Generalized double Pareto shrinkage

A. Armagan, D. Dunson and J. Lee
Submitted to Statistica Sinica

Presented by Esther Salazar
Duke University

March 16, 2012
They propose a generalized double Pareto prior for Bayesian shrinkage estimation in linear models.

Prior obtained via a scale mixture of Laplace or normal distributions.

Some characteristics:
- spike at zero like Laplace density
- Student-$t$ tail behavior
- straightforward computation via Gibbs sampling
Shrinkage estimation and variable selection

- Frequentist framework: LASSO, LARS (Tibshirani, 1996; Efron et al., 2004; Fan and Li, 2001; Zou and Li, 2008; ...)

- Bayesian approaches: Tipping (2001); Figueiredo (2003). For example, considering Student-\(t\) priors for basis coefficients via a scale mixture of normals.

Problems:

- If the scale parameter and d.f. of the Student-\(t\) go to zero we have a Normal-Jeffreys’ prior \(\theta \sim N(0, \tau), \pi(\tau) \propto 1/\tau\)

- Normal-Jeffreys’ prior leads to an improper posterior

- Bayesian LASSO procedure inherit the problem of over-shrinking large coefficients due to the relatively light tails of the Laplace prior

- Recently proposed priors considered those issues, with density near zero, heavy tails, without the improperity posterior problem. For instance, the horseshoe prior (Carvalho et al. 2009)
Generalized double Pareto (GDP) prior

The prior is given by

\[
f(\theta|\xi, \alpha) = \frac{1}{2\xi} \left(1 + \frac{|\theta|}{\alpha \xi}\right)^{-(1+\alpha)}
\]

where \(\xi > 0\) is a scale parameter and \(\alpha > 0\) is a shape parameter.

Also, \(E(\theta) = 0\) for \(\alpha > 1\) and \(V(\theta) = 2\xi^2 \alpha^2 (\alpha - 1)^{-1} (\alpha - 2)^{-1}\) for \(\alpha > 2\)

The dispersion is controlled by \(\xi\) and \(\alpha\) with \(\alpha\) controlling the tail heaviness and \(\alpha = 1\) corresponding to Cauchy-like tails and no finite moments

Notation:

\[\theta \sim GDP(\xi, \alpha)\]
Figure 1: (a) Probability density functions for standard double Pareto (solid line), standard Cauchy (dashed line) and Laplace (dot-dash line) ($\lambda = 1$) distributions. (b) Probability density functions for the generalized double Pareto with $(\xi, \alpha)$ values of (1, 1) (solid line), (0.5, 1) (dashed line), (1, 3) (long-dashed line), and (3, 1) (dot-dash line).
The GDP prior can be represented as a scale mixture of normal distributions leading to computational simplifications

**Proposition 1**

Let \( \theta \sim \mathcal{N}(0, \tau) \), \( \tau \sim \text{Exp}(\lambda^2/2) \) and \( \lambda \sim \text{Ga}(\alpha, \eta) \), where \( \alpha > 0 \) and \( \eta > 0 \). The resulting marginal density is \( \theta \sim GDP(\epsilon = \eta/\alpha, \alpha) \)

Bridge between Laplace and Normal-Jeffreys’ priors

**Proposition 2**

Given the above representation, \( \theta \sim GDP(\epsilon = \eta/\alpha, \alpha) \) implies

1. \( f(\theta) \propto 1/|\theta| \) for \( \alpha = 0 \) and \( \eta = 0 \)
2. \( f(\theta|\lambda') = (\lambda'/2)\exp(-\lambda'|\theta|) \) for \( \alpha \to \infty \), \( \alpha/\eta = \lambda' \) and \( 0 < \lambda' < \infty \)
Default specification: $\alpha = \eta = 1$ (Cauchy-like tails behavior with desirable Bayesian robustness properties)

**Analysis of the prior shrinkage factor $\kappa_i$:**
Suppose we observe a $p$-dimensional vector $y|\theta \sim N(\theta, \sigma^2 I)$, $\theta$ sparse

$$
\theta_i|\tau_i \sim N(0, \tau_i), \quad \tau_i|\lambda \sim f_1(\lambda), \quad \lambda|\alpha, \eta \sim f_2(\alpha, \eta),
$$

then

$$
\mathbb{E}(\theta_i|y) = \int_0^1 (1 - \kappa_i)y_i \pi(\kappa_i|y) d\kappa_i = \{1 - \mathbb{E}(\kappa_i|y)\}y_i
$$

where $\kappa_i = 1/(1 + \tau_i) \in (0, 1)$ and $\mathbb{E}(\kappa_i|y)$, can be interpreted as the amount of shrinkage towards zero, a posteriori. As $\kappa_i \to 0$ the prior does not impose any shrinkage, $\kappa_i \to 1$ it has a strong pull towards zero (Carvalho et al., 2010)

