A Bayesian Method for Sparse Regression IMACS 2013

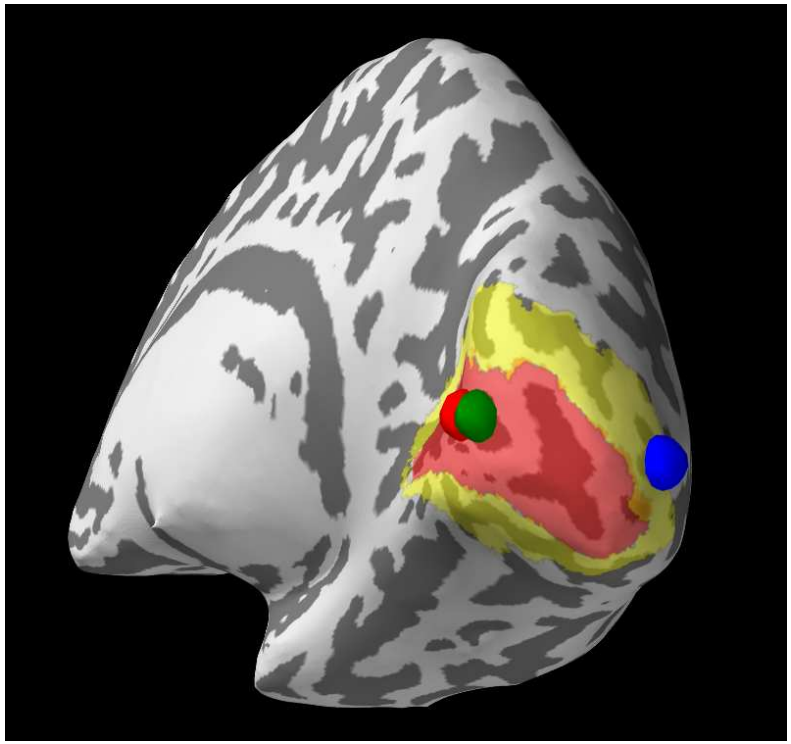
Amandine Schreck

under the supervision of Gersende Fort and Eric Moulines,
joint work with Sylvain Le Corff

