Bayesian uncertainty quantification for sparsity models

Aad van der Vaart

Universiteit Leiden

JdS, Montpellier, May 2016

Contents

Co-authors

Sequence model & Regression

Ismael Castillo

Regression

Johannes Schmidt-Hieber

Horsehoe

Stéphanie van der Pas

Botond Szabo

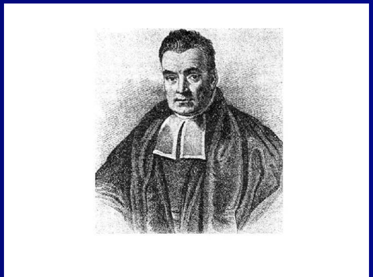
Sparsity

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 talk two simple models:

- Sequence model. Data $Y^n \sim N_n(\theta, I)$.
- Regression model. Data $Y^n \sim N_n(X_{n\times p}\theta, I)$.

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

Bayesian 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

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

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 $\theta_S = (\theta_i \colon i \in S)$ 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 $\theta_S = (\theta_i : i \in S)$ from density g_S on $\mathbb{R}^S$ and set $\theta_{S^c} = 0$.

We are particularly interested in π.

EXAMPLE *(Slab and spike)*

- 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:

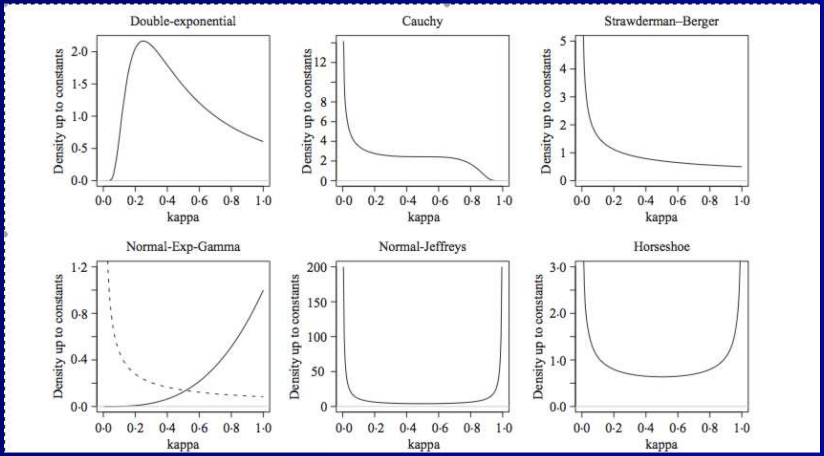
(1) Generate $\tau \sim \mathrm{Cauchy}^+(0, \sigma)$ (?)
(2) Generate $\sqrt{\psi_1}, \ldots, \sqrt{\psi_p}$ iid from $\mathrm{Cauchy}^+(0, \tau)$.
(3) Generate independent $\theta_i \sim N(0, \psi_i)$.

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



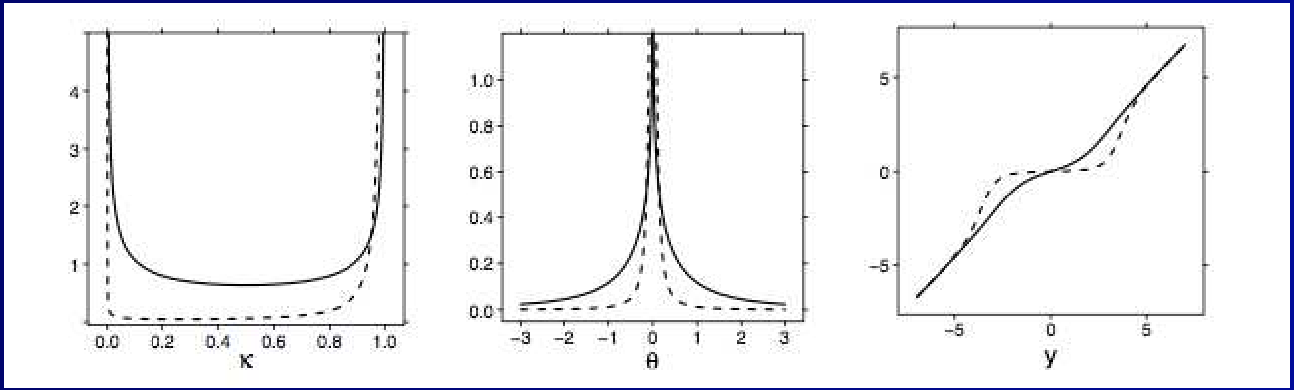
[Carvalho & Polson & Scott, 10.]

