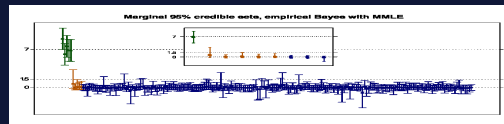


Nonparametric Bayesian Uncertainty Quantification

Lecture 3: High-dimensional Models and Sparsity

Aad van der Vaart

Universiteit Leiden, Netherlands



YES, Eindhoven, January 2017

Contents

Sparsity

Frequentist Bayes

Model Selection Prior

Horseshoe Prior

Co-authors

Sequence model & Regression



Ismael Castillo

Regression



Johannes Schmidt-Hieber

Horsehoe



Stéphanie van der Pas



Botond Szabo

Sparsity

Bayesian sparsity

A **sparse model** has many parameters, but most of them are (nearly) zero.

Bayesian sparsity

A **sparse model** has many parameters, but most of them are (nearly) zero.



We express this in the prior, and apply **the standard (full or empirical) Bayesian machine**.

Bayesian sparsity

A **sparse model** has many parameters, but most of them are (nearly) zero.



We express this in the prior, and apply **the standard (full or empirical) Bayesian machine**.

Parameter with prior $\theta \sim \Pi$, and data $Y^n | \theta \sim p_\theta$, give posterior:

$$d\Pi(\theta | Y^n) \propto p(Y^n | \theta) d\Pi(\theta).$$

Bayesian sparsity

A **sparse model** has many parameters, but most of them are (nearly) zero.



We express this in the prior, and apply **the standard (full or empirical) Bayesian machine**.

In this lecture **sequence model**: $Y^n \sim N_n(\theta, I)$.

Results extend to **regression model**: $Y^n \sim N_n(X_{n \times p}\theta, I)$
under under appropriate conditions on $X_{n \times p}$.

In both cases θ is known to have many (almost) zero coordinates, and p and n are large.

Model selection prior

Constructive definition of prior Π for $\theta \in \mathbb{R}^p$:

- (1) Choose s from prior π on $\{0, 1, 2, \dots, p\}$.
- (2) Choose $S \subset \{0, 1, \dots, p\}$ of size $|S| = s$ at random.
- (3) Choose $\theta_S = (\theta_i: i \in S)$ from density g_S on \mathbb{R}^S and set $\theta_{S^c} = 0$.

Model selection prior

Constructive definition of prior Π for $\theta \in \mathbb{R}^p$:

- (1) Choose s from prior π on $\{0, 1, 2, \dots, p\}$.
- (2) Choose $S \subset \{0, 1, \dots, p\}$ of size $|S| = s$ at random.
- (3) Choose $\theta_S = (\theta_i: i \in S)$ from density g_S on \mathbb{R}^S and set $\theta_{S^c} = 0$.

We are particularly interested in π .

Model selection prior

Constructive definition of prior Π for $\theta \in \mathbb{R}^p$:

- (1) Choose s from prior π on $\{0, 1, 2, \dots, p\}$.
- (2) Choose $S \subset \{0, 1, \dots, p\}$ of size $|S| = s$ at random.
- (3) Choose $\theta_S = (\theta_i: i \in S)$ from density g_S on \mathbb{R}^S and set $\theta_{S^c} = 0$.

We are particularly interested in π .

EXAMPLE (*spike and slab*)

- Choose $\theta_1, \dots, \theta_p$ i.i.d. from $\tau\delta_0 + (1 - \tau)G$.
- Put a prior on τ , e.g. Beta(1, $p + 1$).

This gives binomial π and product densities $g_S = \bigotimes_{i \in S} g$.

[Mitchell & Beachamp (88), George, George & McCulloch, Yuan, Berger, Johnstone & Silverman, Richardson et al., Johnson & Rossell,

Chao Gao, ...]

Horseshoe prior

Horseshoe prior:

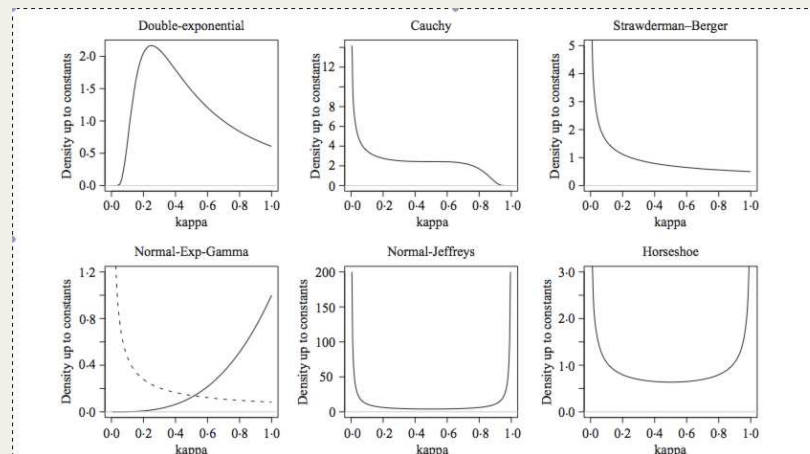
- (1) Generate $\tau \sim \text{Cauchy}^+(0, \sigma)$ (?)
- (2) Generate $\sqrt{\psi_1}, \dots, \sqrt{\psi_p}$ iid from $\text{Cauchy}^+(0, \tau)$.
- (3) Generate independent $\theta_i \sim N(0, \psi_i)$.

Horseshoe prior

Horseshoe prior:

- (1) Generate $\tau \sim \text{Cauchy}^+(0, \sigma)$ (?)
- (2) Generate $\sqrt{\psi_1}, \dots, \sqrt{\psi_p}$ iid from $\text{Cauchy}^+(0, \tau)$.
- (3) Generate independent $\theta_i \sim N(0, \psi_i)$.

