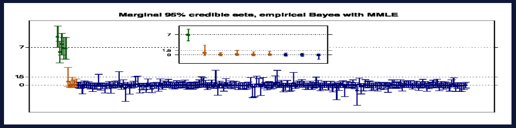
Nonparametric Bayesian Uncertainty Quantification

Lecture 2: High-dimensional Models and Sparsity

Aad van der Vaart

Universiteit Leiden, Netherlands



Hotelling Lectures, UNC, Chapel Hill, March 2017

Contents

Sparsity

Frequentist Bayes

Model Selection Prior

Horseshoe Prior

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 $Y^n | \theta \sim p(\cdot | \theta)$, 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: $Y^n | \theta \sim N_n(\theta, I)$.

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

- $\theta \in \mathbb{R}^n$ (or $\in \mathbb{R}^p$) is known to have many (almost) zero coordinates.
- n (and p) are large.

Sparsity — RNA sequencing

$Y_{i,j}$: RNA expression count of tag $i = 1, \ldots, p$ in tissue $j = 1, \ldots, n$,
x_j: covariate(s) of tissue j, *e.g.* 0 *or* 1 *for normal or cancer*.

$Y_{i,j} \sim$ (zero-inflated) *negative binomial*, with

$$\mathrm{E}Y_{i,j} = e^{\alpha_i + \beta_i x_j}, \qquad \mathrm{var}\, Y_{i,j} = \mathrm{E}Y_{i,j}\big(1 + \mathrm{E}Y_{i,j} e^{-\phi_i}\big).$$

Many tags i are thought to be unrelated to x_j: $\beta_i = 0$ for most i.

[Smyth & Robinson et al., van der Wiel, Leday, van Wieringen, & vdV et al., 12]

Model selection prior

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

(1) Choose s from prior π on $\{0, 1, 2, \ldots, p\}$.

(2) Choose $S \subset \{0, 1, \ldots, p\}$ 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 $\{0, 1, 2, \ldots, p\}$.
(2) Choose $S \subset \{0, 1, \ldots, p\}$ 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.

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

Model selection prior

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

(1) Choose s from prior π on $\{0, 1, 2, \ldots, p\}$.
(2) Choose $S \subset \{0, 1, \ldots, p\}$ of size $|S| = s$ at random.
(3) Choose $\theta_S = (\theta_i \colon i \in S)$ 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 $\theta_1, \ldots, \theta_p$ 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 $g_S = \otimes_{i\in S} g$.

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

Horseshoe prior

Horseshoe prior:

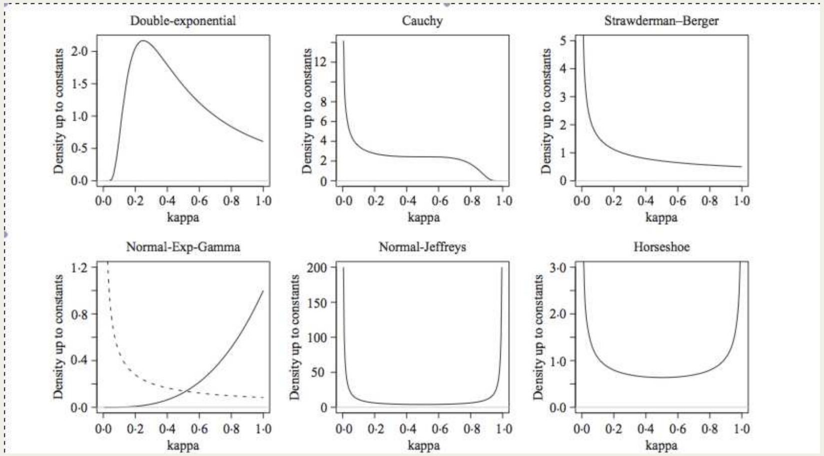
(1) Generate $\tau \sim \mathrm{Cauchy}^+(0, \sigma)$ (?)
(2) Generate $\sqrt{\psi_1}, \ldots, \sqrt{\psi_p}$ iid from $\mathrm{Cauchy}^+(0, \tau)$.
(3) Generate independent $\theta_i \sim N(0, \psi_i)$.

Horseshoe prior

Horseshoe prior:

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

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



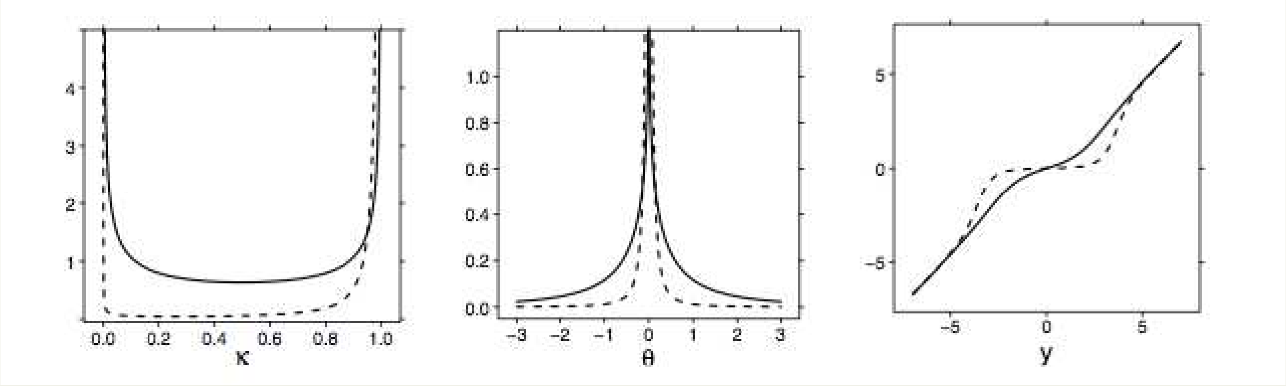
[Carvalho & Polson & Scott, 10.]

