Bayesian estimation of sparse sequences

Ismaël Castillo (CNRS, Paris)

joint work with Aad van der Vaart (VU Amsterdam → Leiden)

INRA Toulouse, September 14th 2012

Introduction

Sparse sequences

Example (Sparse sequences)

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

- $\theta = (\theta_1, \ldots, \theta_n) \in \mathbb{R}^n$
- ε_i i.i.d. Gaussian $\mathcal{N}(0, 1)$

Sparse sequences

Example (Sparse sequences)

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

- $\theta = (\theta_1, \ldots, \theta_n) \in \mathbb{R}^n$
- ε_i i.i.d. Gaussian $\mathcal{N}(0, 1)$

Example (High-dim. linear model)

$$Y \quad = X\theta + \varepsilon$$

- $\theta \in R^M, \quad X \in \mathbb{R}^{n \times M}, \qquad M \gg n$
- $\varepsilon \sim \mathcal{N}(0, I_n)$

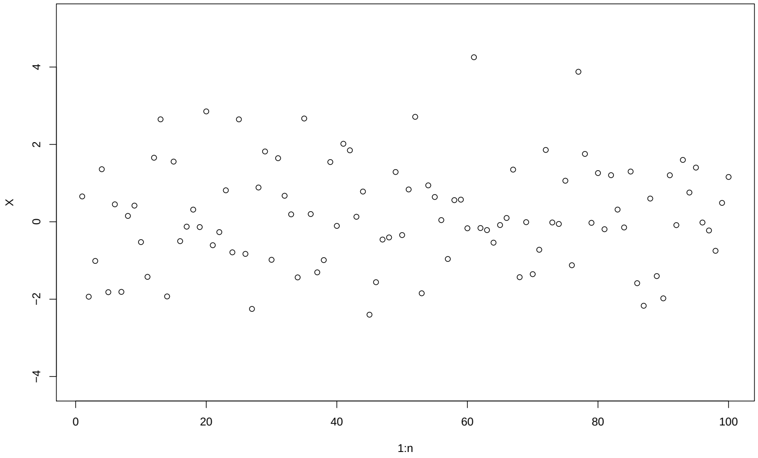
Sparsity assumption. Assume the vector θ is sparse in that

"only a small number of coordinates of θ are significant"

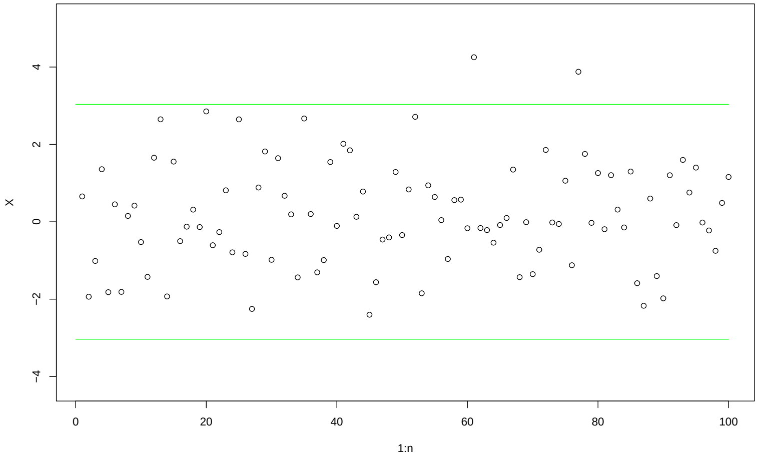
For instance, only at most p_n coefficients of θ are nonzero.

Objective. Estimate θ under sparsity assumption.