MOTIVATION: if $\theta \sim N(0, \psi)$ and $Y | \theta \sim N(\theta, 1)$, then $\theta | Y \sim N((1 - \kappa)Y, 1 - \kappa)$ for $\kappa = 1/(1 + \psi)$. This suggests a prior for κ that concentrates near 0 or 1.

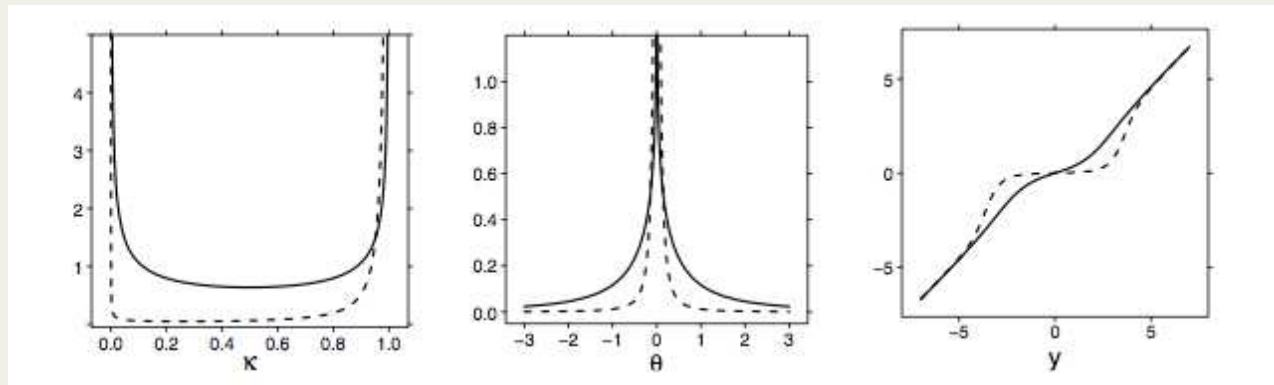


Horseshoe prior

Horseshoe prior:

- (1) Generate $\tau \sim \text{Cauchy}^+(0, \sigma)$ (?)
- (2) Generate $\sqrt{\psi_1}, \dots, \sqrt{\psi_p}$ iid from $\text{Cauchy}^+(0, \tau)$.
- (3) Generate independent $\theta_i \sim N(0, \psi_i)$.

MOTIVATION: if $\theta \sim N(0, \psi)$ and $Y | \theta \sim N(\theta, 1)$, then $\theta | Y \sim N((1 - \kappa)Y, 1 - \kappa)$ for $\kappa = 1/(1 + \psi)$. This suggests a prior for κ that concentrates near 0 or 1.

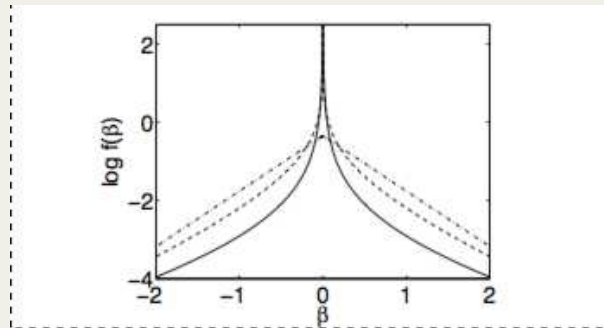


prior shrinkage factor

prior of θ_i

posterior mean of θ_i as function of Y_i

Other sparsity priors



- *Bayesian LASSO*: $\theta_1, \dots, \theta_p$ iid from a mixture of Laplace (λ) distributions over $\lambda \sim \sqrt{\Gamma(a, b)}$.
- *Bayesian bridge*: Same but with Laplace replaced with a density $\propto e^{-|\lambda y|^\alpha}$.
- *Normal-Gamma*: $\theta_1, \dots, \theta_p$ iid from a Gamma scale mixture of Gaussians. *Correlated multivariate normal-Gamma*: $\theta = C\phi$ for a $p \times k$ -matrix C and ϕ with independent normal-Gamma ($a_i, 1/2$) coordinates.
- *Horseshoe*.
- *Horseshoe+*.
- *Normal spike*.
- *Scalar multiple of Dirichlet*.
- *Nonparametric Dirichlet*.
- ...

[Park & Casella 08, Polson & Scott, Griffin & Brown 10, 12, Carvalho & Polson & Scott, 10, George & Rockova 13, Bhattacharya et al.

LASSO is not Bayesian

$$\hat{\theta}_{\text{LASSO}} = \underset{\theta}{\operatorname{argmin}} \left[\|Y - X\theta\|^2 + \lambda_n \sum_{i=1}^p |\theta_i| \right].$$

The LASSO is the *posterior mode* for prior $\theta_i \stackrel{\text{iid}}{\sim} \text{Laplace}(\lambda_n)$,
but the full posterior distribution is useless.

Trouble:

λ must be large to shrink θ_i to 0, but small to model nonzero θ_i .

LASSO is not Bayesian

$$\hat{\theta}_{\text{LASSO}} = \underset{\theta}{\operatorname{argmin}} \left[\|Y - X\theta\|^2 + \lambda_n \sum_{i=1}^p |\theta_i| \right].$$

The LASSO is the *posterior mode* for prior $\theta_i \stackrel{\text{iid}}{\sim} \text{Laplace}(\lambda_n)$,
but the full posterior distribution is useless.

Trouble:

λ must be large to shrink θ_i to 0, but small to model nonzero θ_i .

THEOREM If $\sqrt{n}/\lambda_n \rightarrow \infty$ then

$$\mathbb{E}_0 \Pi_n(\|\theta\|_2 \lesssim \sqrt{n}/\lambda_n | Y^n) \rightarrow 0.$$

LASSO choice $\lambda_n = \sqrt{2 \log n}$ gives almost no “Bayesian shrinkage”.

