A Shrinkage-Thresholding Metropolis adjusted Langevin algorithm for Bayesian Variable Selection

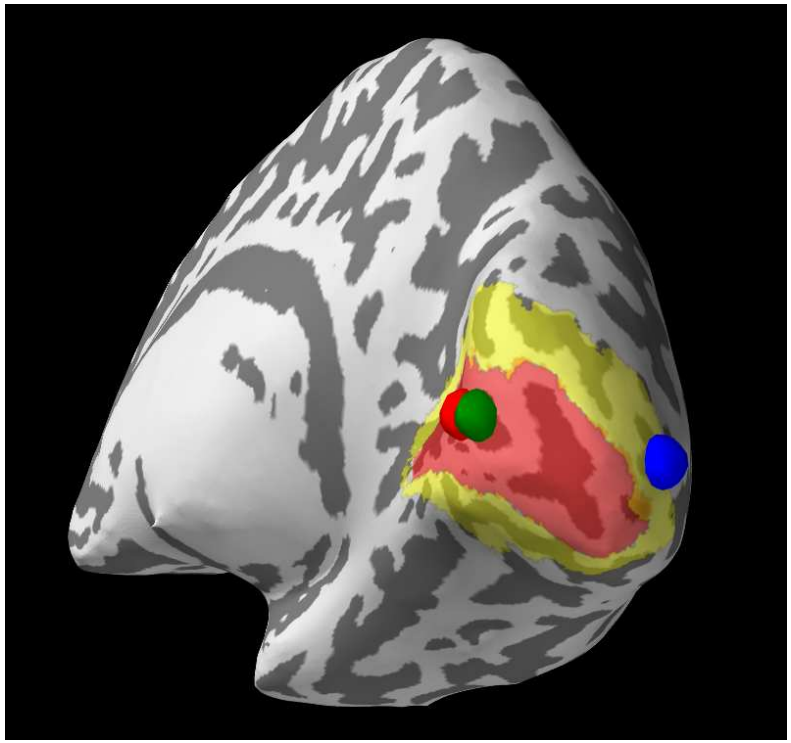
Amandine Schreck

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

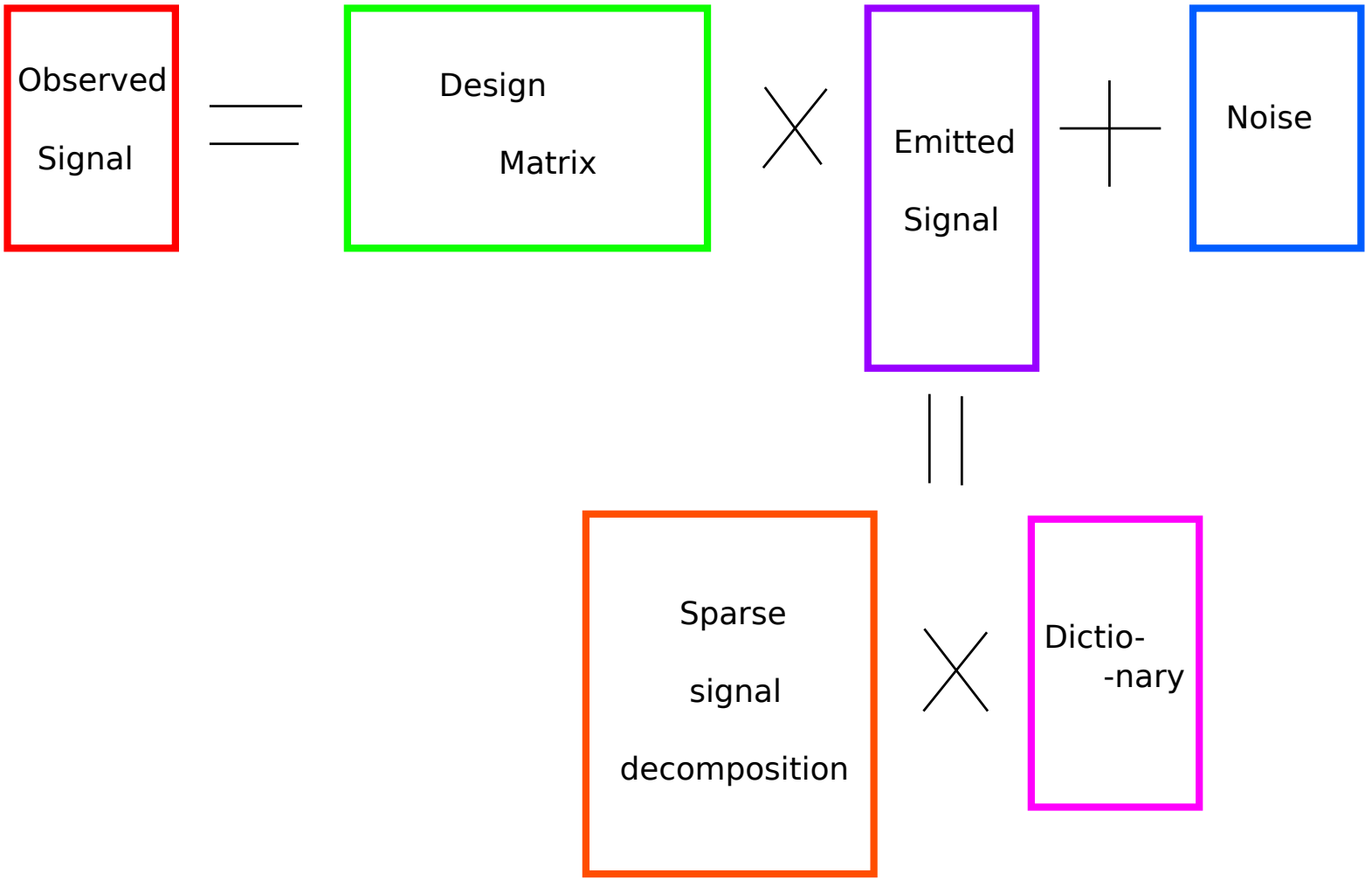
Télécom ParisTech

Paris, March 28th, 2014

Motivation : 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.