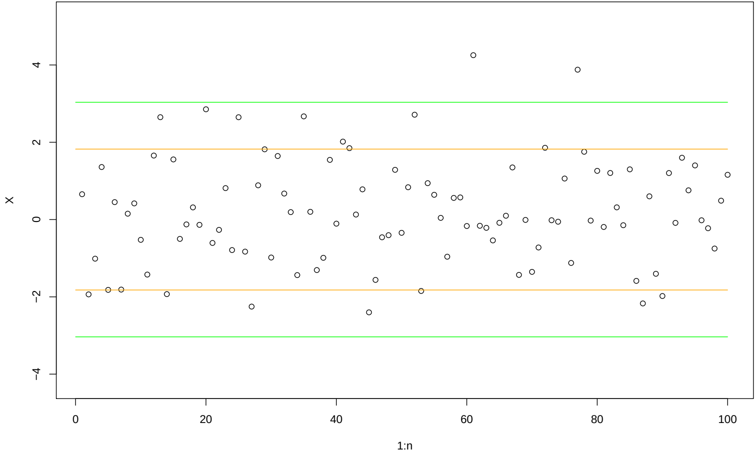
Example of data $n = 100$



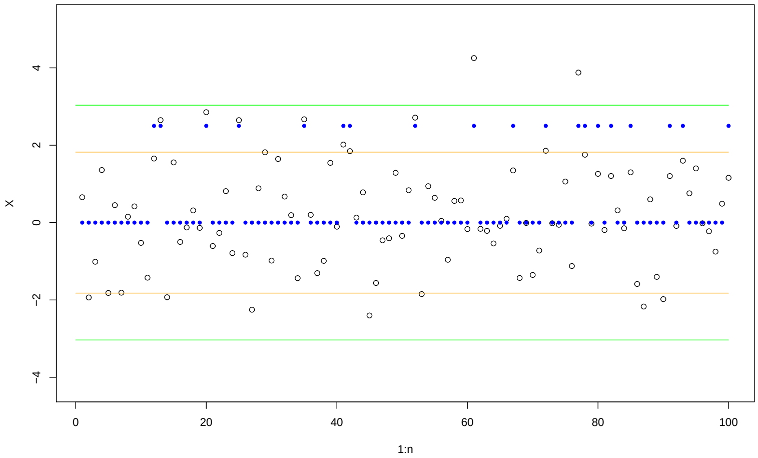
Example of data $n = 100$, Thresholding



Example of data $n = 100$, Oracle thresholding



Example of data $n = 100$, Original data $(p_n = 19)$



Bayes framework

Observations. $X^{(n)} = (X_1, \ldots, X_n)$ independent (but non i.d.)
Parameter space $\Theta = \mathbb{R}^n$, law $dP_\theta^{(n)} = p_\theta^{(n)}(X^{(n)})d\mathscr{L}^{(n)}$ with, here,

$$p_\theta^{(n)}(x_1, \ldots, x_n) = \prod_{i=1}^{n} \phi(x - \theta_i)$$

Bayesian framework. *Prior* Π on $\theta \in \mathbb{R}^n$.
This measure is updated with the data $X^{(n)}$.

The *posterior* given $X^{(n)}$ is the conditional distribution $\Pi(\cdot|X^{(n)})$.

Bayes formula. For any measurable B,

$$\Pi(B|X^{(n)}) = \frac{\int_B p_\theta^{(n)}(X^{(n)})d\pi(\theta)}{\int p_\theta^{(n)}(X^{(n)})d\pi(\theta)}.$$

Posterior converges at rate (at least) $\varepsilon_n \to 0$ for distance d if

$$P_{n,\eta_0}\Pi(\eta : d(\eta, \eta_0) > \varepsilon_n|X^{(n)}) \underset{n\to+\infty}{\longrightarrow} 0.$$

Posterior distribution and aspects of it

Object of interest the posterior distribution $\Pi[\cdot \,|\, X^{(n)}]$

Simulation Sampling from the posterior ! (e.g. via a MCMC method, or any method)
Repeated sampling from the posterior gives an idea of "spread"
Can suggest Credible regions

Aspects of the posterior $\Pi[\cdot \,|\, X^{(n)}]$

- Posterior mean $\int \theta d\Pi(\theta | X^{(n)})$
- Posterior (coordinatewise)-median
- Posterior mode, etc.

Remark Posterior and aspects of it might behave differently
especially in high-dimensional problems

Bayesian method in sparsity context ?

Objectives

- Define a prior distribution Π on $\theta \in \mathbb{R}^n$ which would
  - be adapted to estimation of sparse vectors
  - automatically adapts to the unknown sparsity level p_n
- Find necessary and sufficient conditions on the prior so that the preceding holds.
- One would also like to simulate from the posterior distribution ...