Frequentist Bayes

Frequentist Bayes

Assume data Y^n follows a **given parameter** θ_0 and consider the posterior $\Pi(\theta \in \cdot | Y^n)$ as a *random measure* on the parameter set.

We like $\Pi(\theta \in \cdot | Y^n)$:

- to put “most” of its mass near θ_0 for “most” Y^n .
- to have a spread that expresses “remaining uncertainty”.
- to select the model defined by the nonzero parameters of θ_0 .

We evaluate this by probabilities or expectations, given θ_0 .

Benchmarks for recovery — sequence model

$$Y^n \sim N_n(\theta, I), \text{ for } \theta = (\theta_1, \dots, \theta_n) \in \mathbb{R}^n.$$

$$\|\theta\|_0 = \#\{1 \leq i \leq n: \theta_i \neq 0\},$$

$$\|\theta\|_2^2 = \sum_{i=1}^n |\theta_i|^2.$$

Frequentist benchmark: **minimax rate** relative to $\|\cdot\|_2$ over:

- **black bodies** $\{\theta: \|\theta\|_0 \leq s_n\}$:

$$\sqrt{s_n \log(n/s_n)}.$$

Benchmarks for recovery — sequence model

$$Y^n \sim N_n(\theta, I), \text{ for } \theta = (\theta_1, \dots, \theta_n) \in \mathbb{R}^n.$$

$$\|\theta\|_0 = \#\{1 \leq i \leq n: \theta_i \neq 0\},$$

$$\|\theta\|_q^q = \sum_{i=1}^n |\theta_i|^q, \quad 0 < q \leq 2.$$

Frequentist benchmarks: **minimax rate** relative to $\|\cdot\|_q$ over:

- **black bodies** $\{\theta: \|\theta\|_0 \leq s_n\}$:

$$s_n^{1/q} \sqrt{\log(n/s_n)}.$$

Benchmarks for recovery — sequence model

$$Y^n \sim N_n(\theta, I), \text{ for } \theta = (\theta_1, \dots, \theta_n) \in \mathbb{R}^n.$$

$$\|\theta\|_0 = \#\{1 \leq i \leq n: \theta_i \neq 0\},$$

$$\|\theta\|_q^q = \sum_{i=1}^n |\theta_i|^q, \quad 0 < q \leq 2.$$

Frequentist benchmarks: **minimax rate** relative to $\|\cdot\|_q$ over:

- **black bodies** $\{\theta: \|\theta\|_0 \leq s_n\}$:

$$s_n^{1/q} \sqrt{\log(n/s_n)}.$$

- **weak ℓ_r -balls** $m_r[s_n] := \{\theta: \max_i i |\theta_{[i]}|^r \leq n(s_n/n)^r\}$:

$$n^{1/q} (s_n/n)^{r/q} \sqrt{\log(n/s_n)}^{1-r/q}.$$

Model Selection Prior

Model selection prior

$$Y^n \sim N_n(\theta, I), \text{ for } \theta = (\theta_1, \dots, \theta_n) \in \mathbb{R}^n.$$

Prior Π_n on $\theta \in \mathbb{R}^n$:

- (1) Choose s from prior π_n on $\{0, 1, 2, \dots, n\}$.
- (2) Choose $S \subset \{0, 1, \dots, n\}$ of size $|S| = s$ at random.
- (3) Choose $\theta_S = (\theta_i: i \in S)$ from density g_S on \mathbb{R}^S and set $\theta_{S^c} = 0$.

Model selection prior

$$Y^n \sim N_n(\theta, I), \text{ for } \theta = (\theta_1, \dots, \theta_n) \in \mathbb{R}^n.$$

Prior Π_n on $\theta \in \mathbb{R}^n$:

- (1) Choose s from prior π_n on $\{0, 1, 2, \dots, n\}$.
- (2) Choose $S \subset \{0, 1, \dots, n\}$ of size $|S| = s$ at random.
- (3) Choose $\theta_S = (\theta_i: i \in S)$ from density g_S on \mathbb{R}^S and set $\theta_{S^c} = 0$.