The motivating problem

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

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

Equivalent distribution in $\mathbb{R}^P$: $\pi(x)\mathrm{d}\nu(x)$, where

$$\mathrm{d}\nu(x) = \sum_{m\in\mathcal{M}} \left(\prod_{i\notin I_m} \delta_0(\mathrm{d}x_{i.})\right) \left(\prod_{i\in I_m} \mathrm{d}x_{i.}\right) ,$$

and

$$\pi(X) \propto \omega_{m_X}\, c_\lambda^{-|m_X|} \exp\left(-\frac{1}{2\tau}\|Y - GX\|_2^2 - \lambda\|X\|_{2,1}\right) .$$

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

The STMALA

Goal of the **Shrinkage Thresholding MALA** (STMALA): build a **Markov chain** converging to a target distribution with density with respect to $\mathrm{d}\nu$ of the form

$$\pi(x) \propto \exp(-g(x) - \bar{g}(x)) \ ,$$

where

- g: continuously differentiable, convex, such that ∇g is L_g-Lipschitz,
- $\bar{g}$: contains the non-differentiable part of π.

$\rightarrow$ Applied with $g(x) = \frac{1}{2\tau}\|Y - Gx\|_2^2$ and $\bar{g}(x) = \lambda\|x\|_{2,1} - \log\left(w_m c_\lambda^{-|m|}\right)$.

Base: Metropolis Hastings algorithm (with dominating measure $\mathrm{d}\mu$)

- Goal: sample a distribution $\pi \mathrm{d}\mu$ known **up to a multiplicative constant**.
- Tool: a transition kernel q such that for any x, it is **possible to sample from** $q(x, \cdot)\mathrm{d}\mu$.
- An iteration starting from X^t:
  - Sample Y^{t+1} according to $q(X^t, \cdot)\mathrm{d}\mu$.
  - Compute the acceptance probability

$$\alpha(X^t, Y^{t+1}) = \min\left(1, \frac{\pi(Y^{t+1})q(Y^{t+1}, X^t)}{\pi(X^t)q(X^t, Y^{t+1})}\right).$$

  - Set $X^{t+1} = Y^{t+1}$ with probability $\alpha(X^t, Y^{t+1})$ and $X^{t+1} = X^t$ with probability $1 - \alpha(X^t, Y^{t+1})$.

- Under some assumptions, **convergence** (in some sens) of the Metropolis Hastings algorithm occurs.
- But: if $q(x, \cdot)$ is too far from π, convergence is **too slow**.
- Idea of MALA: **use some knowledge about** π to build q.

Ingredient 1: The **Metropolis Adjusted Langevin Algorithm** (MALA)

Goal: build a Markov chain converging to a target distribution with density $\pi(x) \propto \exp(-g(x))$ with respect to Lebesgue measure, 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**.

$\rightarrow$ We cannot apply directly MALA as our target distribution is **not dominated by Lebesgue measure** and $\bar{g}$ is **not differentiable**.

Ingredient 2: 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

$\rightarrow$ generalisation of the gradient descent for **non differentiable** functions.

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

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

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

(1) **Propose** a new point

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

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

(2) Classical **Acceptation/Rejection step**, with acceptance probability $\alpha(x,y) = 1 \wedge \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}$, where $q(x,y)$ is the density of the proposal distribution (explicitly known).

Examples of operators Ψ

Let $\gamma > 0$ be a fixed threshold.

- **Proximal** (Prox): $(\Psi_1(u))_{i,j} = u_{i,j}\left(1 - \frac{\gamma}{\|u_{i\cdot}\|_2}\right)_+$,
- **Hard thresholding** (HT): $(\Psi_2(u))_{i,j} = u_{i,j}\mathbf{1}_{\|u_{i\cdot}\|_2 > \gamma}$,
- **Soft thresholding with vanishing shrinkage** (STVS): $(\Psi_3(u))_{i,j} = u_{i,j}\left(1 - \frac{\gamma^2}{\|u_{i\cdot}\|_2^2}\right)_+$.

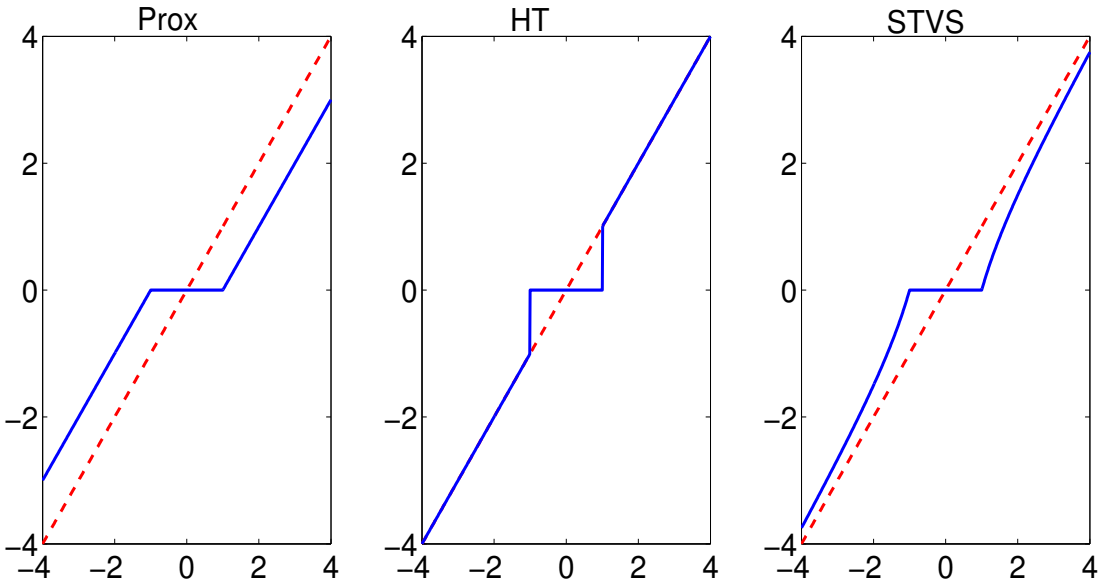


Figure : Shrinkage-Thresholding functions associated with the $L_{2,1}$ proximal operator (Prox - left), the hard thresholding operator (HT - center) and the soft thresholding operator with vanishing shrinkage (STVS - right) in one dimension.

