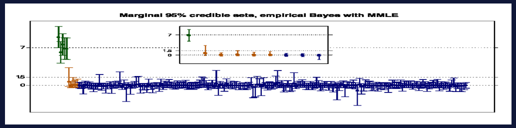
Bayesian Statistics in High Dimensions

Lecture 2: Sparsity

Aad van der Vaart

Universiteit Leiden, Netherlands



47th John H. Barrett Memorial Lectures, Knoxville, Tenessee, May 2017

Contents

Sparsity

Sparsity — sequence model

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

Sparsity — sequence model

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

In this lecture, (Bayesian) theory for:

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

- n independent observations $Y_1^n, \ldots, Y_n^n$.
- n unknowns $\theta_1, \ldots, \theta_n$.
- $Y_i^n = \theta_i + \varepsilon_i$, for standard normal noise ε_i.
- n is large.
- many of $\theta_1, \ldots, \theta_n$ are (almost) zero.

History

Given $Y^n | \theta \sim N_n(\theta, I)$ the obvious estimator of θ is Y^n.

ML, UMVU, best-equivariant, minimax

History

Given $Y^n|\,\theta \sim N_n(\theta, I)$ the obvious estimator of θ is Y^n.

ML, UMVU, best-equivariant, minimax, but *inadmissible*.

Theorem [Stein, 1956]

If $n \geq 3$, then there exists T such that $\forall\theta \in \mathbb{R}^n$,

$$\mathrm{E}_\theta\|T(Y^n) - \theta\|^2 < \mathrm{E}_\theta\|Y^n - \theta\|^2.$$

History

Given $Y^n | \theta \sim N_n(\theta, I)$ the obvious estimator of θ is Y^n.

ML, UMVU, best-equivariant, minimax, but *inadmissible*.

Theorem [Stein, 1956]

If $n \geq 3$, then there exists T such that $\forall \theta \in \mathbb{R}^n$,

$$\mathrm{E}_\theta \|T(Y^n) - \theta\|^2 < \mathrm{E}_\theta \|Y^n - \theta\|^2.$$

Empirical Bayes method [Robbins, 1960s]:

- Working hypothesis: $\theta_1, \ldots, \theta_n \overset{\text{iid}}{\sim} G$.
- Estimate G using Y^n, pretending this is true.
- $T(Y^n) := \mathrm{E}_{\hat{G}(Y^n)}(\theta | Y^n)$.

History

Given $Y^n | \theta \sim N_n(\theta, I)$ the obvious estimator of θ is Y^n.

ML, UMVU, best-equivariant, minimax, but *inadmissible*.

Theorem [Stein, 1956]

If $n \geq 3$, then there exists T such that $\forall \theta \in \mathbb{R}^n$,

$$\mathrm{E}_\theta \|T(Y^n) - \theta\|^2 < \mathrm{E}_\theta \|Y^n - \theta\|^2.$$

Empirical Bayes method [Robbins, 1960s]:

- Working hypothesis: $\theta_1, \ldots, \theta_n \overset{\text{iid}}{\sim} G$.
- Estimate G using Y^n, pretending this is true.
- $T(Y^n) := \mathrm{E}_{\hat{G}(Y^n)}(\theta | Y^n)$.

For large n the gain can be substantial,
("borrowing of strength"),
and be targeted to special subsets of $\mathbb{R}^n$, e.g. sparse vectors.

Sparsity — regression

$$Y^n|\,\theta \sim N_n(X_{n\times p}\theta, \sigma^2 I), \text{ for } \theta = (\theta_1,\ldots,\theta_p) \in \mathbb{R}^p$$

- Y_i^n: measurement on individual $i = 1,\ldots,n$.
- X_{ij} score of individual i on feature $j = 1,\ldots p$.
- θ_j effect of feature j.
- sparse if only few features matter.