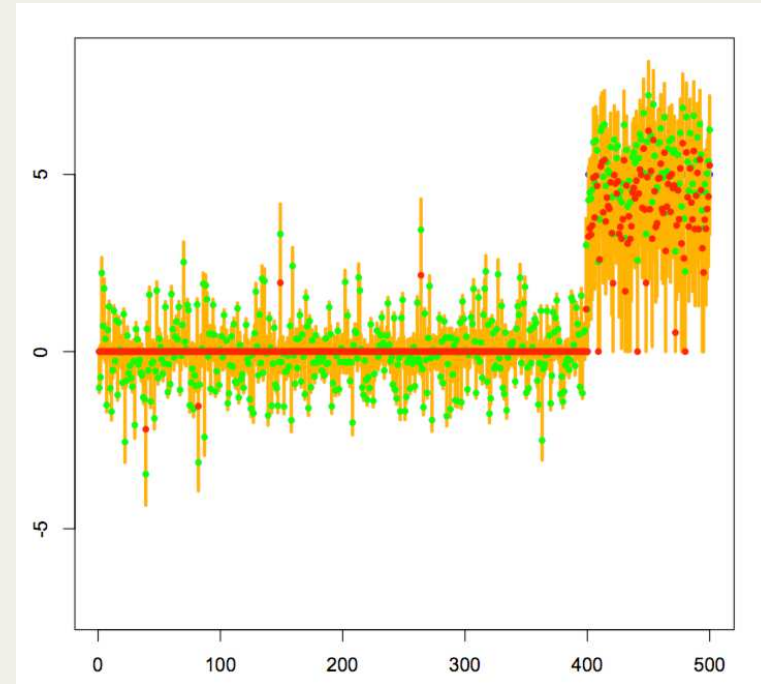
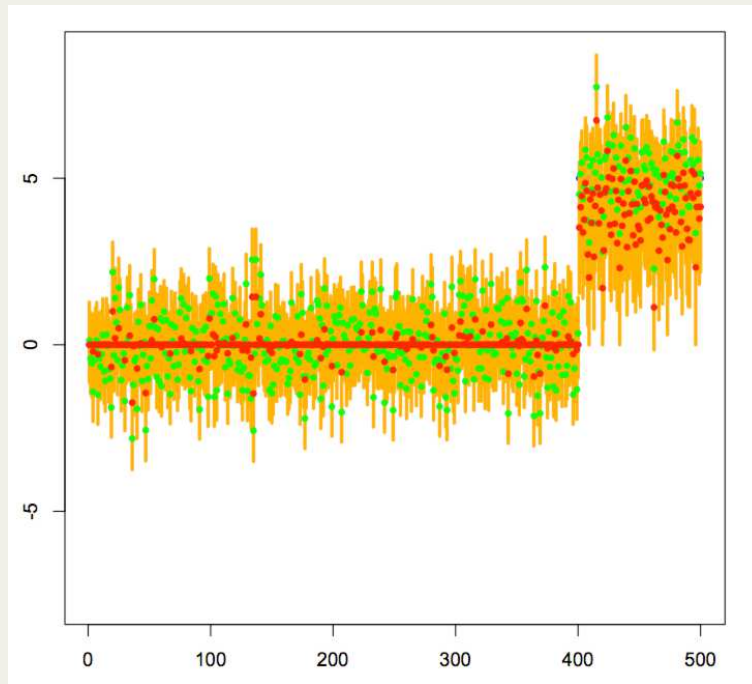
Assume

- $\pi_n(s) \leq c \pi_n(s-1)$ for some $c < 1$, and every (large) s .
- g_S is product of densities e^h for uniformly Lipschitz $h: \mathbb{R} \rightarrow \mathbb{R}$ and with finite second moment.
- $s_n, n \rightarrow \infty, s_n/n \rightarrow 0$. [*true number of nonzero parameters.*]

EXAMPLES:

- *complexity prior*: $\pi_n(s) \propto e^{-as \log(bn/s)}$.
- *spike and slab*: $\theta_i \stackrel{\text{iid}}{\sim} \tau \delta_0 + (1 - \tau)G$ with $\tau \sim B(1, n + 1)$.

Numbers



Single data with $\theta_0 = (0, \dots, 0, 5, \dots, 5)$ and $n = 500$ and $\|\theta_0\|_0 = 100$.

Red dots: marginal posterior medians

Orange: marginal credible intervals

Green dots: data points.

g standard Laplace density.

$$\pi_n(k) \propto \binom{2n-k}{n}^\kappa \text{ for } \kappa_1 = 0.1 \text{ (left) and } \kappa_1 = 1 \text{ (right).}$$

Dimensionality of posterior distribution

THEOREM (*black body*)

There exists M such that

$$\sup_{\|\theta_0\|_0 \leq s_n} \mathbb{E}_{\theta_0} \Pi_n(\theta: \|\theta\|_0 \geq Ms_n | Y^n) \rightarrow 0.$$

Outside the space in which θ_0 lives, the posterior is concentrated in low-dimensional subspaces along the coordinate axes.

Recovery

THEOREM (*black body*)

For every $0 < q \leq 2$ and large M ,

$$\sup_{\|\theta_0\|_0 \leq s_n} \mathbb{E}_{\theta_0} \Pi_n(\theta: \|\theta - \theta_0\|_q > Mr_n s_n^{1/q-1/2} | Y^n) \rightarrow 0,$$

for $r_n^2 = s_n \log(n/s_n) \vee \log(1/\pi_n(s_n))$.

If $\pi_n(s_n) \geq e^{-as_n \log(n/s_n)}$ minimax rate is attained.

Selection

$$S_\theta := \{1 \leq i \leq n: \theta_i \neq 0\}.$$

THEOREM (*No supersets*)

$$\sup_{\|\theta_0\|_0 \leq s_n} \mathbb{E}_{\theta_0} \Pi_n(\theta: S_\theta \supset S_{\theta_0}, S_\theta \neq S_{\theta_0} | Y^n) \rightarrow 0.$$

THEOREM (*Finds big signals*)

$$\inf_{\|\theta_0\|_0 \leq s_n} \mathbb{E}_{\theta_0} \Pi_n(\theta: S_\theta \supset \{i: |\theta_{0,i}| \gtrsim \sqrt{\log n}\} | Y^n) \rightarrow 1.$$

Selection

$$S_\theta := \{1 \leq i \leq n: \theta_i \neq 0\}.$$

THEOREM (*No supersets*)

$$\sup_{\|\theta_0\|_0 \leq s_n} \mathbb{E}_{\theta_0} \Pi_n(\theta: S_\theta \supset S_{\theta_0}, S_\theta \neq S_{\theta_0} | Y^n) \rightarrow 0.$$

THEOREM (*Finds big signals*)

$$\inf_{\|\theta_0\|_0 \leq s_n} \mathbb{E}_{\theta_0} \Pi_n(\theta: S_\theta \supset \{i: |\theta_{0,i}| \gtrsim \sqrt{\log n}\} | Y^n) \rightarrow 1.$$

Corollary: if *all* nonzero $|\theta_{0,i}|$ are suitably big, then posterior probability of true model S_{θ_0} tends to 1.

