Some properties of the Gaussian Scale mixtures prior for Sparse models

Brown Bag Seminar WU

May 2017

J-B. Salomond

Université Paris-Est Créteil

Contents

Contents

Sparse sequence model in a Bayesian setting

Consider the well known Gaussian sequence model

$$X_i = \theta_i + \epsilon_i,\ \ \epsilon_i \sim \mathcal{N}(0,1),\ \ i = 1, \dots, n$$

and assume that the parameter $\theta = (\theta_1, \dots, \theta_n)$ is nearly black

$$p_n = \#\{i, \theta_i \neq 0\} = o(n)$$

Sparse sequence model in a Bayesian setting

Consider the well known Gaussian sequence model

$$X_i = \theta_i + \epsilon_i, \;\; \epsilon_i \sim \mathcal{N}(0,1), \;\; i = 1, \ldots, n$$

and assume that the parameter $\theta = (\theta_1, \ldots, \theta_n)$ is nearly black

$$p_n = \#\{i, \theta_i \neq 0\} = o(n)$$

Applications

Applications for this models are numerous

- Function estimation using wavelets

Sparse sequence model in a Bayesian setting

Consider the well known Gaussian sequence model

$$X_i = \theta_i + \epsilon_i,\ \ \epsilon_i \sim \mathcal{N}(0,1),\ \ i = 1, \ldots, n$$

and assume that the parameter $\theta = (\theta_1, \ldots, \theta_n)$ is nearly black

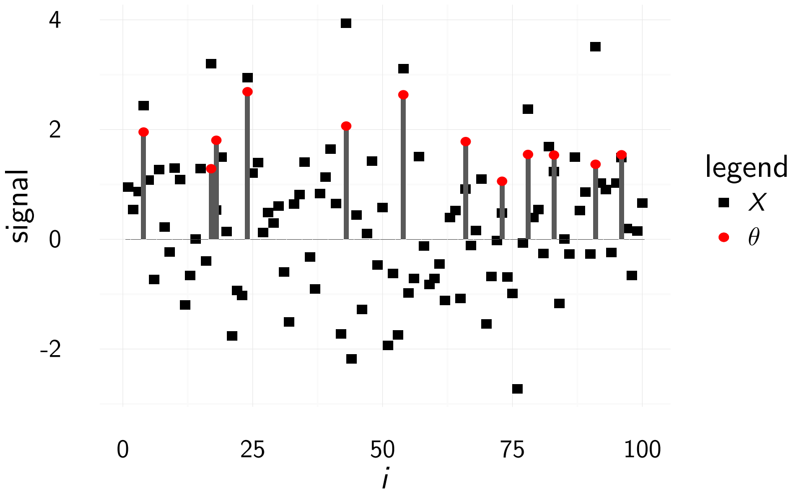
$$p_n = \#\{i, \theta_i \neq 0\} = o(n)$$

Applications

Applications for this models are numerous

- Function estimation using wavelets
- It is also a good way to study the behaviour of more complex sparse models

Example



Sparse sequence model in a Bayesian setting

A wide variety of both frequentist and Bayesian estimator have been proposed in the literature.

Sparse sequence model in a Bayesian setting

A wide variety of both frequentist and Bayesian estimator have been proposed in the literature.

Bayesian framework

In a Bayesian framework, the sparsity is induced through the prior (equivalent of the penalty term).

Sparse sequence model in a Bayesian setting

A wide variety of both frequentist and Bayesian estimator have been proposed in the literature.

Bayesian framework

In a Bayesian framework, the sparsity is induced through the prior (equivalent of the penalty term).

A first approach proposed in the literature is the two components model Spike and Slab

$$\theta_i \sim \lambda_i \delta_0 + (1 - \lambda_i)\pi_1$$

where π_1 has some heavy tails properties.

