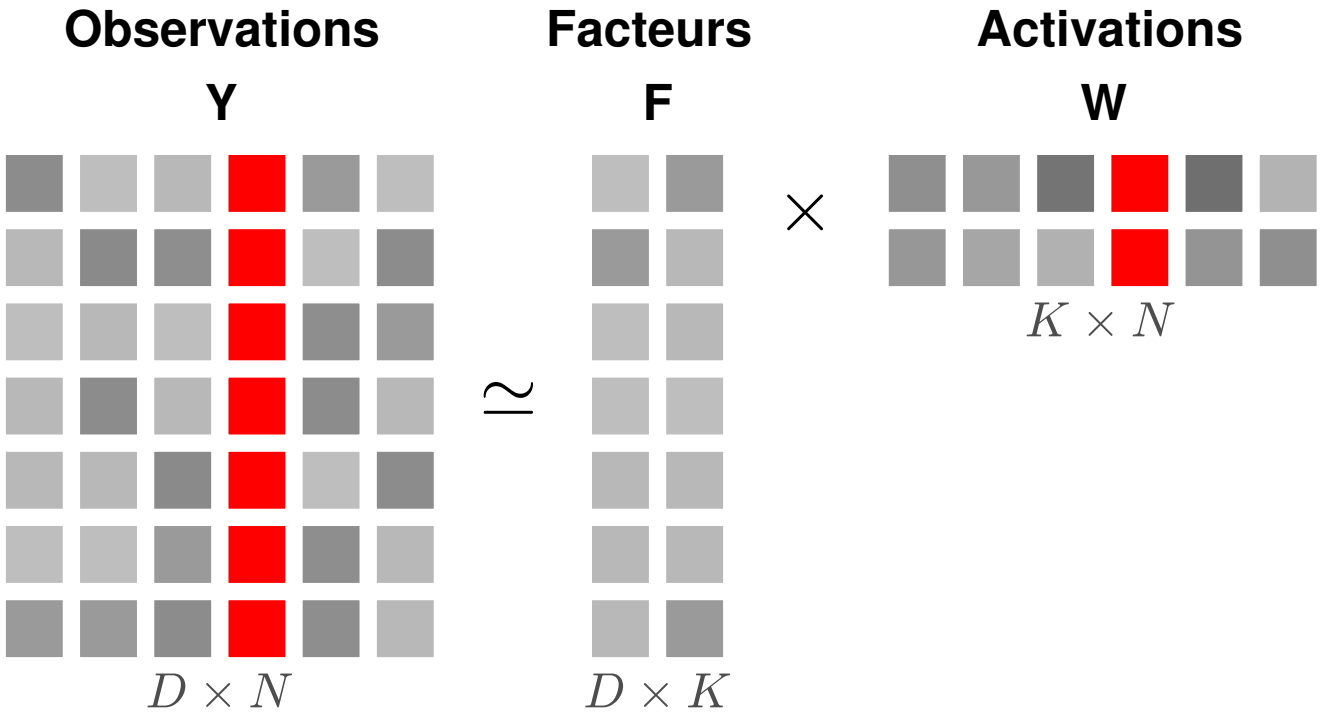
Modèles bayésiens pour l'identification de représentations antiparcimonieuses et l'analyse en composantes principales non paramétrique

Clément Elvira

Directeurs de thèse :
Pierre Chainais et Nicolas Dobigeon

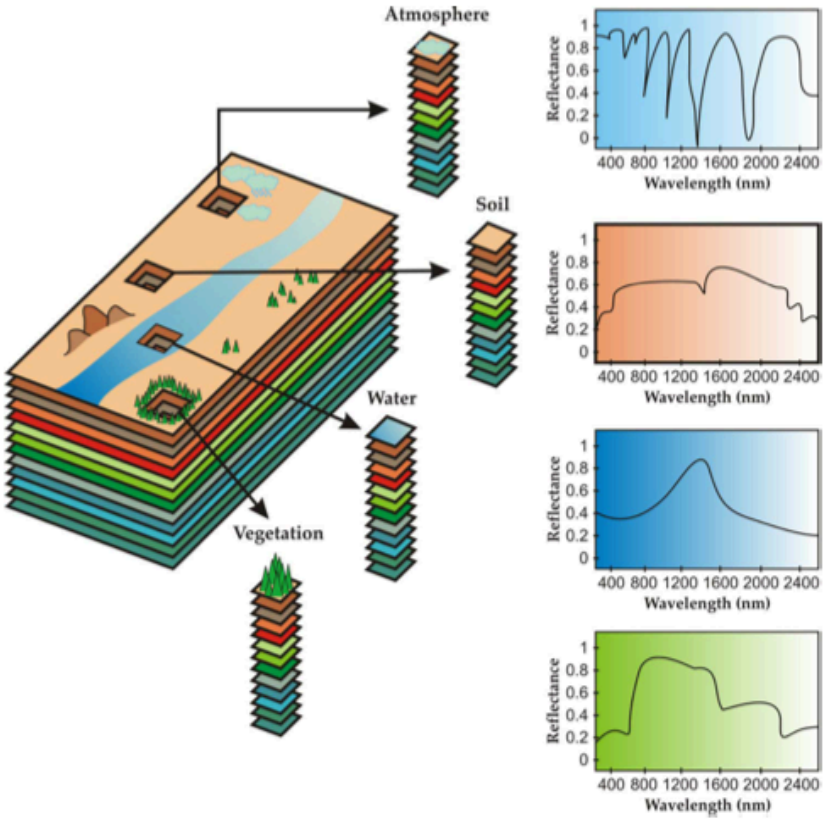
10 novembre 2017

Modèles linéaires et factorisation de matrices



Modèles linéaires et factorisation de matrices

Exemple : imagerie hyperspectrale & hypothèse de linéarité



$$\text{pixel}_n = \sum_{k=1}^{K} \text{signature}_k \times \text{prop}_{n,k}$$

$$\mathbf{Y} = \mathbf{F} \times \mathbf{W}$$

Bioucas-Dias & al (2012)

