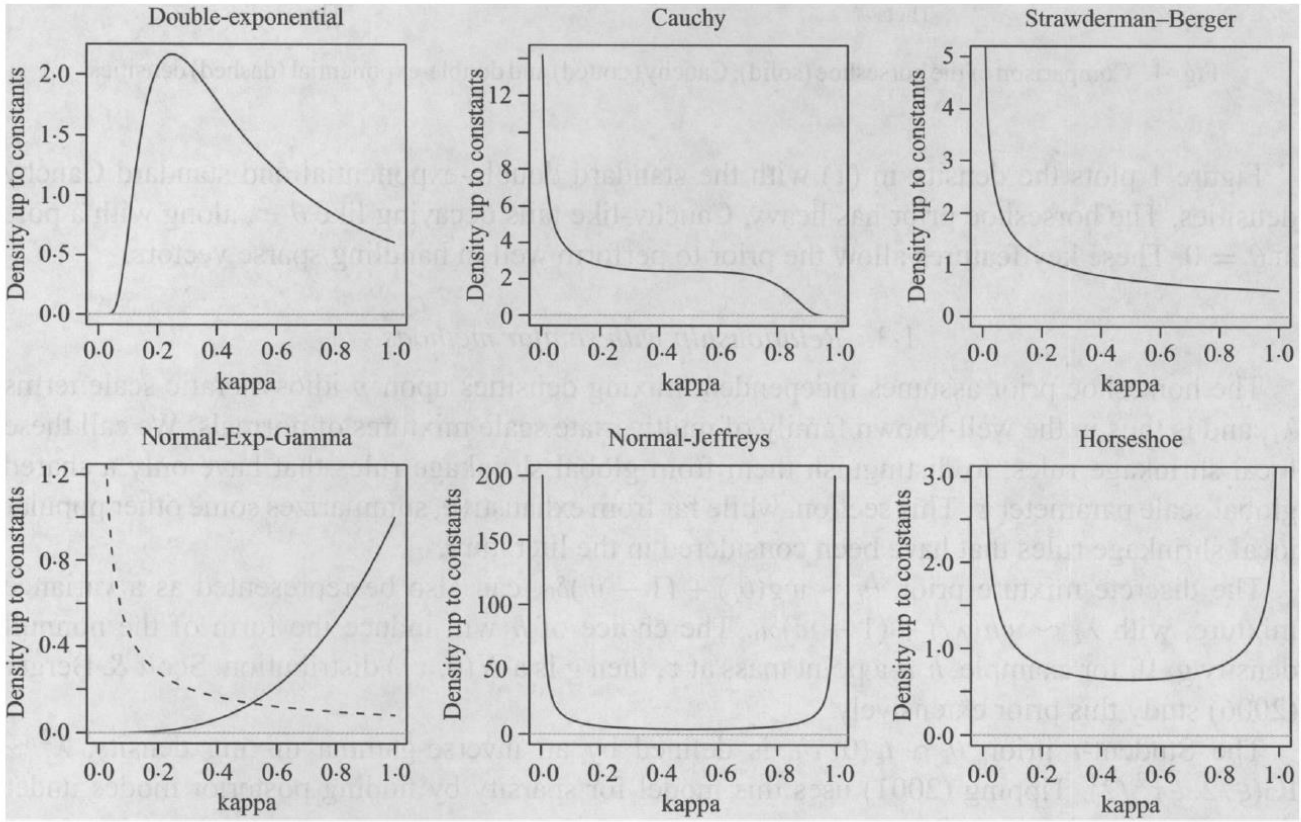

Fig. 2. The implied densities $p(\kappa_i)$ up to proportionality for six priors: the double exponential, Cauchy, Strawderman–Berger, normal-exponential-gamma, normal-Jeffreys and horseshoe. In the bottom-left panel for the normal-exponential prior, the solid line is for $c = 4$ and $d = 1$, while the dashed line is for $c = 1/4$ and $d = 1$.

Vanguard mutual funds

Carvalho & Scott (2009) Objective Bayesian model selection in Gaussian graphical models. Biometrika 96, 497-512.

n = 86 weekly returns for p = 59 funds

Out-of-sample predictive performance against four different approaches for estimating the covariance matrix:

- **MLE**
- **Lasso AND/OR** (Meinshausen & Buhlmann, 2006, High dimensional graphs and variable selection with the lasso. AOS, 34, 1436-62)
- **BMA**

Table 4. *Covariance-estimation example. The table entries are risk ratios versus Bayesian model averaging in the out-of-sample prediction exercise*

	MLE	Lasso AND	Lasso OR	Horseshoe	BMA
Risk ratio (SE)	10·63	1·25	2·12	1·07	1·00
Risk ratio (AE)	3·51	1·22	1·47	1·04	1·00

SE, squared-error loss; AE, absolute-error loss; MLE, maximum likelihood estimator; BMA, Bayesian model averaging.