Numbers: mean square errors

p_n	25			50			100			
	A	3	4	5	3	4	5	3	4	5
PM1		111	96	94	176	165	154	267	302	307
PM2		106	92	82	169	165	152	269	280	274
EBM		103	96	93	166	177	174	271	312	319
PMed1		129	83	73	205	149	130	255	279	283
PMed2		125	86	68	187	148	129	273	254	245
EBMed		110	81	72	162	148	142	255	294	300
HT		175	142	70	339	284	135	676	564	252
HTO		136	92	84	206	159	139	306	261	245

Average $\|\hat{\theta} - \theta\|^2$ over 100 data experiments.

$n = 500; \theta_0 = (0, \dots, 0, A, \dots, A).$

PM1, PM2: posterior means for priors $\pi_n(k) \propto e^{-k \log(3n/k)/10}, \binom{2n-k}{n}^{0.1}$.

PMed1, PMed2 marginal posterior medians for the same priors

EBM, EBMed: empirical Bayes mean, median for Laplace prior (Johnstone et al.)

HT, HTO: thresholding at $\sqrt{2 \log n}, \sqrt{2 \log(n/\|\theta_0\|_0)}$.

Short Summary: Bayesian reconstruction is neither better nor worse.

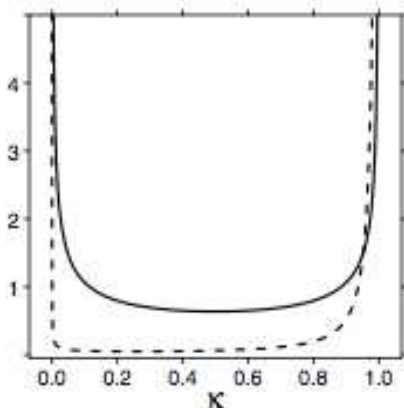
Horseshoe Prior

Horseshoe prior

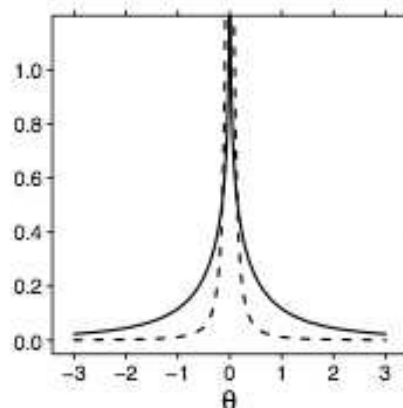
$$Y^n \sim N_n(\theta, I), \text{ for } \theta = (\theta_1, \dots, \theta_n) \in \mathbb{R}^n.$$

Prior Π_n on \mathbb{R}^n :

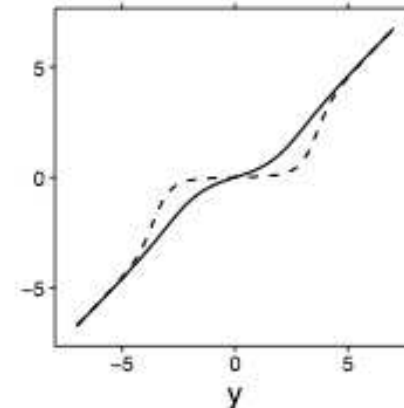
- (1) Choose “sparsity level” $\hat{\tau}$.
- (2) Generate $\sqrt{\psi_1}, \dots, \sqrt{\psi_n}$ iid from $\text{Cauchy}^+(0, \hat{\tau})$.
- (3) Generate independent $\theta_i \sim N(0, \psi_i)$.



prior shrinkage factor



prior of θ_i



posterior mean of θ_i as function of Y_i

Recovery — prechosen τ

THEOREM (*black body*)

If $(s_n/n)^c \leq \tau_n \leq C(s_n/n)\sqrt{\log(n/s_n)}$ for some $c, C > 0$, then for every $M_n \rightarrow \infty$,

$$\sup_{\|\theta_0\|_0 \leq s_n} \mathbb{E}_{\theta_0} \Pi_n(\theta: \|\theta - \theta_0\|_2 > M_n s_n \log(n/s_n) | Y^n, \tau_n) \rightarrow 0.$$

Minimax rate $s_n \log(n/s_n)$ is attained,
 τ can be interpreted as sparsity level.

Credible balls — prechosen τ

For $\hat{\theta}(\tau) = \mathbb{E}(\theta | Y^n, \tau)$ the **posterior mean**, set

$$\hat{C}_n(L, \tau) = \left\{ \theta : \|\theta - \hat{\theta}(\tau)\|_2 \leq L\hat{r}(\tau) \right\},$$

with $\hat{r}(\tau)$ satisfying $\Pi(\theta : \|\theta - \hat{\theta}(\tau)\|_2 \leq \hat{r}(\tau) | Y^n, \tau) = 0.95$.

THEOREM

If $\tau_n \rightarrow 0$ such that $\tau_n \geq (s_n/n) \sqrt{\log(n/s_n)}$, then for large enough $L > 0$

$$\inf_{\|\theta_0\|_0 \leq s_n} P_{\theta_0}(\theta_0 \in \hat{C}_n(L, \tau_n)) \geq 0.95.$$

Coverage provided **shrinkage is not too big.**

Credible intervals — prechosen τ

For $\hat{\theta}_i(\tau) = \mathbb{E}(\theta_i | Y_i, \tau)$ the posterior mean of θ_i , set

$$\hat{C}_{ni}(L, \tau) = \left\{ \theta_i : |\theta_i - \hat{\theta}_i(\tau)| \leq L \hat{r}_i(\tau) \right\},$$

with $\hat{r}_i(\tau)$ satisfying $\Pi(\theta_i : |\theta_i - \hat{\theta}_i(\tau)| \leq \hat{r}_i(\tau) | Y_i, \tau) = 0.95$.