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*

$$\sum_{m \in \mathcal{M}} \left(\prod_{i \notin I_m} p_1(\boldsymbol{\mu}_i)\, \delta_0(\mathrm{d}\boldsymbol{z}_i) \right) \left(\prod_{i \in I_m} f_1(\boldsymbol{\mu}_i, \boldsymbol{z}_i) \mathrm{d}\boldsymbol{z}_i \right) ,$$

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

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

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

The proposal mechanism of STMALA (with $\Psi = \Psi_1$) starting from x is equivalent to:

(i) sample $m' = (m'_1, \cdots, m'_P)$ with $(m'_i, i \in \{1, \cdots, P\})$ independent 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

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

A variant: STMALA with partial updating

For a fixed **block size** η, an iteration from X^t becomes:

(1) **Select a block** at random, i.e. a set b of η indices in $\{1, \ldots, P\}$.

(2) Propose a new point Y^{t+1} given by $Y^{t+1}_{-b} = X^t_{-b}$ and

$$Y^{t+1}_b = Z_b \quad \text{where} \quad Z = \Psi\left(X^t - \frac{\sigma^2}{2}\nabla g(X^t) + \sigma W^{t+1}\right) .$$

(3) Acceptation/Rejection step

Under some classical assumptions, i.e.

- regularity of the target density π,
- super-exponential behavior of π,
- positive measure of the acceptance set,

geometric ergodicity holds for STMALA (with $\Psi = \Psi_1$ and truncated gradient).

Example: π defined by

$$\pi(X) \propto \omega_{m_X} \, c_\lambda^{-|m_X|} \, \exp\left(-\frac{1}{2\tau}\|Y - GX\|_2^2 - \lambda\|X\|_{2,1} - \nu\|X\|_2^2\right) ,$$

satisfies these assumptions.

Theorem

Under some "classical assumptions", for any $\beta \in (0,1)$*, there exist* $C > 0$ *and* $\rho \in (0,1)$ *such that for any* $n \geq 0$ *and any* $x \in \mathbb{R}^P$*,*

$$\| P_{\Psi_1}^n(x,.) - \pi \|_V \leq C\, \rho^n\, V(x)\ ,$$

where $V(x) \propto \pi(x)^{-\beta}$ *and for any signed measure* η*,*
$\|\eta\|_V = \sup_{f, |f| \leq V} |\int f \mathrm{d}\eta|$.

Sketch of proof (1): expression of the kernel

Transition kernel:

$$P(x, A) = \int_A q(x, y)\alpha(x, y)\mathrm{d}\nu(y) + \mathbf{1}_A(x) \int q(x, y)(1 - \alpha(x, y))\mathrm{d}\nu(y) \ ,$$

where

$$q(x, y) = \prod_{i \notin I_m} p\left(\tilde{\mu}_i(x)\right) \quad \prod_{i \in I_m} f\left(\tilde{\mu}_i(x), y_i\right) \ ,$$

and (**truncated gradient**)

$$\tilde{\mu}(x) = x - \frac{\sigma^2}{2} \frac{D\,\nabla g(x)}{\max\left(D, \|\nabla g(x)\|_2\right)} \ .$$

Sketch of proof (2): main ingredients

- By construction, π is **invariant** with respect to P (*i.e.* $\pi(A) = \int \pi(dx)P(x, A)$).
- The chain is **aperiodic** (*i.e.* no k-cycle for $k \geq 2$) and **psi-irreducible** (*i.e.* for any x, A there exists n such that $P^n(x, A) > 0$).
- C such that $C \cap S_m$ is compact for any m are **small sets** for P (*i.e.* there exists a measure $\tilde{\nu}$ on $\mathbb{R}^P$ such that $P_{trunc}(x, A) \geq \tilde{\nu}(A)\mathbf{1}_C(x)$).
- **Drift condition**: there exist $C_1 \in (0, 1)$, $C_2 < \infty$ and a small set C such that $PV(x) \leq C_1 V(x) + C_2 \mathbf{1}_C(x)$.

Sketch of proof (3): results for the drift

Final step for the drift:

$$\limsup_{\|x\|\to\infty} \frac{\int P(x,dy)V(y)}{V(x)} < 1 \ .$$

Indeed

$$\frac{PV(x)}{V(x)} \leq \int \alpha(x,y)\frac{V(y)}{V(x)} q(x,y)\mathrm{d}\nu(y) + 1 - \int_{A(x)} q(x,y)\mathrm{d}\nu(y) \ .$$

And

$$\int_{A(x)} q(x,y)\mathrm{d}\nu(y) \geq C \ , \quad \limsup_{\|x\|\to\infty} \int \alpha(x,y)\frac{\pi^{-\beta}(y)}{\pi^{-\beta}(x)} q(x,y)\mathrm{d}\nu(y) = 0 \ .$$

