The horseshoe+ estimator of sparse signals

Anindya Bhadra
bhadra@purdue.edu
www.stat.purdue.edu/∼bhadra

Purdue University

January 21, 2015

Overview

- The one and two groups models
- The horseshoe+ prior
- Theoretical properties
- Simulations and applications on a prostate cancer data set
- Joint work with Jyotishka Datta (SAMSI and Duke University); Nick Polson and Brandon Willard (The University of Chicago).

The high-dimensional normal means problem

- **Data:** n conditionally independent continuous observations $y = (y_1, y_2, \ldots, y_n)$; $y_i|\theta_i \sim \text{Normal}(\theta_i, 1)$.
- **Testing and null hypotheses:** $H_{0,i} : \theta_i = 0$, $i = 1, \ldots, n$. Goal: provide a decision rule with good error rates.
- **Multiplicity problem:** Number of hypotheses (n) very large $\Rightarrow$ Higher chance of false positives.
- **Estimation:** Goal: provide an estimator for $(\theta_1, \ldots, \theta_n)$ with good MSE properties.

Nearly black vectors: Global shrinkage estimators

- Nearly black vectors: $\#(\theta_i \neq 0) \leq p_n$ where $p_n = o(n)$ as $p_n, n \to \infty$.
- Pure "global" shrinkage estimators (e.g, James-Stein) do poorly in this setting.
- Johnstone and Silverman (2004) show that a simple thresholding-based estimator has better risk bounds than the James-Stein estimator in this case.

Nearly black vectors: the two groups model

- Use indicators $\nu_i, i = 1, \ldots, n$ such that $\nu_i = 0$ indicates $\theta_i = 0$ and $\nu_i = 1$ indicates $\theta_i \neq 0$.

- Given μ, suppose, θ_i's are conditionally independent and:

$$\theta_i|\mu \sim (1-\mu)\underbrace{\delta_{\{0\}}}_{\textit{Spike}} + \mu \overbrace{\text{Normal}(0, \psi^2)}^{\textit{Slab}} \tag{1}$$

- Recall the likelihood is given by $y_i|\theta_i \sim \text{Normal}(\theta_i, 1)$. Thus, the marginal distribution of $y_i|\mu$ is then a mixture of normals:

$$y_i|\mu \sim (1-\mu)\text{Normal}(0,1) + \mu\text{Normal}(0, 1+\psi^2) \tag{2}$$

Posterior mean under the two groups model: the global-local shrinkage

- The posterior mean $E(\theta_i|y_i)$ under the two groups model is:

$$E(\theta_i|y_i) = \omega_i \frac{\psi^2}{1+\psi^2} y_i = \omega_i^* y_i$$

  where, ω_i is the posterior inclusion probability $P(\theta_i \neq 0|y_i)$.

- If $\psi^2 \to \infty$ as the number of tests $n \to \infty$, we have

$$\boxed{E(\theta_i|y_i) \approx \omega_i y_i}$$

- The ψ^2 term provides "global" shrinkage. The ω_i terms act "locally" to adapt to the sparsity level of the data.

Towards the one group model

- Instead of starting with a classification scheme to arrive at an estimator $\omega_i^* y_i$, the one-group approach *directly models the posterior inclusion probability* ω_i^* without making use of the discrete mixture.

- Carvalho, Polson and Scott (2009) observed a suitable "global-local" mixture prior leads to the same form of the posterior mean.

- The "global" term should provide substantial shrinkage towards zero.

- The local terms should have heavy tails so that "signals" are not shrunk too much.

Advantages of the "One-group model"

- "*Sparsity can be construed in a weaker sense, where all of the entries in θ are nonzero, yet most are small compared to a handful of large signals.*" [Polson & Scott (2010), Stephens and Balding (2009)]

- The one-group model yields Analytically Tractable Marginal and under suitable choices of the hyper-parameters, behaves like a two-groups model.

- Normal scale mixture allows for block-updating the local and global shrinkage parameters $\Rightarrow$ Fast Computation.