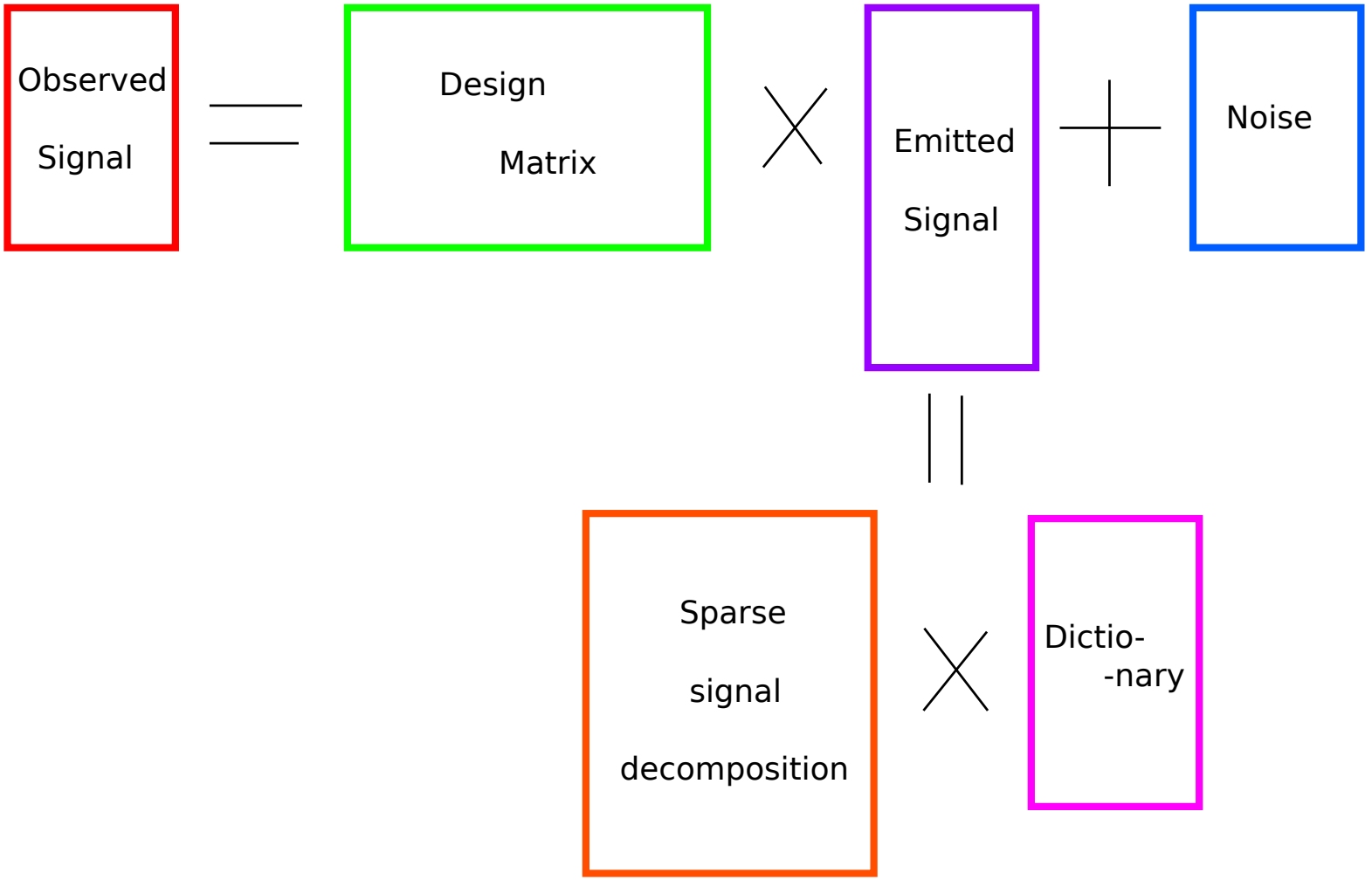
Télécom ParisTech

Thursday, July 18th, 2013

Goal : brain imaging - locate activated zones in a brain



Collaboration with Alexandre Gramfort on brain imaging problems.


Observed
Signal
=
Design
Matrix
×
Emitted
Signal
+
Noise
||
Sparse
signal
decomposition
×
Dictio-
-nary

- Goal: find the **active** (i.e. non-zero) components of the sparse signal decomposition.
- Difficulty: **high dimensional** setting, potentially low number of observations, high number of regressors.
- Existing solutions: deterministic methods (e.g. ISTA), transdimensional MCMC methods (Reversible Jump, Metropolised Carlin and Chib).

Simplified model :

$$Y = GX + \sqrt{\tau}E \ ,$$

where

- $Y \in \mathbb{R}^{N \times T}$ is the observed signal
- $G \in \mathbb{R}^{N \times P}$ is the design matrix (known)
- $X \in \mathbb{R}^{P \times T}$ is the emitted signal, directly assumed to be **sparse**
- $E \in \mathbb{R}^{N \times T}$ is a standard Gaussian noise

For concision of notations: $T = 1$.

X can be equivalently defined by (m, X_m) where

- $m = (m_1, \cdots, m_P) \in \mathcal{M} = \{0,1\}^P$ is the **model**, with $m_i = 0$ iff $X_i = 0$,
- $X_m \in \mathbb{R}^{|m|}$ collects the **active** rows of X, where $|m| = \sum_i m_i$.

$\rightarrow$ Sampling set:

$$\Theta = \bigcup_{m \in \mathcal{M}} \left(\{m\} \times \mathbb{R}^{|m|} \right) .$$

Likelihood and prior distributions:

- $\pi(Y|m, X_m) = (2\pi\tau)^{-N/2} \exp\left(-\frac{1}{\tau}\|Y - G_{\cdot m}X_m\|_2^2\right)$.
- $\pi(X_m|m) = \exp(-\lambda\|X_m\|_1 - |m|\log(c_\lambda))$, where $\lambda \geq 0$.
- $\pi(m) = w_m$, where $\sum_{m\in\mathcal{M}} w_m = 1$.

Posterior distribution on $\Theta = \bigcup_{m\in\mathcal{M}} \left(\{m\} \times \mathbb{R}^{|m|}\right)$:

$$\pi(m, X_m|Y) \propto w_m \, c_\lambda^{-|m|} \exp\left(-\frac{1}{2\tau}\|Y - G_{\cdot m}X_m\|_2^2 - \lambda\|X_m\|_1\right) .$$

Goal : propose a transdimensional MCMC method to sample the posterior distribution.

- Robust in **high dimensional settings**
- Can deal with non-differentiability in the penalization function
- In harmony with **sparsity** assumption

Goal of the **Proximal MALA** algorithm (PMALA): build a **Markov chain** converging to a target distribution with density of the form

$$\pi(x) \propto \exp(-g(x) - h(x)) \ ,$$

where

- g: continuously differentiable, convex, such that ∇g is L_g-Lipschitz,
- h: convex.

$\rightarrow$ Applied with $g(x) = \frac{1}{2\tau}\|Y - Gx\|_2^2$ and $h(x) = \lambda\|x\|_1$.