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 $\theta_i \sim N(0, \psi_i)$.

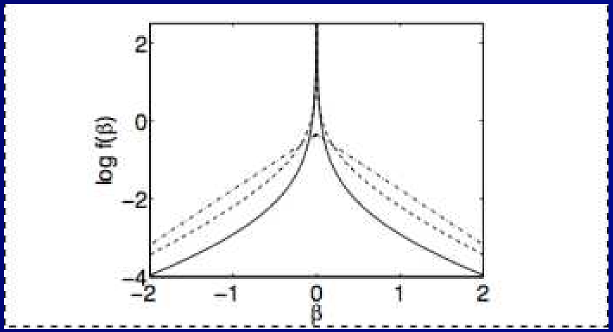
MOTIVATION: if $\theta \sim N(0, \psi)$ and $Y|\,\theta \sim N(\theta, 1)$,
then $\theta|\,Y \sim N\big((1-\kappa)Y, 1-\kappa\big)$ for $\kappa = 1/(1+\psi)$.
This suggests a prior for κ that concentrates near 0 or 1.



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

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

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

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

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

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

Usual 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.

Frequentist Bayes

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

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

- 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.

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.

Frequentist Bayes — recovery — contraction rate

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.

The *posterior contraction rate* for a given metric d is smallest r_n with

$$\mathrm{E}_{\theta_0}\Pi\big(\theta\colon d(\theta, \theta_0) \leq M r_n \,|\, Y^n\big) \to 1.$$

Frequentist Bayes — recovery — contraction rate

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.

The *posterior contraction rate* for a given metric d is smallest r_n with

$$\mathrm{E}_{\theta_0}\Pi\big(\theta\colon d(\theta,\theta_0) \leq M r_n \,|\, Y^n\big) \to 1.$$

Is r_n of same order as the optimal rate?

Frequentist Bayes — uncertainty quantification — credible sets

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.

A *credible set* is a set $C_n(Y^n)$ of parameters such that

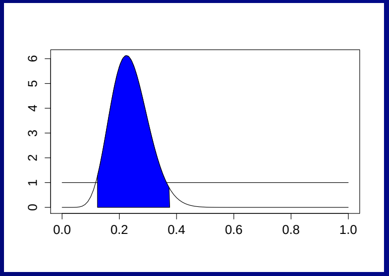
$$\Pi\big(\theta \in C_n(Y^n) \,|\, Y^n\big) = 0.95.$$

Frequentist Bayes — uncertainty quantification — credible sets

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.

A *credible set* is a set $C_n(Y^n)$ of parameters such that

$$\Pi\big(\theta \in C_n(Y^n) \,|\, Y^n\big) = 0.95.$$



Frequentist Bayes — uncertainty quantification — credible sets

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.

A *credible set* is a set $C_n(Y^n)$ of parameters such that

$$\Pi\big(\theta \in C_n(Y^n) \,|\, Y^n\big) = 0.95.$$

Is such a set a confidence set? i.e.

$$\mathrm{P}_{\theta_0}\big(C_n(Y^n) \ni \theta_0\big) \geq 0.95.$$

Frequentist Bayes — uncertainty quantification — credible sets

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.

A *credible set* is a set $C_n(Y^n)$ of parameters such that

$$\Pi\big(\theta \in C_n(Y^n) \,|\, Y^n\big) = 0.95.$$

Is such a set a confidence set? i.e.

$$\mathrm{P}_{\theta_0}\big(C_n(Y^n) \ni \theta_0\big) \geq 0.95.$$

Is the order of magnitude of $C_n(Y^n)$ correct?

Benchmarks for recovery — sequence model

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

$$\|\theta\|_0 = \#(1 \le i \le n \colon \theta_i \ne 0),$$

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

Benchmarks for recovery — sequence model

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

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

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

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

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

[(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 $\theta = (\theta_1, \ldots, \theta_n) \in \mathbb{R}^n$.

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

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

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

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

Benchmarks for recovery — sequence model

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

$$\|\theta\|_0 = \#(1 \le i \le n\colon \theta_i \ne 0),$$

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

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

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

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

- weak ℓ_r-balls

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

Benchmarks for recovery — regression model

$\boxed{Y^n \sim N_n(X_{n\times p}\theta, I)\text{, for } \theta = (\theta_1, \ldots, \theta_p) \in \mathbb{R}^p \text{ and known } X_{n\times p}.}$

Benchmarks for recovery — regression model

$$\boxed{Y^n \sim N_n(X_{n\times p}\theta, I),\ \text{for } \theta = (\theta_1, \ldots, \theta_p) \in \mathbb{R}^p \text{ and known } X_{n\times p}.}$$

Frequentist benchmarks: minimax rate depends on the sparsity of θ
in combination with "sparse invertibility" of $X_{n\times p}$.