The horseshoe prior (Carvalho, Polson and Scott (2010))

- The horseshoe prior falls in a class of "global-local" shrinkage priors:

$$
\begin{aligned}
y_i|\theta_i, \lambda_i, \tau &\sim \text{Normal}(\theta_i, 1), \\
\theta_i|\lambda_i, \tau &\sim \text{Normal}\left(0, \lambda_i^2\right), \\
\lambda_i|\tau &\sim C^+(0, \tau). \text{ (Heavy Tailed Prior)}
\end{aligned}
$$

- **Posterior mean:** $\mathbb{E}(\theta_i|y_i, \tau) = (1 - \mathbb{E}(\overbrace{\frac{1}{1+\lambda_i^2\tau^2}}^{\kappa_i}|y_i, \tau))y_i.$

Two-groups Model	One-group Model
$E(\theta_i\|y_i) \approx \omega_i y_i;\ \omega_i = P(\theta_i \neq 0\|y_i)$	$\mathbb{E}(\theta_i\|y_i, \tau) = (1 - \mathbb{E}(\kappa_i\|y_i, \tau))y_i$

- $1 - \mathbb{E}(\kappa_i|y_i, \tau)$ mimics the posterior inclusion probability ω_i.

Global-local shrinkage priors: other examples

- Since horseshoe, there has been considerable work in this area.
  - The horseshoe Prior (Carvalho, Polson and Scott, 2010).
  - The hypergeometric Inverted-beta priors (Polson and Scott, 2010)
  - The generalized double Pareto priors (Armagan, Dunson and Lee, 2011)
  - The three parameter beta priors (Armagan, Dunson and Clyde, 2011) (which includes the NEG priors (Griffin and Brown, 2011) and the half-t prior).
  - The Dirichlet-Laplace prior (Bhattacharya et al., 2012). (resembles the joint distribution of θ, unlike the other shrinkage priors that mimic the marginal behavior).
- "Global-local" shrinkage priors because they shrink small observations (encourages sparsity) but leave the tails unshrunk (helps detect signals)

Properties of the horseshoe prior

- The horseshoe prior has a number of attractive features for the sparse signal recovery problem :
  1. Datta and Ghosh (2013): horseshoe estimator attains Bayes Oracle in testing.
  2. Polson and Scott (2012): Horseshoe estimator $\hat{\theta}_{HS}$ uniformly dominates the traditional sample mean estimator in MSE.
  3. van der Pas et al. (2014): Horseshoe estimator has good posterior concentration properties for nearly black objects (Donoho et al., 1992). Specifically, the horseshoe estimator attains the asymptotically minimax risk rate.

$$\sup_{\theta \in l_0[p_n]} \mathbb{E}_\theta ||\hat{\theta}_{HS} - \theta||^2 \asymp p_n \log(n/p_n)$$

Our Contributions: The horseshoe+ prior

- We (Bhadra, Datta, Polson, and Willard, 2014) propose a *new* prior (horseshoe+) that sharpens the ability of the horseshoe estimator to extract signals from sparsity.

- horseshoe+ is a 'natural' extension of the horseshoe model.

- **Advantages:**
  1. Lower mean squared error compared to existing sparse estimators (horseshoe, Dirichlet-Laplace).
  2. Achieves lower misclassification probability.
  3. Better posterior concentration properties in the Kullback-Liebler (K-L) sense.

Hierarchical Models: horseshoe and horseshoe+

The horseshoe hierarchical model is defined by

$$\begin{aligned}(y_i|\theta_i,\lambda_i,\tau) &\sim N(\theta_i,1),\\ (\theta_i|\lambda_i,\tau) &\sim N\left(0,\lambda_i^2\right),\\ (\lambda_i|\tau) &\sim C^+\left(0,\tau\right).\end{aligned}$$

We define the horseshoe+ model similarly by

$$\begin{aligned}(y_i|\theta_i,\lambda_i,\eta_i,\tau) &\sim N(\theta_i,1),\\ (\theta_i|\lambda_i,\eta_i,\tau) &\sim N\left(0,\lambda_i^2\right),\\ (\lambda_i|\eta_i,\tau) &\sim C^+\left(0,\tau\eta_i\right),\\ \eta_i &\sim C^+\left(0,1\right).\end{aligned}$$