Ingredient 1: The **proximal gradient algorithm** (also known as the Iterative Shrinkage Thresholding Algorithm)

Goal: minimize $g + h$ where

- g: continuously differentiable, convex, such that ∇g is L_g-Lipschitz,
- h: convex

An iteration of the proximal gradient algorithm starting from x^t:

(1) Define a **local approximation** of $g + h$ at x^t by

$$Q_L(x^t, x) = h(x) + g(x^t) + \left\langle x - x^t, \nabla g(x^t) \right\rangle + \frac{L}{2}\|x - x^t\|_2^2 \ .$$

(2) Set $x^{t+1} = \operatorname{argmin}_x Q_L(x^t, x) = \operatorname{prox}_{h/L}\left(x^t - \frac{1}{L}\nabla g(x^t)\right)$, where

$$\operatorname{prox}_{\gamma \mathbf{h}}(\mathbf{u}) = \operatorname{argmin}_x \left(\gamma h(x) + \frac{1}{2}\|x - u\|_2^2\right) \ .$$

Ingredient 2: The **Metropolis Adjusted Langevin Algorithm** (MALA)

Goal: build a Markov chain converging to a target distribution with density $\pi(x) \propto \exp(-g(x))$, where g is differentiable.

An iteration of MALA starting from X^t:

(1) **Propose** a new point

$$Y^{t+1} = X^t - \frac{\sigma^2}{2}\nabla g(X^t) + \sigma W^{t+1} \ ,$$

where W^{t+1} is a random vector with i.i.d. entries from $\mathcal{N}(0,1)$.

(2) Classical **Acceptation/Rejection step**.

An iteration of **PMALA** starting from X^t:

(1) **Propose** a new point

$$Y^{t+1} = \text{prox}_{\sigma^2 \mathbf{h}/2}\left(X^t - \frac{\sigma^2}{2}\nabla g(X^t) + \sigma W^{t+1}\right),$$

where W^{t+1} is a random vector with i.i.d. entries from $\mathcal{N}(0,1)$.

(2) Classical **Acceptation/Rejection step**, with acceptance probability $\alpha(x,y) = \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}$, where $q(x,y)$ is the density of the proposal distribution (explicitly known).

Lemma

Let $\boldsymbol{\mu} \in \mathbb{R}^P$ *and* $\gamma, \sigma > 0$. *Set* $\boldsymbol{Y} = prox_{\gamma\|\cdot\|_1}(\boldsymbol{\mu} + \sigma \boldsymbol{W})$ *where* $\boldsymbol{W} \in \mathbb{R}^P$ *is a matrix of i.i.d random variables* $\sim \mathcal{N}(0,1)$. *The distribution of* $\boldsymbol{Y} \in \mathbb{R}^P$ *is given by*

$$\sum_{m \in \mathcal{M}} \left(\prod_{i \notin I_m} p_1(\mu_i)\, \delta_0(\mathrm{d}z_i) \right) \left(\prod_{i \in I_m} f_1(\mu_i, z_i) \mathrm{d}z_i \right) ,$$

where for any $c, z \in \mathbb{R}$,

$$p_1(c) = \mathbb{P}\left\{ |c + \xi| \leq \gamma \right\} \text{ , with } \xi \sim \mathcal{N}(0, \sigma^2) ,$$

$$f_1(c, z) = \left(2\pi\sigma^2\right)^{-1/2} \exp\left(-\frac{1}{2\sigma^2} \left| \left(1 + \frac{\gamma}{|z|}\right) z - c \right|_2^2 \right) .$$

An iteration of PMALA starting from x is equivalent to:

(i) sample $m' = (m'_1, \cdots, m'_P)$ with $(m'_i, i \in \{1, \cdots, P\})$ i.i.d. and such that m'_i is a Bernoulli r.v. with success parameter

$$1 - \mathbb{P}\left(\left| \left(x - \frac{\sigma^2}{2} \nabla g(x) \right)_i + \xi \right|^2 \leq \gamma \right) \qquad \xi \sim \mathcal{N}(0, \sigma^2) \, .$$

(ii) sample $y = (y_i)_{1 \leq i \leq P}$ in $\mathbb{R}^{|m'|}$ with independant components such that for any $i \in I_{m'}$, the distribution of y_i is proportional to

$$\exp\left(-\frac{1}{2\sigma^2} \left| \left(1 + \frac{\gamma}{|y_i|} \right) y_i - \left(x - \frac{\sigma^2}{2} \nabla g(x) \right)_i \right|^2 \right) \, .$$

The data: $Y = GX + \sqrt{\tau}E$

- The components of E are samples of $\mathcal{N}(0,1)$
- $X = (X_i)_{1 \leq i \leq P}$ with $X_i = \mathbf{1}_{i \leq S}$ with S depending on the example.
- Columns of $G \in \mathbb{R}^{N \times P}$:
  - uncorrelated designs: independant Gaussian samples.
  - correlated designs: independant Gaussian samples or linear combinations of other columns plus Gaussian vectors.