Bayesian Analysis (2010) **5**, Number 1, pp. 171–188

Inference with normal-gamma prior distributions in regression problems

Jim. E. Griffin* and Philip. J. Brown†

Abstract. This paper considers the effects of placing an absolutely continuous prior distribution on the regression coefficients of a linear model. We show that the posterior expectation is a matrix-shrunken version of the least squares estimate where the shrinkage matrix depends on the derivatives of the prior predictive density of the least squares estimate. The special case of the normal-gamma prior, which generalizes the Bayesian Lasso (Park and Casella 2008), is studied in depth. We discuss the prior interpretation and the posterior effects of hyperparameter choice and suggest a data-dependent default prior. Simulations and a chemometric example are used to compare the performance of the normal-gamma and the Bayesian Lasso in terms of out-of-sample predictive performance.

Keywords: Multiple regression, $p > n$, Normal-Gamma prior, "Spike-and-slab" prior, Bayesian Lasso, Posterior moments, Shrinkage, Scale mixture of normals, Markov chain Monte Carlo

The standard multiple linear regression model assumes that a vector of responses $y = (y_1, y_2, \ldots, y_n)$ can be represented as

$$y = \alpha \mathbf{1} + \mathbf{X}\beta + \epsilon \tag{1}$$

where $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$ are independent, $p(\epsilon_i) = \mathrm{N}(\epsilon_i|0, \sigma^2)$ and X is an $n \times p$ matrix of explanatory variables. Here, $\mathrm{N}(x|\mu, \sigma^2)$ denotes the density of a normal distribution with mean μ and variance σ^2. The scalar α is the intercept and $\mathbf{1}$ a $n \times 1$ unit vector. This paper is concerned with the Bayesian analysis of this model and, in particular, the choice of the prior distribution of the $(p \times 1)$-dimensional vector of regression coefficients β. A zero mean normal prior leads to the ridge estimator as posterior mean. This estimator performs poorly if there are large differences in the size of regression coefficients. Alternatively, we could perform variable selection and assume that only a subset of the variables have non-zero regression coefficients which mitigate the problems associated with the normal prior. The standard approach is the "spike-and-slab" prior (Mitchell and Beauchamp 1988). An indicator variable z_i is introduced to identify whether the i-th variable is included in the model ($z_i = 1$) or excluded ($z_i = 0$). The prior for β_i can be written as

$$\pi(\beta_i) = z_i \mathrm{N}(\beta_i|0, \sigma_\beta^2) + (1 - z_i)\delta_{\beta_i = 0}\,, \qquad p(z_i = 1) = w,$$

3 The normal-gamma prior

A wide and natural class of prior densities for regression coefficients is the scale mixtures of normals (SMN) (see *e.g.* West (1987)), which we write as

$$\pi(\beta_i) = \int \mathrm{N}(\beta_i|0, \Psi_i) dG(\Psi_i)$$

where G is a mixing distribution. The prior can be expressed in a hierarchical form as

$$\beta_i|\Psi_i \sim \mathrm{N}(0, \Psi_i), \qquad \Psi_i \sim G. \tag{2}$$

This hierarchical form for the model shows that the i-th regression coefficient has a normal prior distribution conditional on an idiosyncratic variance (or scale), Ψ_i. This allows for larger differences in the sizes of the regression coefficients than would be possible under a normal prior. The marginal prior distribution for $\hat{\beta}_i$ has heavier than normal tails (apart from the degenerate case where G places all its mass at a single point). The "spike-and-slab" prior can be represented in this way by choosing

$$G(\Psi_i) = z_i \delta_{\Psi_i = \sigma^2_\beta} + (1 - z_i)\delta_{\Psi_i = 0}.$$

The double exponential prior of the Bayesian Lasso arises if G is an exponential distribution.

An interesting choice of absolutely continuous prior is the normal-gamma distribution, which includes the double exponential prior as a special case. Let $\mathrm{Ga}(x|c,d)$ represent the density of a gamma distribution with shape c and rate d so that

$$\mathrm{Ga}(x|c,d) = \frac{d^c}{\Gamma(c)} x^{c-1} \exp\{-dx\}.$$

We refer to the distribution as $\mathrm{Ga}(c,d)$. The normal-gamma distribution arises by assuming that the mixing distribution in a SMN has the density $g(x) = \mathrm{Ga}(x|\lambda, 1/(2\gamma^2))$. The density function is expressible as

$$\pi(\beta_i) = \frac{1}{\sqrt{\pi} 2^{\lambda-1/2} \gamma^{\lambda+1/2} \Gamma(\lambda)} |\beta_i|^{\lambda-1/2} K_{\lambda-1/2}(|\beta_i|/\gamma), \tag{3}$$

where K is the modified Bessel function of the third kind. The variance of β_i is $v_\beta = 2\lambda\gamma^2$ and the excess kurtosis is $\frac{3}{\lambda}$. The gamma distribution can represent a wide-range of shapes. As the shape parameter λ decreases these include distributions that place a lot of mass close to zero but at the same time have heavy tails. Figure 1 shows the

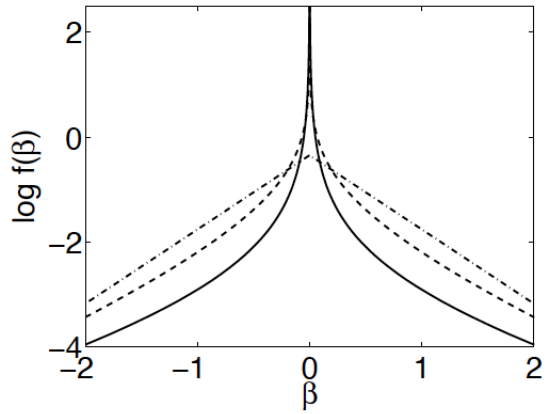

Figure 1: The log density of the normal-gamma prior with a variance of 2 and different values of λ. $\lambda = 0.1$ (solid line), $\lambda = 0.333$ (dot-dashed line) and $\lambda = 1$ (dashed line).