If $p > n$, then *sparsity is necessary* to recover θ,
and X must be *sparse-invertible*, e.g.:

Compatibility:
$$\inf_{\theta:|S_\theta|\leq 5s_n} \frac{\|X\theta\|_2\sqrt{|S_\theta|}}{\|X\|\,\|\theta\|_1} \gg 0.$$

Mutual coherence:
$$s_n \max_{i\neq j}\bigl|\mathrm{cor}(X_{.i}, X_{.j})\bigr| \ll 1.$$

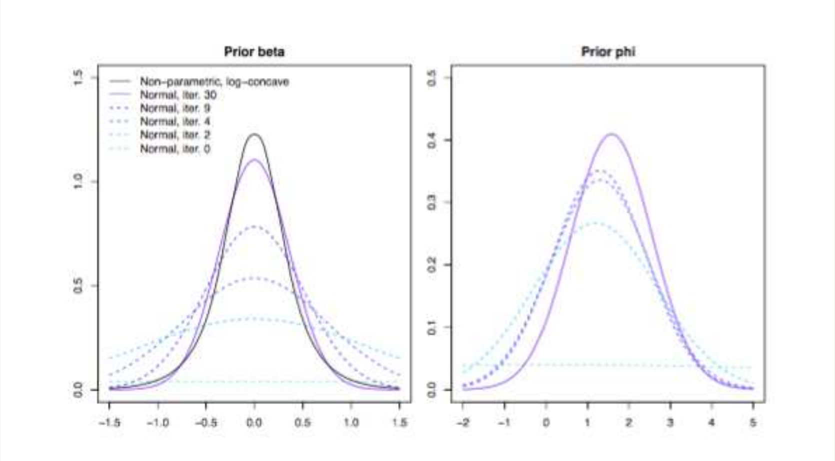
$s_n =$ true number of nonzero coordinates.

Sparsity — RNA sequencing

- $Y_{i,j}$: RNA expression count of tag $j = 1, \ldots, p$ in tissue $i = 1, \ldots, n$,
- x_i: covariate(s) of tissue i, *e.g.* 0 *or* 1 *for normal or cancer*.
- sparse if only few tags (genes) matter.

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

$$\mathrm{E}Y_{i,j} = e^{\alpha_j + \beta_j x_i}, \qquad \operatorname{var} Y_{i,j} = \mathrm{E}Y_{i,j}\bigl(1 + \mathrm{E}Y_{i,j} e^{-\phi_j}\bigr).$$



distribution of β_j's and ϕ_j's estimated by Empirical Bayes

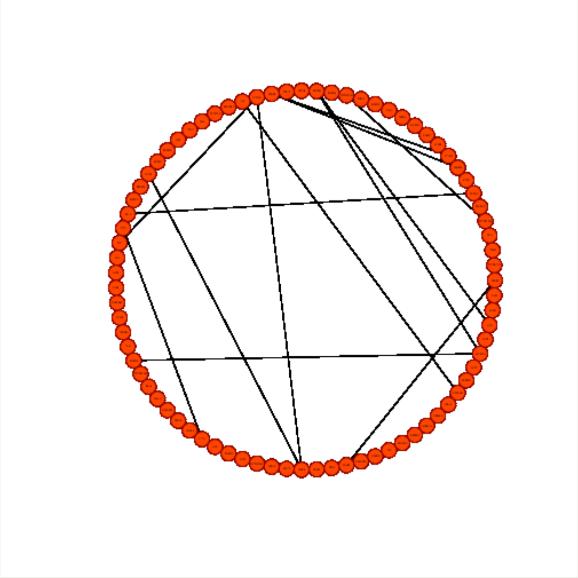
[Smyth & Robinson et al., van der Wiel, Leday, van Wieringen, & vdV et al., 12]

Sparsity — Gaussian graphical model

Data n iid copies of $Y^p|\,\theta \sim N_p(0, \theta^{-1})$

- Y_j^p = value of individual on feature j.
- precision matrix θ gives partial correlations:

$$\mathrm{cor}(Y_i^p, Y_j^p|\,Y_k^p\colon k \neq i, j) = -\frac{\theta_{i,j}}{\sqrt{\theta_{i,i}\theta_{j,j}}}.$$



Apoptosis network

- nodes $1, 2, \ldots, p$
- edge (i, j) present iff $\theta_{i,j} \neq 0$
- sparse if few edges

Bayesian 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

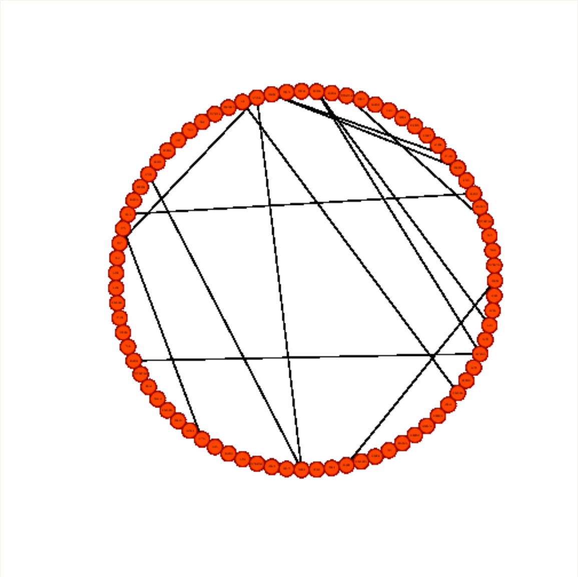
A sparse model has many parameters, but most of them are (nearly) zero.