**A prior for $\kappa_i$, $\pi(\kappa_i)$ has a horseshoe shaped**
For the standard double Pareto with $\alpha = 1$ and $\eta = 1$, $\pi(\kappa)$ is given by

$$
\pi(\kappa) = \frac{1}{2(1 - \kappa)^2} \left[ \sqrt{\pi} \exp \left\{ \frac{\kappa}{2(1 - \kappa)} \right\} \text{Erfc} \left\{ \sqrt{\frac{\kappa}{2(1 - \kappa)}} \right\} - 1 \right],
$$

For general $\alpha > 0$ and $\eta > 0$ values, $\pi(\kappa)$ is given by

$$
\pi(\kappa|\alpha, \eta) = \frac{2^{\alpha/2-1} \eta^{\alpha} \kappa^{(\alpha-1)/2} (1 - \kappa)^{-(\alpha+3)/2}}{\Gamma(\alpha)} \left\{ \left( \frac{1}{\kappa} - 1 \right)^{1/2} \Gamma \left( \frac{\alpha}{2} + 1 \right) {}_1F_1 \left( \frac{\alpha}{2} + 1, \frac{1}{2}, \frac{\eta^2 \kappa}{2(1 - \kappa)} \right) 
- \sqrt{2} \eta \Gamma \left( \frac{\alpha + 3}{2} \right) {}_1F_1 \left( \frac{\alpha + 3}{2}, \frac{3}{2}, \frac{\eta^2 \kappa}{2(1 - \kappa)} \right) \right\}.
$$

For $\alpha = 0$ and $\eta = 0$, $\pi(\kappa|\alpha, \eta)$ takes a horseshoe shape.
Figure 2: Prior density of $\kappa$ implied by the standard double Pareto prior (solid line), Strawderman–Berger prior (dashed line), horseshoe prior (dot-dash line) and standard Cauchy prior (dotted line).

Figure 3: Prior density of $\kappa$ (a) when $\alpha = 1$ and $\eta = 0.5$ (dashed), $\eta = 1$ (solid), $\eta = 2$ (dot-dash) (b) when $\eta = 1$ and $\alpha = 1$ (solid), $\alpha = 2$ (dashed), $\alpha = 3$ (dot-dash).
Bayesian inference in linear models

Consider the linear regression model, \( y = X\beta + \epsilon \), where \( y \) is an \( n \)-dimensional vector of responses, \( X \) is the \( n \times p \) design matrix and \( \epsilon \sim N(0, \sigma^2I_n) \). Letting \( \beta_j|\sigma \sim GDP(\xi = \sigma\eta/\alpha, \alpha) \) independently for \( j = 1, \ldots, p \),

\[
\pi(\beta|\sigma) = \prod_{j=1}^{p} \frac{1}{2\sigma\eta/\alpha} \left( 1 + \frac{|\beta_j|}{\alpha\sigma\eta/\alpha} \right)^{-(\alpha+1)}.
\] (4)

From Proposition 1 this prior is equivalent to \( \beta_j|\sigma \sim N(0, \sigma^2\tau_j) \), with \( \tau_j \sim \text{Exp}(\lambda_j^2/2) \) and \( \lambda_j \sim \text{Ga}(\alpha, \eta) \). We place the Jeffreys’ prior on the error variance, \( \pi(\sigma) \propto 1/\sigma \).