- Horseshoe: λ_i's are conditionally independent with $p(\lambda_i|\tau)$:

$$p(\lambda_i|\tau) = \frac{2}{\pi\tau(1+(\lambda_i/\tau)^2)}.$$

- Horseshoe+: λ_i's are conditionally independent with $p(\lambda_i|\tau)$:

$$p(\lambda_i|\tau) = \frac{4}{\pi^2\tau}\frac{\ln\{(\lambda_i/\tau)\}}{(\lambda_i/\tau)^2-1}.$$

- An extra $\ln(\lambda_i/\tau)$ term in the numerator.

The horseshoe+ Jacobian

- The horseshoe prior:

$$\lambda_i \sim C^+(0,\tau)$$

$$\Rightarrow p(\kappa_i|\tau) \propto \frac{1}{\sqrt{\kappa_i(1-\kappa_i)}} \underbrace{\frac{1}{(1-\kappa_i(1-\tau^2))}}_{\text{Jacobian}}$$

- The horseshoe+ prior:

$$(\lambda_i|\tau) \sim C^+(0,\tau\eta_i)\,;\ \eta_i \sim C^+(0,1)$$

$$\Rightarrow p(\kappa_i|\tau) \propto \frac{1}{\sqrt{\kappa_i(1-\kappa_i)}} \underbrace{\frac{\ln\left\{(1-\kappa_i)/\kappa_i\tau^2\right\}}{(1-\kappa_i(\tau^2+1))}}_{\text{Jacobian}}$$

- The horseshoe+ Jacobian pushes the posterior mass to the extremes ($\kappa = 0, 1$)

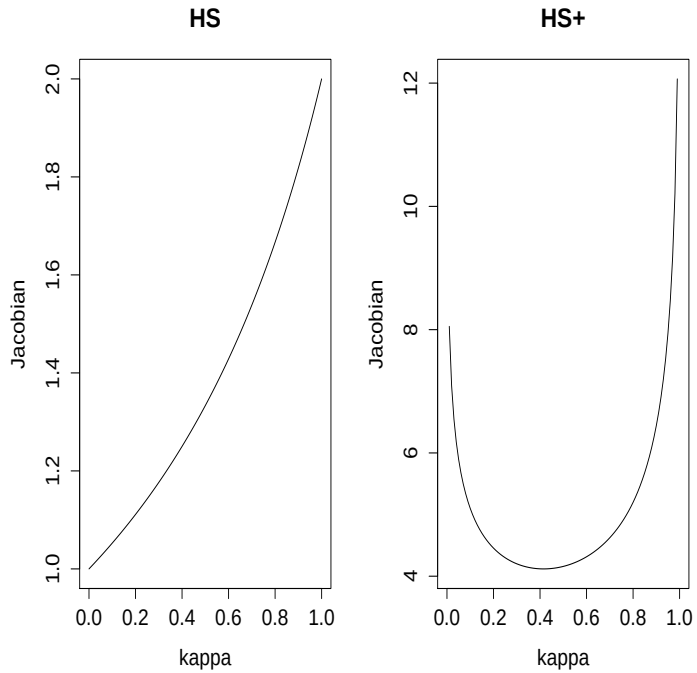
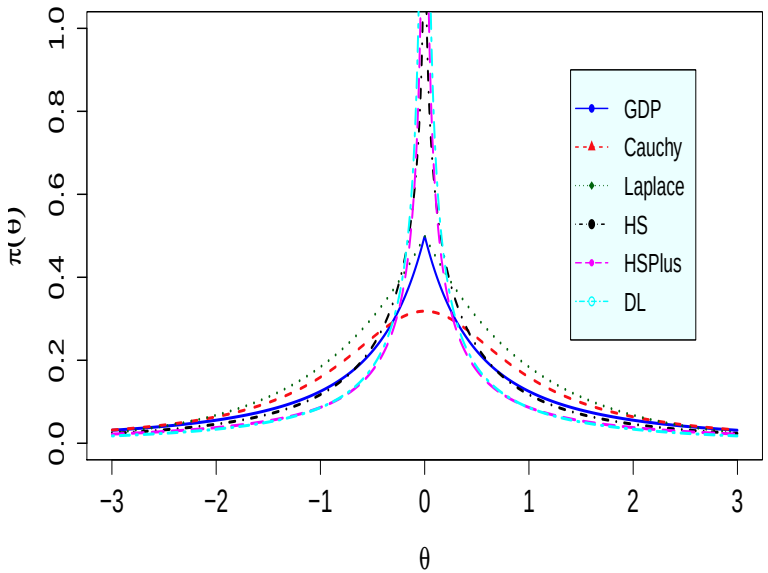