Succint Bibliography

- Thresholding methods [Donoho & Johnstone] (90's), ...
- Penalization methods [Birgé & Massart] (90's), [Golubev] (2000), ...
- False Discovery Rate (FDR) [Abramovich et al.] (2006)
- Empirical Bayes method [Johnstone & Silverman] (2004)
  - Prior distribution
    $$\bigotimes_{i=1}^{n}(1-\alpha_n)\delta_0+\alpha_n\gamma,$$
    for some continuous distribution γ.
  - Leads to some posterior depending on α_n
  - Estimate α_n from the data : $\hat{\alpha}_n$
  - Plug-in $\hat{\alpha}_n$ into the expression of posterior expectation
- Bayesian t-estimation [Abramovich et al.] (2007)
- NP Empirical Bayes [W. Jiang & C.-H. Zhang] (2009)

Succint Bibliography

- Thresholding methods [Donoho & Johnstone] (90's), ...
- Penalization methods [Birgé & Massart] (90's), [Golubev] (2000), ...
- False Discovery Rate (FDR) [Abramovich et al.] (2006)
- Empirical Bayes method [Johnstone & Silverman] (2004)
  - Prior distribution
    $$\bigotimes_{i=1}^{n}(1-\alpha_n)\delta_0+\alpha_n\gamma,$$
    for some continuous distribution γ.
  - Leads to some posterior depending on α_n
  - Estimate α_n from the data : $\hat{\alpha}_n$
  - Plug-in $\hat{\alpha}_n$ into the expression of posterior expectation
- Bayesian t-estimation [Abramovich et al.] (2007)
- NP Empirical Bayes [W. Jiang & C.-H. Zhang] (2009)

What about a fully Bayes method ?

Prior and assumptions

A class of sparse priors Π

Definition

1. Pick an integer k under $\pi_n(\cdot)$ law on $\{0, \ldots, n\}$
2. Given k pick *uniformly* at random $S \subset \{1, \ldots, n\}$ of cardinality k
   for $|S| = k$, $\Pi_n(S|\, k) = 1/\binom{n}{k}$
3. Given S, define $\theta_S = (\theta_i)_{i \in S}$ and $\theta_{S^c} = (\theta_i)_{i \notin S}$ by

$$\begin{aligned} \theta_S &\sim g_S \quad \text{density on } \mathbb{R}^S \\ \theta_{S^c} &= 0 \end{aligned}$$

The resulting prior Π on $\theta = (\theta_1, \ldots, \theta_n) \in \mathbb{R}^n$ is completely determined by

- the law of the size k of the picked subset $S \sim \pi_n(\cdot)$
- the collection of densities $\{g_S\}_{S \subset \{1,\ldots,n\}}$

Sparse prior Π, examples

Example (α_n-Coin-flipping prior)

$$
\begin{aligned}
k &\sim \mathscr{B}(n, \alpha_n) \\
g_S &= g^{\otimes |S|}
\end{aligned}
$$

$$\Updownarrow$$

$$\Pi \sim \bigotimes_{i=1}^{n} (1-\alpha_n)\delta_0 + \alpha_n g$$

Problem How does one choose α_n ??

"Bayesian Thresholding"
at level α_n

Sparse prior Π, examples

Example (α_n-Coin-flipping prior)

$$\begin{aligned} k &\sim \mathscr{B}(n, \alpha_n) \\ g_S &= g^{\otimes |S|} \end{aligned}$$

$$\Updownarrow$$

$$\Pi \sim \bigotimes_{i=1}^{n} (1 - \alpha_n)\delta_0 + \alpha_n g$$

Problem How does one choose α_n ??

Example (Bayes Coin-flipping)

$$\begin{aligned} \alpha &\sim \text{Beta}(1, n) \\ k \mid \alpha &\sim \mathscr{B}(n, \alpha) \\ g_S &= g^{\otimes |S|} \end{aligned}$$

$$\Updownarrow$$

$$\begin{aligned} \alpha &\sim \text{Beta}(1, n) \\ \Pi | \alpha &\sim \bigotimes_{i=1}^{n} (1 - \alpha)\delta_0 + \alpha g \end{aligned}$$

"Bayesian Thresholding"
at level α_n

"Bayesian Thresholding"
with automatic threshold choice

Sparse prior Π, examples

Remark "Bayesian Thresholding" induces a Beta-Binomial prior on dimension which behaves like $\pi_n(k = p) \approx e^{-p}$

Sparse prior Π, examples

Remark "Bayesian Thresholding" induces a Beta-Binomial prior on dimension which behaves like $\pi_n(k = p) \approx e^{-p}$

Example (Many other possibilities !)