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

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 — Gaussian graphical model



Apoptosis network

- nodes $1, 2, \ldots, p$
- edge (i, j) present iff $\theta_{i,j} \neq 0$
- sparse if few edges

For given *incidence matrix* $(P_{i,j})$ use different priors for

$$(\theta_{i,j}\colon P_{i,j} = 0) \qquad \text{and} \qquad (\theta_{i,j}\colon P_{i,j} = 1).$$

[Leday, Kpogbezan, vdV, van Wieringen, van der Wiel (2016, 17).]

Model selection prior

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

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

(2) Choose $S \subset \{0, 1, \ldots, p\}$ of size $|S| = s$ at random.

(3) Choose $\theta_S = (\theta_i\colon i \in S) \sim g_S$ and set $\theta_{S^c} = 0$.

[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 on $\{0, 1, 2, \ldots, p\}$.
(2) Choose $S \subset \{0, 1, \ldots, p\}$ of size $|S| = s$ at random.
(3) Choose $\theta_S = (\theta_i : i \in S) \sim g_S$ and set $\theta_{S^c} = 0$.

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$(1, p+1)$.

Then $s \sim$ binomial and $g_S = \otimes_{i \in S} g$.

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

Horseshoe prior

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

(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

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

(1) Generate $\tau \sim \mathrm{Cauchy}^+(0, \sigma)$ (?)

(2) Generate $\sqrt{\psi_1}, \ldots, \sqrt{\psi_p}$ iid from $\mathrm{Cauchy}^+(0, \tau)$.

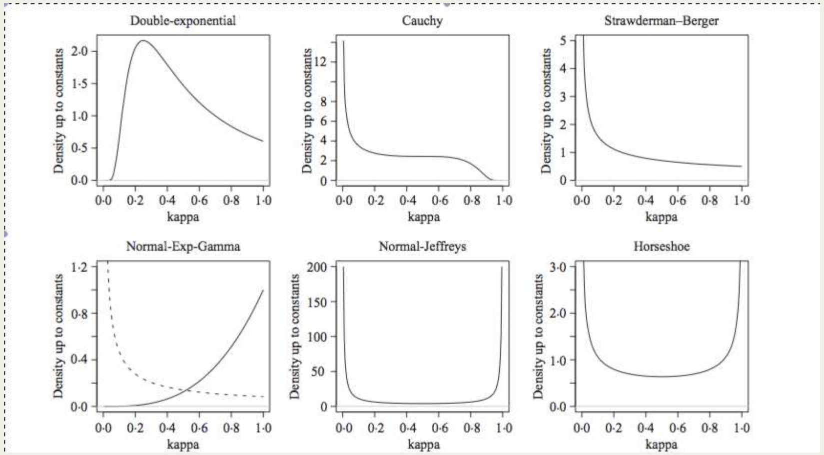
(3) Generate independent $\theta_i \sim N(0, \psi_i)$.

Motivation

if $\theta \sim N(0, \psi)$ and $Y \,|\, \theta \sim N(\theta, 1)$,

then $\theta | \, Y, \psi \sim N\big((1-\kappa)Y, 1-\kappa\big)$ for $\kappa = 1/(1+\psi)$.

This suggests a prior for κ that concentrates near 0 or 1.



[Carvalho & Polson & Scott, 10.]

Horseshoe prior

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

(1) Generate $\tau \sim \mathrm{Cauchy}^+(0, \sigma)$ (?)

(2) Generate $\sqrt{\psi_1}, \ldots, \sqrt{\psi_p}$ iid from $\mathrm{Cauchy}^+(0, \tau)$.