5.3 Example: NIR spectroscopy data

The data consists of 215 near-infrared absorbance spectra of meat samples, recorded on a Tecator Infratec Food Analyzer (represented as a 100-channel absorbance spectrum in the wavelength range 850-1050nm) and the composition of each sample in terms of water, fat and protein content. We consider predicting fat content on the basis of its infrared spectrum using the 100 channels. The data is split in a training/monitoring/testing set

of 129/43/43 samples. The data, originally used by Borggaard and Thodberg (1992), is available at http://lib.stat.cmu.edu/datasets/tecator. More recently it was analysed in Eilers et al. (2009). We used the training and monitoring data comprising $n = 172 = 129 + 43$ samples (*all data*) as our main data set and also took a random subset of 60 of the training data to create a p larger than n data set (*small*). The RMSEs for

	All data	Small
normal-gamma	1.94	2.59
Lasso	3.54	3.09

Table 3: RMSEs for fat prediction

prediction of the 43-sample test set, using the normal-gamma and Lasso priors, are given in Table 3. The difference in predictive performance is not surprising when one looks at the posterior median of λ in the normal-gamma prior which is 0.020 with a 95% credibility interval of $(0.016, 0.026)$ on the full data.

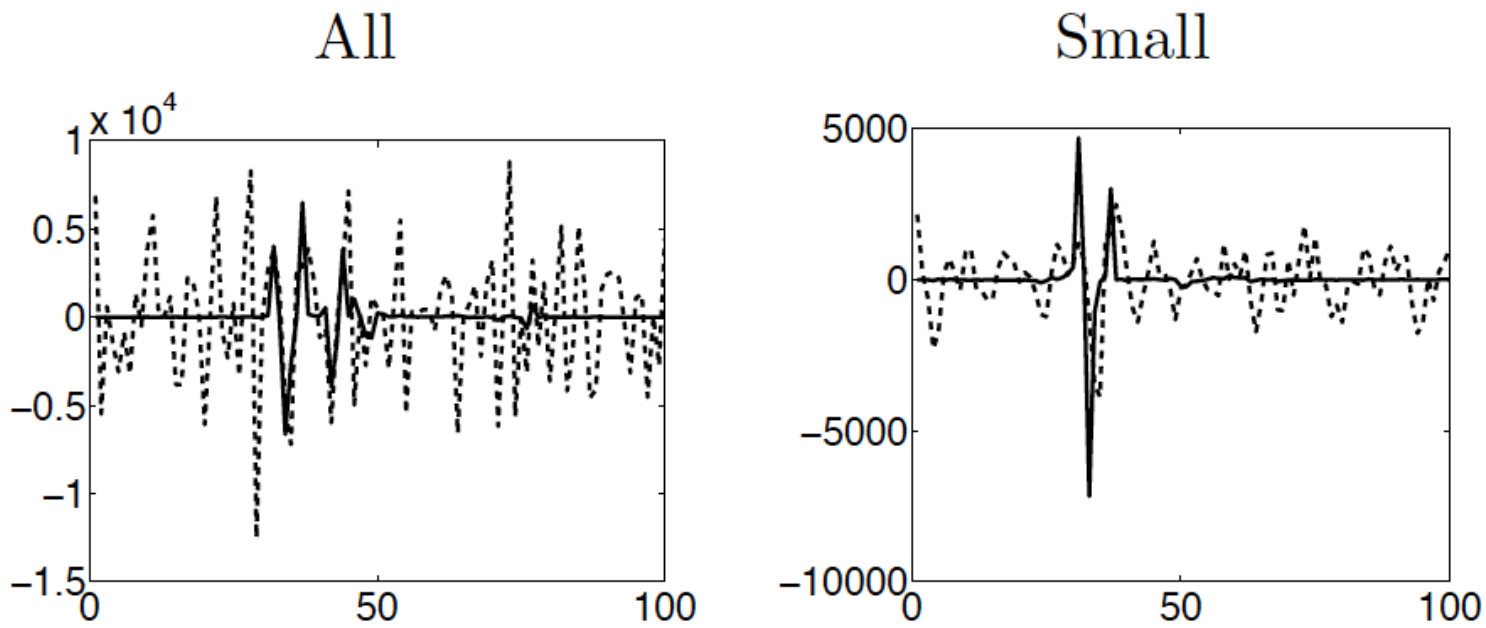

Figure 7: The posterior means of the β for the normal-gamma (solid) and Lasso (dashed) for two datasets.