Horseshoe prior

Horseshoe prior:

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

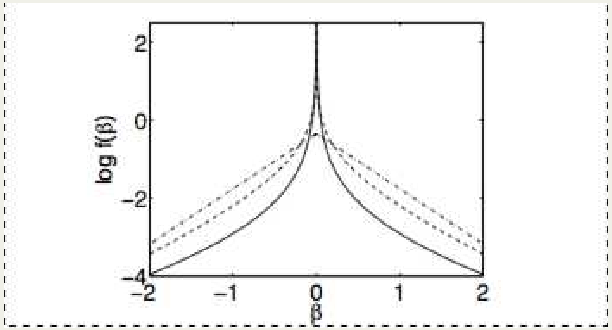
MOTIVATION: if $\theta \sim N(0, \psi)$ and $Y \mid \theta \sim N(\theta, 1)$,
then $\theta \mid Y, \psi \sim N\big((1-\kappa)Y, 1-\kappa\big)$ for $\kappa = 1/(1+\psi)$.
This suggests a prior for κ that concentrates near 0 or 1.



prior shrinkage factor   prior of θ_i posterior mean of θ_i as function of Y_i

[Carvalho & Polson & Scott, 10.]

Other sparsity priors



- *Bayesian LASSO*: $\theta_1, \ldots, \theta_p$ 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*: $\theta_1, \ldots, \theta_p$ 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. 12,...]

LASSO is not Bayesian

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

The LASSO is the *posterior mode* for prior $\theta_i \overset{\text{iid}}{\sim} \mathrm{Laplace}(\lambda_n)$,
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

$$\hat{\theta}_{\mathsf{LASSO}} = \operatorname*{argmin}_{\theta} \Big[\|Y^n - X\theta\|^2 + \lambda_n \sum_{i=1}^{p} |\theta_i| \Big].$$

The LASSO is the *posterior mode* for prior $\theta_i \overset{\text{iid}}{\sim} \mathrm{Laplace}(\lambda_n)$,
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 $\sqrt{n}/\lambda_n \to \infty$ then

$$\mathrm{E}_0 \Pi_n \big(\|\theta\|_2 \lesssim \sqrt{n}/\lambda_n \,|\, Y^n \big) \to 0.$$

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

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

$$\|\theta\|_0 = \#(1 \leq i \leq n\colon \theta_i \neq 0),$$
$$\|\theta\|_2^2 = \sum_{i=1}^n |\theta_i|^2.$$

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

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

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

[(if $s_n \to \infty$ with $s_n/n \to 0$.) Donoho & Johnstone, Golubev, Johnstone and Silverman, Abramovich et al.,. . .]

Benchmarks for recovery — sequence model

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

$$\|\theta\|_0 = \#(1 \leq i \leq n \colon \theta_i \neq 0),$$
$$\|\theta\|_q^q = \sum_{i=1}^n |\theta_i|^q, \qquad 0 < q \leq 2.$$

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

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

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

Benchmarks for recovery — sequence model

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

$$\|\theta\|_0 = \#(1 \leq i \leq n \colon \theta_i \neq 0),$$
$$\|\theta\|_q^q = \sum_{i=1}^n |\theta_i|^q, \qquad 0 < q \leq 2.$$

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

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

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

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

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

[(if $s_n \to \infty$ with $s_n/n \to 0$.) Donoho & Johnstone, Golubev, Johnstone and Silverman, Abramovich et al.,. . .]

Model Selection Prior

Model selection prior

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

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

(1) Choose s from prior π_n on $\{0, 1, 2, \ldots, n\}$.
(2) Choose $S \subset \{0, 1, \ldots, n\}$ 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