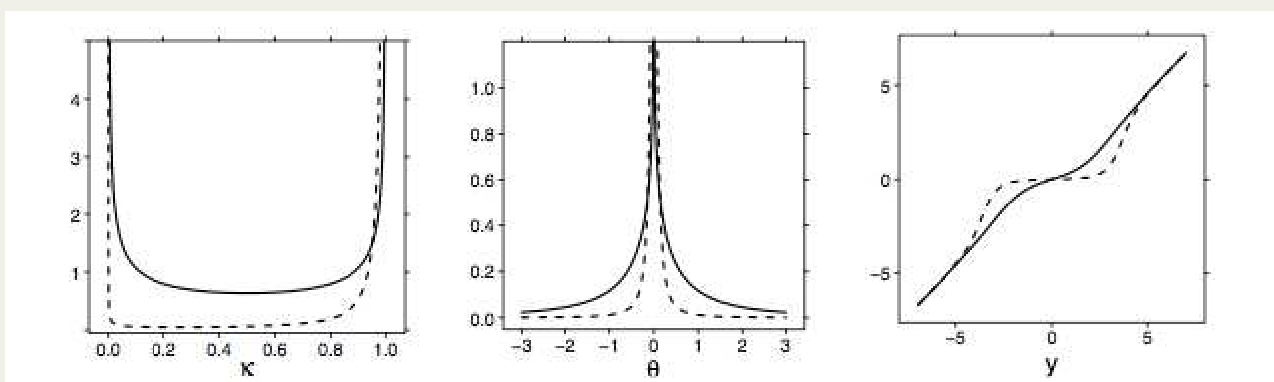
(3) Generate independent $\theta_i \sim N(0, \psi_i)$.

Motivation

if $\theta \sim N(0, \psi)$ and $Y | \theta \sim N(\theta, 1)$,

then $\theta | Y, \psi \sim N\big((1-\kappa)Y, 1-\kappa\big)$ for $\kappa = 1/(1+\psi)$.

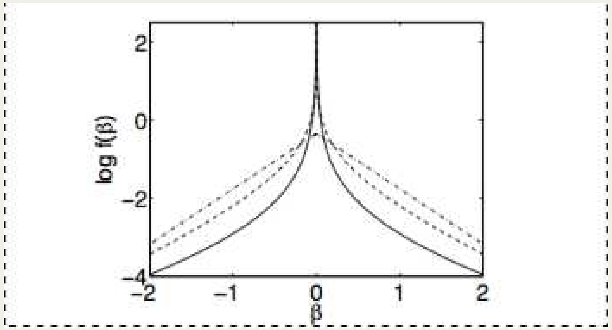
This suggests a prior for κ that concentrates near 0 or 1.



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

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

posterior mode for prior $\theta_i \overset{\text{iid}}{\sim} \mathrm{Laplace}(\lambda_n)$,
works great,
but the full posterior distribution is useless.

LASSO is not Bayesian

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

posterior mode for prior $\theta_i \overset{\text{iid}}{\sim} \mathrm{Laplace}(\lambda_n)$,
works great,
but the full posterior distribution is useless.

Theorem If $\sqrt{n}/\lambda_n \to \infty$ then

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

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

Trouble: λ_n must be large to shrink θ_i to 0, but small to model nonzero θ_i.

Frequentist Bayes

Frequentist Bayes

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

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

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

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

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

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

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

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

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

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

[(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

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

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

(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 $\theta_{S^c} = 0$.

Model selection prior

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

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

(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 $\theta_{S^c} = 0$.

