High-dimensional Bayes

Debdeep Pati
Florida State University

November 10, 2016

Robust shrinkage of the sparse signals

- Consider $y \sim N(\theta, 1)$
- A representation of the posterior mean.
- A normal likelihood of known variance $p(y - \theta)$.
- The prior for the mean is $\pi(\theta)$.
- Marginal density $m(y) = \int p(y - \theta)\pi(\theta)d\theta$.
- For one sample of y,

$$E(\theta \mid y) = y + \frac{d}{dy} \log m(y)$$

Why Horseshoe is robust to outlying signals?

- The following result speaks to the horseshoe's robustness to large outlying signals.

Theorem

Suppose $y \sim N(\theta, 1)$. *Let* $m(y)$ *denote the predictive density under the horseshoe prior for known scale parameter* $\tau < \infty$, *i.e. where* $(\theta \mid \lambda) \sim N(0, \tau^2\lambda^2)$ *and* $\lambda \sim Ca^+(0, 1)$. *Let* $E(\theta \mid y)$ *denote the posterior mean. Then* $\lim_{|y| \to \infty} d \log m(y)/dy = 0$.

- This is NOT true for Bayesian Lasso given by

$$\theta_j \mid \tau \sim DE(\tau) \Leftrightarrow \theta_j \mid \tau, \psi_j \sim N(0, \psi_j\tau^2), \psi_j \sim \mathrm{Exp}(1/2)$$

Comparison with Other Bayes Estimators

- In sparse situations, posterior learning τ allows most noise observations to be shrunk very near zero.
- Yet this small value of τ will not inhibit the estimation of large signals
- Under the double-exponential prior, for example, small values of τ can also lead to strong shrinkage near the origin.
- This shrinkage, however, can severely compromise performance in the tails.

Double exponential score function

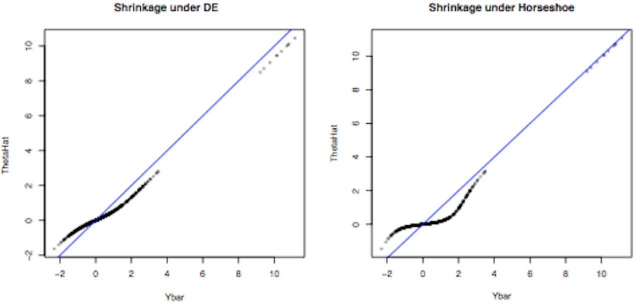
- For DE, smaller value of τ may reduce the risk at the origin,
- But do so at the expense of increased risk in the tails $|E(\theta_i \mid y_i) - y_i| \approx \sqrt{2}/\tau$ for large y_i.

Simulation study

- Ten standard normal observations were simulated for each of 1000 means: 10 signals of mean 10, 90 signals of mean 2 and 900 noise of mean 0.
- Two models were then fit to this data: one that used independent horseshoe priors and one that used independent double-exponential priors.

Simulation study

The double-exponential prior tends to shrink small observations not enough, and the larger observations too much.



Dirichlet-Laplace prior - motivation

- A large subclass of global-local (GL) priors fail to concentrate sufficiently well around sparse vectors
- Horseshoe can perform well in highly sparse situations, but not that well when the number of signals is relatively large compared to the dimension
- Global-local priors mimic point mass mixtures marginally
- Investigate analogy jointly
- Point mass priors equiv. to (i) draw $s \sim \text{Bino}(p, \pi_0)$ (ii) draw a subset S of size s uniformly (iii) set $\theta_j = 0$ for all $j \notin S$ and (iv) draw $\theta_j, j \in S$ i.i.d. from $g(\cdot)$
- Aim to mimic the joint structure implied by point mass priors

Dirichlet Laplace prior & properties

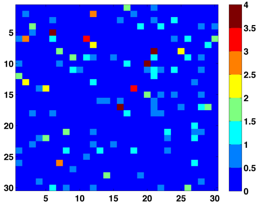
- We propose a simple dependent modification leading to optimal concentration & efficient computation

$$\boxed{\theta_j \sim \mathrm{DE}(\phi_j \tau), \quad \phi = (\phi_1, \ldots, \phi_p)^{\mathrm{T}} \in \mathcal{S}^{p-1}, \quad \tau > 0}$$

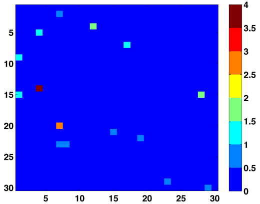
- Constraining ϕ to the simplex crucial - allows for dependence
- We let $\phi \sim \mathrm{Diri}(\alpha, \ldots, \alpha)$ - $\alpha < 1$ favors small # dominant values with remaining ≈ 0
- Normal scale mixture rep: $\theta_j \sim \mathrm{N}(0, \psi_j \phi_j^2 \tau^2)$, $\psi_j \sim \mathrm{Exp}(1/2)$
- Spike at zero controlled by α - use $U(0, 1)$ prior or fix at $1/p$

