The Horseshoe and More General Sparsity Priors

S.L. van der Pas

Joint with A.W. van der Vaart, B.J.K. Kleijn, J.-B. Salomond, and J. Schmidt-Hieber

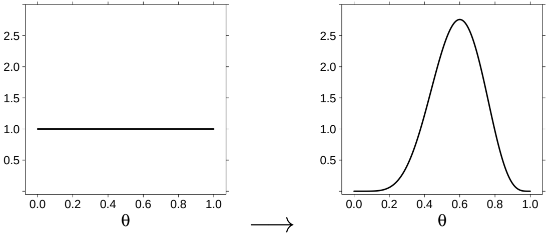
EYSM 2015

Outline

Frequentist Bayes

Frequentist Bayes

Example: coin toss, θ = probability of heads. Data: n coin tosses, h heads, t tails.



- Estimator: $\mathbb{E}[\theta \mid \text{data}] = \frac{h+1}{n+2}$.

Goal

Assumption: there is some true θ_0 generating the data (*frequentist* Bayes).

Goal: recovery and uncertainty quantification.

The normal means model

We observe a vector $Y \in \mathbb{R}^n$ such that

$$Y_i = \theta_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

where

- $\varepsilon_i \sim \mathcal{N}(0, 1)$ i.i.d.
- $\theta \in \ell_0[p_n]$, i.e. $\#(1 \leq i \leq n : \theta_i \neq 0) \leq p_n$.

Assumption: $p_n = o(n)$, $p_n \to \infty$ as $n \to \infty$.

Minimax risk

[Donoho, Johnstone, Hoch and Stern (1992)]

As $n, p_n \to \infty$:

$$\inf_{\widehat{\theta}\in\mathbb{R}^n} \sup_{\theta\in\ell_0[p_n]} \mathbb{E}_\theta\|\theta - \widehat{\theta}\|^2 = 2p_n \log \frac{n}{p_n}(1 + o(1)),$$

where $\|.\|$ denotes the ℓ_2 norm.

Spike-and-slab

Mixture of a Dirac measure at zero and a univariate distribution:

$$\theta_i \sim (1-\alpha)\delta_0 + \alpha g.$$

- good posterior concentration for several combinations of g and priors on α. (Castillo and Van der Vaart (2012))
- Requires exploration of model space of size 2^n.

Lasso

- MAP estimator, Laplace prior with common parameter on each θ_i.
- ℓ_2 risk of order $p_n \log n$ (Bickel, Ritov and Tsybakov (2009)).
- Full posterior distribution contracts at much slower rate than the mode (Castillo, Schmidt-Hieber and Van der Vaart (2014)).

The horseshoe prior

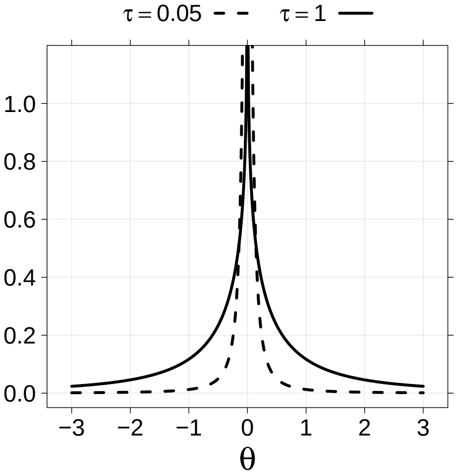
[Carvalho, Polson and Scott (2010)]

$$\theta_i \mid \lambda_i, \tau \sim \mathcal{N}\left(0, \tau^2\lambda_i^2\right), \quad \lambda_i \sim C^+(0,1), \quad i = 1, \ldots, n.$$

The horseshoe prior

[Carvalho, Polson and Scott (2010)]

$$\theta_i \mid \lambda_i, \tau \sim \mathcal{N}\left(0, \tau^2\lambda_i^2\right), \quad \lambda_i \sim C^+(0,1), \quad i = 1, \ldots, n.$$