Prior - Normal scale mixture

Normal scale mixture

Consider a product prior on $\theta = (\theta_1, \dots, \theta_n)$

$$\sigma_i^2 \sim \pi$$
$$\theta_i \sim \mathcal{N}(0, \sigma_i^2)$$

Prior - Normal scale mixture

Normal scale mixture

Consider a product prior on $\theta = (\theta_1, \dots, \theta_n)$

$$\sigma_i^2 \sim \pi$$
$$\theta_i \sim \mathcal{N}(0, \sigma_i^2)$$

Examples of such priors :

- Horseshoe (Carvalho et al., 2010; van der Pas et al., 2014)
- Normal-Gamma (Caron and Doucet, 2008)
- Global-local scale mixtures (Ghosh and Chakrabarti, 2015)
- Spike and Slab Lasso (Ročková, 2015)
- ...

Prior - Normal scale mixture cnt'd

We are interested in the asymptotic properties of the posterior distribution and simultaneous testing procedures.

Questions

For the Normal scale mixture class of priors

$$p(\theta_i) = \int_{\mathbb{R}^+} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\theta_i^2}{2\sigma^2}} \pi(\sigma^2) d\sigma^2$$

what are the conditions on π such that our procedures have optimal asymptotic properties ?

Prior - Normal scale mixture cnt'd

We are interested in the asymptotic properties of the posterior distribution and simultaneous testing procedures.

Questions

For the Normal scale mixture class of priors

$$p(\theta_i) = \int_{\mathbb{R}^+} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\theta_i^2}{2\sigma^2}} \pi(\sigma^2) d\sigma^2$$

what are the conditions on π such that our procedures have optimal asymptotic properties ?

Qualitative answer :

- A lot of mass in a neighbourhood of 0 shrinkage effect
- Heavy tails counteract the shrinkage for large θ_i

Contents

Non-zeros coefficients I

Regular varying functions at infinity

We say that L is uniformly regular varying at infinity if there exist $R, u_0 > 1$ such that

$$\frac{1}{R} \leq \frac{L(au)}{L(u)} \leq R, \quad \forall a \in [1,2], \quad u > u_0$$

- Some examples : u^b, $\log^b(u)$
- Not uniformly varying : e^{au}

Non-zeros coefficients II

Condition 1

For some $b \geq 0$, $\pi(u) = L_n(u)e^{-bu}$ where L_n is uniformly regularly varying at 0, and

$$\pi(u) \gtrsim \left(\frac{p_n}{n}\right)^K e^{-b'u}, \quad \forall u > u_*$$

This condition assure the recovery of non-zeros coefficients

- The tails of π can decay exponentially fast
- The dependence on n of the prior should behave roughly as a power of p_n/n

Non-zeros coefficients III

Often practitioners are considering the following prior model

$$\theta|\sigma^2, \tau^2 \sim \mathcal{N}(0, \tau^2\sigma^2)$$
$$\sigma^2 \sim \pi'$$

and τ is an hyper-parameter. In this case the following condition implies condition 1

Condition 1'

π' is an uniformly regularly varying function and $\tau = (p_n/n)^K$

Zeros coefficients

A first condition to recover the 0 coefficients is

Condition 2

For some constant $c > 0$ we have $\int_0^1 \pi(u)du \geq c$

We need sufficient mass around 0

- This condition will induce a shrinkage of the posterior
- Form a modelling point of view, it makes sense since we assume that most of the coefficients are 0

Still zeros coefficients

A more surprising condition is the following

Condition 3

Let $s_n = \frac{p_n}{n}\sqrt{\log(n/p_n)}$ and let $b_n = \sqrt{\log(n/p_n)}$ then there exists $C > 0$ such that

$$\int_{s_n}^{\infty} \left(u \wedge \frac{b_n^3}{\sqrt{u}}\right) \pi(u)du + b_n \int_1^{b_n^2} \frac{\pi(u)}{\sqrt{u}} du \leq Cs_n$$