$$\mathbf{S} := \{1 \leq i \leq n : |\theta_{0,i}| \leq \tau\},$$

$$\mathbf{M} := \{1 \leq i \leq n : \tau \ll |\theta_{0,i}| \leq 0.99 \sqrt{2 \log(1/\tau)}\},$$

$$\mathbf{L} := \{1 \leq i \leq n : 1.01 \sqrt{2 \log(1/\tau)} \leq |\theta_{0,i}|\}.$$

Credible intervals — prechosen τ

For $\hat{\theta}_i(\tau) = \mathbb{E}(\theta_i | Y_i, \tau)$ the posterior mean of θ_i , set

$$\hat{C}_{ni}(L, \tau) = \left\{ \theta_i : |\theta_i - \hat{\theta}_i(\tau)| \leq L \hat{r}_i(\tau) \right\},$$

with $\hat{r}_i(\tau)$ satisfying $\Pi(\theta_i : |\theta_i - \hat{\theta}_i(\tau)| \leq \hat{r}_i(\tau) | Y_i, \tau) = 0.95$.

$$\mathbf{S} := \{1 \leq i \leq n : |\theta_{0,i}| \leq \tau\},$$

$$\mathbf{M} := \{1 \leq i \leq n : \tau \ll |\theta_{0,i}| \leq 0.99 \sqrt{2 \log(1/\tau)}\},$$

$$\mathbf{L} := \{1 \leq i \leq n : 1.01 \sqrt{2 \log(1/\tau)} \leq |\theta_{0,i}|\}.$$

THEOREM For $\tau \rightarrow 0$ and any $\gamma > 0$,

$$P_{\theta_0} \left(\frac{1}{\#\mathbf{S}} \#\{i \in \mathbf{S} : \theta_{0,i} \in \hat{C}_{ni}(L_S, \tau)\} \geq 1 - \gamma \right) \rightarrow 1,$$

$$P_{\theta_0}(\theta_{0,i} \notin \hat{C}_{ni}(L, \tau)) \rightarrow 1, \quad \text{for any } L > 0 \text{ and } i \in \mathbf{M},$$

$$P_{\theta_0} \left(\frac{1}{\#\mathbf{L}} \#\{i \in \mathbf{L} : \theta_{0,i} \in \hat{C}_{ni}(L_L, \tau)\} \geq 1 - \gamma \right) \rightarrow 1.$$

Credible intervals — prechosen τ

For $\hat{\theta}_i(\tau) = \mathbb{E}(\theta_i | Y_i, \tau)$ the posterior mean of θ_i , set

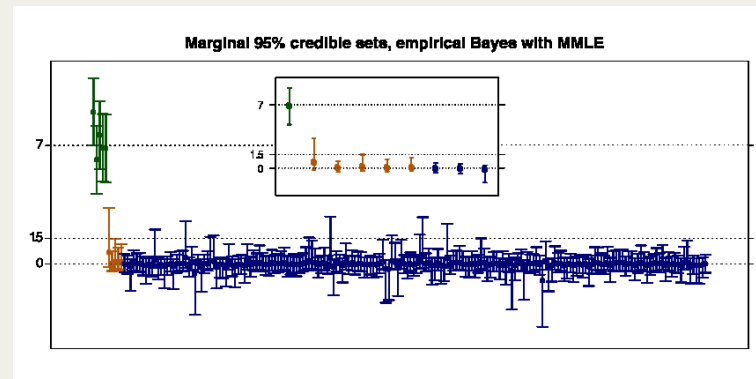
$$\hat{C}_{ni}(L, \tau) = \left\{ \theta_i : |\theta_i - \hat{\theta}_i(\tau)| \leq L \hat{r}_i(\tau) \right\},$$

with $\hat{r}_i(\tau)$ satisfying $\Pi(\theta_i : |\theta_i - \hat{\theta}_i(\tau)| \leq \hat{r}_i(\tau) | Y_i, \tau) = 0.95$.

$$\mathbf{S} := \{1 \leq i \leq n : |\theta_{0,i}| \leq \tau\},$$

$$\mathbf{M} := \{1 \leq i \leq n : \tau \ll |\theta_{0,i}| \leq 0.99 \sqrt{2 \log(1/\tau)}\},$$

$$\mathbf{L} := \{1 \leq i \leq n : 1.01 \sqrt{2 \log(1/\tau)} \leq |\theta_{0,i}|\}.$$



marginal credible intervals for a single Y^n with $n = 200$ and $s_n = 10$.

$\theta_1 = \dots = \theta_5 = 7$ (green), $\theta_6 = \dots = \theta_{10} = 1.5$ (orange). Insert: credible sets 5 to 13.

Estimating τ

Ad-hoc:

$$\hat{\tau}_n = \frac{\#\{|Y_i^n| \geq \sqrt{2 \log n}\}}{1.1n}.$$

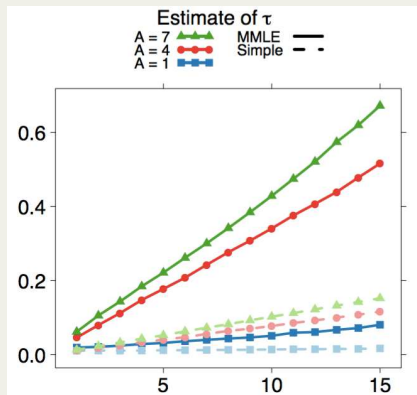
Empirical Bayes: For g_τ the prior of θ_i ,

$$\hat{\tau}_n = \operatorname{argmax}_{\tau \in [1/n, 1]} \prod_{i=1}^n \int \phi(y_i - \theta) g_\tau(\theta) d\theta.$$

Full Bayes: τ set by a “hyper prior” (supported on $[1/n, 1]$).