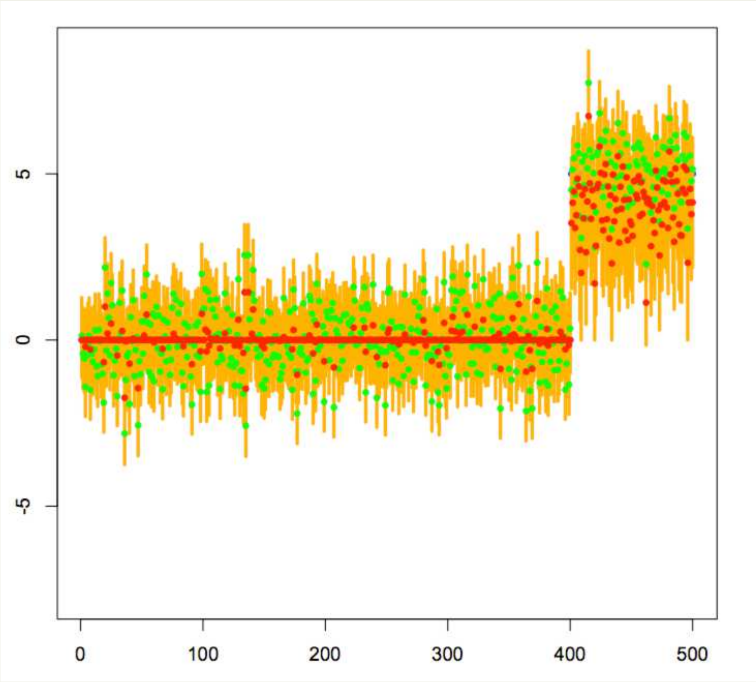
Assume

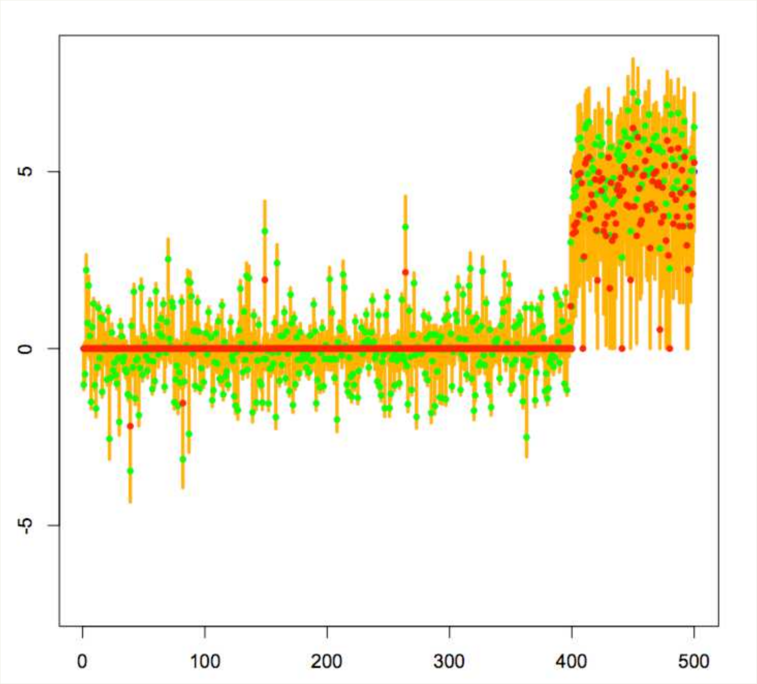
- $\pi_n(s) \leq c\,\pi_n(s-1)$, for some $c < 1$ and every s.
- $g_S = \otimes_{i \in S} e^h$, for uniformly Lipschitz $h \colon \mathbb{R} \to \mathbb{R}$.
- $s_n \colon= \|\theta_0\|_0 \to \infty$, $n \to \infty$, $s_n/n \to 0$.

Examples:

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

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

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

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

Selection

$S_\theta\colon= \{1 \le i \le n\colon \theta_i \ne 0\}$.

Theorem *[No supersets]*

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

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

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

Theorem *[No supersets]*

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

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 $\hat{w}_S$,

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

Theorem

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

> **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.