Details

- A fair part of the mass is in $[0, s_n]$

Still zeros coefficients

A more surprising condition is the following

Condition 3

Let $s_n = \frac{p_n}{n}\sqrt{\log(n/p_n)}$ and let $b_n = \sqrt{\log(n/p_n)}$ then there exists $C > 0$ such that

$$\int_{s_n}^{\infty} \left(u \wedge \frac{b_n^3}{\sqrt{u}} \right) \pi(u) du + b_n \int_1^{b_n^2} \frac{\pi(u)}{\sqrt{u}} du \leq C s_n$$

Details

- A fair part of the mass is in $[0, s_n]$
- π decays sufficiently fast outside $[0, s_n]$

Stronger conditions

Under the assumption that $p_n = o(n)$ the following two conditions implies conditions 2 and 3

Condition A

There exists C such that

$$\pi(u) \leq \frac{C}{u^{3/2}} \frac{p_n}{n} \sqrt{\log(p_n/n)}, \quad \forall u > s_n$$

Condition B

There exists C such that

$$\int_{s_n}^{\infty} \pi(u) \leq \frac{C p_n}{n}$$

Recovery

Non-Zero Coefficients

Under condition 1

$$\sup_{\theta_0 \in l_0(p_n)} \Pi\left(\sum_{i, \theta_{0,i} \neq 0} (\theta_i - \theta_{0,i})^2 > M_n p_n \log(n/p_n) | \mathbf{X}^n \right) \to 0$$

and

$$\sup_{\theta_0 \in l_0(p_n)} \sum_{i, \theta_{0,i} \neq 0} \mathbb{E}_0^n (\hat{\theta}_i - \theta_{0,i})^2 \lesssim p_n \log(n/p_n)$$

Recovery

Zero Coefficients

Under condition 2 and 3

$$\sup_{\theta_0 \in l_0(p_n)} \Pi\left(\sum_{i, \theta_{0,i}=0} (\theta_i - \theta_{0,i})^2 > M_n p_n \log(n/p_n) | \mathbf{X}^n \right) \to 0$$

and

$$\sup_{\theta_0 \in l_0(p_n)} \sum_{i, \theta_{0,i}=0} \mathbb{E}_0^n (\hat{\theta}_i - \theta_{0,i})^2 \lesssim p_n \log(n/p_n)$$

Sketch of the proof I

Using the hierarchical form of the prior we have that

$$\theta_i | X_i, \sigma_i^2 \sim \mathcal{N}\left(X_i \frac{\sigma_i^2}{1+\sigma_i^2}, \frac{\sigma_i^2}{1+\sigma_i^2} \right)$$

$$\pi(\sigma_i^2 | X_i) \propto (1+\sigma_i)^{-1/2} e^{X_i^2 \frac{\sigma_i}{1+\sigma_i}} \pi(\sigma_i)$$

To control the posterior mass of a set $B_n = \{||\theta - \theta_0||^2 \geq M_n p_n \log(n/p_n)\}$ we will simply use a Markov inequality

$$\Pi(B_n | X^n) \leq \frac{\mathbb{E}(||\theta - \theta_0||^2)}{M_n p_n \log(n/p_n)} = \frac{\sum_{i=1}^n \left(X_i \mathbb{E}(\frac{\sigma_i^2}{1+\sigma_i^2} | X_i) - \theta_{0,i} \right)^2 + \mathbb{V}(\theta_i | X_i)}{M_n p_n \log(n/p_n)}$$

Sketch of the proof II

We see that

1. We can separate the case $\theta_i = 0$ and $\theta_i \neq 0$
2. We only have to control $\mathbb{E}(\frac{\sigma_i^2}{1+\sigma_i^2}|X_i) := m_{X_i}$