Numbers

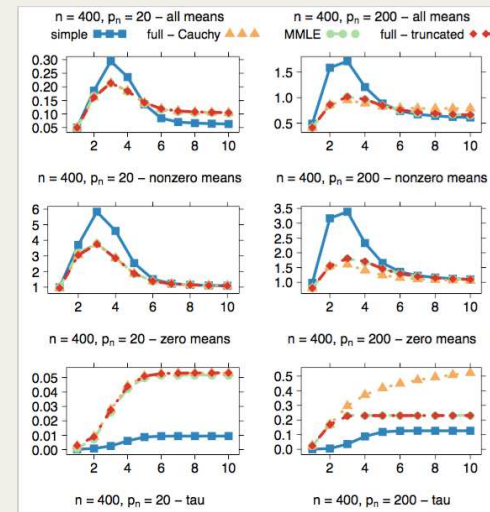
estimating τ



$s_n \rightarrow$

$n = 100$, s_n coordinates from $N(0, 1/4)$,
 $n - s_n$ coordinates from $N(A, 1)$.

MSE of posterior mean
 as function of nonzero parameter



“ $p_n = s_n$ ”

Empirical Bayes and Full Bayes are similar and outperform ad-hoc estimator

Recovery

THEOREM (*black body*)

For the likelihood based empirical Bayes $\hat{\tau}_n$,

$$\sup_{\|\theta_0\|_0 \leq s_n} \mathbb{E}_{\theta_0} \left[\Pi \left(\theta: \|\theta_0 - \theta\|_2 \geq M_n \sqrt{s_n \log n} \mid Y^n, \tau \right) \Big|_{\tau = \hat{\tau}_n} \right] \rightarrow 0.$$

For the full Bayes choice of τ (under mild conditions on hyper prior),

$$\sup_{\|\theta_0\|_0 \leq s_n} \mathbb{E}_{\theta_0} \Pi \left(\theta: \|\theta_0 - \theta\|_2 \geq M_n \sqrt{s_n \log n} \mid Y^n \right) \rightarrow 0.$$

Credible intervals

For $\hat{\theta}_i(\tau) = \mathbb{E}(\theta_i | Y_i, \tau)$ the **posterior mean** of θ_i

$$\hat{C}_{ni}(L, \tau) = \left\{ \theta_i : |\theta_i - \hat{\theta}_i(\tau)| \leq L r_i(\tau) \right\},$$

for $r_i(\tau)$ satisfying $\Pi(\theta_i : |\theta_i - \hat{\theta}_i(\tau)| \leq \hat{r}_i(\tau) | Y_i, \tau) = 0.95$.

$$\mathbf{S}_a := \{1 \leq i \leq n : |\theta_{0,i}| \leq 1/n\},$$

$$\mathbf{M}_a := \{1 \leq i \leq n : (s_n/n) \sqrt{\log(n/s_n)} \ll |\theta_{0,i}| \leq 0.99 \sqrt{2 \log(n/s_n)}\}.$$

$$\mathbf{L}_a := \{1 \leq i \leq n : 1.01 \sqrt{2 \log n} \leq |\theta_{0,i}|\}.$$

THEOREM For any $\gamma > 0$ and $\|\theta_0\|_0 \leq s_n$,

$$P_{\theta_0} \left(\frac{1}{\#\mathbf{S}_a} |\#\{i \in \mathbf{S}_a : \theta_{0,i} \in \hat{C}_{ni}(L_{S,\gamma}, \hat{\tau}_n)\}| \geq 1 - \gamma \right) \rightarrow 1,$$

$$P_{\theta_0}(\theta_{0,i} \notin \hat{C}_{ni}(L, \hat{\tau}_n)) \rightarrow 1, \quad \text{for any } L > 0 \text{ and } i \in \mathbf{M}_a,$$

$$P_{\theta_0} \left(\frac{1}{\#\mathbf{L}_a} |\#\{i \in \mathbf{L}_a : \theta_{0,i} \in \hat{C}_{ni}(L_{L,\gamma}, \hat{\tau}_n)\}|/l \geq 1 - \gamma \right) \rightarrow 1.$$

If $s_n \gtrsim \log n$, the analogous is true for the hierarchical Bayes intervals.

Credible sets — impossibility of adaptation

General principle: the size of an **honest confidence set** is determined by the **biggest model**. [Cai and Low, Juditzky & Lambert-Lacroix, 2003; Robins & van der Vaart, 2006]

THEOREM [Li, 1987]

If $C_n(Y)$ satisfies $P_{\theta_0}(C_n(Y) \ni \theta_0) \geq 0.95$ for all $\theta_0 \in \mathbb{R}^n$, then $\text{diam}(C_n(Y)) \gtrsim n^{-1/4}$, for some θ_0 .

THEOREM [Nickl, van de Geer, 2013]

If $s_{1,n} \ll s_{2,n}$ and $\text{diam}(C_n(Y))$ is of order $\left((s_{i,n}/n) \log(n/s_{i,n})\right)^{1/2}$, uniformly in $\|\theta_0\|_0 \leq s_{i,n}$ for $i = 1, 2$, then $C_n(Y)$ cannot have uniform coverage over $\{\theta_0: \|\theta_0\|_0 \leq s_{2,n}\}$.

Credible sets — impossibility of adaptation

General principle: the size of an **honest confidence set** is determined by the **biggest model**. [Cai and Low, Juditzky & Lambert-Lacroix, 2003; Robins & van der Vaart, 2006]

THEOREM [Li, 1987]

If $C_n(Y)$ satisfies $P_{\theta_0}(C_n(Y) \ni \theta_0) \geq 0.95$ for all $\theta_0 \in \mathbb{R}^n$, then $\text{diam}(C_n(Y)) \gtrsim n^{-1/4}$, for some θ_0 .