Implementation parameters:

- Prior on the models: uniform.
- Starting point: empty model.

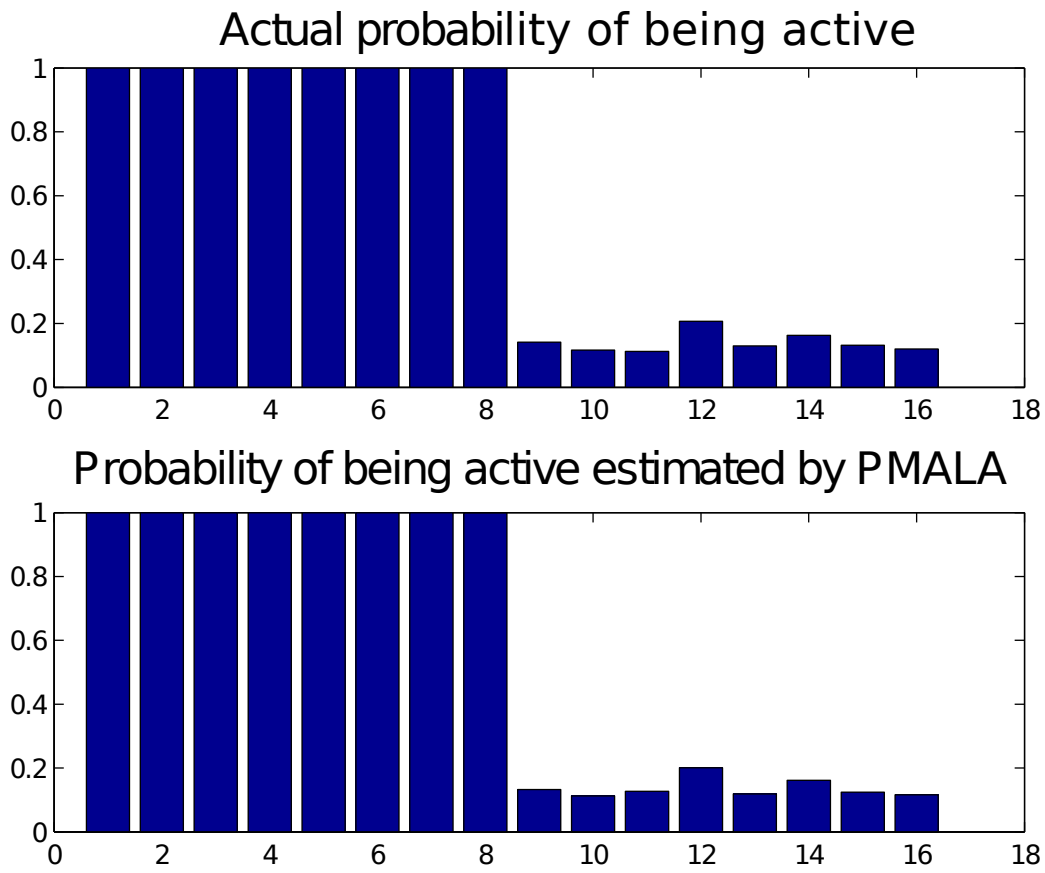


Figure: Posterior probability of activation in 16 dimensions, with uncorrelated design (i.e. $\mathbb{P}(x_i \neq 0|M)$ for $1 \leq i \leq 16$); S=8, P=16, N=100.