We first consider the case $\theta_i = 0$. We show that under Conditions 1 and 2, we have the following bound for m_x

$$m_x \leq s_n \left(1 + \frac{\sqrt{2}C}{c} e^{\frac{x^2}{4}}\right) + q_n \frac{2\sqrt{2}C}{c} e^{\frac{x^2}{2}}$$

where $s_n = \frac{p_n}{n} \log(n/p_n)$ and $q_n = s_n(\log(n/p_n))^{-1/2}$. With this we can show that

$$\mathbb{E}(Xm_X)^2 \leq \frac{p_n}{n} \log(n/p_n)$$

Sketch of the proof III

We now consider $\theta_i \neq 0$. Note that because we only have p_n of them, we simply need to bound the bias and the variance by something of the order of $\log(n/p_n)$. We show that under condition 3 we have for $|x| > c_0 + \sqrt{2K(u_0 \vee 1)\log(n/p_n)}$

$$1 - m_x \leq \frac{C}{|x|}$$

Now note that

$$\mathbb{E}_{\theta_{0,i}}\left(X_i m_{X_i} - \theta_{0,i}\right) = \mathbb{E}_{\theta_{0,i}}\left(X_i(m_{X_i} - 1)\right).$$

This is enough to control the bias and the variance.

Contents

Multiple testing problems

We consider now the problem of selecting which components θ_i are non zero.

Questions

1. How to select the non-zero coefficient
2. How to assess the quality of the decision rule ?

An answer to 1 has been proposed in Carvalho et al. (2010). Recall that our prior is defined as

$$\sigma^2 \sim \pi$$
$$\theta | \sigma^2 \sim \mathcal{N}(0, \sigma^2)$$

Define $\kappa_i = \sigma_i^2/(1+\sigma_i^2)$ the shrinkage coefficient.

Shrinkage Coefficient

Recall that

$$\theta_i | \sigma_i^2, X_i \overset{ind}{\sim} \mathcal{N}(X_i \kappa_i, \kappa_i).$$

$\kappa_i = \frac{\sigma_i^2}{1+\sigma_i^2}$ is thus the coefficient that shrinks the MLE X_i. Carvalho et al. (2010) proposed the following selection rule : Chose θ_i to be non zero if

$$\mathbb{E}_i^\pi(\kappa_i | X_i) > 1/2$$

Multiple testing risk

We thus have the following decision rule $\delta_i = \mathbb{I}_{\mathbb{E}_i^\pi(\kappa_i|X_i) > \tau}$.

Multiple testing risk

We thus have the following decision rule $\delta_i = \mathbb{I}_{\mathbb{E}_i^\pi(\kappa_i|X_i) > \tau}$.We will consider a Bayesian classification risk to assess the quality of the multiple testing rule $\delta = (\delta_1, \dots, \delta_n)$
Bayesian Risk associated with a 2 group prior
$\mu : \theta_i \sim (1 - \frac{p_n}{n})\delta_0 + \frac{p_n}{n}\mathcal{N}(0, \psi^2)$ Thus

$$R_n^\psi(\delta) = \sum_{i=1}^{n} \left\{ (1 - \frac{p_n}{n})\mathcal{P}^{\mathcal{N}(0,1)}(\delta_i = 1) + \frac{p_n}{n}\mathcal{P}^{\mathcal{N}(0,1+\psi^2)}(\delta_i = 0) \right\}$$

Multiple testing risk

We thus have the following decision rule $\delta_i = \mathbb{I}_{\mathbb{E}_i^\pi(\kappa_i|X_i)>\tau}$.We will consider a Bayesian classification risk to assess the quality of the multiple testing rule $\delta = (\delta_1, \dots, \delta_n)$

Bayesian Risk associated with a 2 group prior $\mu : \theta_i \sim (1 - \frac{p_n}{n})\delta_0 + \frac{p_n}{n}\mathcal{N}(0, \psi^2)$ Thus

$$R_n^\psi(\delta) = \sum_{i=1}^n \left\{(1 - \frac{p_n}{n})\mathcal{P}^{\mathcal{N}(0,1)}(\delta_i = 1) + \frac{p_n}{n}\mathcal{P}^{\mathcal{N}(0,1+\psi^2)}(\delta_i = 0)\right\}$$

How does the decision rule behave for this risk under the previous conditions ?