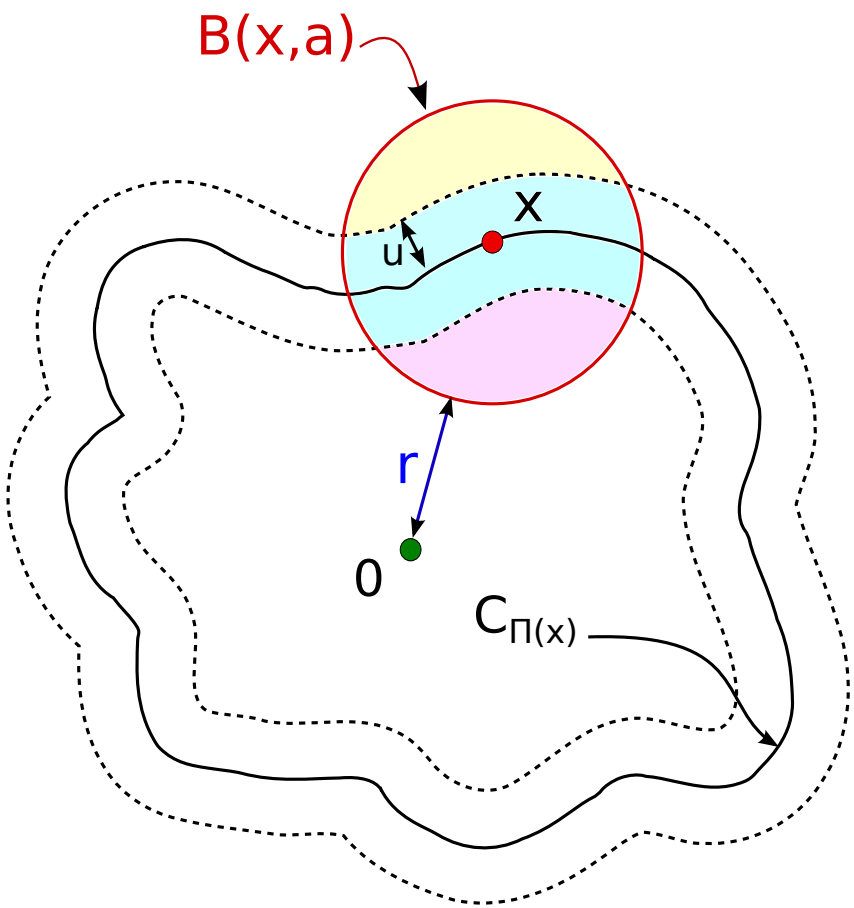


Figure : How to cut the integral

Numerical illustrations

Competitor: Reversible Jump MCMC

An iteration of RJMCMC starting from (m, X):

- Sample a **new model** $m' \in \{0,1\}^P$ uniformly **among the neighbors of** m (by adding, deleting or replacing an active component of m or by keeping m).
- Sample a **new point** $X' \in \mathbb{R}^P$ such that $X'_{-m'} = 0$ and that $X'_i = X_i$ for any $i \in \{1, \ldots, P\}$ such that the i-th component is active in m' and in m.
- Acceptation/Rejection step.

Main drawbacks of RJMCMC

As **only local moves** occur:

- slow mixing
- slow convergence
- problems with high dimension (2^P models)
- problems with correlated designs (possible moves limited)
- difficulties to escape from local maxima

The data: $Y = GX + E$

- $N = 100$, $P = 16$.
- The components of E are samples of $\mathcal{N}(0, 1)$
- $X = (X_i)_{1 \leq i \leq P}$ with $X_i = \mathbf{1}_{i \leq 8}$.
- Columns of $G \in \mathbb{R}^{N \times P}$: independant Gaussian samples (uncorrelated designs).

Implementation parameters:

- Prior on the models: m_k are i.i.d. Bernoulli with success parameter 0.1.
- Starting point: empty model.

Interest of this model:

- The posterior activation probabilities $\mathbb{P}(X_i \neq 0)$, defined by

$$\mathbb{P}(X_i \neq 0) = \sum_{m \in \mathcal{M}} \pi(m|Y)\ m_i\ ,$$

  can be computed.
- Error:

$$\mathcal{E} = \sum_{i=1}^{P} \left| \mathbb{P}(X_i \neq 0) - \frac{1}{N_{it}} \sum_{n=B}^{N_{it}+B} \mathbf{1}_{X_i^n \neq 0} \right| .$$

Comparison of the thresholding operators:

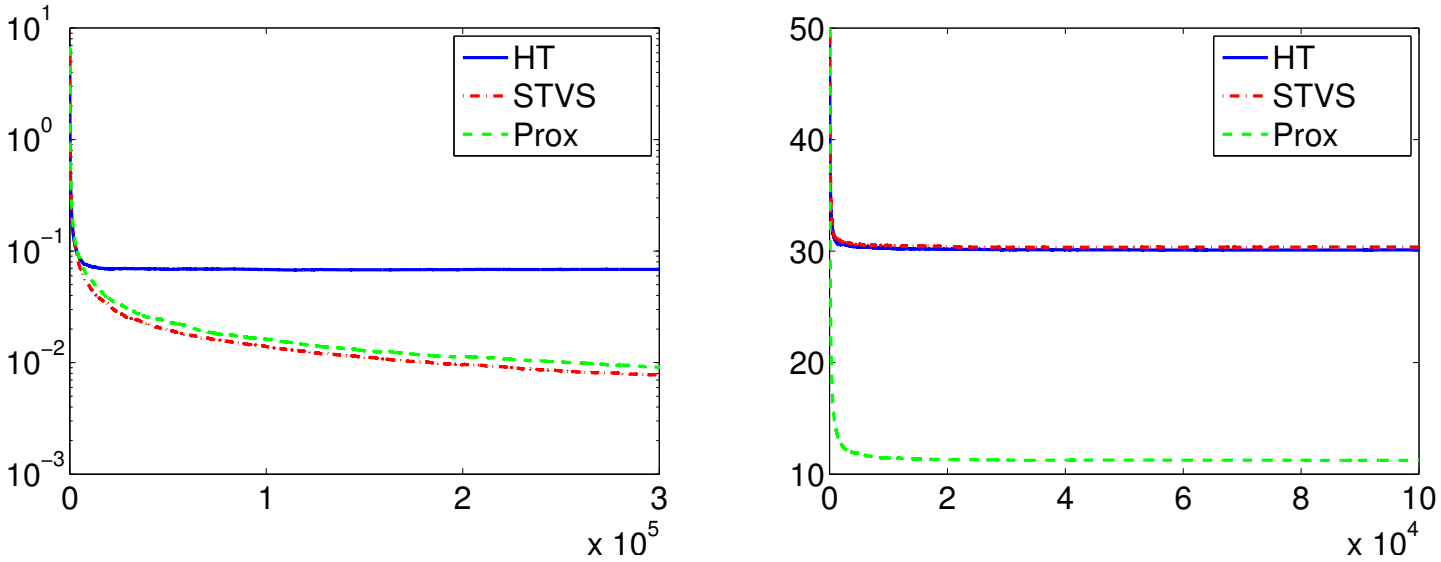


Figure : (left) Evolution of the **mean estimation error** of the activation probabilities for block-STMALA as a function of the number of iterations, when $\Psi = \Psi_1$ (Prox), $\Psi = \Psi_2$ (HT) and $\Psi = \Psi_3$ (STVS) as shrinkage-thresholding operator. (right) Evolution of the **mean acceptance rate**.