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

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

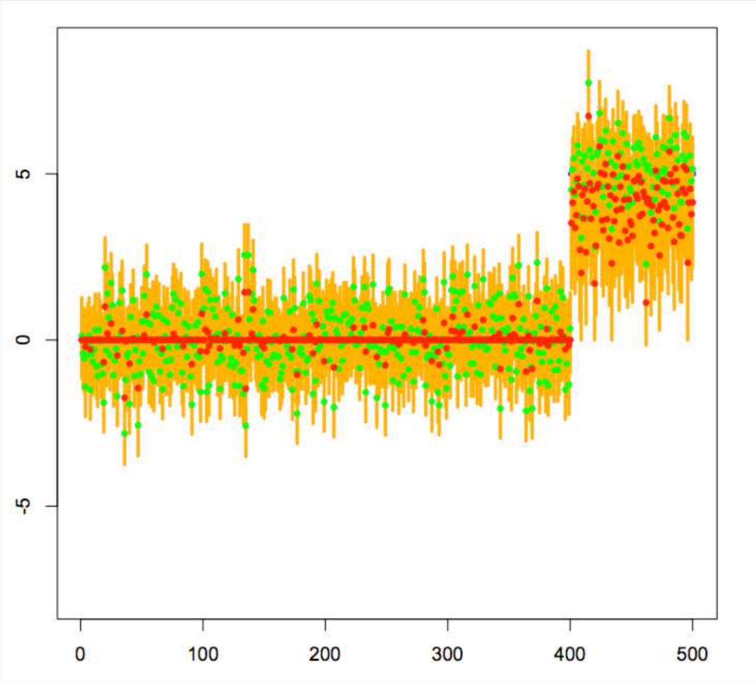
(1) Choose s from prior π_n on $\{0, 1, 2, \ldots, n\}$.
(2) Choose $S \subset \{0, 1, \ldots, n\}$ of size $|S| = s$ at random.
(3) Choose $\theta_S = (\theta_i\colon i \in S)$ from density g_S on $\mathbb{R}^S$ and set $\theta_{S^c} = 0$.

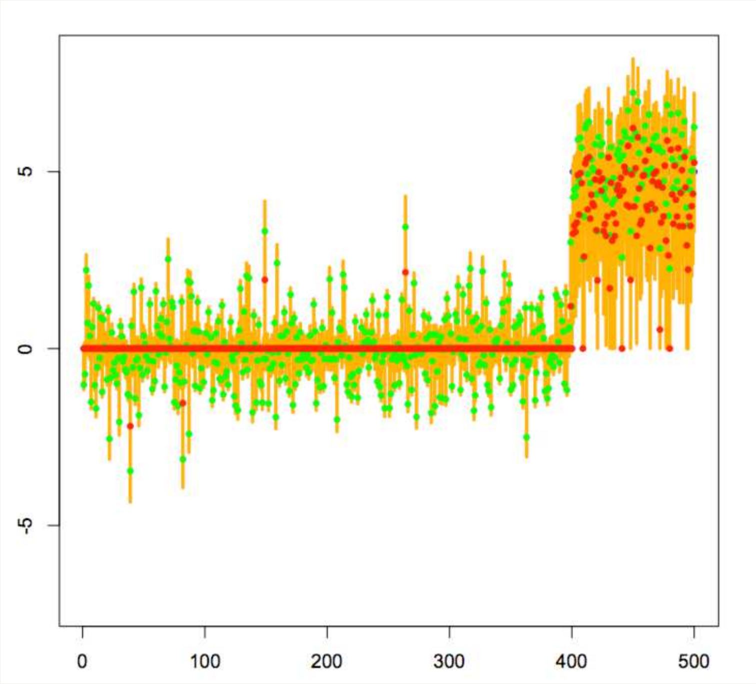
Assume

- $\pi_n(s) \leq c\,\pi_n(s-1)$ for some $c < 1$, and every (large) s.
- g_S is product of densities e^h for uniformly Lipschitz $h\colon \mathbb{R} \to \mathbb{R}$ and with finite second moment.
- $s_n\colon = \|\theta_0\|_0 \to 0$, $n \to \infty$, $s_n/n \to 0$.

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

Numbers





Single data with $\theta_0 = (0, \ldots, 0, 5, \ldots, 5)$ 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.

$\pi_n(k) \propto \binom{2n-k}{n}^{0.1}$ (left) and $\pi_n(k) \propto \binom{2n-k}{n}$ (right).

Dimensionality of posterior distribution

THEOREM *(black body)*
There exists M such that

$$\sup_{\|\theta_0\|_0 \le s_n} \mathrm{E}_{\theta_0} \Pi_n\bigl(\theta: \|\theta\|_0 \ge M s_n \,|\, Y^n\bigr) \to 0.$$

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,

$$\sup_{\|\theta_0\|_0 \leq s_n} \mathrm{E}_{\theta_0} \Pi_n\big(\theta\colon \|\theta - \theta_0\|_q > M r_n s_n^{1/q-1/2} \,|\, Y^n\big) \to 0,$$

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

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

Selection

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

THEOREM *(No supersets)*

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

Selection

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

THEOREM *(No supersets)*

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

THEOREM *(Finds big signals)*

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

Selection

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

THEOREM *(No supersets)*

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

THEOREM *(Finds big signals)*

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

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

Bernstein-von Mises theorem

THEOREM

For spike-and-Laplace(λ_n)-slab prior with $\lambda_n\sqrt{\log n}/s_n \to 0$, there are random weights $\hat{w}_S$,

$$\mathrm{E}_{\theta_0}\Big\|\Pi_n(\cdot\,|\,Y^n) - \sum_S \hat{w}_S N_{|S|}(Y_S^n, I)\otimes\delta_{S^c}\Big\| \to 0.$$

THEOREM

Given consistent model selection, mixture can be replaced by $N_{|S_0|}(Y_{S_{\theta_0}}, I)\otimes\delta_{S^c_{\theta_0}}$.