Remainder of the talk: sequence model only.
Results on model selection prior extend to regression model, under appropriate conditions on $X_{n\times p}$.

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 $\theta \in \mathbb{R}^n$:

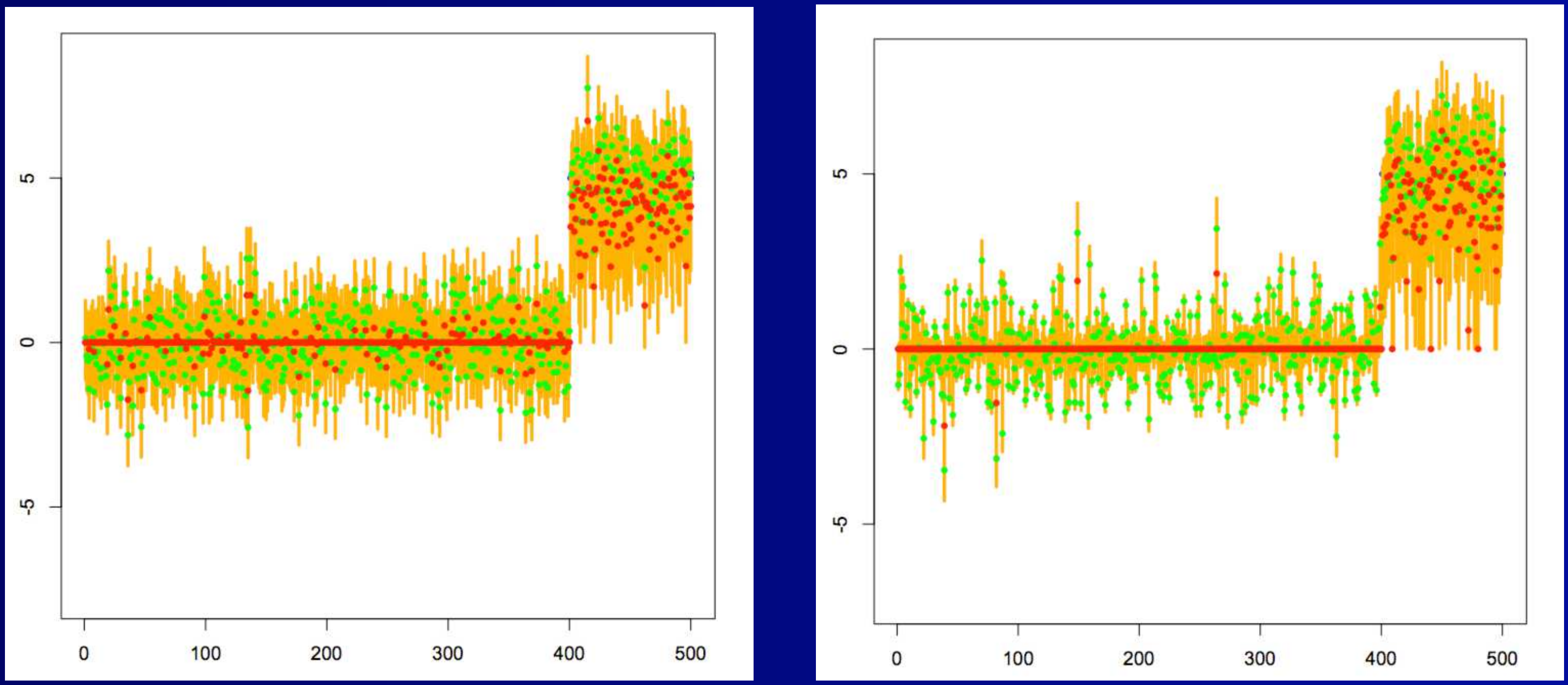
(1) Choose s from prior π_n on $\{0, 1, 2, \ldots, n\}$.
(2) Choose $S \subset \{0, 1, \ldots, n\}$ of size $|S| = s$ at random.
(3) Choose $\theta_S = (\theta_i : i \in S)$ from density g_S on $\mathbb{R}^S$ and set $\theta_{S^c} = 0$.

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 $h : \mathbb{R} \to \mathbb{R}$ and with finite second moment.
- $s_n, n \to \infty$, $s_n/n \to 0$. [*true number of nonzero parameters*.]