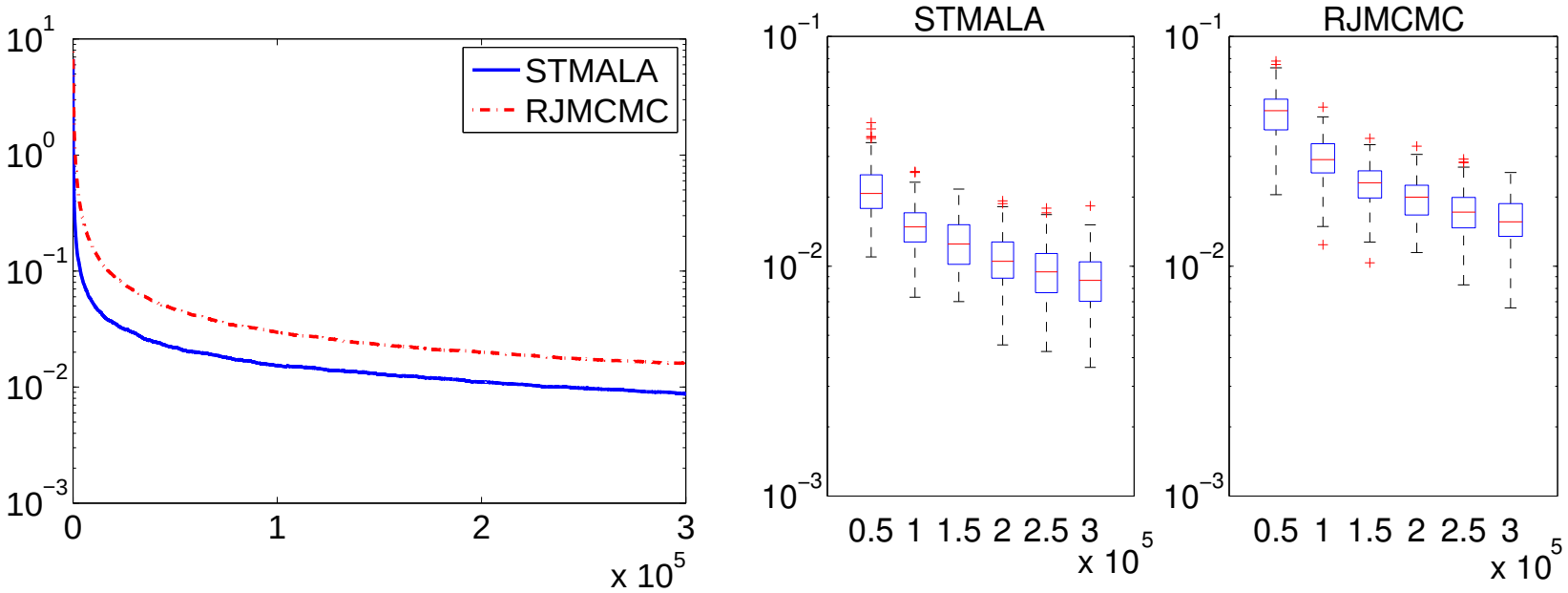


Figure : Evolution of the **mean error** for block-STMALA and RJMCMC as a function of the number of iterations (left) and the **associated boxplots** (right).

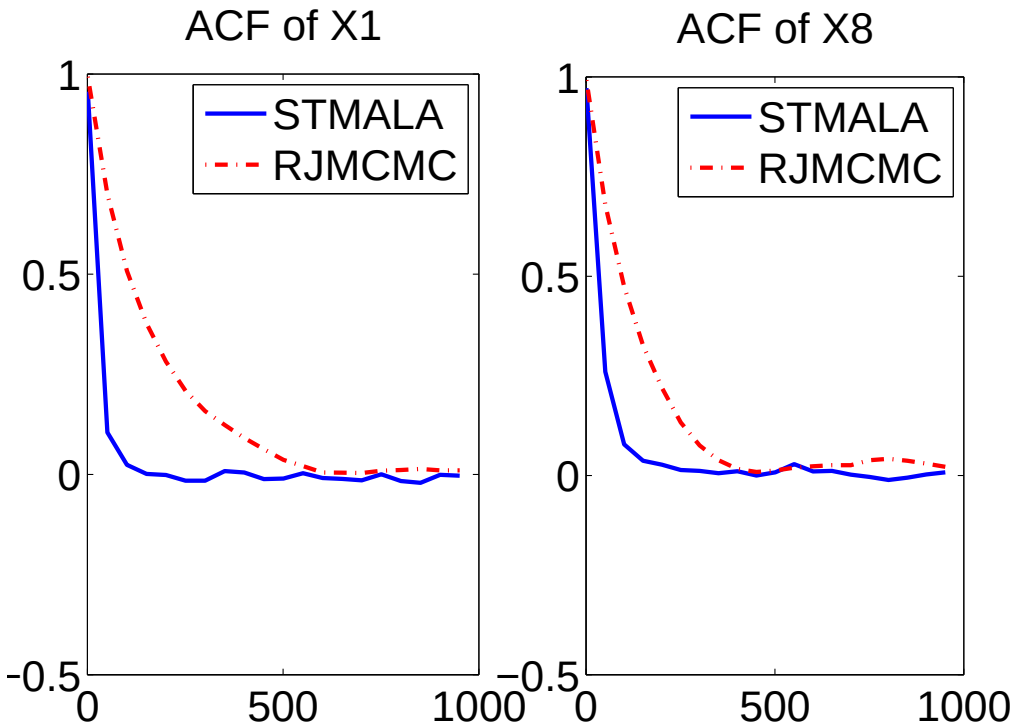


Figure : Empirical **autocorrelation function** of X_1 and X_8 of block-STMALA and RJMCMC.