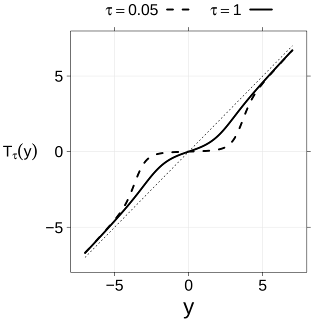
Posterior mean

By Tweedie's formula:

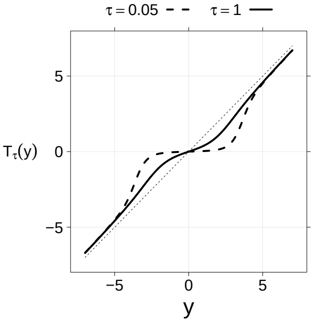
$$T_\tau(y_i) := \mathbb{E}[\theta \mid y_i, \tau] = y_i \left(1 - \frac{2\Phi_1\left(\frac{1}{2}, 1, \frac{5}{2}; \frac{y_i^2}{2}, 1 - \frac{1}{\tau^2}\right)}{3\Phi_1\left(\frac{1}{2}, 1, \frac{3}{2}; \frac{y_i^2}{2}, 1 - \frac{1}{\tau^2}\right)} \right),$$

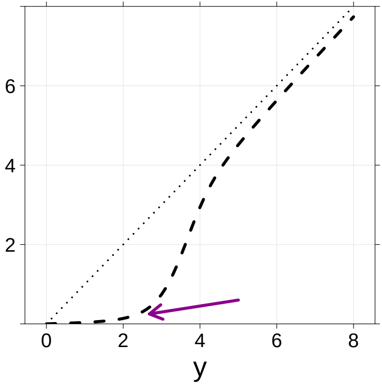
where Φ_1 is the degenerate hypergeometric function of two variables.

Posterior mean - threshold effect



Posterior mean - threshold effect

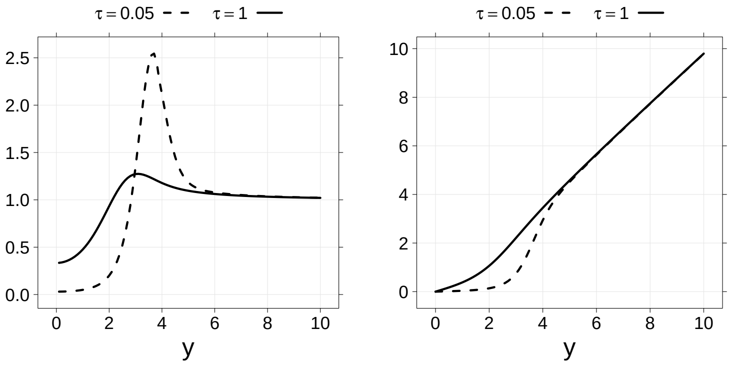




'Threshold' at $\sqrt{2 \log \frac{1}{\tau}}$.

Posterior variance

$$\mathrm{Var}(\theta_i|y_i) = \frac{1}{y_i} T_\tau(y_i) - (T_\tau(y_i) - y_i)^2 + y_i^2 \frac{8\Phi_1\left(\frac{1}{2}, 1, \frac{7}{2}; \frac{y_i^2}{2}, 1 - \frac{1}{\tau^2}\right)}{15\Phi_1\left(\frac{1}{2}, 1, \frac{3}{2}; \frac{y_i^2}{2}, 1 - \frac{1}{\tau^2}\right)}.$$



Posterior concentration

Theorem
With $\tau = \frac{p_n}{n}$:

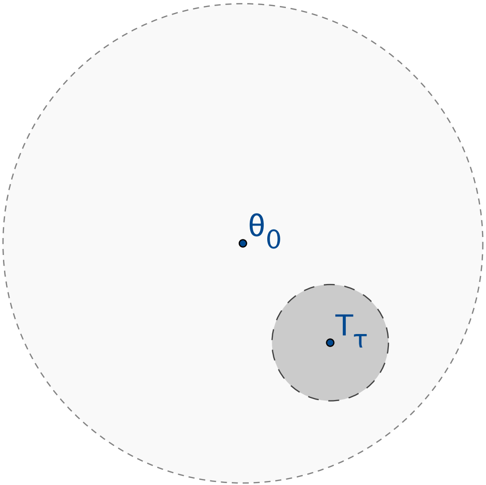
$$\sup_{\theta_0 \in \ell_0[p_n]} \mathbb{E}_{\theta_0} \Pi_\tau \left(\theta : \|\theta - \theta_0\|^2 > M_n p_n \log \frac{n}{p_n} \,\middle|\, Y \right) \to 0,$$