- For the law $\pi_n(\cdot)$
  - $\pi_n(k = p) \propto e^{-p \log p}$
  - $\pi_n(k = p) \propto e^{-p \log n/p}$ ...
- For the continuous density g_S, a possibility is $g = \otimes_S g_S$, with
  - $g(x) \propto e^{-x^2}$ (Gaussian)
  - $g(x) \propto e^{-|x|}$ (Laplace)
  - $g(x) \propto (1 + x^2)^{-1}$ (Cauchy) ...
- Another possibility for g_S is mixing densities (i.e. g_S is not a coordinatewise product)

Which ones of all these priors work ?

Convergence rates

Classes of sparse signals

Definition

Nearly-black class of vectors

$$\ell_0[p_n] = \{\theta \in \mathbb{R}^n,\ \#(1 \le i \le n : \theta_i \ne 0) \le p_n\}.$$

Sparsity coefficient $\eta_n = p_n/n$

Distance on $\mathbb{R}^n$: euclidian norm $\|\cdot\|_2 = \|\cdot\|$

$$\|\theta - \psi\|^2 = \sum_{i=1}^{n} (\theta_i - \psi_i)^2.$$

Minimax rate in $\ell_0[p_n]$ for squared $\|\cdot\|_2$-norm, as $n \to +\infty$

$$\inf_{\hat\theta} \sup_{\theta \in \ell_0[p_n]} P_{n,\theta} \|\hat\theta - \theta\|_2^2 = 2 p_n \log(n/p_n)(1 + o(1)).$$

ℓ_q-type distances $0 < q < 2$ can also be considered

$$d_q(\theta, \psi) = \sum_{i=1}^{n} |\theta_i - \psi_i|^q.$$

Classes of sparse signals

Strong and weak ℓ_r-balls $r \in (0,2)$. Let $\theta_{(1)} \geq \theta_{(2)} \geq \cdots \geq \theta_{(n)}$

$$\begin{aligned} \ell_r[p_n] &= \left\{\theta \in \mathbb{R}^n,\ \sum_{i=1}^n |\theta_i|^r \leq n\left(\frac{p_n}{n}\right)^r\right\} \\ m_r[p_n] &= \left\{\theta \in \mathbb{R}^n,\ |\theta_{(i)}|^r \leq \frac{n}{i}\left(\frac{p_n}{n}\right)^r, \quad i = 1, \ldots, n\right\}. \end{aligned}$$

Minimax rates for $0 < r < 2$ for $\|\cdot\|_2$-norm, with $\eta_n = p_n/n$

- for $\ell_r[\eta_n]$ minimax rate is $\sim n\eta_n^r(\sqrt{2\log\eta_n^{-r}})^{2-r}$ $(n \to +\infty)$
- for $m_r[\eta_n]$ minimax rate is $\sim \frac{2}{2-r}R_n(\ell_r[\eta_n])$ $(n \to +\infty)$

Assumptions (P) on the prior Π

Our prior Π is defined by specifying

- The discrete law $\pi_n(k = \cdot)$ of k = number of coefficients chosen
- The continuous law g_S on the chosen subspace $\mathbb{R}^S$

Assumption (P)

We assume that g_S is positive, $g_S(\theta) = e^{-h_S(\theta)}$ and

- does not have too light tails in that

$$\log g_S(\theta) - \log g_S(\theta') \lesssim |S| + \sqrt{|S|}\|\theta - \theta'\|, \qquad \forall S, \forall \theta, \theta' \in \mathbb{R}^S,$$

- has some approximate subspace compatibility in the sense

$$\left|\log g_S(\theta) - \log g_{S'}(\pi_{S'}\theta)\right| \lesssim |S| + \sqrt{|S|}\|\pi_{S-S'}\theta\|, \qquad \forall S' \subset S, \forall \theta \in \mathbb{R}^S,$$

with $\pi_S\theta = \theta_S = (\theta_i : i \in S)$.