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

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

Dimensionality of posterior distribution

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

$$\sup_{\|\theta_0\|_0 \leq s_n} \mathrm{E}_{\theta_0} \Pi_n\big(\theta \colon \|\theta\|_0 \geq M s_n \,|\, Y^n\big) \to 0.$$

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

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

THEOREM *(No supersets)*

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

THEOREM *(Finds big signals)*

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

Selection

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

THEOREM *(No supersets)*

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

THEOREM *(Finds big signals)*

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

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 $\hat{w}_S$,

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

THEOREM

Given consistent model selection, mixture can be replaced by $N_{|S_0|}(Y_{S_{\theta_0}}, I)\otimes\delta_{S^c_{th_0}}$.

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 $\sqrt{2\log n}$, $\sqrt{2\log(n/\|\theta_0\|_0)}$.

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 reconstruction is neither better nor worse.

Horseshoe Prior

Horseshoe prior

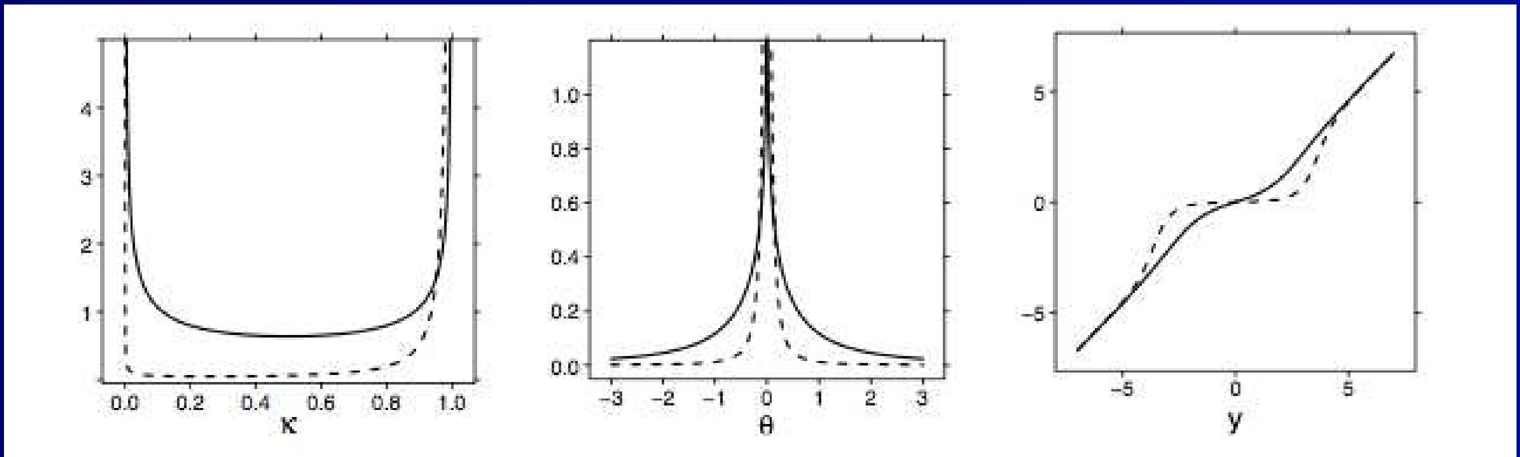
$$\boxed{Y^n \sim N_n(\theta, I), \text{ for } \theta = (\theta_1, \ldots, \theta_n) \in \mathbb{R}^n.}$$

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

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

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

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

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

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

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

Coverage provided shrinkage is not too big.

Credible intervals — prechosen τ

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

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

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

$$\begin{aligned}
S &:= \big\{1 \leq i \leq n \colon |\theta_{0,i}| \leq \tau\big\},\\
M &:= \big\{1 \leq i \leq n \colon \tau \ll |\theta_{0,i}| \leq 0.99\sqrt{2\log(1/\tau)}\big\},\\
L &:= \big\{1 \leq i \leq n \colon 1.01\sqrt{2\log(1/\tau)} \leq |\theta_{0,i}|\big\}.
\end{aligned}$$

Credible intervals — prechosen τ

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

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

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

$$S\colon = \{1 \le i \le n\colon |\theta_{0,i}| \le \tau\},$$
$$M\colon = \{1 \le i \le n\colon \tau \ll |\theta_{0,i}| \le 0.99\sqrt{2\log(1/\tau)}\},$$
$$L\colon = \{1 \le i \le n\colon 1.01\sqrt{2\log(1/\tau)} \le |\theta_{0,i}|\}.$$