and

$$\sup_{\theta_0 \in \ell_0[p_n]} \mathbb{E}_{\theta_0} \Pi_\tau \left(\theta : \|\theta - T_\tau(Y)\|^2 > M_n p_n \log \frac{n}{p_n} \,\middle|\, Y \right) \to 0,$$

for every $M_n \to \infty$ as $n \to \infty$.

Mismatch?



Posterior variance

Theorem

Suppose $Y \sim \mathcal{N}(\theta_0, \sigma^2 I_n)$, $\theta_0 \in \ell_0[p_n]$. Then:

$$\mathbb{E}_{\theta_0} \sum_{i=1}^{n} \mathrm{Var}(\theta_{0i} | Y_i) \asymp p_n \log \frac{n}{p_n}$$

if $\tau = \frac{p_n}{n}\sqrt{\log \frac{n}{p_n}}$, and $p_n = o(n)$, as $n \to \infty$.

Posterior variance

Theorem
Suppose $Y \sim \mathcal{N}(\theta_0, \sigma^2 I_n)$, $\theta_0 \in \ell_0[p_n]$. Then:

$$\mathbb{E}_{\theta_0} \sum_{i=1}^{n} \mathrm{Var}(\theta_{0i} | Y_i) \asymp p_n \log \frac{n}{p_n}$$

if $\tau = \frac{p_n}{n}\sqrt{\log \frac{n}{p_n}}$, and $p_n = o(n)$, as $n \to \infty$.

For smaller values of τ, there are θ_0 for which $\mathbb{E}_{\theta_0} \| T_\tau(Y) - \theta_0 \|^2$ is of larger order than $\mathbb{E}_{\theta_0} \sum_{i=1}^{n} \mathrm{Var}(\theta_{0i} \mid Y_i)$.

Empirical Bayes - conditions on $\widehat{\tau}$

Theorem

If $\widehat{\tau} \in (0,1)$ satisfies, for $\tau = \frac{p_n}{n}$ or $\tau = \frac{p_n}{n}\sqrt{\log(n/p_n)}$:

1. $\widehat{\tau}$ does not overestimate τ by too much.
   $\mathbb{P}_\theta(\widehat{\tau} > c\tau) \lesssim \frac{p_n}{n}$ for a constant $c \geq 1$ such that $\tau \leq 1/c$
2. $\widehat{\tau}$ does underestimate τ by too much.
   $\exists\, g : \mathbb{N} \times \mathbb{N} \to (0,1)$ such that $\widehat{\tau} \geq g(n, p_n)$ w.p. 1, and
   $-\log(g(n, p_n))\mathbb{P}_\theta\left(\widehat{\tau} \leq \tau\right) \lesssim \log \frac{n}{p_n}$,

then:

$$\sup_{\theta_0 \in \ell_0[p_n]} \mathbb{E}_{\theta_0} \| T_{\widehat{\tau}}(Y) - \theta_0 \|^2 \asymp p_n \log \frac{n}{p_n}.$$

Empirical Bayes - conditions on $\widehat{\tau}$

Theorem

If $\widehat{\tau} \in (0,1)$ satisfies, for $\tau = \frac{p_n}{n}$ or $\tau = \frac{p_n}{n}\sqrt{\log(n/p_n)}$:

1. $\widehat{\tau}$ does not overestimate τ by too much.

   $\mathbb{P}_\theta(\widehat{\tau} > c\tau) \lesssim \frac{p_n}{n}$ for a constant $c \geq 1$ such that $\tau \leq 1/c$

2. $\widehat{\tau}$ does underestimate τ by too much.