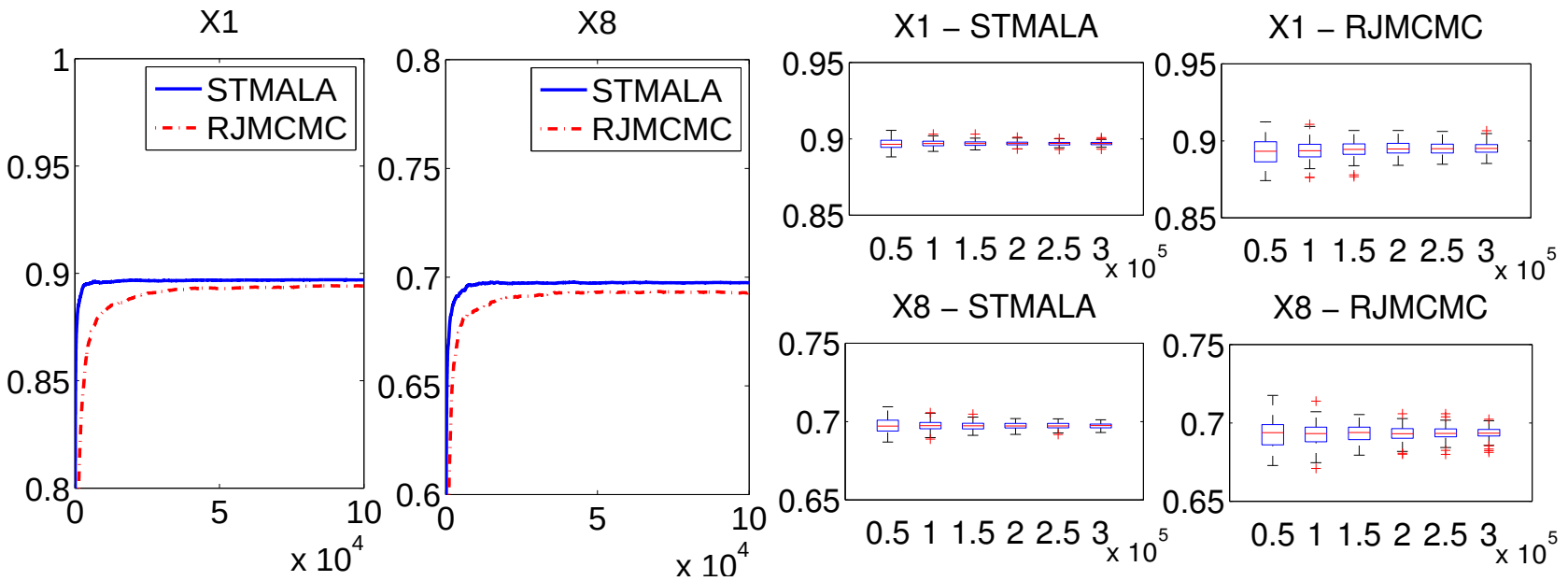


Figure : (left) Evolution of the **mean estimators** (over 100 independent runs) of $\int x_i \, \pi(x|Y)\mathrm{d}\nu(x)$ for $i = 1$ and $i = 8$ computed by block-STMALA and RJMCMC as a function of the number of iterations. (right) Associated boxplots.

Model for the observations $Y \in \mathbb{R}^N$:

$$Y = GX + E \ .$$

Spike and slab prior:

- $$(X_k | m, \vartheta_1, \cdots, \vartheta_P) \sim \begin{cases} \delta_0(X_k) & \text{if } m_k = 0, \\ \mathcal{N}(0, 1/\vartheta_k) & \text{if } m_k = 1. \end{cases}$$
- $(\vartheta_\ell)_{1 \leq \ell \leq P}$ are i.i.d. with Gamma distribution $\mathrm{Ga}(a, aK)$, where $a = 2$, $K = 0.08$.
- The components of $m \in \mathcal{M}$ are i.i.d. Bernoulli with parameter $\omega_\star = 0.1$.

Here,

- $N = 100$, $P = 200$.
- $(G_{:,i})_{1 \leq i \leq P}$ are Gaussian with $\mathbb{E}[G_{.i}] = 0$ and $\mathbb{E}[G_{ji} G_{ki}] = 0.3^{|j-k|}$.
- The nonzero coefficients of X are such that, for all $k \in \{1, 2, 3, 4\}$ and all $j \in \{1, 2, 3, 4, 5\}$, $X_{50*(k-1)+j} = (-1)^{k+1} j^{1/k}$.

The design parameters are chosen so that STMALA and RJMCMC have similar acceptance rates.