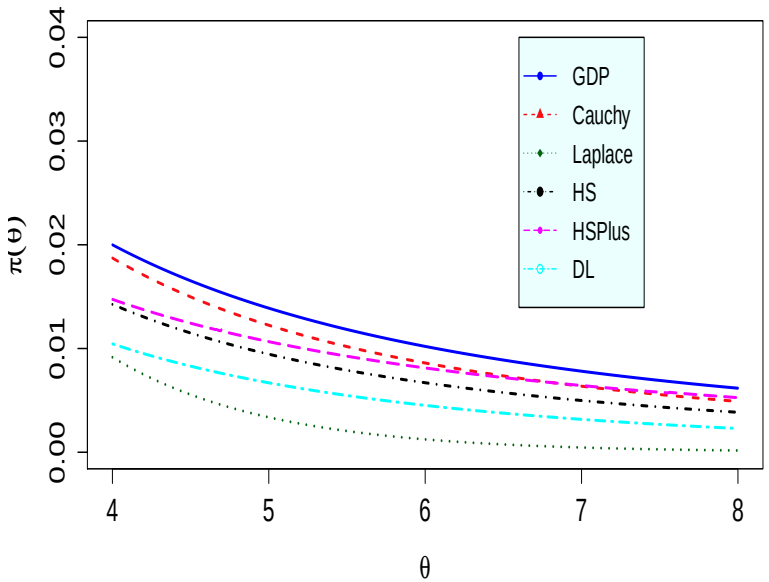
Figure : Comparison of the Jacobian terms for the HS and the HS+ priors with $\tau = 1/2$.

Comparison of different priors: near origin and at the tails

- The order of peakedness near zero: $HS+ \approx DL > HS > GDP = Laplace > Cauchy$



- The order of tail heaviness: $GDP > Cauchy > HS+ > HS > DL > Laplace$.



- $p_{HS+}(\theta) > p_{HS}(\theta)$ in a neigbourhood around near zero and also at infinity.
- Dirichlet Laplace has spike at zero, but exponentially decaying tail.

Properties of Horsehoe+ (I): the marginal prior density

- The horseshoe+ density satisfies the following:

Theorem

1. $$\frac{1}{\pi^2\sqrt{2\pi}} \log\left(1+\frac{4}{\theta^2}\right) < p_{HS+}(\theta) \leq \frac{1}{\pi^2|\theta|}$$
2. $$\lim_{|\theta|\to 0} p_{HS+}(\theta) = \infty$$

- The horseshoe+ prior has unbounded mass near the origin and polynomially decaying tails.

Properties of horseshoe+ (II): mean squared error

Theorem

(Barndorff-Nielsen et al., 1982). Consider the Gaussian scale mixture $y|u \sim \text{Normal}(0, u)$. *If we can write* $f(u) \propto u^{\lambda-1}L(u)$ *as* $u \to \infty$, *then* $m(y) \propto |y|^{2\lambda-1}L(y^2)$ *as* $|y| \to \infty$, *where* $L(\cdot)$ *is a slowly varying function, defined as* $\lim_{y\to\infty} L(ty)/L(y) = 1$ *for any* $t \in (0, \infty)$.