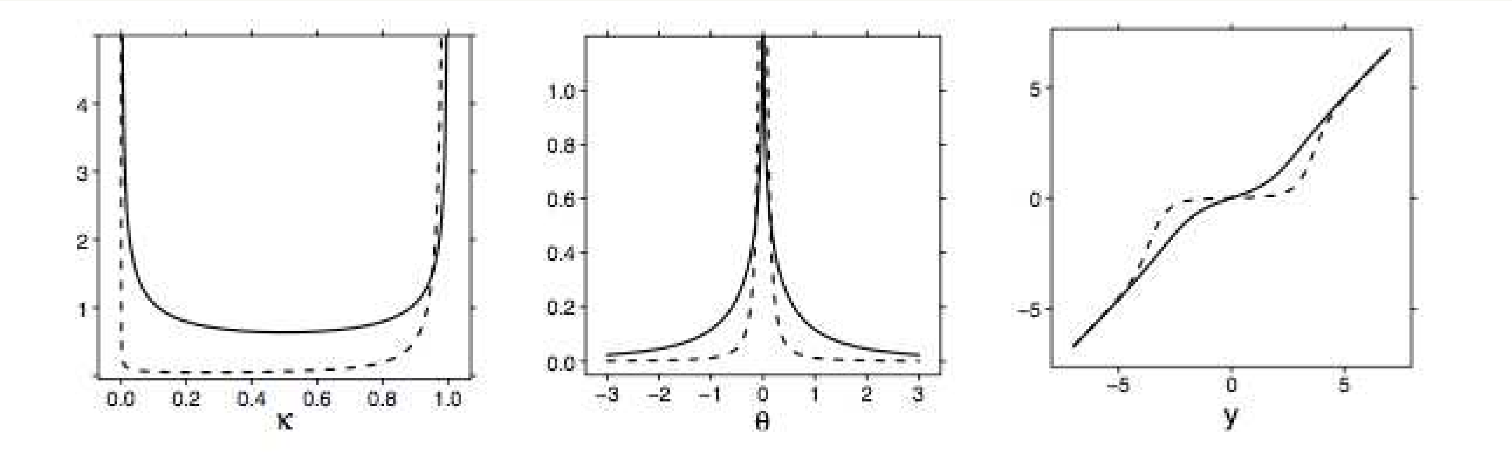
Horseshoe Prior

Horseshoe prior

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

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

(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.]

Estimating τ

Ad-hoc:

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

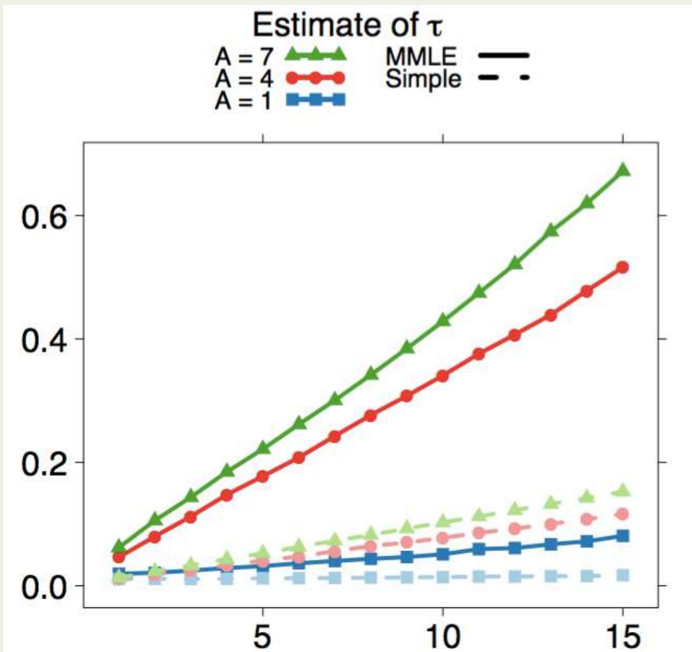
Empirical Bayes: For g_τ the prior of θ_i,

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

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

Numbers

estimating τ

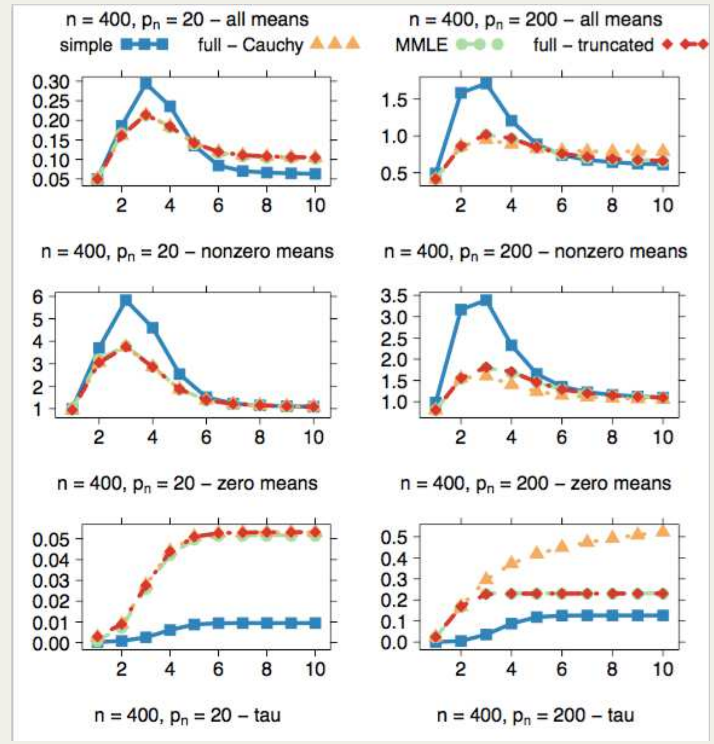


$s_n \rightarrow$

$n = 100$, 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



"$p_n = s_n$"

Short summary:
Empirical Bayes and Full Bayes outperform ad-hoc estimator.