   $\exists\, g : \mathbb{N} \times \mathbb{N} \to (0,1)$ such that $\widehat{\tau} \geq g(n, p_n)$ w.p. 1, and

   $-\log(g(n,p_n))\mathbb{P}_\theta\left(\widehat{\tau} \leq \tau\right) \lesssim \log \frac{n}{p_n}$,

then:

$$\sup_{\theta_0 \in \ell_0[p_n]} \mathbb{E}_{\theta_0} \| T_{\widehat{\tau}}(Y) - \theta_0 \|^2 \asymp p_n \log \frac{n}{p_n}.$$

If only the first condition can be verified for an estimator $\widehat{\tau}$, then $\sup\{\frac{1}{n}, \widehat{\tau}\}$ will have an ℓ_2 risk of at most order $p_n \log n$.

Example

$$\widehat{\tau} = \frac{\#\{|y_i| \geq \sqrt{c_1 \log n}, i = 1, \ldots, n\}}{c_2 n}$$

- Satisfies first condition if $c_1 > 2, c_2 > 1$: $p_n \to \infty$.

The horseshoe is not special.

Conditions

$$\theta_i \mid \sigma_i^2 \sim \mathcal{N}(0, \sigma_i^2), \quad \sigma_i^2 \sim \pi(\sigma_i^2), \quad i = 1, \ldots, n.$$

Conditions:

1. The tails of π decay at most exponentially fast.
2. The tails of π are not too heavy.
3. π should have a large amount of mass around zero relative to the tail, and it should increase as p_n decreases.

Main result

$X^n \sim \mathcal{N}(\theta_0, I_n)$. Under Conditions 1-3:

$$\sup_{\theta_0 \in \ell_0[p_n]} \Pi\left(\theta \,:\, \|\theta - \theta_0\|^2 > M_n p_n \log(n/p_n) \,\Big|\, X^n\right) \to 0 \qquad (1)$$

and

$$\sup_{\theta_0 \in \ell_0[p_n]} \Pi\left(\theta \,:\, \|\theta - \widehat{\theta}\|^2 > M_n p_n \log(n/p_n) \,\Big|\, X^n\right) \to 0 \qquad (2)$$

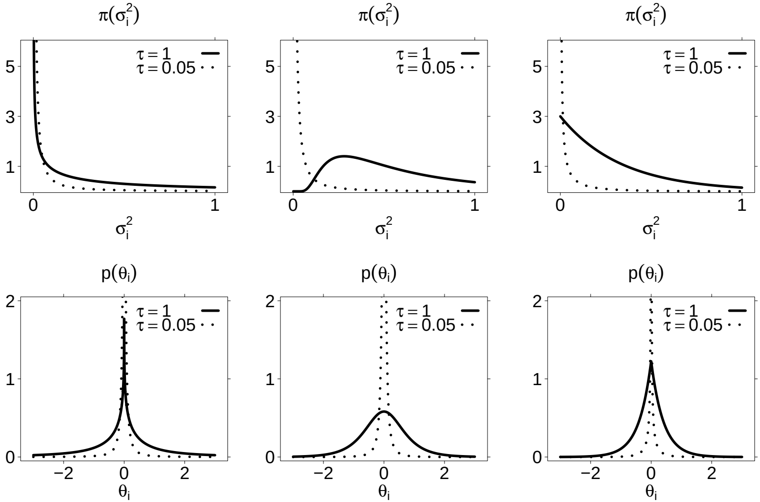
for every $M_n \to \infty$ as $p_n, n \to \infty$, $p_n = o(n)$.

Examples

- Class of priors of Ghosh and Chakrabarti (2015), including the horseshoe and normal-exponential gamma prior (Griffin and Brown (2005)).
- Horseshoe+ (Bhadra et al. (2015).
- Inverse-Gaussian prior (Caron and Doucet (2008)).
- Normal-gamma prior (Caron and Doucet (2008), Griffin and Brown (2010)).
- Spike-and-slab Lasso (Ročkova (forthcoming)).

Examples

Horseshoe, Inverse-Gaussian, and normal gamma.



Conclusions

- Continuous shrinkage priors enjoy good posterior concentration properties, as well as computational convenience.
- The horseshoe is not special; many shrinkage priors lead to posterior contraction at the minimax rate around the truth.
- Or is the horseshoe special after all? $\rightarrow$ lower bound on posterior variance.