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

$$P_{\theta_0}\Big(\frac{1}{|S|}|\{i \in S\colon \theta_{0,i} \in \hat C_{ni}(L_S,\tau)\}| \ge 1-\gamma\Big) \to 1,$$
$$P_{\theta_0}\big(\theta_{0,i} \notin \hat C_{ni}(L,\tau)\big) \to 1, \quad \text{for any } L > 0 \text{ and } i \in M,$$
$$P_{\theta_0}\Big(\frac{1}{|L|}|\{i \in L\colon \theta_{0,i} \in \hat C_{ni}(L_L,\tau)\}| \ge 1-\gamma\Big) \to 1.$$

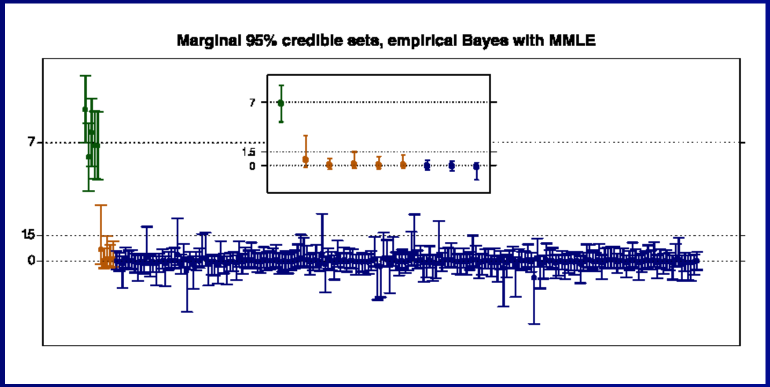
Credible intervals — prechosen τ

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

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

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

$$\begin{aligned} S &\colon= \{1 \leq i \leq n\colon |\theta_{0,i}| \leq \tau\}, \\ M &\colon= \{1 \leq i \leq n\colon \tau \ll |\theta_{0,i}| \leq 0.99\sqrt{2\log(1/\tau)}\}, \\ L &\colon= \{1 \leq i \leq n\colon 1.01\sqrt{2\log(1/\tau)} \leq |\theta_{0,i}|\}. \end{aligned}$$



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

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

Estimating τ

Ad-hoc:

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

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,

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

Estimating τ

Ad-hoc:

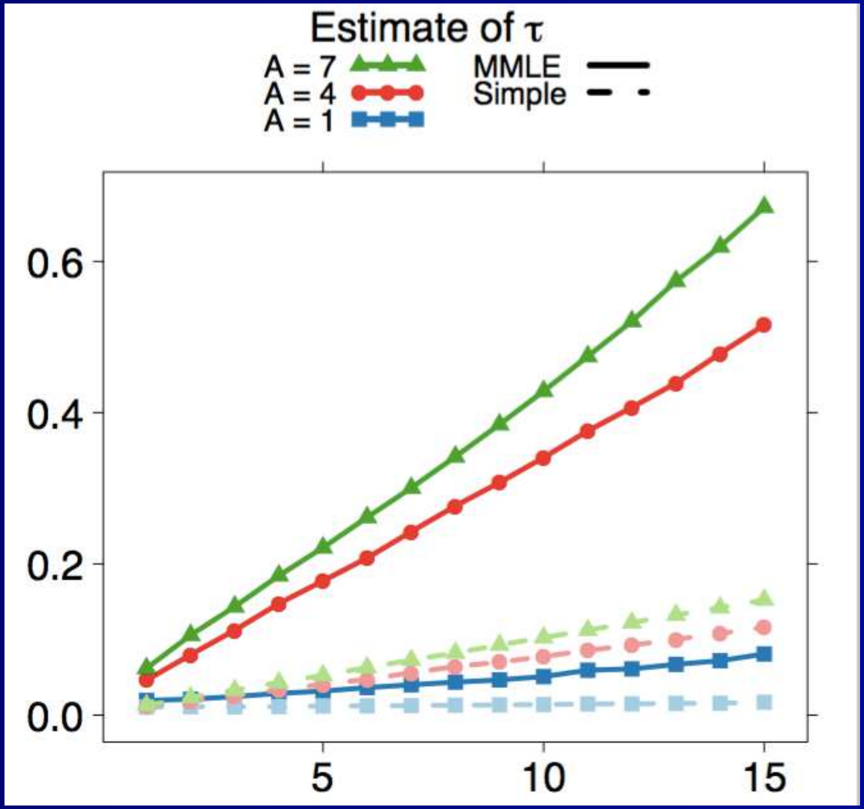
$$\hat{\tau}_n = \frac{\#\{|Y_i^n| \geq \sqrt{2\log n}\}}{1.1n}.$$

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 $[1/n, 1]$).