- The heavy tail of the prior scale translates to heavy tail of the marginal. Can show $m_{HS+}(y) = m_{HS}(y)\log(|y|)(1 + o(1))$ as $|y| \to \infty$.

Properties of horseshoe+ (II): mean squared error

- Tweedie formula relates the marginal $m(y_i)$ to the posterior moments

$$\begin{aligned} \mathrm{E}(\theta_i|y_i) &= y + \frac{d}{dy_i}[\log m(y_i)] \\ \mathrm{Var}(\theta_i|y_i) &= 1 + \frac{d^2}{dy_i^2}[\log m(y_i)]. \end{aligned}$$

- After some simple calculations, it follows that

$$MSE_{HS+}(\hat{\theta}_i) \leq MSE_{HS}(\hat{\theta}_i) - \frac{1}{y_i^2 \log |y_i|} + O(1/y_i^3).$$

as $|y_i| \to \infty$.

Properties of horseshoe+ (III): posterior concentration

- Fact 1: The posterior distribution of the shrinkage coefficient κ_i given τ and the observation y_i would converge to the degenerate distribution at one if $\tau \to 0$.

- Fact 2: The posterior distribution of the shrinkage coefficient κ_i given τ and an observation y_i would converge to the degenerate distribution at zero if $|y_i| \to \infty$.

Properties of horseshoe+ (III): posterior concentration

Theorem

$P(\kappa_i < \epsilon | y_i, \tau) \leq e^{\frac{y_i^2}{2}} \tau^2 \epsilon (1-\epsilon)^{-2}$ *for any fixed* $\epsilon \in (0,1)$, *and any* $\tau \in (0,1)$ *uniformly in* $y_i \in \mathcal{R}$.

Theorem

$P(\kappa_i > \eta | y_i, \tau) \leq e^{-\eta(1-\delta)\frac{y_i^2}{2}} \frac{1}{\tau^2} C(\eta, \delta)$ *for any fixed* $\eta \in (0,1)$, *any* δ *such that* $\eta\delta < \frac{1}{1+\tau^2}$ *and uniformly in* $y_i \in \mathcal{R}$, *where* $C(\eta, \delta)$ *is a constant independent of* y_i.

Properties of horseshoe+ (III): posterior concentration

- Decision rule :

$$\text{Reject } H_{0i} \text{ if } \omega_i = 1 - \mathbb{E}(\kappa_i | y_i, \tau) > 1/2.$$

Theorem

Probability of type-I error: $t_1 = \left\{ \sqrt{2}\tau^2 / \sqrt{\pi \ln(1/2\tau)} \right\} (1 + o(1))$

Theorem

Probability of type-II error: $t_2 \leq (2\Phi(\sqrt{\frac{2}{\eta(1-\delta)}}\sqrt{C}) - 1)(1 + o(1))$

- Using these two results, we can show that horseshoe+ achieves the optimal Bayes risk in testing.

Properties of horseshoe+ (IV): Kullback-Leibler (KL) risk

- θ_0 = true parameter, $p_\theta = p(y|\theta)$ = sampling model.
- $L(p_1, p_2) = E_{p_1}(\log(p_1/p_2))$ = KL divergence of p_2 from p_1.
- $A_\epsilon = \{\theta : L(p_{\theta_0}, p_\theta) < \epsilon\}$ KL information nbd. of size ϵ.
- $\hat{p}_n = \int p_\theta \mu_n(d\theta)$ = posterior predictive density.

Lemma (Clarke and Barron, 1990)

Suppose $\mu(A_\epsilon) > 0$, then R_n, Cesaro-average risk of $\hat{p}_n$ will satisfy:

$$R_n = n^{-1} \sum_{j=1}^{n} L(p_{\theta_0}, \hat{p}_j) \leq \epsilon - n^{-1} \log(\mu(A_\epsilon))$$

Characterizes KL-risk in terms of $\mu(A_\epsilon)$ = prior mass around θ_0.

- The $\mu(A_\epsilon)$ bound is larger for the horseshoe+ prior. R_n converges at a faster rate.