Corollary: Given consistent model selection, credible sets for individual parameters are asymptotic confidence sets.

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.**
$n = 500; \theta_0 = (0, \ldots, 0, A, \ldots, A)$.
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 method is neither better nor worse.

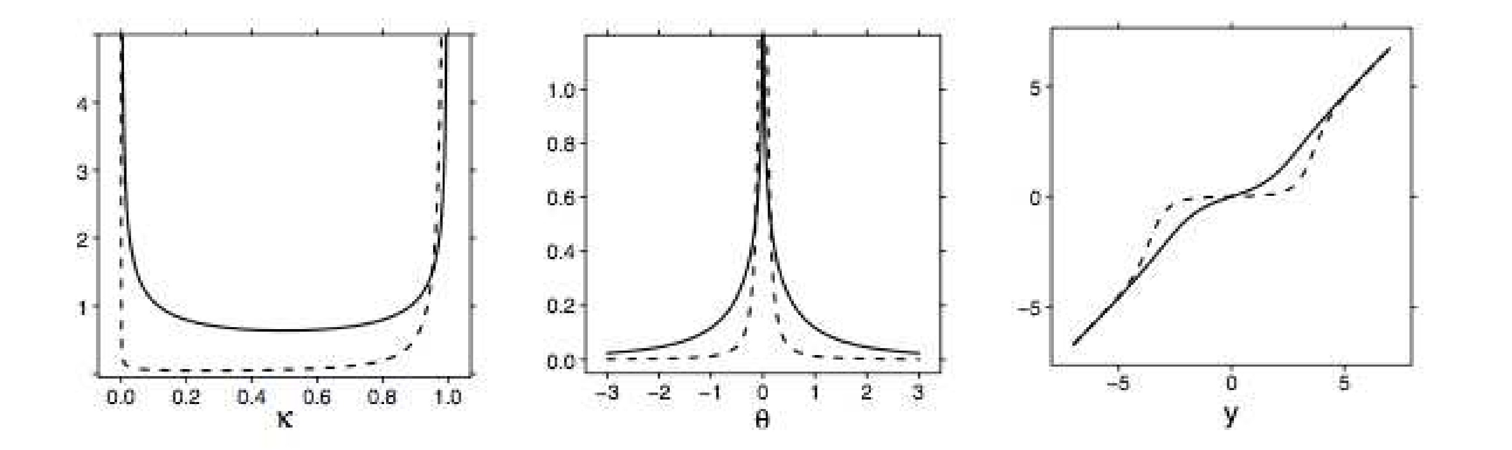
Horseshoe Prior

Horseshoe prior

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

Prior Π_n on $\mathbb{R}^n$:

(1) Choose "sparsity level" $\hat{\tau}$.
(2) Generate $\sqrt{\psi_1}, \ldots, \sqrt{\psi_n}$ iid from $\mathrm{Cauchy}^+(0, \hat{\tau})$.
(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

[Carvalho & Polson & Scott, 10.]

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 $M_n \to \infty$,

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

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

Credible balls — prechosen τ

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

$$\hat{C}_n(L, \tau) = \Big\{\theta\colon \|\theta - \hat{\theta}(\tau)\|_2 \leq L\hat{r}(\tau)\Big\},$$

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

Credible balls — prechosen τ

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

$$\hat C_n(L,\tau) = \Big\{\theta\colon \|\theta - \hat\theta(\tau)\|_2 \le L\hat r(\tau)\Big\},$$

for $\hat r(\tau)$ satisfying $\Pi\big(\theta\colon \|\theta - \hat\theta(\tau)\|_2 \le \hat r(\tau)|\, Y^n, \tau\big) = 0.95$.

THEOREM

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

$$\inf_{\|\theta_0\|_0 \le s_n} P_{\theta_0}\big(\theta_0 \in \hat C_n(L,\tau_n)\big) \ge 0.95.$$

Coverage, provided shrinkage is not *too* big.

Credible intervals — prechosen τ

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

$$\hat{C}_{ni}(L, \tau) = \Big\{\theta_i \colon \big|\theta_i - \hat{\theta}_i(\tau)\big| \leq L \hat{r}_i(\tau)\Big\},$$

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

Credible intervals — prechosen τ

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

$$\hat C_{ni}(L,\tau) = \Big\{\theta_i \colon \big|\theta_i - \hat\theta_i(\tau)\big| \le L\hat r_i(\tau)\Big\},$$

for $\hat r_i(\tau)$ satisfying $\Pi\big(\theta_i \colon |\theta_i - \hat\theta_i(\tau)| \le \hat r_i(\tau) \,|\, Y_i, \tau\big) = 0.95$.

$$\begin{aligned}
\mathbf{S} &:= \big\{1 \le i \le n \colon |\theta_{0,i}| \le \tau\big\},\\
\mathbf{M} &:= \big\{1 \le i \le n \colon \tau \ll |\theta_{0,i}| \le 0.999\sqrt{2\log(1/\tau)}\big\},\\
\mathbf{L} &:= \big\{1 \le i \le n \colon 1.001\sqrt{2\log(1/\tau)} \le |\theta_{0,i}|\big\}.
\end{aligned}$$