Rate theorem I

Theorem

Assume the prior satisfies (P) then for any $n \geq 1$ and any $r > 1$,

$$\sup_{\theta_0 \in \ell_0[p_n]} P_{n,\theta_0} \Pi_n\left(\theta : \|\theta - \theta_0\| > 10r \,|\, X^{(n)}\right) \leq e^{-r^2/9}\left(C_n(r, \pi_n, p_n) + 1\right).$$

$$C_n(r; \pi_n, p_n) = \kappa e^{cp_n} \frac{\sum_{p=1}^n \left(\pi_n(p) \binom{n}{p} \left(1 \vee r^2/p\right)^{p/2}\right)^{1/2}}{\left(\sum_{p=p_n}^n \frac{\binom{n-p_n}{p-p_n}}{\binom{n}{p}} \pi_n(p)(dr^2/p)^{p/2}\right)^{1/2}},$$

Rate theorem I, asymptotics

Denote ${r_n^*}^2 = p_n \log(n/p_n)$

Corollary

Assume $\{g_S\}_S$ satisfies (P) and π_n satisfies

$$\sum_{p=1}^{n} \sqrt{\pi_n(p)\binom{n}{p} C_1^p} \leq e^{C_2 {r_n^*}^2}$$

$$\pi_n(p_n) \geq e^{-C_3 {r_n^*}^2}$$

Then for M large enough, as $n \to +\infty$.

$$\sup_{\theta_0 \in \ell_0[p_n]} P_{n,\theta_0} \Pi(\|\theta - \theta_0\| > M r_n^* \,|\, X^{(n)}) \to 0$$

Examples of application

The following choices lead to the optimal rate on $\ell_0[p_n]$

- For the prior $\pi_n(k = \cdot)$ a natural choice is
  - $\pi_n(k = p) = \binom{n}{p}^{-1}$
  - or $\pi_n(k = p) = e^{-p \log(nc/p)}$ for some $c > 0$.
- For the continuous part g_S, product priors $g^{\otimes S}$ with $g = e^{-h}$ and
  - $|h(x) - h(y)| \lesssim 1 + |x - y| \quad \forall x, y \in \mathbb{R}$

  For instance, as soon as

  - the tails of g are at least as heavy as Laplace

  then conditions (P) holds.

Mixing priors

Example

1. few mixing

$$g_{|S|}(\theta) = a_{|S|} \frac{e^{-\|\theta_S\|_1}}{1 + \|\theta_S\|_2^2} \qquad \text{satisfies (P)}$$

2. rotationally symmetric priors
   Set $p = |S|$. Let r_p a density on $\mathbb{R}$

$$g_p(\theta) = \frac{r_p\left(\|\theta\|\right)}{p v_p \|\theta\|^{p-1}},$$

   The Gamma$(p,1)$-density r_p leads to

$$g_p(\theta) = \frac{e^{-\|\theta\|}\Gamma(p/2+1)}{\pi^{p/2}\Gamma(p+1)}, \quad \text{satisfies (P) with extra } \log p$$

Aspects of posterior under complexity prior

- $\hat{\theta}^{PM} = \int \theta d\Pi_n(\theta | X^{(n)})$ posterior mean
- $m(X^{(n)})$ posterior coordinatewise median

Corollary

Assume $\{g_S\}_S$ satisfies (P) and π_n satisfies $\pi_n(p) \lesssim e^{-ap\log(bn/p)}$ for large constants a, b. Then it holds, as $n \to +\infty$, with $r_n^* = p_n \log(n/p_n)$,

$$\sup_{\theta_0 \in \ell_0[p_n]} P_{n,\theta_0} \left\| \hat{\theta}^{PM} - \theta_0 \right\|^2 \lesssim r_n^{*2}$$

$$\sup_{\theta_0 \in \ell_0[p_n]} P_{n,\theta_0} \left\| m(X^{(n)}) - \theta_0 \right\|^2 \lesssim r_n^{*2}$$

Rate Theorem II

Can we go beyond $\pi_n(k) = \exp(-k \log(n/k))$?

Rate Theorem II

Can we go beyond $\pi_n(k) = \exp(-k \log(n/k))$?

Yes if slightly more stringent conditions on the mixing of g_S ...