Gibbs sampler:

Using the scale mixture of normals representation, we obtain a simple data augmentation Gibbs sampler having the following conditional posteriors: 

\((\beta|\sigma^2, T, y) \sim N\{(X'X+T^{-1})^{-1}X'y, \sigma^2(X'X+T^{-1})^{-1}\}, \)

\((\sigma^2|\beta, T, y) \sim IG\{(n+p)/2, (y-X\beta)'(y-X\beta)/2 + \beta'T^{-1}\beta/2\}, (\lambda_j|\beta_j, \sigma^2) \sim \text{Ga}(\alpha+1, |\beta_j|/\sigma+\eta), \)

\((\tau_j^{-1}|\beta_j, \lambda_j, \sigma^2) \sim \text{Inv-Gauss}\{\mu = (\lambda_j^2\sigma^2/\beta_j^2)^{1/2}, \rho = \lambda^2\} \) where \( T = \text{diag}(\tau_1, \ldots, \tau_p) \) and Inv-Gauss denotes the inverse Gaussian distribution with location and scale parameters \( \mu \) and \( \rho \). In our experience, this Gibbs sampler is efficient having fast rates of convergence and mixing.
Bayesian inference in linear models

Consider the linear regression model, \( y = X\beta + \epsilon \), where \( y \) is an \( n \)-dimensional vector of responses, \( X \) is the \( n \times p \) design matrix and \( \epsilon \sim N(0, \sigma^2 I_n) \). Letting \( \beta_j|\sigma \sim \text{GDP}(\xi = \sigma \eta / \alpha, \alpha) \) independently for \( j = 1, \ldots, p \),

\[
\pi(\beta|\sigma) = \prod_{j=1}^{p} \frac{1}{2\sigma \eta / \alpha} \left(1 + \frac{|\beta_j|}{\alpha \sigma \eta / \alpha}\right)^{-(\alpha+1)}.
\]  (4)

From Proposition 1 this prior is equivalent to \( \beta_j|\sigma \sim N(0, \sigma^2 \tau_j) \), with \( \tau_j \sim \text{Exp}(\lambda_j^2 / 2) \) and \( \lambda_j \sim \text{Ga}(\alpha, \eta) \). We place the Jeffreys’ prior on the error variance, \( \pi(\sigma) \propto 1/\sigma \).

Gibbs sampler:

Using the scale mixture of normals representation, we obtain a simple data augmentation Gibbs sampler having the following conditional posteriors: \( (\beta|\sigma^2, T, y) \sim N\{(X'X+T^{-1})^{-1}X'y, \sigma^2 (X'X+T^{-1})^{-1}\}, (\sigma^2|\beta, T, y) \sim \text{IG}\{(n+p)/2, (y-X\beta)'(y-X\beta)/2 + \beta'T^{-1}\beta/2\}, (\lambda_j|\beta_j, \sigma^2) \sim \text{Ga}(\alpha+1, |\beta_j|/\sigma + \eta), (\tau_j^{-1}|\beta_j, \lambda_j, \sigma^2) \sim \text{Inv-Gauss}\{\mu = (\lambda_j^2 \sigma^2 / \beta_j^2)^{1/2}, \rho = \lambda^2\} \) where \( T = \text{diag}(\tau_1, \ldots, \tau_p) \) and Inv-Gauss denotes the inverse Gaussian distribution with location and scale parameters \( \mu \) and \( \rho \). In our experience, this Gibbs sampler is efficient having fast rates of convergence and mixing.

Priors for \( \alpha \) and \( \eta \):

One may either set \( \alpha = 1 \) and \( \eta = 1 \) or choose hyper-priors

\[
\pi(\alpha) = 1/(1 + \alpha)^2, \quad \pi(\eta) = 1/(1 + \eta)^2
\]

which correspond to generalized Pareto hyper-priors with location 0, scale 1 and shape 1.

Sampling scheme: embedded griddy Gibbs (Ritter and Tanner, 1992)
Sparse Maximum a Posteriori Estimation

The generalized double Pareto distribution can be used not only as a prior in a Bayesian analysis but also to induce a sparsity-favoring penalty in regularized least squares,

$$
\hat{\beta} = \arg \min_{\beta} \left\{ \frac{1}{2\sigma^2} \| y - X\beta \|^2 + \sum_{j=1}^{p} p(|\beta_j|) \right\},
$$

where $X$ is initially assumed to have orthonormal columns and $p(.)$ denotes the penalty function implied by the prior on the regression coefficients. Following Fan and Li (2001), let $\hat{\beta} = X'y$ and denote the minimization problem in (5) for a component of $\beta$ as

$$
\tilde{\beta}_j = \arg \min_{\beta_j} \left\{ \frac{1}{2} \left( \hat{\beta}_j - \beta_j \right)^2 + \sigma^2 p(|\beta_j|) \right\},
$$

with the penalty function $p(|\beta_j|) = (\alpha+1) \log (\sigma \eta + |\beta_j|)$ which simply retains the term in $-\log \pi(\beta_j|\alpha, \eta)$ that depends on $\beta_j$.