Recovery

$\boxed{\text{Horseshoe prior gives similar recovery as model selection prior.}}$

Recovery

$\boxed{\text{Horseshoe prior gives similar recovery as model selection prior.}}$

$\boxed{\tau \text{ can be interpreted as } (s_n/n)\sqrt{\log(n/s_n)}.}$

Credible intervals

Credible interval:

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

$$\boxed{\begin{array}{c}\hat{\theta} = \mathrm{E}(\theta|\, Y^n) \\ \Pi\big(\theta_i \colon |\theta_i - \hat{\theta}_i| \leq \hat{r}_i|\, Y^n\big) = 0.95\end{array}}$$

Credible intervals

Credible interval:

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

$$\boxed{\begin{gathered}\hat{\theta} = \mathrm{E}(\theta|\, Y^n)\\ \Pi\big(\theta_i \colon |\theta_i - \hat{\theta}_i| \leq \hat{r}_i|\, Y^n\big) = 0.95\end{gathered}}$$

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

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

$$\mathbf{L}_a \colon = \{1 \leq i \leq n \colon 1.001\sqrt{2\log n} \leq |\theta_{0,i}|\}.$$

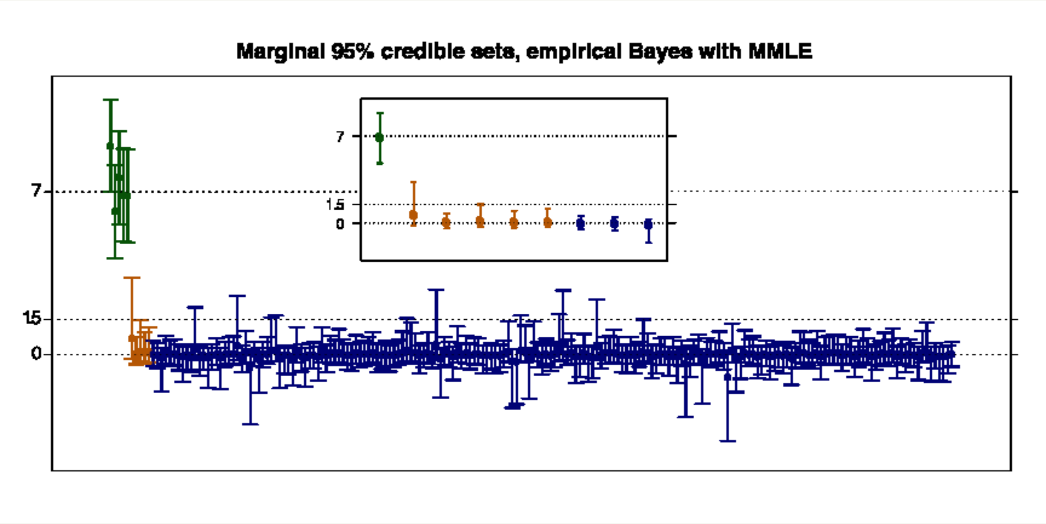
Credible intervals

Credible interval:

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

$$\hat{\theta} = \mathrm{E}(\theta |\, Y^n)$$
$$\Pi\big(\theta_i : |\theta_i - \hat{\theta}_i| \leq \hat{r}_i |\, Y^n\big) = 0.95$$

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



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

Credible interval:

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

$$\hat{\theta} = \mathrm{E}(\theta|\, Y^n)$$
$$\Pi\big(\theta_i \colon |\theta_i - \hat{\theta}_i| \leq \hat{r}_i|\, Y^n\big) = 0.95$$

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

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

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

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

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

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

Simultaneous credible balls — impossibility of adaptation

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

Simultaneous credible balls — impossibility of adaptation

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

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

Simultaneous credible balls — impossibility of adaptation

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

Theorem [Li, 1987]
If $P_{\theta_0}(C_n(Y^n) \ni \theta_0) \geq 0.95$, all $\theta_0 \in \mathbb{R}^n$, then $\mathrm{diam}(C_n(Y^n)) \gtrsim n^{-1/4}$, 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 optimal size, uniformly in $\|\theta_0\|_0 \leq s_{i,n}$ for $i = 1, 2$,
then $C_n(Y^n)$ cannot have uniform coverage over $\{\theta_0\colon \|\theta_0\|_0 \leq s_{2,n}\}$.

Simultaneous credible balls — impossibility of adaptation

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

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

[Optimal size is $\big((s_{i,n}/n)\log(n/s_{i,n})\big)^{1/2}$.]