Case of product $g_S = g \otimes \ldots \otimes g$

Definition

- $S_\theta = \{i,\ \theta_i \neq 0\}$ support of $\theta \in \ell_0[p_n]$. Denote $S_0 = S_{\theta_0}$

$$\{1, \ldots, n\} = S_0 \cup S_0^c$$

- $\pi_{n,k}$ prior on dimensions induced on S_0^c, given that $|S_\theta \cap S_0| = k$

$$\nu_k := \sum_{p=0}^{n-k} p\pi_{n,k}(p)$$

Rate Theorem II

Condition (M)

Assume π_n is such that for some $d < 1$, for any $p > Cp_n$ ($C > 1$),

$$\pi_n(p) \leq d\pi_n(p-1)$$

("exponentially decreasing")

Lemma (Dimension reduction)

Assume condition (M). Then for large enough C, as $n \to +\infty$,

$$P_{n,\theta_0}\Pi_n(\theta : \ |S_\theta| \geq Cp_n | X) \to 0.$$

Rate Theorem II

Theorem

Assume the prior satisfies (P)+(M) and that $g_S = \otimes_S g$. Set $r_n^* = p_n \log(n/p_n)$. Then for M large enough, as $n \to +\infty$,

$$\sup_{\theta_0 \in \ell_0[p_n]} P_{n,\theta_0} \Pi_n(\theta : \|\theta - \theta_0\| > M r_n^* | X) \to 0.$$

Idea of the proof

- Small k's : argue as in Theorem 1
- Large k : use Lemma 1 to get $\Pi_n(k > C p_n | X) \to 0$

Remark. Can be extended to mixing priors up to extra condition on marginals of g_S.

Rate Theorem II

Theorem

Assume the prior satisfies (P)+(M) and that $g_S = \otimes_S g$*. Set* $r_n^* = p_n \log(n/p_n)$*. Then for* M *large enough, as* $n \to +\infty$*,*

$$\sup_{\theta_0 \in \ell_0[p_n]} P_{n,\theta_0} \Pi_n(\theta : \|\theta - \theta_0\| > M r_n^* \,|\, X) \to 0.$$

Corollary

Bayesian Hard Theresholding defined by

$$\alpha \sim Beta(1, n) \quad \text{and} \quad \Pi|\alpha \sim \bigotimes_{i=1}^{n} (1-\alpha)\delta_0 + \alpha g$$

with g *the Laplace density (for instance) is rate optimal*

Indeed, the induced π_n verifies $\pi_n(p) \propto \binom{2n-p}{n}$ and $\binom{2n-p}{n} \approx e^{-p/2}$. Satisfies (M)

Lower bounds

Theorem

Take $g_S = g^{\otimes S}$ *with* $g(y) \propto e^{-|y|^\alpha}$ *and* $\pi_n(p_n) \geq e^{-cp_n \log(n/p_n)}$.

- *if* $\alpha \geq 2$ *and* $\|\theta_0^n\| \to \infty$ *fast enough, then for small universal* $\eta > 0$,

$$P_{n,\theta_0^n}\Pi_n(\theta : \|\theta - \theta_0^n\| \leq \eta\|\theta_0^n\| \,|\, X^n) \to 0.$$

- *if* $1 < \alpha < 2$ *set* $\rho_{0,\alpha}^n = \left(\frac{\|\theta_0^n\|_\alpha^\alpha}{\|\theta_0^n\|_2^2} \wedge 1\right) \|\theta_0^n\|_\alpha p_n^{\frac{1}{2}-\frac{1}{\alpha}}$. *If* $\rho_{0,\alpha}^n \to \infty$ *fast enough*

$$P_{n,\theta_0^n}\Pi_n(\theta : \|\theta - \theta_0^n\| \leq \eta\rho_{0,\alpha}^n \,|\, X^n) \to 0,$$

for $\eta > 0$ *small enough*

Consequence Tails of g should be as least as heavy as Laplace.

Example

Taking $g = \varphi$ standard Gaussian is suboptimal. Tails are too light.

The mean/median phenomenon, a surprise ?

Consider estimation of $\theta \in \ell_0[p_n]$ for the d_q-distance, for some $0 < q < 2$.

$$d_q(\theta, \psi) = \sum_{i=1}^{n} |\theta_i - \psi_i|^q.$$

Minimax risk $r^*_{n,q} := \inf_{\hat\theta} \sup_{\theta \in \ell_0[p_n]} P_{n,\theta} d_q(\hat\theta, \theta) = O(p_n \log^{q/2}(n/p_n))$

Johnstone-Silverman (04) show that

- Their posterior *median* plug-in $\hat\theta^{med}(\hat\alpha_n)$ converges at rate $r^*_{n,q}$, any $0 < q < 2$
- Their posterior *mean* plug-in $\hat\theta^{mean}(\hat\alpha_n)$ has suboptimal rate if $q < 1$.