The authors show that the penalty function $p(|\beta_j|)$ induced by the GDP prior should results in an estimator with the following properties:
(1) nearly unbiased,
(2) a thresholding rule, automatically sets small coefficients to zero,
(3) continuous in data to avoid instabilities in prediction.
MAP estimation via EM algorithm

- Exploiting the Normal Mixture Representation: Use the EM algorithm to update $\beta_j$ and $\sigma$ (called the GDP(MAP) estimator)

- Exploiting the Laplace Mixture Representation and the One-step Estimator: Integration over $\tau$ leads to a Laplace mixture representation of the prior. After that use the LARS algorithm to obtain the one-step estimator (called the GDP(OS) estimator)
Normal vs. Laplace Representations in Computation

Figure 5: Number of iterations until convergence of the EM algorithms under normal and Laplace representations.
Experiments: Simulated data

We generate $n = \{50, 400\}$ observations from $y_i = x_i'\beta^* + \epsilon_i$, where $x_{ij}$ are generated as standard normals with Cov$(x_j, x_{j'}) = 0.5^{j-j'}$, $\epsilon_i \sim N(0, \sigma^2)$ and $\sigma = 3$. We use the following five $\beta^*$ configurations:

Model 1: 5 randomly chosen components of $\beta^*$ are set to 1 and rest to 0.
Model 2: 5 randomly chosen components of $\beta^*$ are set to 3 and rest to 0.
Model 3: 10 randomly chosen components of $\beta^*$ are set to 1 and rest to 0.
Model 4: 10 randomly chosen components of $\beta^*$ are set to 3 and rest to 0.
Model 5: $\beta^* = (0.85, \ldots, 0.85)'$

<table>
<thead>
<tr>
<th>Method</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
<th>Model 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>2.299,0.085</td>
<td>4.879,0.203</td>
<td>2.585,0.134</td>
<td>4.972,0.385</td>
<td>2.886,0.150</td>
</tr>
<tr>
<td>Laplace</td>
<td>2.634,0.137</td>
<td>3.662,0.233</td>
<td>2.837,0.126</td>
<td>4.326,0.211</td>
<td>3.458,0.120</td>
</tr>
<tr>
<td>Horseshoe</td>
<td>2.264,0.086</td>
<td>2.316,0.167</td>
<td>3.205,0.140</td>
<td>3.929,0.218</td>
<td>4.409,0.130</td>
</tr>
<tr>
<td>BMA</td>
<td>2.451,0.123</td>
<td>1.647,0.126</td>
<td>4.043,0.233</td>
<td>3.062,0.194</td>
<td>6.015,0.301</td>
</tr>
<tr>
<td>GDP(PM)$^1$</td>
<td>2.306,0.114</td>
<td>2.405,0.192</td>
<td>3.193,0.215</td>
<td>4.123,0.304</td>
<td>4.283,0.142</td>
</tr>
<tr>
<td>GDP(PM)$^2$</td>
<td>2.303,0.095</td>
<td>2.309,0.195</td>
<td>3.124,0.153</td>
<td>3.910,0.237</td>
<td>4.451,0.199</td>
</tr>
<tr>
<td>GDP(PM)</td>
<td>2.271,0.085</td>
<td>2.606,0.167</td>
<td>3.047,0.147</td>
<td>4.348,0.171</td>
<td>3.640,0.134</td>
</tr>
<tr>
<td>GDP(MAP)$^1$</td>
<td>3.414,0.148</td>
<td>1.619,0.150</td>
<td>5.605,0.298</td>
<td>2.970,0.188</td>
<td>8.769,0.403</td>
</tr>
<tr>
<td>GDP(MAP)$^2$</td>
<td>4.250,0.354</td>
<td>1.618,0.153</td>
<td>6.331,0.300</td>
<td>3.040,0.163</td>
<td>9.308,0.377</td>
</tr>
<tr>
<td>GDP(MAP)</td>
<td>4.876,0.355</td>
<td>2.091,0.182</td>
<td>4.299,0.222</td>
<td>3.740,0.284</td>
<td>5.724,0.177</td>
</tr>
<tr>
<td>LASSO</td>
<td>2.183,0.124</td>
<td>2.618,0.152</td>
<td>3.258,0.194</td>
<td>3.531,0.172</td>
<td>5.646,0.229</td>
</tr>
<tr>
<td>SCAD</td>
<td>3.792,0.214</td>
<td>2.132,0.229</td>
<td>5.249,0.239</td>
<td>3.179,0.193</td>
<td>8.505,0.387</td>
</tr>
</tbody>
</table>