Credible intervals — prechosen τ

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

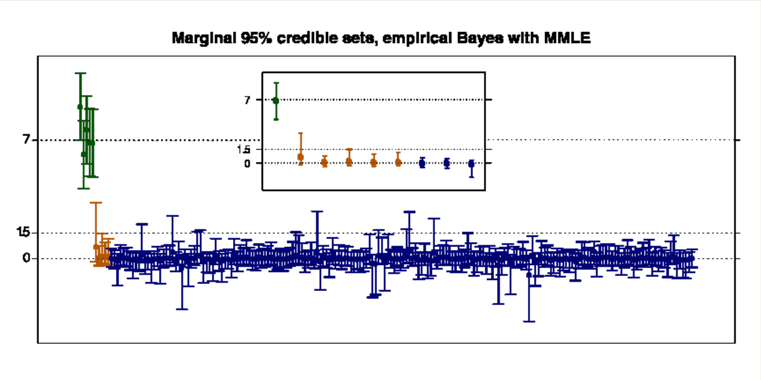
$$\hat C_{ni}(L,\tau) = \Big\{\theta_i \colon \big|\theta_i - \hat\theta_i(\tau)\big| \le L\hat r_i(\tau)\Big\},$$

for $\hat r_i(\tau)$ satisfying $\Pi\big(\theta_i\colon |\theta_i - \hat\theta_i(\tau)| \le \hat r_i(\tau) |\, Y_i, \tau\big) = 0.95$.

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

$$\mathbf{M} := \{1 \le i \le n\colon \tau \ll |\theta_{0,i}| \le 0.999\sqrt{2\log(1/\tau)}\},$$

$$\mathbf{L} := \{1 \le i \le n\colon 1.001\sqrt{2\log(1/\tau)} \le |\theta_{0,i}|\}.$$



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

$\theta_1 = \cdots = \theta_5 = 7$, $\theta_6 = \cdots = \theta_{10} = 1.5$. Insert: credible sets 5 to 13.

Credible intervals — prechosen τ

For $\hat\theta_i(\tau) = \mathrm{E}(\theta_i \mid Y_i, \tau)$ the posterior mean of θ_i, set

$$\hat C_{ni}(L,\tau) = \Bigl\{\theta_i \colon \bigl|\theta_i - \hat\theta_i(\tau)\bigr| \le L\hat r_i(\tau)\Bigr\},$$

for $\hat r_i(\tau)$ satisfying $\Pi\bigl(\theta_i \colon |\theta_i - \hat\theta_i(\tau)| \le \hat r_i(\tau) \mid Y_i, \tau\bigr) = 0.95$.

$$\begin{aligned}
\mathbf{S} &:= \bigl\{1 \le i \le n \colon |\theta_{0,i}| \le \tau\bigr\},\\
\mathbf{M} &:= \bigl\{1 \le i \le n \colon \tau \ll |\theta_{0,i}| \le 0.999\sqrt{2\log(1/\tau)}\bigr\},\\
\mathbf{L} &:= \bigl\{1 \le i \le n \colon 1.001\sqrt{2\log(1/\tau)} \le |\theta_{0,i}|\bigr\}.
\end{aligned}$$

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

$$P_{\theta_0}\Bigl(\frac{1}{\#\mathbf{S}}\#\{i \in \mathbf{S} \colon \theta_{0,i} \in \hat C_{ni}(L_S,\tau)\} \ge 1-\gamma\Bigr) \to 1,$$

$$P_{\theta_0}\bigl(\theta_{0,i} \notin \hat C_{ni}(L,\tau)\bigr) \to 1, \quad \text{for any } L > 0 \text{ and } i \in \mathbf{M},$$

$$P_{\theta_0}\Bigl(\frac{1}{\#\mathbf{L}}\#\{i \in \mathbf{L} \colon \theta_{0,i} \in \hat C_{ni}(L_L,\tau)\} \ge 1-\gamma\Bigr) \to 1.$$

Few false discoveries; most easy discoveries made.
Intermediate discoveries not made.

Estimating τ

Ad-hoc:

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

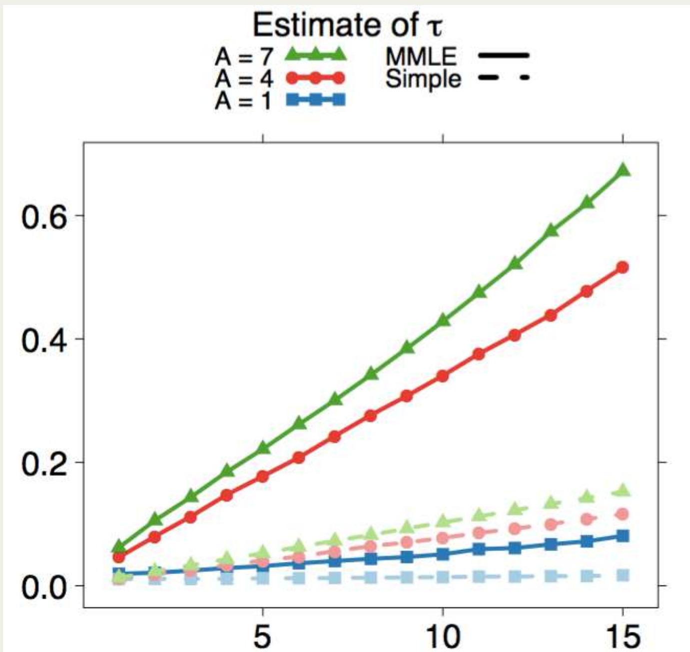
Empirical Bayes: For g_τ the prior of θ_i,