Upper bound on the risk

Results

Under Conditions 1-3' we have for the decision rule $\delta_i = \mathbb{I}_{\mathbb{E}^{\pi}(\kappa_i|X_i)>\tau}$

$$R_n^{\psi_n}(\delta) \leq p_n \left(\frac{8\sqrt{\pi}C}{c\tau} + 2\Phi\left(\sqrt{2K(u_0 \vee 1)C_\psi}\right) - 1 \right)(1 + o(1))$$

if $\psi_n^2 = C_\psi \log(n/p_n)(1 + o(1))$

Upper bound on the risk

Results

Under Conditions 1-3' we have for the decision rule $\delta_i = \mathbb{I}_{\mathbb{E}^\pi(\kappa_i|X_i)>\tau}$

$$R_n^{\psi_n}(\delta) \leq p_n \left(\frac{8\sqrt{\pi}C}{c\tau} + 2\Phi\left(\sqrt{2K(u_0 \vee 1)C_\psi}\right) - 1 \right)(1 + o(1))$$

if $\psi_n^2 = C_\psi \log(n/p_n)(1 + o(1))$

Where K and u_0 are the constants from condition 1 and c and C are the constants in condition 2 and 3

Remarks

The constants for the Bayesian risk is almost sharp !

Bogdan et al. (2011) derived an Oracle and computed the optimal Bayes Risk

$$p_n \left(2\Phi(\sqrt{C_\psi}) - 1\right)(1 + o(1)),$$

here the best possible constant is $p_n \left(2\Phi(2\sqrt{C_\psi}) - 1\right)(1 + o(1))$ (but for a large class of priors !)

Sketch of the proof

Because the observations are independent, we simply have to control the Types I $t_1 = \mathcal{P}^{\mathcal{N}(0,1)}(\delta_i = 1)$ and Type II $t_2^{\psi} = P^{\mathcal{N}(0,1+\psi^2)}(\delta_i = 0)$ error for each test. Using the same notations as before we have

$$t_1 = \mathcal{P}^{\mathcal{N}(0,1)}(m_X \geq \tau)$$
$$t_2^{\psi} = P^{\mathcal{N}(0,1+\psi^2)}((1 - m_X) \geq 1 - \tau)$$

The proofs uses the same bounds presented before.

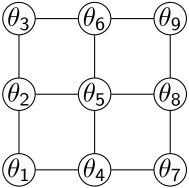
Contents

Extension - Known structure

In many cases, we have additional information on the structure of the parameter $(\theta_1, \dots, \theta_n)$.

Extension - Known structure

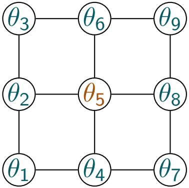
In many cases, we have additional information on the structure of the parameter $(\theta_1, \dots, \theta_n)$.
There is some way of taking advantage of this structure (e.g. fused lasso)



Example of a grid structure

Extension - Known structure

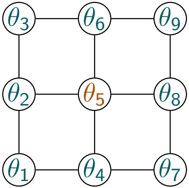
In many cases, we have additional information on the structure of the parameter $(\theta_1, \dots, \theta_n)$.
There is some way of taking advantage of this structure (e.g. fused lasso)



Example of a grid structure
If θ_5 is non zero, then there is high chances that $(\theta_1, \dots, \theta_9)$ are also non-zero.

Extension - Known structure

In many cases, we have additional information on the structure of the parameter $(\theta_1, \ldots, \theta_n)$.
There is some way of taking advantage of this structure (e.g. fused lasso)



Example of a grid structure
If θ_5 is non zero, then there is high chances that $(\theta_1, \ldots, \theta_9)$ are also non-zero.

This additional information can be easily introduced through the prior π on $(\sigma_1, \ldots, \sigma_n)$

A dependent prior

We consider the following depend prior

$$
\begin{aligned}
s_i &\sim \pi(s_i) \\
\sigma &= As \\
\theta &\sim \mathcal{N}_n(0, \mathrm{diag}(\sigma))
\end{aligned}
$$

where A is the adjacency matrix of the underlying graph.