Simultaneous credible balls — 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

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

Simultaneous credible balls — 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 $\mathbf{S}, \mathbf{M}, \mathbf{L}$.)

Simultaneous credible balls

Credible ball:

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

$$\boxed{\begin{gathered}\hat{\theta} = \mathrm{E}(\theta|\, Y^n)\\ \Pi\big(\theta\colon \|\theta - \hat{\theta}\| \leq \hat{r}|\, Y^n\big) = 0.95\end{gathered}}$$

Simultaneous credible balls

Credible ball:

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

$$\hat{\theta} = \mathrm{E}(\theta |\, Y^n)$$
$$\Pi\big(\theta\colon \|\theta - \hat{\theta}\| \leq \hat{r} |\, Y^n\big) = 0.95$$

Theorem

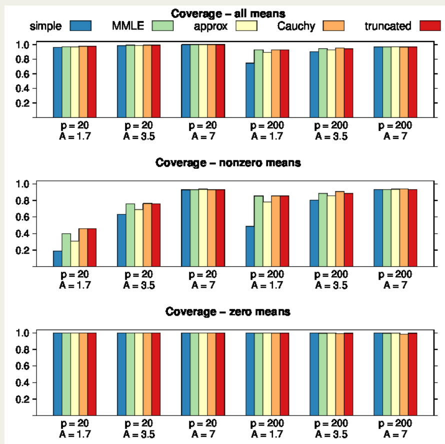
If $s_n/n \to 0$, for sufficiently large L,

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

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

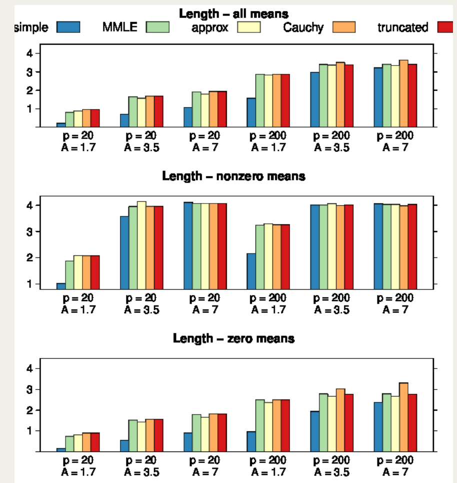
Numbers

coverage



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

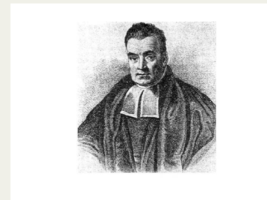
average interval length



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

Short summary:
empirical and full Bayes work well

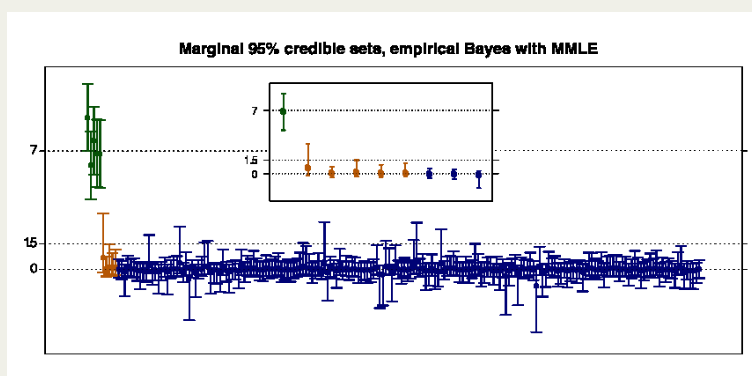
Conclusions



Bayesian sparse estimation gives excellent recovery.

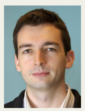
For valid simultaneous credible sets need
a fraction of nonzero parameters above the “universal threshold”.

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



Co-authors

 Subhashis Ghoshal

 Ismael Castillo

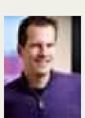
 Stéphanie van der Pas

 Willem Kruijer

 Bartek Knapik

 Suzanne Sniekers

 Gwenael Leday

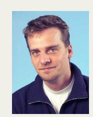
 Mark van de Wiel

 Harry van Zanten

 Johannes Schmidt-Hieber

 Botond Szabo

 Magnus Münch

 Bas Kleijn

 Fengnan Gao

 Gino Kpogbezan

 Wessel van Wieringen

Compatibility and coherence

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

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

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

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

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

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.