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

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

Numbers

estimating τ



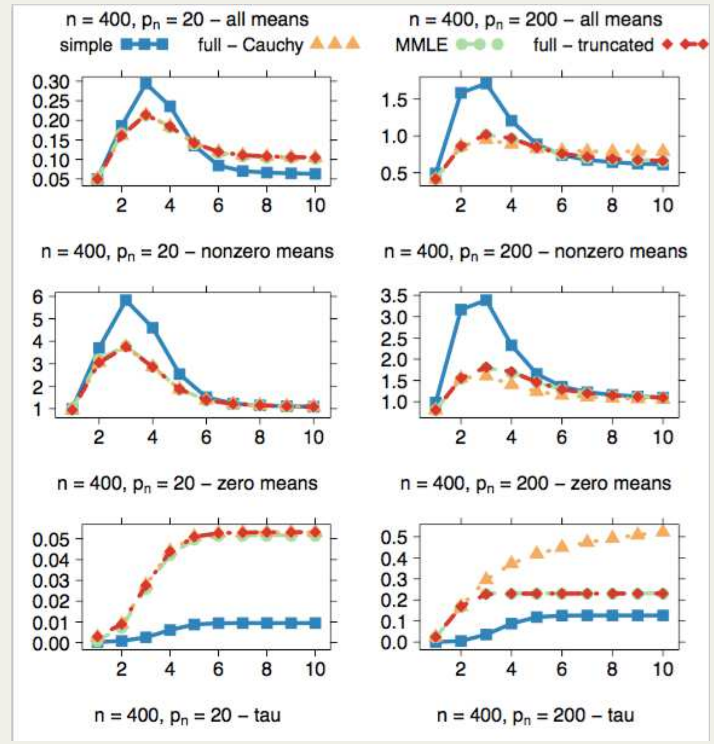
$s_n \to$

$n = 100$, 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



“$p_n = s_n$”

Short summary:

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

Recovery

THEOREM *(black body)*
For the likelihood based empirical Bayes $\hat\tau_n$,

$$\sup_{\|\theta_0\|_0\le s_n} \mathrm{E}_{\theta_0}\Big[\Pi\Big(\theta\colon \|\theta_0-\theta\|_2 \ge M_n\sqrt{s_n\log n} \,|\, Y^n,\tau\Big)_{|\tau=\hat\tau_n}\Big] \to 0.$$

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

$$\sup_{\|\theta_0\|_0\le s_n} \mathrm{E}_{\theta_0}\Pi\Big(\theta\colon \|\theta_0-\theta\|_2 \ge M_n\sqrt{s_n\log n} \,|\, Y^n\Big) \to 0.$$

Credible intervals

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

$$\hat C_{ni}(L,\tau) = \Big\{\theta_i\colon \big|\theta_i - \hat\theta_i(\tau)\big| \leq L r_i\hat{(\tau)}\Big\},$$

for $r_i(\tau)$ satisfying $\Pi\big(\theta_i\colon |\theta_i - \hat\theta_i(\tau)| \leq \hat r_i(\tau) |\, Y_i, \tau\big) = 0.95$.

$$\mathbf{S}_a\colon = \big\{1 \leq i \leq n\colon |\theta_{0,i}| \leq 1/n\big\},$$
$$\mathbf{M}_a\colon = \big\{1 \leq i \leq n\colon (s_n/n)\sqrt{\log(n/s_n)} \ll |\theta_{0,i}| \leq 0.99\sqrt{2\log(n/s_n)}\big\}.$$
$$\mathbf{L}_a\colon = \big\{1 \leq i \leq n\colon 1.001\sqrt{2\log n} \leq |\theta_{0,i}|\big\}.$$

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

$$P_{\theta_0}\Big(\frac{1}{\#\mathbf{S}_a}\#\{i \in \mathbf{S}_a\colon \theta_{0,i} \in \hat C_{ni}(L_{S,\gamma}, \hat\tau_n)\} \geq 1-\gamma\Big) \to 1,$$

$$P_{\theta_0}\big(\theta_{0,i} \notin \hat C_{ni}(L, \hat\tau_n)\big) \to 1, \quad \text{for any } L > 0 \text{ and } i \in \mathbf{M}_a,$$

$$P_{\theta_0}\Big(\frac{1}{\#\mathbf{L_a}}\#\{i \in \mathbf{L}_a\colon \theta_{0,i} \in \hat C_{ni}(L_{L,\gamma}, \hat\tau_n)\} \geq 1-\gamma\Big) \to 1.$$

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

Credible sets — impossibility of adaptation

General principle:
size of honest confidence set is determined by biggest model.

[Cai and Low, Juditzkyv& Lambert-Lacroix, 2003; Robins & van der Vaart, 2006]

Credible sets — impossibility of adaptation

General principle:
size of honest confidence set is determined by biggest model.