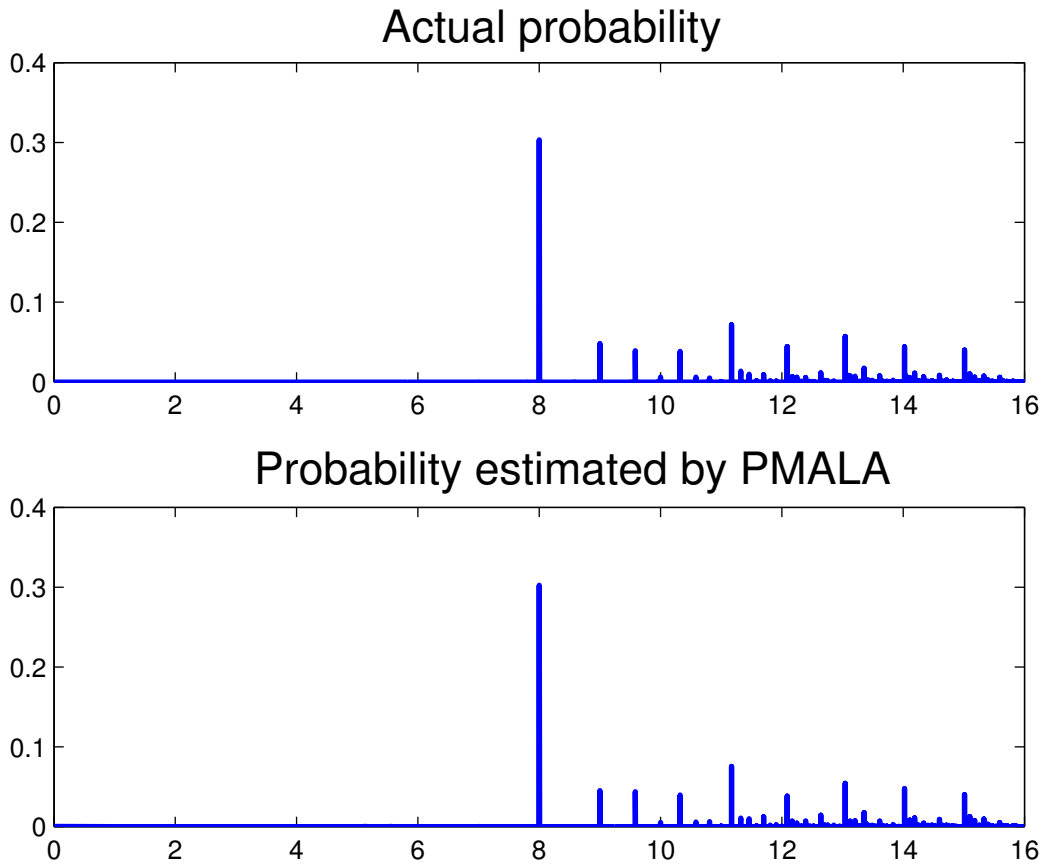
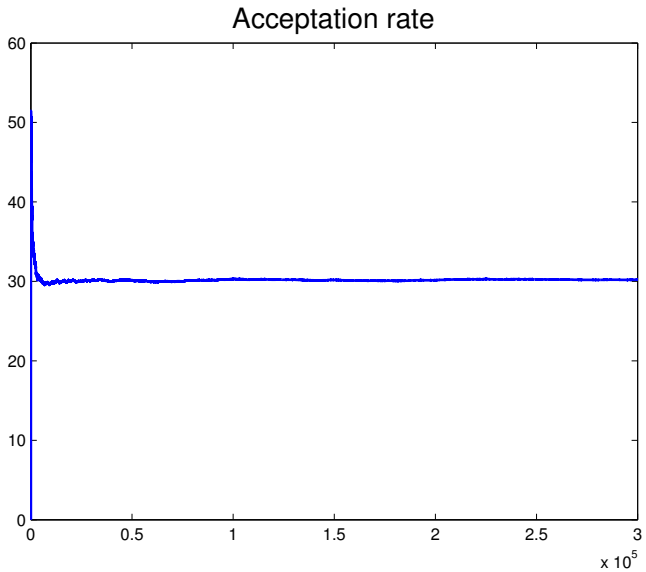


Figure: Posterior probability of the models in 16 dimensions, with uncorrelated design (i.e. $\mathbb{P}(m|M)$ for $1 \leq m \leq 2^{16}$; logarithmic scale); S=8, P=16, N=100.



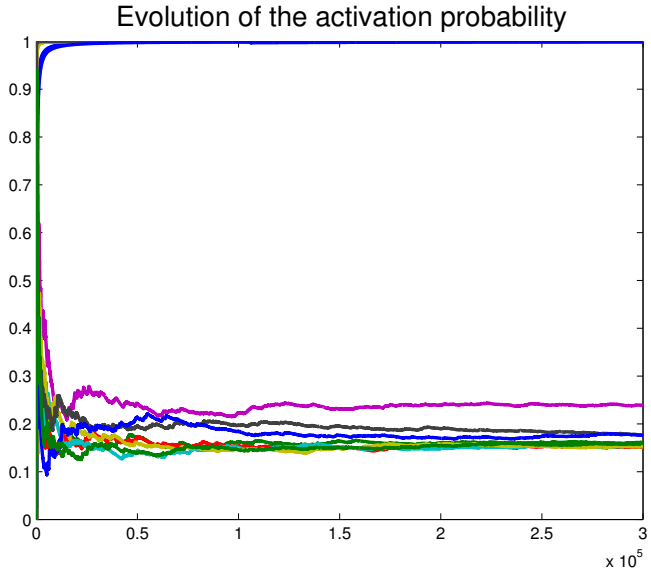


Figure: Dimension 16, uncorrelated design. Right: evolution of the acceptance rate; Left: evolution of the probability of activation.