THEOREM [Nickl, van de Geer, 2013]

If $s_{1,n} \ll s_{2,n}$ and $\text{diam}(C_n(Y))$ is of order $\left((s_{i,n}/n) \log(n/s_{i,n})\right)^{1/2}$, uniformly in $\|\theta_0\|_0 \leq s_{i,n}$ for $i = 1, 2$, then $C_n(Y)$ cannot have uniform coverage over $\{\theta_0: \|\theta_0\|_0 \leq s_{2,n}\}$.

Since the Bayesian procedure adapts to sparsity, its credible sets *cannot* be honest confidence sets.

Credible sets — impossibility of adaptation — restricting the parameter

Coverage only when θ_0 does not cause too much shrinkage.

DEFINITION [self-similarity]

For $s = \|\theta_0\|_0$ at least $0.02s$ coordinates of θ_0 satisfy

$$|\theta_{0,i}| \geq 1.01 \sqrt{2 \log(n/s)}.$$

Credible sets — impossibility of adaptation — restricting the parameter

Coverage only when θ_0 does not cause too much shrinkage.

DEFINITION [self-similarity]

For $s = \|\theta_0\|_0$ at least $0.02s$ coordinates of θ_0 satisfy

$$|\theta_{0,i}| \geq 1.01\sqrt{2\log(n/s)}.$$

DEFINITION [excessive-bias restriction, Belitser & Nurushev, 2015]

$\|\theta\|_0 \leq s$ and $\exists \tilde{s}$ with $\tilde{s} \asymp \#\{i: |\theta_{0,i}| \geq 1.01\sqrt{2\log(n/\tilde{s})}\}$ and

$$\sum_{i: |\theta_{0,i}| \leq 1.01\sqrt{2\log(n/\tilde{s})}} \theta_{0,i}^2 \lesssim \tilde{s} \log(n/\tilde{s}).$$

Excessive-bias restriction implies self-similarity.
Self-similarity allows to tighten up the sets **S**, **M**, **L**.

Credible balls

Empirical Bayes: Plug-in: $\hat{C}_n(L) := \hat{C}_n(L, \hat{\tau}_n)$.

Credible balls

Empirical Bayes: Plug-in: $\hat{C}_n(L) := \hat{C}_n(L, \hat{\tau}_n)$.

Hierarchical Bayes: For $\hat{\theta} = \mathbb{E}(\theta | Y^n)$ the **posterior mean**,

$$\hat{C}_n(L) = \left\{ \theta : \|\theta - \hat{\theta}\|_2 \leq L\hat{r} \right\},$$

for \hat{r} satisfying $\Pi(\theta : \|\theta - \hat{\theta}\|_2 \leq \hat{r} | Y^n) = 0.95$.

Credible balls

Empirical Bayes: Plug-in: $\hat{C}_n(L) := \hat{C}_n(L, \hat{\tau}_n)$.

Hierarchical Bayes: For $\hat{\theta} = \mathbb{E}(\theta | Y^n)$ the **posterior mean**,

$$\hat{C}_n(L) = \left\{ \theta : \|\theta - \hat{\theta}\|_2 \leq L\hat{r} \right\},$$

for \hat{r} satisfying $\Pi(\theta : \|\theta - \hat{\theta}\|_2 \leq \hat{r} | Y^n) = 0.95$.

THEOREM

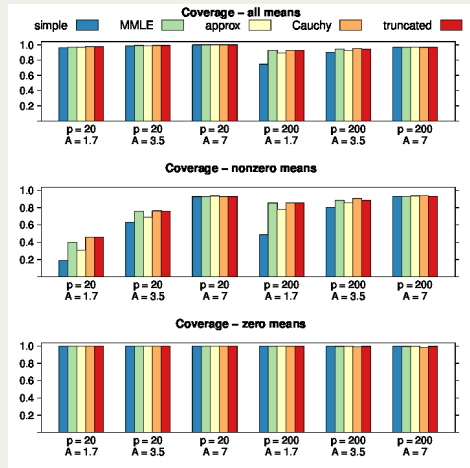
If $s_n \rightarrow 0$, for sufficiently large L ,

$$\liminf_{n \rightarrow \infty} \inf_{\theta_0 \in \text{EBR}[s_n]} P_{\theta_0} \left(\theta_0 \in \hat{C}_n(L) \right) \geq 1 - \alpha.$$

EBR[s]: vectors θ_0 that satisfy excessive bias restriction.

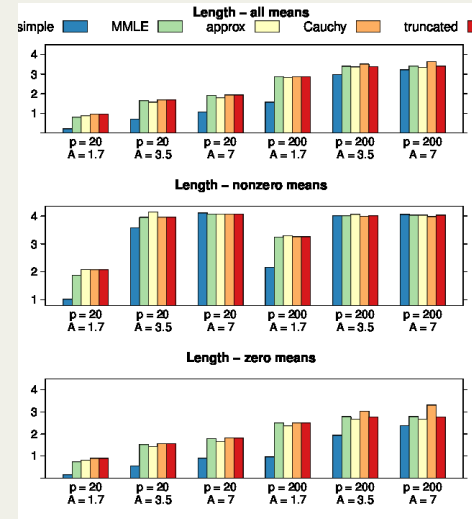
Numbers

coverage



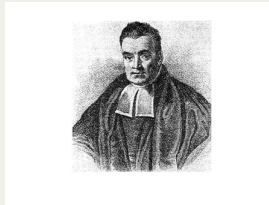
$n = 400$. s_n ("= p ") nonzero means from $\mathcal{N}(A, 1)$.

average interval length



$n = 400$. s_n ("= p ") nonzero means from $\mathcal{N}(A, 1)$.

Conclusions



Bayesian sparse estimation gives excellent recovery.

For valid simultaneous credible sets roughly need a fraction of nonzero parameters above the “universal threshold”.

The danger of failing uncertainty quantification is **not** finding nonzero coordinates.