$n = 50$

<table>
<thead>
<tr>
<th>Method</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
<th>Model 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>0.395,0.014</td>
<td>0.455,0.019</td>
<td>0.426,0.016</td>
<td>0.455,0.024</td>
<td>0.412,0.013</td>
</tr>
<tr>
<td>Laplace</td>
<td>0.315,0.016</td>
<td>0.374,0.014</td>
<td>0.388,0.016</td>
<td>0.422,0.015</td>
<td>0.457,0.014</td>
</tr>
<tr>
<td>Horseshoe</td>
<td>0.219,0.016</td>
<td>0.205,0.010</td>
<td>0.341,0.014</td>
<td>0.346,0.009</td>
<td>0.514,0.023</td>
</tr>
<tr>
<td>BMA</td>
<td>0.151,0.011</td>
<td>0.125,0.006</td>
<td>0.240,0.016</td>
<td>0.211,0.009</td>
<td>0.646,0.037</td>
</tr>
<tr>
<td>GDP(PM)$^1$</td>
<td>0.233,0.016</td>
<td>0.206,0.009</td>
<td>0.326,0.015</td>
<td>0.284,0.014</td>
<td>0.625,0.031</td>
</tr>
<tr>
<td>GDP(PM)$^2$</td>
<td>0.228,0.017</td>
<td>0.215,0.009</td>
<td>0.332,0.013</td>
<td>0.303,0.010</td>
<td>0.579,0.027</td>
</tr>
<tr>
<td>GDP(PM)</td>
<td>0.248,0.017</td>
<td>0.182,0.007</td>
<td>0.377,0.016</td>
<td>0.362,0.012</td>
<td>0.466,0.016</td>
</tr>
<tr>
<td>GDP(MAP)$^1$</td>
<td>0.154,0.014</td>
<td>0.111,0.011</td>
<td>0.286,0.016</td>
<td>0.210,0.011</td>
<td>0.739,0.043</td>
</tr>
<tr>
<td>GDP(MAP)$^2$</td>
<td>0.161,0.013</td>
<td>0.111,0.010</td>
<td>0.284,0.016</td>
<td>0.210,0.009</td>
<td>0.652,0.035</td>
</tr>
<tr>
<td>GDP(MAP)</td>
<td>0.185,0.017</td>
<td>0.119,0.010</td>
<td>0.326,0.016</td>
<td>0.336,0.010</td>
<td>0.478,0.020</td>
</tr>
<tr>
<td>LASSO</td>
<td>0.251,0.014</td>
<td>0.276,0.014</td>
<td>0.339,0.020</td>
<td>0.348,0.011</td>
<td>0.485,0.021</td>
</tr>
<tr>
<td>SCAD</td>
<td>0.121,0.010</td>
<td>0.118,0.008</td>
<td>0.233,0.011</td>
<td>0.206,0.017</td>
<td>0.469,0.019</td>
</tr>
</tbody>
</table>

$^1\alpha = 1, \eta = 1; ^2\eta = 1$
## Inferences on Hyper-parameters $\alpha, \eta$

Table 2: Posterior means of the hyper-parameters and the resulting model error.

<table>
<thead>
<tr>
<th>Model 2</th>
<th></th>
<th></th>
<th>Model 5</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(n = 50)</td>
<td>(n = 400)</td>
<td>(n = 50)</td>
<td>(n = 400)</td>
<td></td>
</tr>
<tr>
<td>GDP(PM)</td>
<td>2.464</td>
<td>1.165</td>
<td>0.688</td>
<td>0.870</td>
<td>5.262</td>
</tr>
<tr>
<td>GDP(PM)$^2$</td>
<td>0.784</td>
<td>0.784</td>
<td>0.784</td>
<td>0.784</td>
<td>0.784</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>2.464</td>
<td>1.165</td>
<td>0.688</td>
<td>0.870</td>
<td>5.262</td>
</tr>
<tr>
<td>ME</td>
<td>2.443</td>
<td>2.219</td>
<td>0.149</td>
<td>0.181</td>
<td>6.290</td>
</tr>
</tbody>
</table>

$\eta^2 = 1$