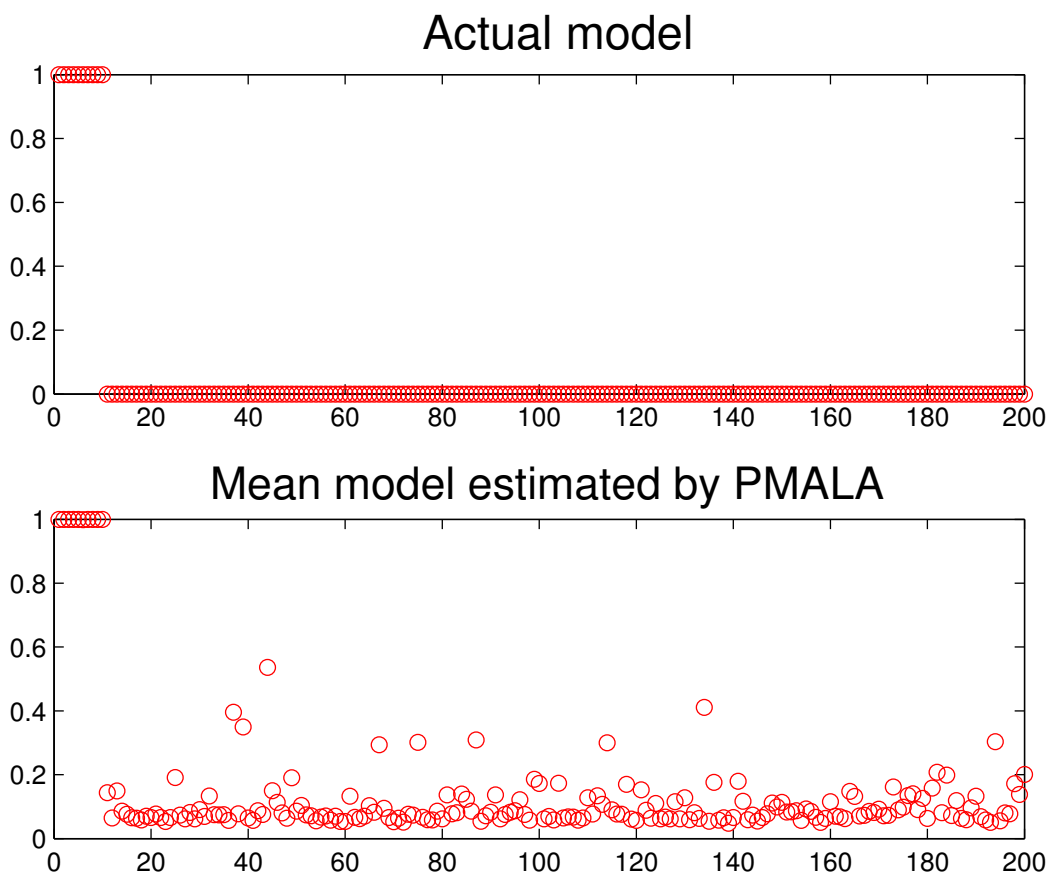


Figure: Posterior probability of activation in 200 dimensions, with uncorrelated design; S=10, P=200, N=100.