Even taking the "oracle" level $\alpha_n = \alpha_n^{oracle} = p_n/n$, one can check that

- $\hat\theta^{mean}(\alpha_n^{oracle})$ converges at suboptimal rate for any $q < 1$

The mean/median phenomenon, a surprise ?

Theorem

Under the conditions of Rate Theorem II, the posterior measure does converge *at optimal rate* $r^*_{n,q}$, *any* $0 < q < 2$

$$P_{n,\theta_0^n}\Pi(\theta : \ d_q(\theta, \theta_0) > Mr^*_{n,q} \,|\, X) \to 0$$

In particular, applying the result for instance to the oracle estimator $\hat{\theta}^{mean}(\alpha_n^{oracle})$,

- Its posterior measure converges at optimal rate $r^*_{n,q}$ over $\ell_0[p_n]$. $\leq p_n \log^{q/2}(n/p_n)$
- Its posterior mean converges at suboptimal rate, any $q < 1$ $\geq n(p_n/n)^q$

Posterior measure and posterior mean
have fairly different behaviors in this case

Algorithm

Algorithm

The posterior probability $\Pi_n(B| X^{(n)})$ of a Borel set B is

$$\frac{\sum_{p=0}^{n} \pi_n(p) \binom{n}{p}^{-1} \sum_{|S|=p} \prod_{i \notin S} \phi(X_i) \int_{(\theta_S,0)\in B} \prod_{i\in S} \phi(X_i - \theta_i)\, g_S(\theta_S) \prod_{i\in S} d\theta_i}{\sum_{p=0}^{n} \pi_n(p) \binom{n}{p}^{-1} \sum_{|S|=p} \prod_{i \notin S} \phi(X_i) \int \prod_{i\in S} \phi(X_i - \theta_i)\, g_S(\theta_S) \prod_{i\in S} d\theta_i}.$$

The posterior mean is the vector

$$\hat{\theta}^{PM} = \left(\int \theta_1 d\Pi_n(\theta| X^{(n)}), \ldots, \int \theta_n d\Pi_n(\theta| X^{(n)}) \right)$$

At first sight, the number of computations is of the order of 2^n ...

Algorithm

Assume g_S is of the product form $g^{\otimes S}$. Then

$$\hat{\theta}_1^{PM} = \frac{\sum_{p=0}^{n} \pi_n(S_p)\zeta(X_1) \sum_{|S|=p,\, 1\in S} \prod_{i\notin S,\, i\neq 1} \phi(X_i) \prod_{i\in S,\, i\neq 1} \psi(X_i)}{\sum_{p=0}^{n} \pi_n(S_p) \sum_{|S|=p} \prod_{i\notin S} \phi(X_i) \prod_{i\in S} \psi(X_i)},$$

with

- $\pi_n(S_p) = \pi_n(p)\binom{n}{p}^{-1}$ prior mass of any model of size p
- $\psi(X_i) = \int \phi(X_i - \theta_i)\, g(\theta_i)\, d\theta_i$
- $\zeta(X_1) = \int \theta_1 \phi(X_1 - \theta_1) g(\theta_1) d\theta_1$

Algorithm

Remark that

$$\sum_{|S|=p} \prod_{i \notin S} \phi(X_i) \prod_{i \in S} \psi(X_i)$$

is nothing but the coefficient in front of Z^p in the polynomial

$$\prod_{i=1}^{n} (\phi(X_i) + \psi(X_i) Z)$$

and, similarly,

$$\sum_{|S|=p,\, 1 \in S} \prod_{i \notin S,\, i \neq 1} \phi(X_i) \prod_{i \in S,\, i \neq 1} \psi(X_i)$$

is the coefficient in front of Z^p in the polynomial

$$\prod_{i=2}^{n} (\phi(X_i) + \psi(X_i) Z)$$

It is thus possible to

- Compute explicitly the posterior (mean)
- by just computing the product of polynomials