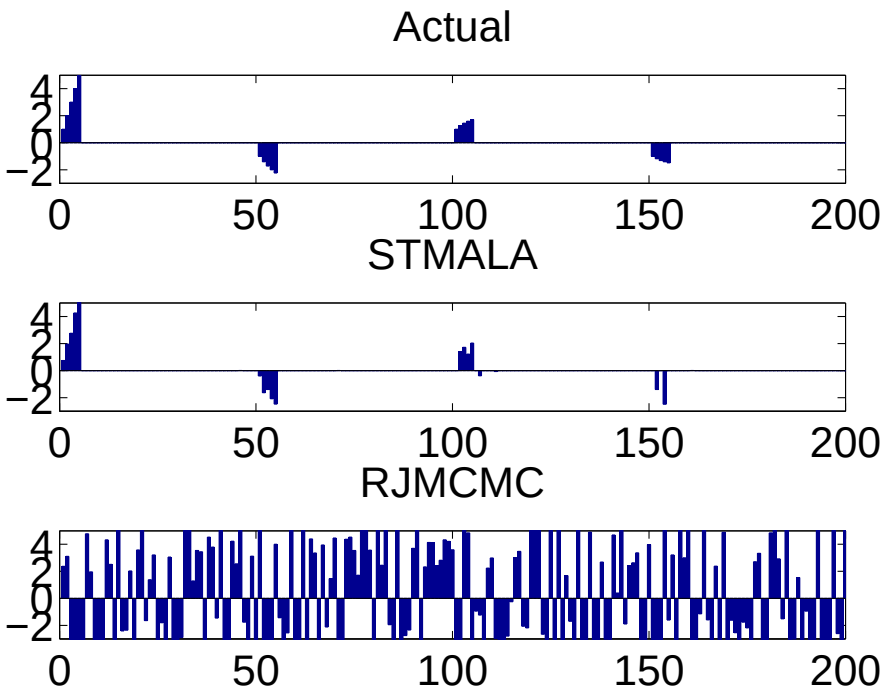


Figure : **Regression vectors** estimated by block-STMALA and RJMCMC.

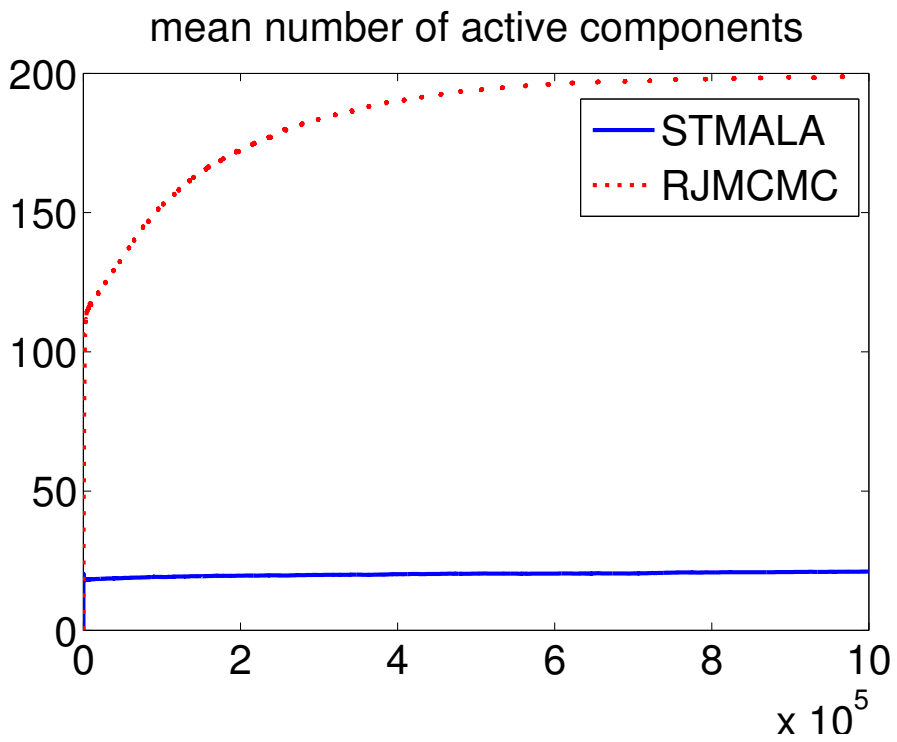


Figure : Evolution of the **mean number of active components** for STMALA and RJMCMC.

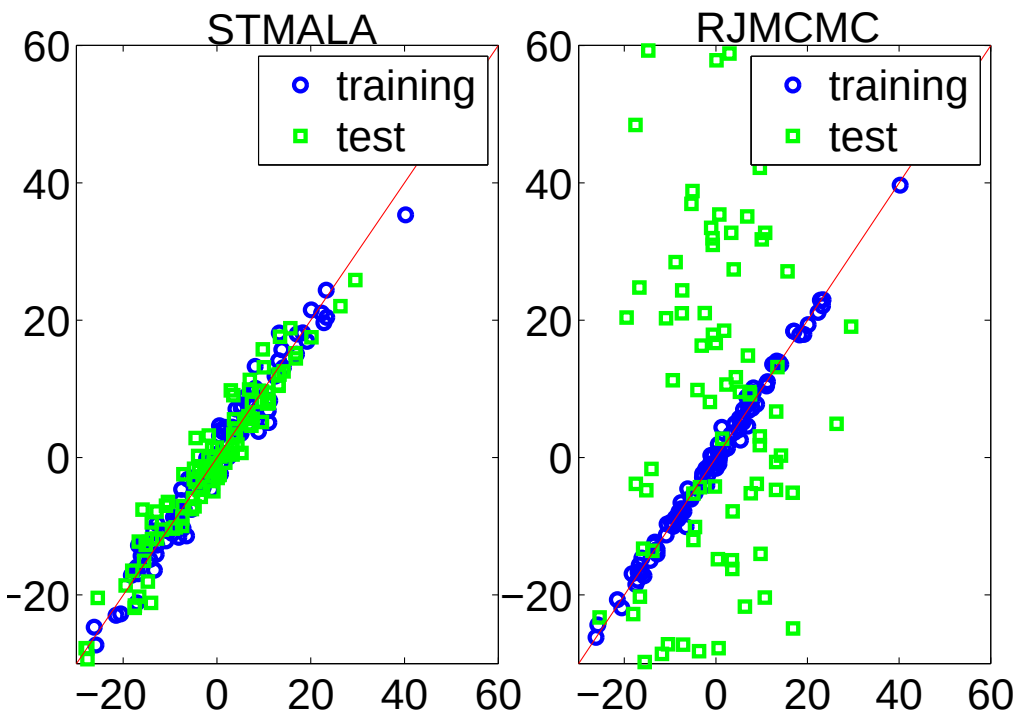


Figure : Emitted signal $G\hat{X}$ estimated by block-STMALA and RJMCMC versus actual emitted signal GX.

- The dataset:
  - Y: fat content of **70 different cookies**.
  - G: each row of G contains $P = 300$ **spectroscopy measurements**.
- The dataset is cut in a training set of $N = 39$ cookies and a test set of 31 cookies.
- Goal: **predict fat content**.
- A spike is expected at 1726 nm (fat absorbance region).