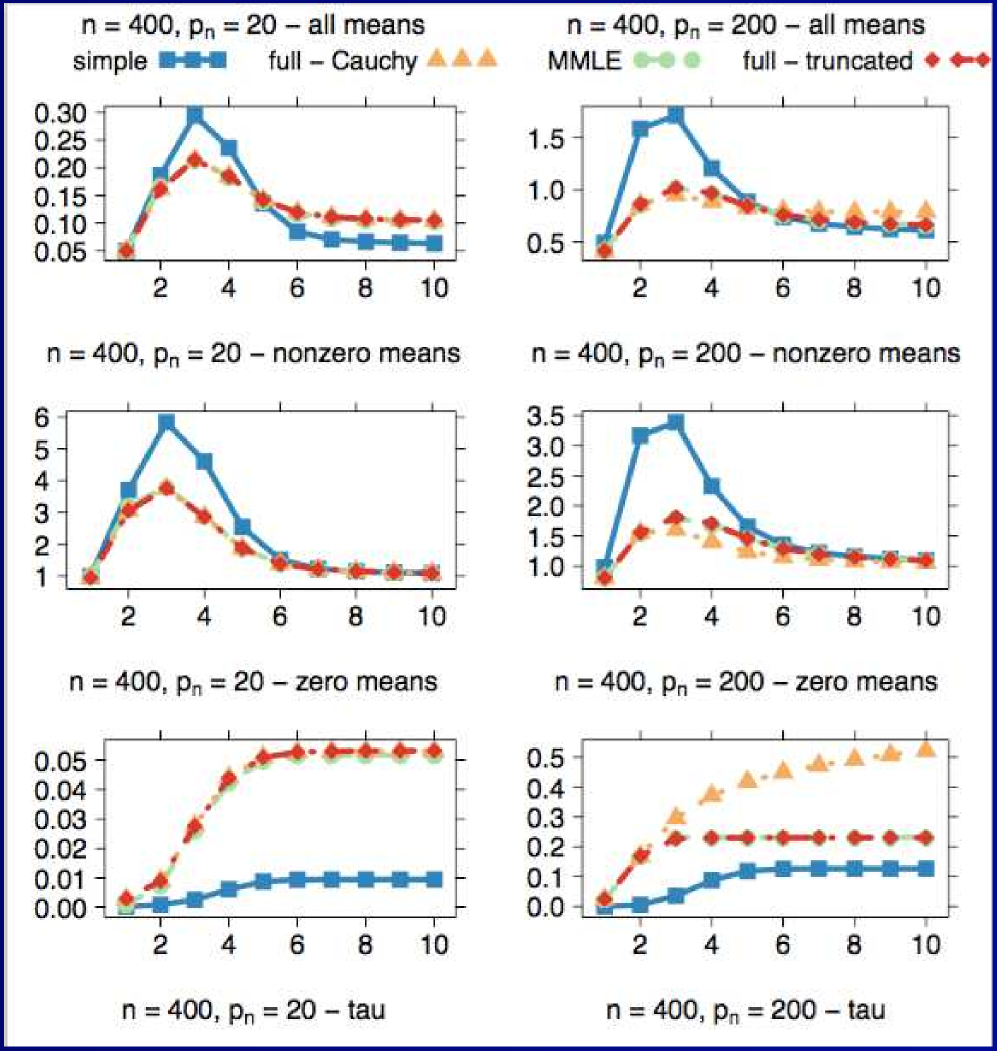
Numbers: estimating τ



$s_n \longrightarrow$

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

Numbers: MSE of posterior mean, as function of nonzero parameter



“$p_n = s_n$”

Recovery

THEOREM *(black body)*

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

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

Recovery

THEOREM *(black body)*

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

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

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 sets — impossibility of adaptation

General principle: the size of an honest confidence set is determined by the biggest model. [Cai and Low, Juditzkyv& 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 $\mathrm{diam}(C_n(Y)) \gtrsim n^{-1/4}$, for some θ_0.

Credible sets — impossibility of adaptation

General principle: the size of an honest confidence set is determined by the biggest model. [Cai and Low, Juditzkyv& 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 $\mathrm{diam}(C_n(Y)) \gtrsim n^{-1/4}$, for some θ_0.

THEOREM [Nickl, van de Geer, 2013]

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

Credible sets — impossibility of adaptation

General principle: the size of an honest confidence set is determined by the biggest model. [Cai and Low, Juditzkyv& 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 $\mathrm{diam}(C_n(Y)) \gtrsim n^{-1/4}$, for some θ_0.