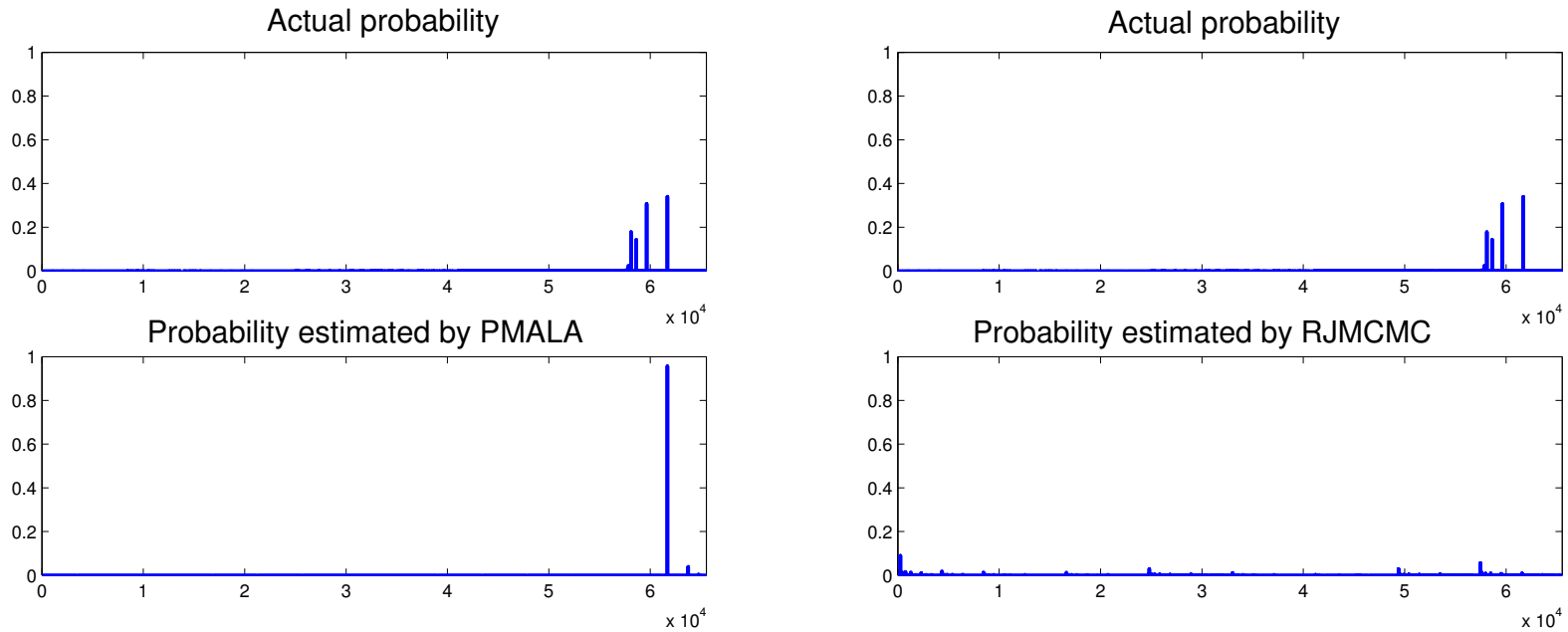


Figure: Comparison of posterior probabilities of the models in 16 dimensions, with correlated design; S=8, P=16, N=100.