$$\prod_{i=1}^{n}(\phi(X_i) + \psi(X_i)Z)$$

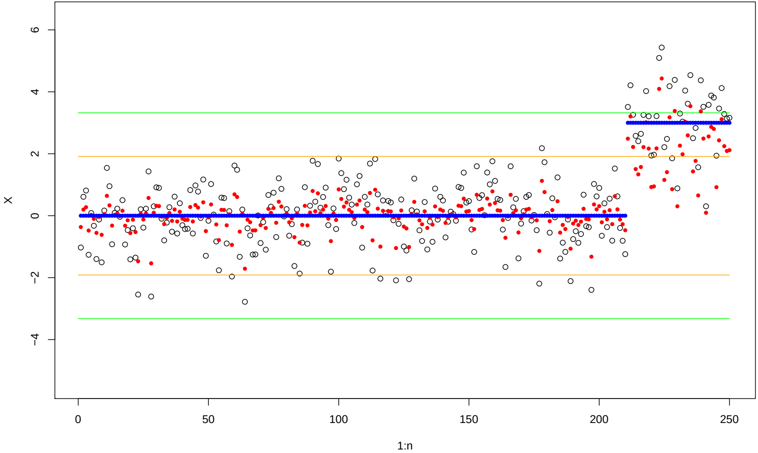
- assuming that g_S is of product form (and $\pi_n(S)$ only depend on $|S|$)

Remark The posterior is not of product form in general.

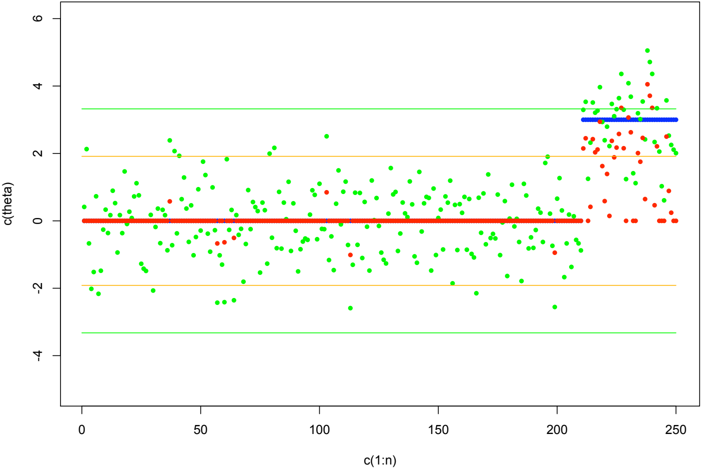
Simulation results For not too large n's, ($n \lesssim 800$), one can easily implement the method. The resulting estimator $\hat{\theta}^{PM}$

- is significantly better than Hard Thresholding
- is competitive with EBayesThresh algorithm
  from J-S 04 using Empirical Bayes

Posterior mean $n = 250,\ p_n = 40,\ A = 3$



Posterior coordinatewise-median $n = 250,\ p_n = 40,\ A = 3$



High dimensional linear model

[Work in progress with Johannes Schmidt-Hieber & Aad van der Vaart]

Let $\theta \in R^M$, $X \in \mathbb{R}^{n \times M}$, $M \gg n$

$$Y = X\theta + \varepsilon, \quad \text{with } \varepsilon \sim \mathcal{N}(0, I_n)$$

Sparsity Suppose θ has at most $s_n \ll n$ nonzero coefficients

Prior

- $\pi_M(k) = e^{-ak\log(M/k)}$ complexity-type prior on dimension
- $g_S = \otimes g$, with g Laplace, otherwise Dirac mass at 0

Concentration for θ under compatibility condition on X

$$\sup_{\theta_0 \in \ell_0[s_n]} P_{n,\theta_0} \Pi(\|\theta - \theta_0\|^2 > M s_n \log(M/s_n) | \, Y) \to 0.$$

Prediction result without compatibility with mild growth condition on θ

Conclusion

We propose a general Bayes method for the study of sparse sequences

We have identified

- some sufficient conditions for optimal convergence (upper bounds)
- some necessary conditions for optimality (lower bounds)

The method

- is flexible : lot of priors are optimal or nearly optimal
- allows non-independent priors
- can be implemented for some functionals of the posterior measure (more work needed for very large n's ...)