Prior draws from $|\theta_j|$, $j = 1, \dots, 900$

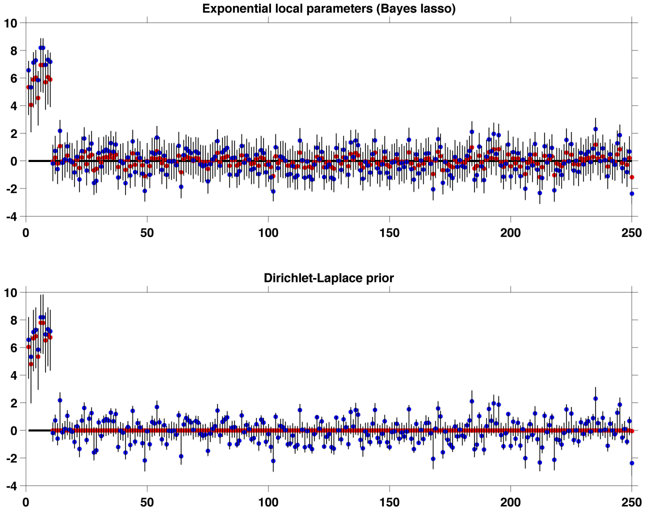
- $\alpha = 1/2$



- $\alpha = 1/10$

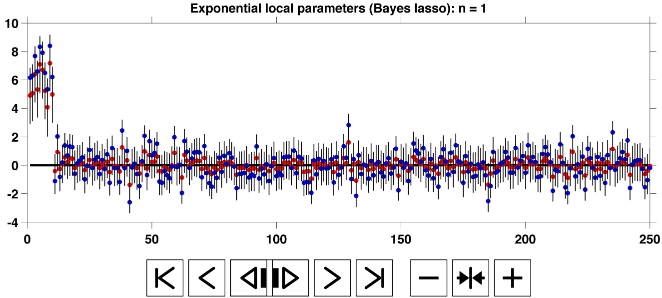


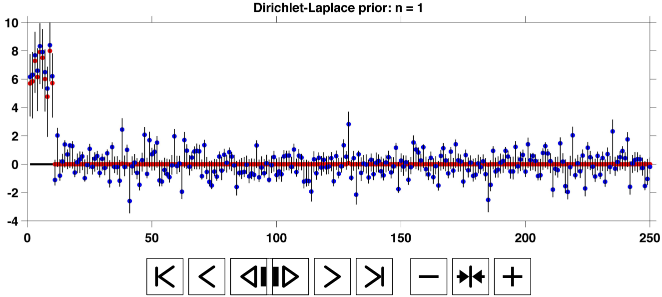
Improved prior concentration reflected in the posterior



Draw $y \sim \mathrm{N}_{250}(\theta_0, \mathrm{I}_{250})$ with $\theta_0[1:10] = 7, \theta_0[11:250] = 0$. Blue dots: entries of y, red dots: posterior median of θ, bars: point wise 95% credible intervals

Increase sample size





Simulation study

- We ran a simulation to assess the performance of our new approach vs Bayes lasso & a host of other methods
- 100 simulation replicates
- Each replicate - one observation $y \sim \mathsf{N}_p(\theta_0, \mathrm{I}_p)$
- θ_0 sparse: $\theta_0[1:q] = A$, $\theta_0[q+1, p] = 0$
- Show results for $q = 10$, $A = 7$
- We cheated on behalf of the frequentist Lasso & selected the penalty that produced the lowest MSE

Simulation study

Table 1: Squared error comparison over 100 replicates. Average squared error across replicates reported for BL (Bayesian lasso), DL (Dirichlet-Laplace), Lasso, EBMed (Empirical Bayes median), PM (Point mass prior) and HS (horseshoe).

p	**100**						**200**					
$\frac{q}{p}\%$	**5**		**10**		**20**		**5**		**10**		**20**	
A	**7**	**8**	**7**	**8**	**7**	**8**	**7**	**8**	**7**	**8**	**7**	**8**
BL	33.05	33.63	49.85	50.04	68.35	68.54	64.78	69.34	99.50	103.15	133.17	136.83
DL	8.20	7.19	17.29	15.35	32.00	29.40	16.07	14.28	33.00	30.80	65.53	59.61
LS	21.25	19.09	38.68	37.25	68.97	69.05	41.82	41.18	75.55	75.12	137.21	136.25
EBMed	13.64	12.47	29.73	27.96	60.52	60.22	26.10	25.52	57.19	56.05	119.41	119.35
PM	12.15	10.98	25.99	24.59	51.36	50.98	22.99	22.26	49.42	48.42	101.54	101.62
HS	8.30	7.93	18.39	16.27	37.25	35.18	15.80	15.09	35.61	33.58	72.15	70.23