Simulations: MSE comparison

- We provide a comprehensive MSE comparison with existing sparse estimators.
- Performance Criterion: mean squared error for the posterior median over 100 replicates.
- Candidates: Dirichlet Laplace $DL_{1/n}$ (Bhattacharya et al., 2014), horseshoe and horseshoe+.

- Data generating scheme in Bhattacharya et al. (2014)

$$\mathbf{Y}_{n\times 1} \sim \text{Normal}(\theta_{n\times 1}, \mathbf{I}_n)$$
$$\theta = (\underbrace{A, \ldots, A}_{[qn]}, \underbrace{0, \ldots, 0}_{n-[qn]})$$

- Multiplicity $n = 200$. For each n, $q = 1, 5, 10, 20, 30\%$ of n and $A = 7, 8$.

Simulations: MSE comparison

	Proportion of signals					
	0.01		0.05		0.1	
	A		A		A	
Prior	7	8	7	8	7	8
D-L	3.88 (8.57)	**2.63** (2.97)	16.46 (16.81)	14.8 (14.98)	29.12 (19.74)	31.8 (21.7)
HS Cauchy	2.57 (2.93)	2.93 (3.47)	13.93 (6.54)	14.39 (12.23)	26.77 (9.39)	27.73 (9.27)
HS+ Cauchy	**2.43** (2.7)	2.98 (3.47)	**13.26** (6.51)	**12.58** (5.83)	**25.55** (9.29)	**26.14** (9.12)
HS Unif	2.97 (3.2)	3.39 (4.08)	14.23 (6.53)	13.9 (6.79)	27.22 (9.8)	28.54 (10.18)
HS+ Unif	2.86 (3.07)	3.3 (4.06)	13.5 (6.47)	13.07 (6.03)	26.83 (10.98)	26.92 (9.73)

	Proportion of signals			
	0.2		0.3	
	A		A	
Prior	7	8	7	8
D-L	53.32 (23.23)	51.08 (23.98)	**72.32** (20.76)	79.19 (40.48)
HS Cauchy	57.15 (14.52)	56.62 (16.59)	82 (14.65)	84.8 (25.57)
HS+ Cauchy	**51.5** (11.03)	**50.37** (13.41)	73.74 (11.93)	**76.85** (20.15)
HS Unif	56.98 (13)	55.06 (13.19)	83.15 (21.05)	81.66 (16.25)
HS+ Unif	53.19 (14.27)	52.27 (19.32)	74.43 (12.84)	77.89 (26.77)

Table : Average mean squared error (and s.d.) about the posterior median computed over 100 simulated data sets.