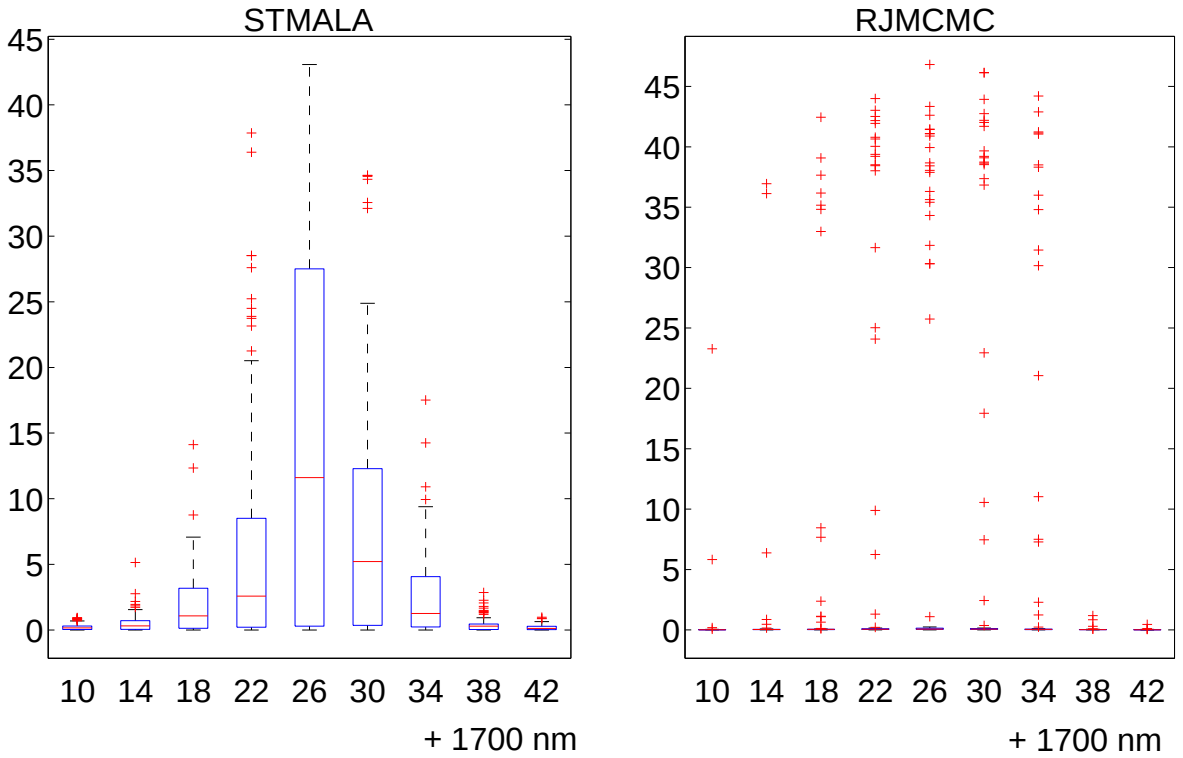


Figure : Boxplots of the 100 independent values of the **components of the regression vectors** estimated by block-STMALA and RJMCMC associated to 9 wavelengths close to 1726 nm.

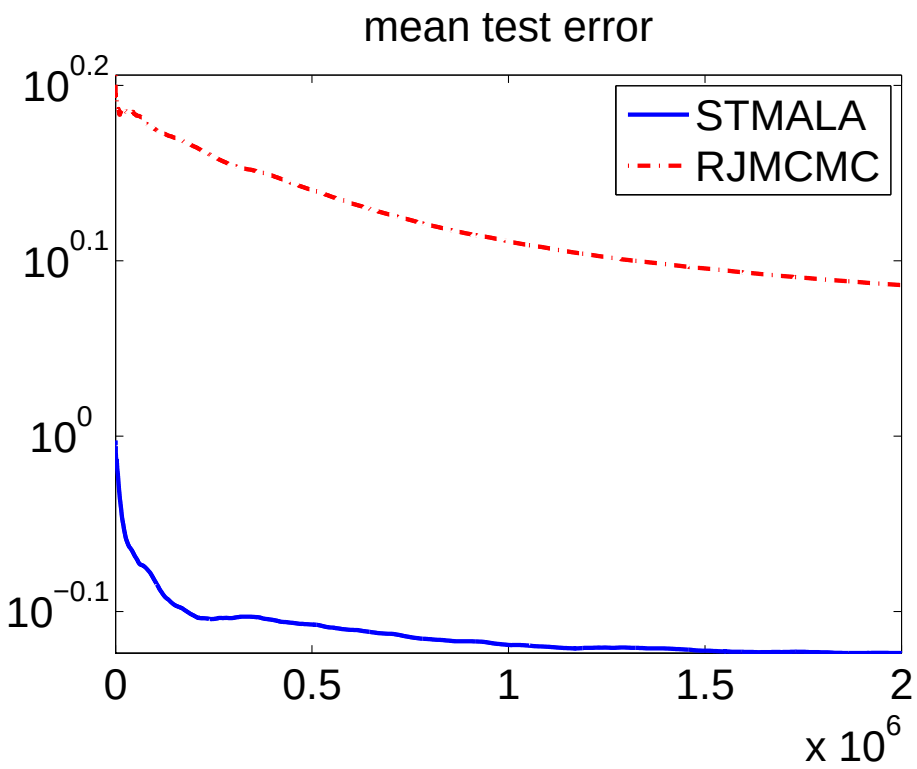


Figure : Evolution of the **mean MSE** (over 100 independent trajectories) on the test data set for RJMCMC and block-STMALA.

Future directions

- tempering (to deal with multimodality)
- **adaptation** (automatic choice of design parameters)
- real data (back to **brain imaging**)...

Thank you !

A. Schreck, G. Fort, S. Le Corff and E. Moulines.

A shrinkage-thresholding Metropolis adjusted Langevin algorithm for Bayesian variable selection.

on ArXiv, 2013.