THEOREM [Nickl, van de Geer, 2013]

If $s_{1,n} \ll s_{2,n}$ and $\mathrm{diam}(C_n(Y))$ 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 $C_n(Y)$ cannot have uniform coverage over $\{\theta_0 : \|\theta_0\|_0 \leq s_{2,n}\}$.

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

Credible sets — impossibility of adaptation — restricting the parameter

Coverage pertains only for θ_0 that *do 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]
For $\tilde{s} := \#\bigl(i\colon |\theta_{0,i}| \geq 1.01\sqrt{2\log(n/s)}\bigr)$,

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

Excessive-bias restriction implies self-similarity.

Credible balls

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

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

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

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

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

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

Credible balls

THEOREM

For sufficiently large L,

$$\inf_{1\leq \tilde{s}\leq \tilde{s}_n} \inf_{\theta_0\in \mathsf{EBR}[\tilde{s}]} P_{\theta_0}\Big(\theta_0 \in \hat{C}_n(L, \hat{\tau}_n)\Big) \to 1,$$

$$\inf_{\log n\lesssim \tilde{s}\leq \tilde{s}_n} \inf_{\theta_0\in \mathsf{EBR}[\tilde{s}]} P_{\theta_0}\Big(\theta_0 \in \hat{C}_n(L)\Big) \to 1.$$

EBR$[\tilde{s}]$: vectors θ_0 that satisfy excessive bias restriction.

Credible intervals

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

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

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

$$\begin{aligned}
S_a &:= \bigl\{1 \le i \le n\colon |\theta_{0,i}| \le 1/n\bigr\},\\
M_a &:= \bigl\{1 \le i \le n\colon (s_n/n)\sqrt{\log(n/s_n)} \ll |\theta_{0,i}| \le 0.99\sqrt{2\log(n/s_n)}\bigr\}.\\
L_a &:= \bigl\{1 \le i \le n\colon 1.01\sqrt{2\log n} \le |\theta_{0,i}|\bigr\}.
\end{aligned}$$

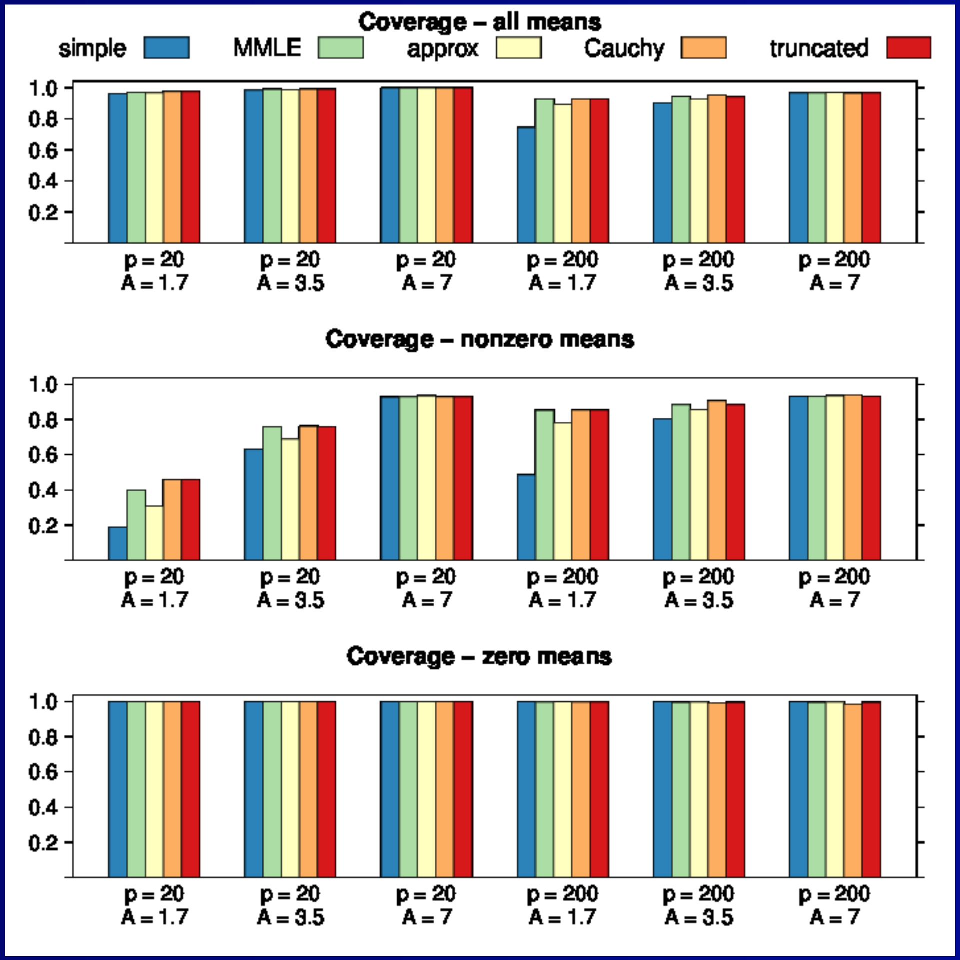
THEOREM For any $\gamma > 0$ and $\|\theta_0\|_0 \le s_n$,