[Cai and Low, Juditzkyv& Lambert-Lacroix, 2003; Robins & van der Vaart, 2006]

THEOREM [Li, 1987]
If $P_{\theta_0}(C_n(Y^n) \ni \theta_0) \geq 0.95$ for all $\theta_0 \in \mathbb{R}^n$, then
$\mathrm{diam}(C_n(Y^n)) \gtrsim n^{-1/4}$, for some θ_0.

Credible sets — impossibility of adaptation

General principle:
size of honest confidence set is determined by biggest model.

[Cai and Low, Juditzkyv& Lambert-Lacroix, 2003; Robins & van der Vaart, 2006]

THEOREM [Li, 1987]
If $P_{\theta_0}(C_n(Y^n) \ni \theta_0) \geq 0.95$ for all $\theta_0 \in \mathbb{R}^n$, then
$\mathrm{diam}(C_n(Y^n)) \gtrsim n^{-1/4}$, for some θ_0.

THEOREM [Nickl, van de Geer, 2013]
If $s_{1,n} \ll s_{2,n}$ and
$\mathrm{diam}(C_n(Y^n))$ is of order $\big((s_{i,n}/n)\log(n/s_{i,n})\big)^{1/2}$, uniformly in $\|\theta_0\|_0 \leq s_{i,n}$
for $i = 1, 2$, then
$C_n(Y^n)$ cannot have uniform coverage over $\{\theta_0 : \|\theta_0\|_0 \leq s_{2,n}\}$.

Credible sets — impossibility of adaptation

General principle:
size of honest confidence set is determined by biggest model.

[Cai and Low, Juditzkyv& Lambert-Lacroix, 2003; Robins & van der Vaart, 2006]

THEOREM [Li, 1987]
If $P_{\theta_0}(C_n(Y^n) \ni \theta_0) \geq 0.95$ for all $\theta_0 \in \mathbb{R}^n$, then
$\mathrm{diam}(C_n(Y^n)) \gtrsim n^{-1/4}$, for some θ_0.

THEOREM [Nickl, van de Geer, 2013]
If $s_{1,n} \ll s_{2,n}$ and
$\mathrm{diam}(C_n(Y^n))$ is of order $\big((s_{i,n}/n)\log(n/s_{i,n})\big)^{1/2}$, uniformly in $\|\theta_0\|_0 \leq s_{i,n}$
for $i = 1, 2$, then
$C_n(Y^n)$ cannot have uniform coverage over $\{\theta_0 \colon \|\theta_0\|_0 \leq s_{2,n}\}$.

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 $0.001s$ coordinates of θ_0 satisfy

$$|\theta_{0,i}| \geq 1.001\sqrt{2\log(n/s)}.$$

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 $0.001s$ coordinates of θ_0 satisfy

$$|\theta_{0,i}| \geq 1.001\sqrt{2\log(n/s)}.$$

DEFINITION [excessive-bias restriction, Belitser & Nurushev, 2015]
$\|\theta\|_0 \leq s$ and $\exists \tilde{s}$ with $\tilde{s} \asymp \#\big(i\colon |\theta_{0,i}| \geq 1.001\sqrt{2\log(n/\tilde{s})}\big)$ and

$$\sum_{i:|\theta_{0,i}|\leq 1.001\sqrt{2\log(n/\tilde{s})}} \theta_{0,i}^2 \lesssim \tilde{s}\log(n/\tilde{s}).$$

Excessive-bias restriction implies self-similarity.
Self-similarity allows to tighten up the sets $\mathbf{S}, \mathbf{M}, \mathbf{L}$.

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 $\hat{\theta} = \mathrm{E}(\theta | Y^n)$ the posterior mean,

$$\hat{C}_n(L) = \Big\{\theta\colon \|\theta - \hat{\theta}\|_2 \leq L\hat{r}\Big\},$$

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

Credible balls

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

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

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

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

THEOREM

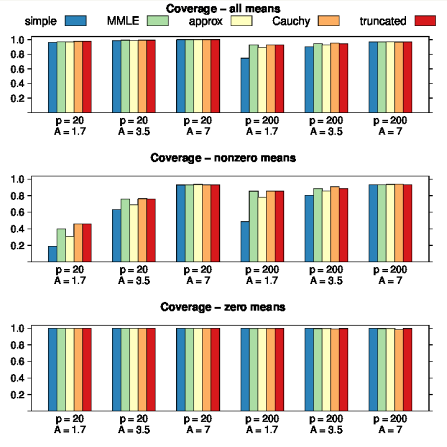
If $s_n \to 0$, for sufficiently large L,