Simulations: misclassification probability

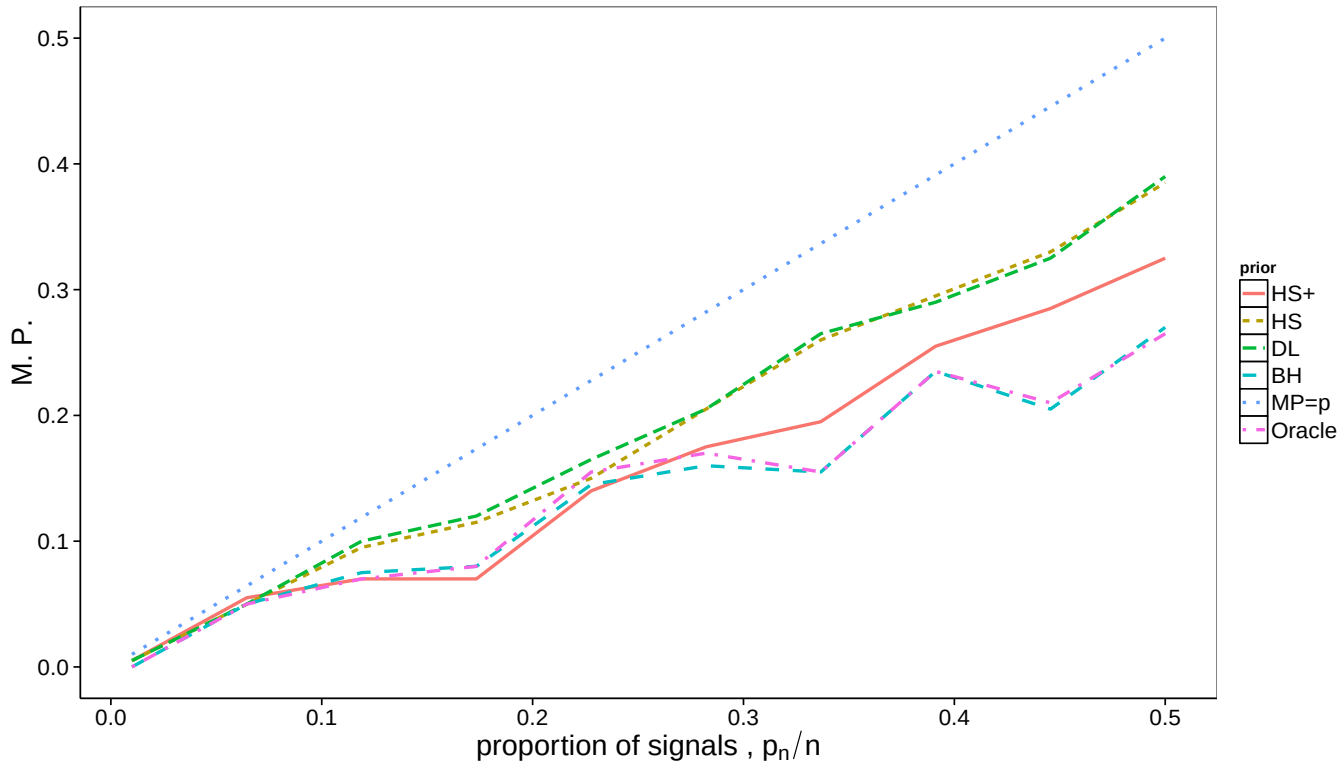


Figure : Misclassification probability plots for the horseshoe+, horsesshoe, and the Dirichlet-Laplace ($DL_{1/n}$) shrinkage priors, Benjamini-Hochberg and the Bayes oracle for $p \in (0.1, 0.5)$.

Real data application: prostate cancer data

- We use the prostate cancer data set from Efron (2008).
- The data are inverse cdf transform for 6033 genes for two-sample t-test statistic (computed over 52 cancer patients and 50 normal subjects).
- Specifically, $y_i = \Phi^{-1}(F_{t,df=100}(t_i))$. Natural to model $y_i = \text{Normal}(\theta_i, 1)$; $i = 1, \ldots, 6033$.
- Histogram of y_i displays heavy tails, suggesting a few regulatory genes.

Real data application: prostate cancer data

Table : The test statistics (y-values) and the effect-size estimates for the top 10 genes selected by Efron (2008) by the horseshoe, horseshoe+ models, and Efron's two-groups model estimates.

Gene	y-value	$\hat{\theta}_i^{HS+}$	$\hat{\theta}_i^{HS}$	$\hat{\theta}_i^{\text{Efron}}$
610	5.29	5.20	5.12	4.11
1720	4.83	4.77	4.54	3.65
332	4.47	3.24	4.11	3.24
364	-4.42	-4.43	-4.14	-3.57
914	4.40	4.40	3.89	3.16
3940	-4.33	-3.78	-3.77	-3.52
4546	-4.29	-3.88	-3.46	-3.47
1068	4.25	3.71	3.03	2.99
579	4.19	3.99	2.88	2.92
4331	-4.14	-3.48	-3.26	-3.30

- HS+ shrinks the large signals less compared to HS and Efron.

Conclusions

- We proposed a new prior for estimation/testing in the sparse normal means problem.
- Theoretical and empirical results suggest considerable improvements over existing alternatives.
- Open question: Necessary conditions for "global-local" shrinkage priors?
- Open question: Extending the hierarchy? What about HS ++ ...?

Thank you!