$$P_{\theta_0}\Bigl(\frac{1}{|S_a|}|\{i \in S_a\colon \theta_{0,i} \in \hat C_{ni}(L_{S,\gamma}, \hat\tau_n)\}| \ge 1-\gamma\Bigr) \to 1,$$

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

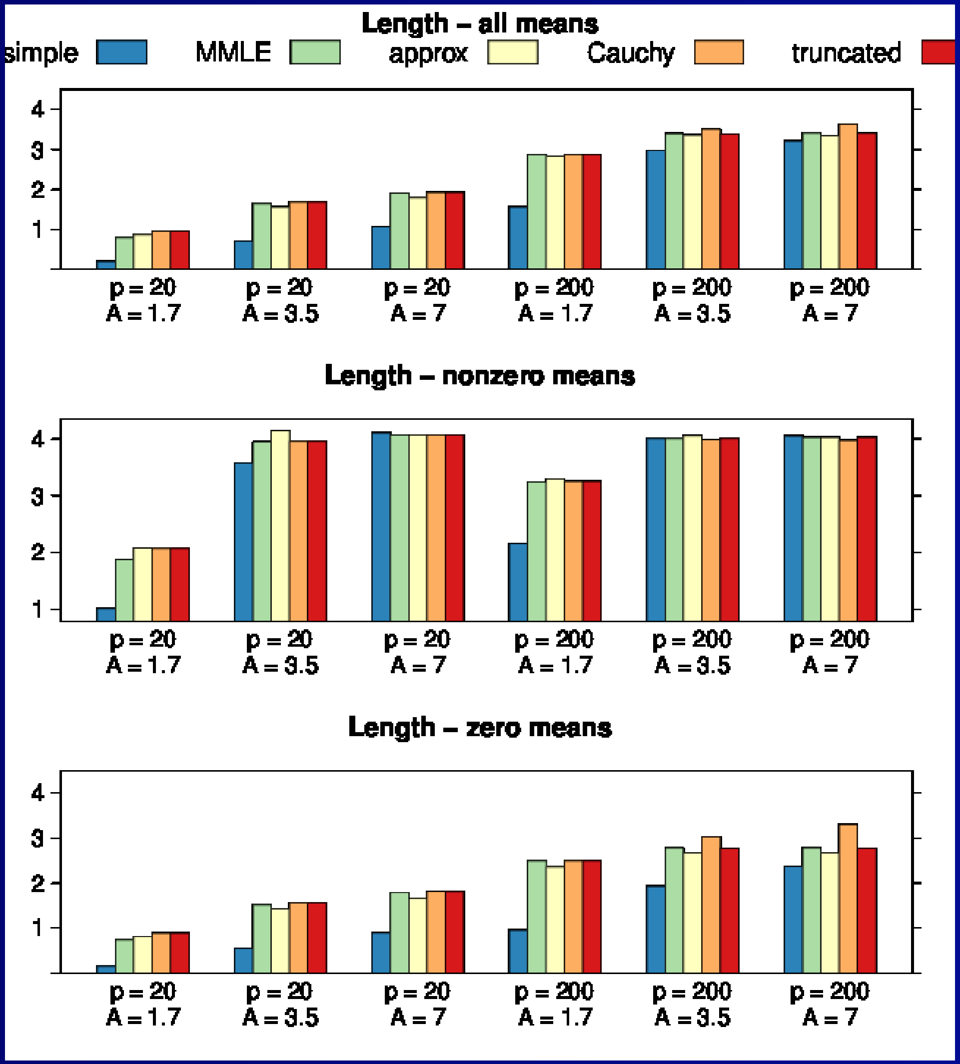
$$P_{\theta_0}\Bigl(\frac{1}{|L_a|}|\{i \in L_a\colon \theta_{0,i} \in \hat C_{ni}(L_{L,\gamma}, \hat\tau_n)\}|/l \ge 1-\gamma\Bigr) \to 1.$$

Numbers — coverage



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

Numbers — average interval length



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