A dependent prior

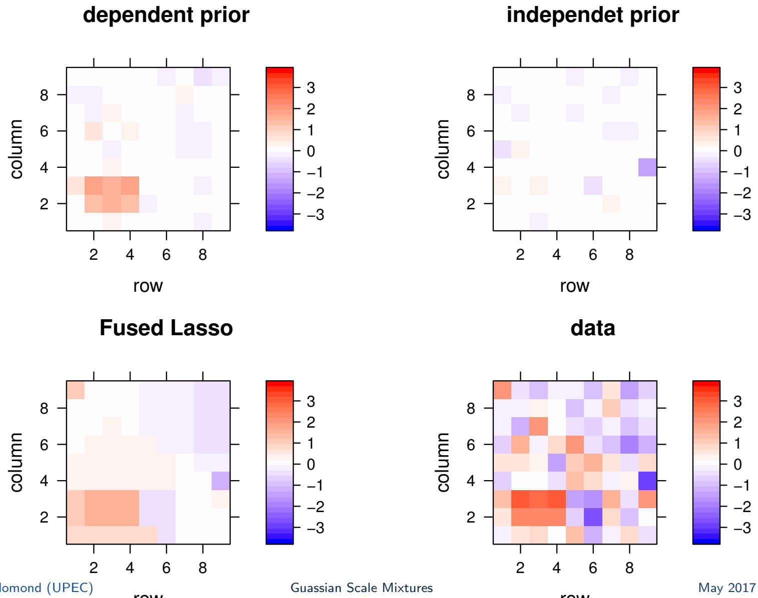
We consider the following depend prior

$$s_i \sim \pi(s_i)$$
$$\sigma = As$$
$$\theta \sim \mathcal{N}_n(0, \mathrm{diag}(\sigma))$$

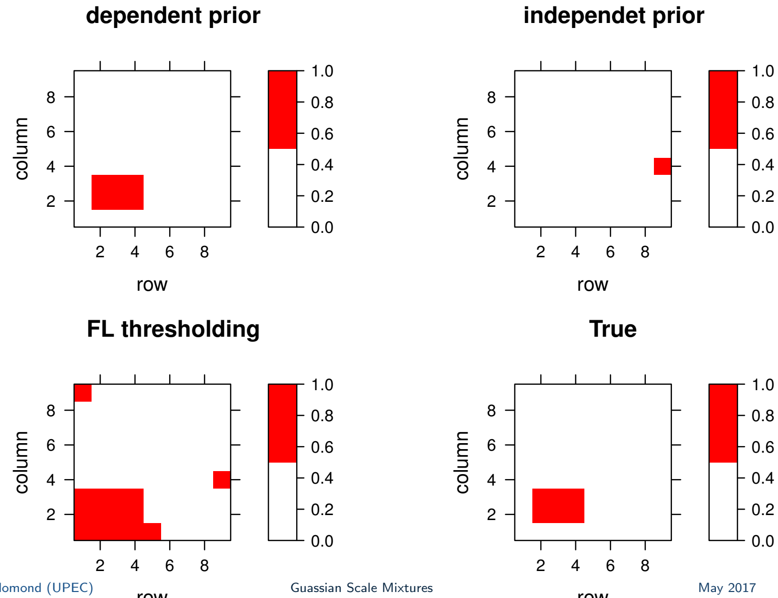
where A is the adjacency matrix of the underlying graph. We thus get the posterior

$$\pi(s_i|\mathbf{X}^n, s_{-i}) \propto \frac{1}{\prod_{i=1}^n \left(1 + \sum_{j=1}^n a_{i,j}s_j\right)^{1/2}} \exp\left(\frac{1}{2}\sum_{i=1}^n X_i^2 \frac{\sum_{j=1}^n a_{i,j}s_j}{1 + \sum_{j=1}^n a_{i,j}s_j}\right)\pi(s)$$