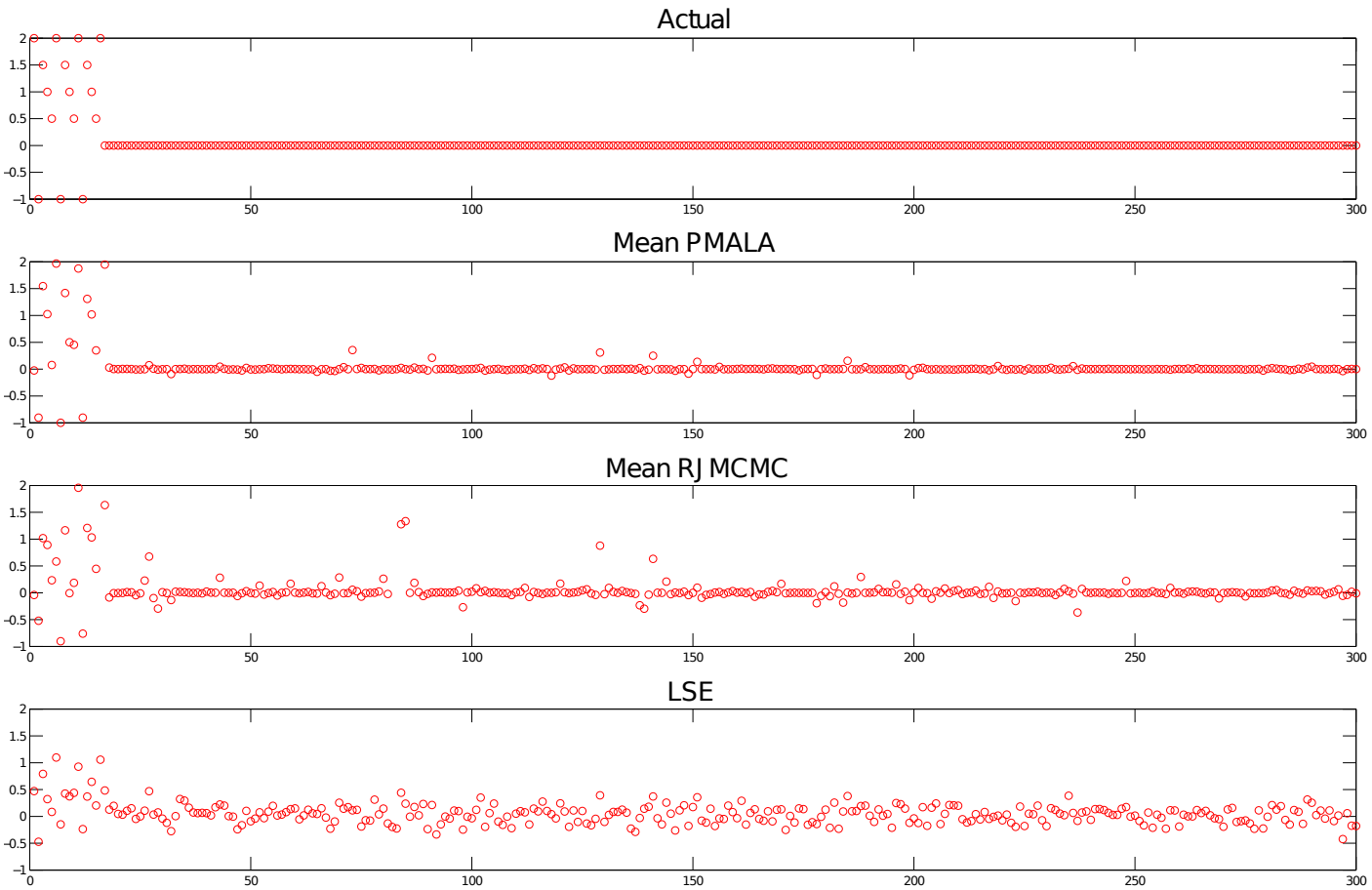


Figure: Comparison of mean estimated regression vectors in 300 dimensions, with correlated design; S=16, P=300, N=100.

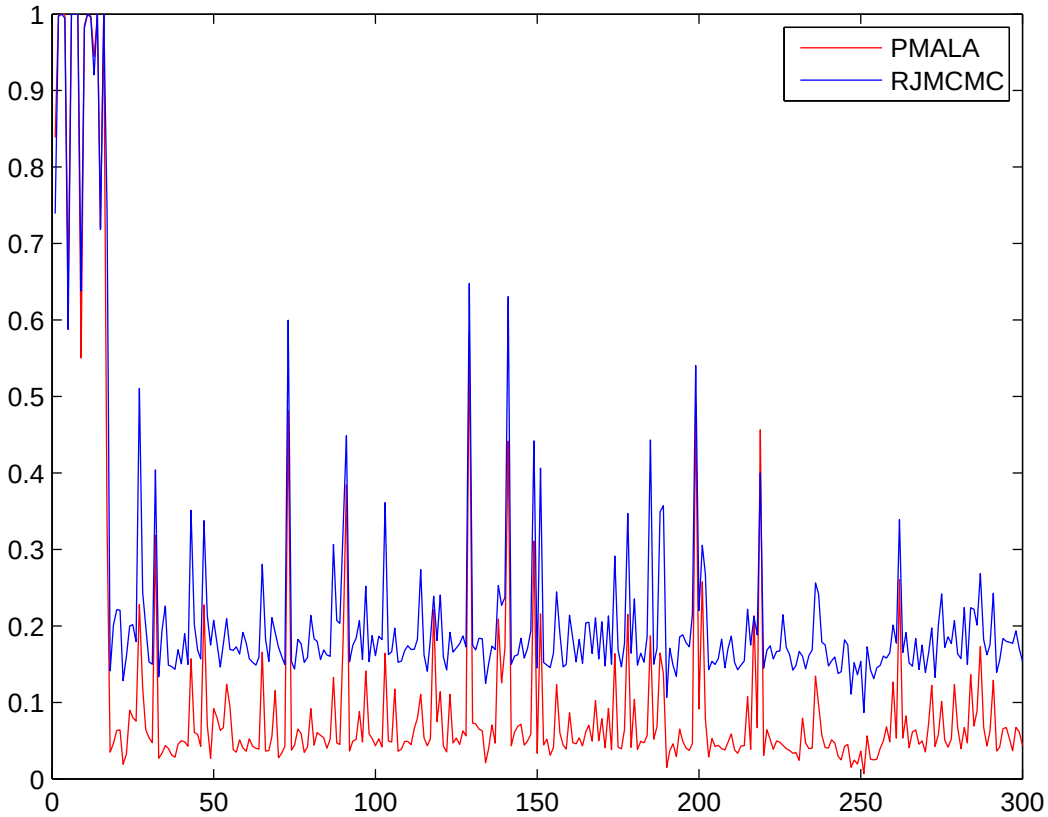


Figure: Comparison of posterior probabilities of activation in 300 dimensions, with correlated design; S=16, P=300, N=100.

Futur directions

- partial updating (higher control on the acceptance rate)
- tempering (to deal with multimodality)
- hard thresholding or other thresholding functions (to avoid shrinkage for regression applications)
- theory : geometric ergodicity, ...
- real data (back to brain imaging, regression problems)

Thank you !