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

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

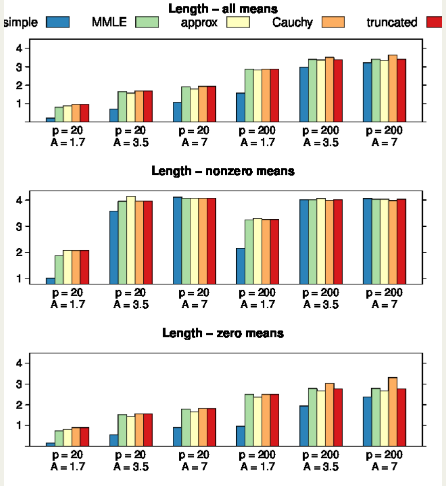
Numbers

coverage



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

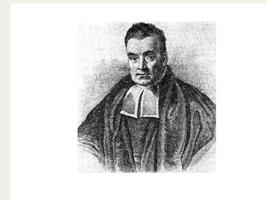
average interval length



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

Short summary:
empirical Bayes works well

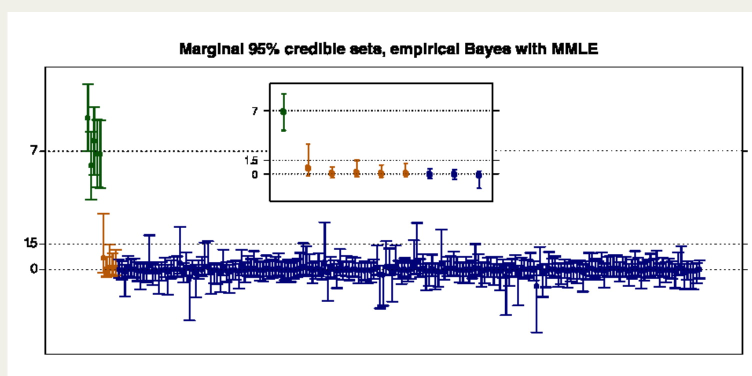
Conclusions



Bayesian sparse estimation gives excellent recovery.

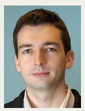
For valid simultaneous credible sets need a fraction of nonzero parameters above the “universal threshold”.

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



Co-authors

 Subhashis Ghoshal

 Ismael Castillo

 Stéphanie van der Pas

 Willem Kruijer

 Bartek Knapik

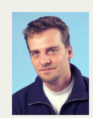
 Suzanne Sniekers

 Harry van Zanten

 Johannes Schmidt-Hieber

 Botond Szabo

 Judith Rousseau

 Bas Kleijn

 Fengnan Gao

Compatibility and coherence

$$\|X\| := \max_j \|X_{.,j}\|.$$

Compatibility number $\phi(S)$ for $S \subset \{1, \ldots, p\}$ is: $\displaystyle\inf_{\|\theta_{S^c}\|_1 \le 7\|\theta\|_1} \frac{\|X\theta\|_2 \sqrt{|S|}}{\|X\| \|\theta_S\|_1}.$

Compatibility in s_n-sparse vectors means: $\displaystyle\inf_{\theta: \|\theta\|_0 \le 5s_n} \frac{\|X\theta\|_2 \sqrt{|S_\theta|}}{\|X\| \|\theta\|_1} \gg 0.$

Strong compatibility in s_n-sparse vectors means: $\displaystyle\inf_{\theta: \|\theta\|_0 \le 5s_n} \frac{\|X\theta\|_2}{\|X\| \|\theta\|_2} \gg 0.$

Mutual coherence means: $\displaystyle s_n \max_{i \ne j} \left|\mathrm{cor}(X_{.i}, X_{.j})\right| \ll 1.$

Compatibility and coherence — examples

Mutual coherence $\Rightarrow$ Strong compatibility $\Rightarrow$ Compatibility.

Mutual coherence is easy to understand and gives best recovery results, but is very restrictive.

Compatibility and coherence — examples

Mutual coherence $\Rightarrow$ Strong compatibility $\Rightarrow$ Compatibility.

Mutual coherence is easy to understand and gives best recovery results, but is very restrictive.

Sequence model:

Completely compatible, with zero mutual coherence number.

Compatibility and coherence — examples

Mutual coherence $\Rightarrow$ Strong compatibility $\Rightarrow$ Compatibility.

Mutual coherence is easy to understand and gives best recovery results, but is very restrictive.

Sequence model:

Completely compatible, with zero mutual coherence number.

Response model:

If $X_{i,j}$ are i.i.d. random variables, then coherence if $s_n \lesssim \sqrt{n/\log p}$.

- if $\log p = o(n)$ and $X_{i,j}$ are bounded.
- if $\log p = o(n^{\alpha/(4+\alpha)})$ and $\mathrm{E}^{tX_{i,n}^{\alpha}} < \infty$.

Compatibility and coherence — examples

Mutual coherence $\Rightarrow$ Strong compatibility $\Rightarrow$ Compatibility.

Mutual coherence is easy to understand and gives best recovery results, but is very restrictive.

Sequence model:
Completely compatible, with zero mutual coherence number.

Response model:
If $X_{i,j}$ are i.i.d. random variables, then coherence if $s_n \lesssim \sqrt{n/\log p}$.

- if $\log p = o(n)$ and $X_{i,j}$ are bounded.
- if $\log p = o(n^{\alpha/(4+\alpha)})$ and $\mathrm{E}^{tX_{i,n}^{\alpha}} < \infty$.

$C = X^T X/n$: Compatibility, but no coherence if

- $C_{i,j} = \rho^{|i-j|}$, for $0 < \rho < 1$, and $p = n$.
- C is block diagonal with fixed block sizes.