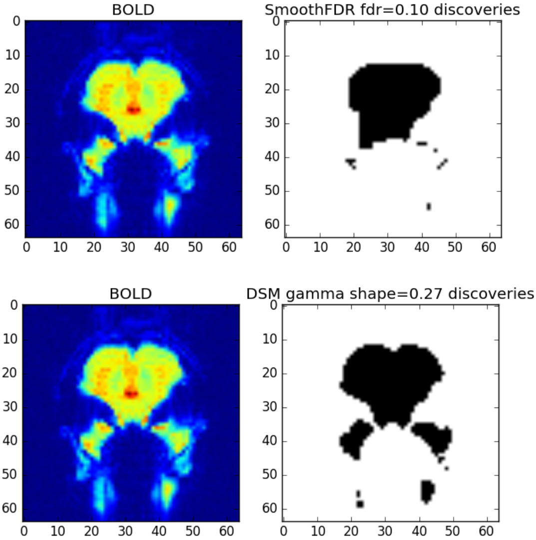
Numerical results - estimation



Numerical results - testing



Real data example



Contents

When considering multiple testing, one could also want to consider False Discovery rates.

False Discovery Rate

Recall that FDR is given by

$$FDR_n = \mathbb{E}\left(\frac{FD_n}{TD_n + FD_n}\right)$$

Similarly one could consider the False Non-discovery rate

$$FND_n = \mathbb{E}\left(\frac{FN_n}{p_n}\right)$$

Extensions and Perspectives II

Recently Rabinovich et al. (2017) studied a new risk defined as

$$R_n = FDR_n + FNR_n$$

Question

- Can we get an upper bound for this risk for the considered testing procedure ?
- Can we ensure that the Risk will tend to 0 uniformly over a certain set ?

Extensions and Perspectives III

One can also want to consider Gaussian linear model

$$X = Z\theta + \epsilon$$

where Z is a $m \times n$ matrix with $m \gg n$. In this case the proofs techniques developed so far cannot be used. Can we get contraction rates under similar conditions such as

Conditions for sparse linear model

$$\pi([s_p, \infty[) \leq s_p, \ \forall u > u_0, \pi(u) \geq \left(\frac{s}{p}\right)^K e^{-bu}$$

It seems that we can get the minimax contraction rate in this case work in progress...

Thank you for your attention !

References I

Bogdan, M., Chakrabarti, A., Frommlet, F., and Ghosh, J. K. (2011). Asymptotic bayes-optimality under sparsity of some multiple testing procedures. *Ann. Statist.*, 39(3) :1551–1579.

Caron, F. and Doucet, A. (2008). Sparse Bayesian nonparametric regression. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 88–95, New York, NY, USA. ACM.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97(2) :465–480.

Ghosh, P. and Chakrabarti, A. (2015). Posterior concentration properties of a general class of shrinkage estimators around nearly black vectors. arXiv :1412.8161v2.

Rabinovich, M., Ramdas, A., Jordan, M. I., and Wainwright, M. J. (2017). Optimal rates and tradeoffs in multiple testing. *arXiv preprint arXiv :1705.05391*.

Ročková, V. (2015). Bayesian estimation of sparse signals with a continuous spike-and-slab prior. submitted manuscript, available at `http://stat.wharton.upenn.edu/~vrockova/rockova2015.pdf`.

van der Pas, S., Kleijn, B., and van der Vaart, A. (2014). The horseshoe estimator : Posterior concentration around nearly black vectors. *Electron. J. Stat.*, 8 :2585–2618.

Condition 3 can be re-written as

$$\int_{s_n}^{1} u\pi(u)du + \int_{1}^{b_n^2} \left(u + \frac{b_n}{\sqrt{u}}\right) \pi(u)du + b_n^3 \int_{b_n^2}^{\infty} \frac{\pi(u)}{\sqrt{u}} du \leq Cs_n$$

Back