Méthodologie bayésienne

- Formulé comme un problème inverse
- Généralement mal posé $\Longrightarrow$ régularisation, *lois a priori*

Formule de Bayes $\mathbf{Y} \simeq \mathbf{F}\,\mathbf{W}$

$$p(\mathbf{F}, \mathbf{W}|\mathbf{Y}) \propto \underbrace{p(\mathbf{Y}|\mathbf{F}, \mathbf{W})}_{\text{Vraisemblance}} \times \underbrace{p(\mathbf{F}, \mathbf{W})}_{\textit{a priori}}$$

- Liberté : *a priori*
- Algorithmes d'échantillonnage et estimateurs
- **Aucun réglage de paramètres**
- Aide à la décision : intervalle de confiance, …

Plan de la présentation

Partie 1 : codage antiparcimonieux

- $D \ll K \longrightarrow$ robustesse
- échantillonnage efficace

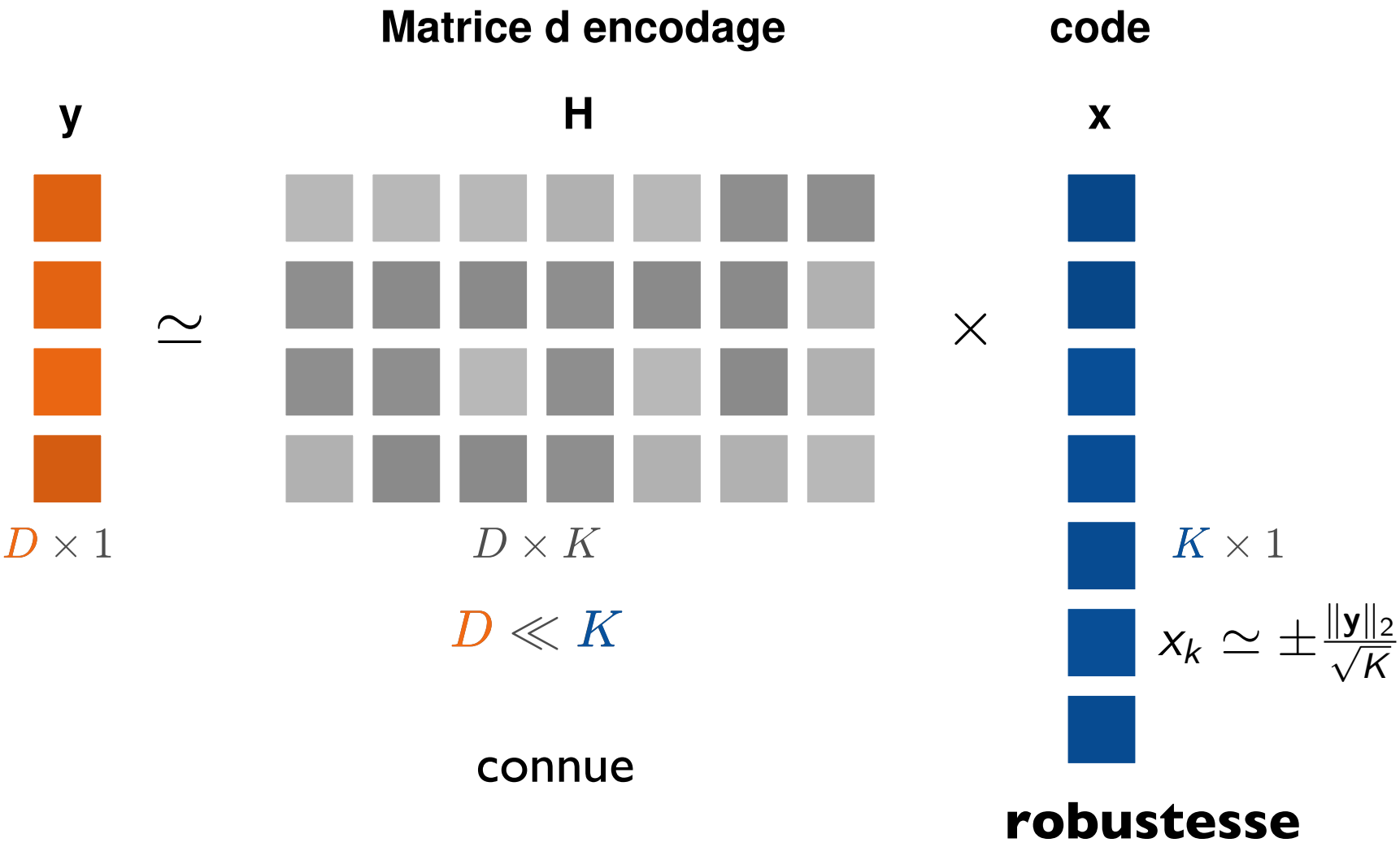
Partie 2 : estimation de sous-espaces

- $D \gg K \longrightarrow$ modélisation
- Étude des estimateurs de la dimension

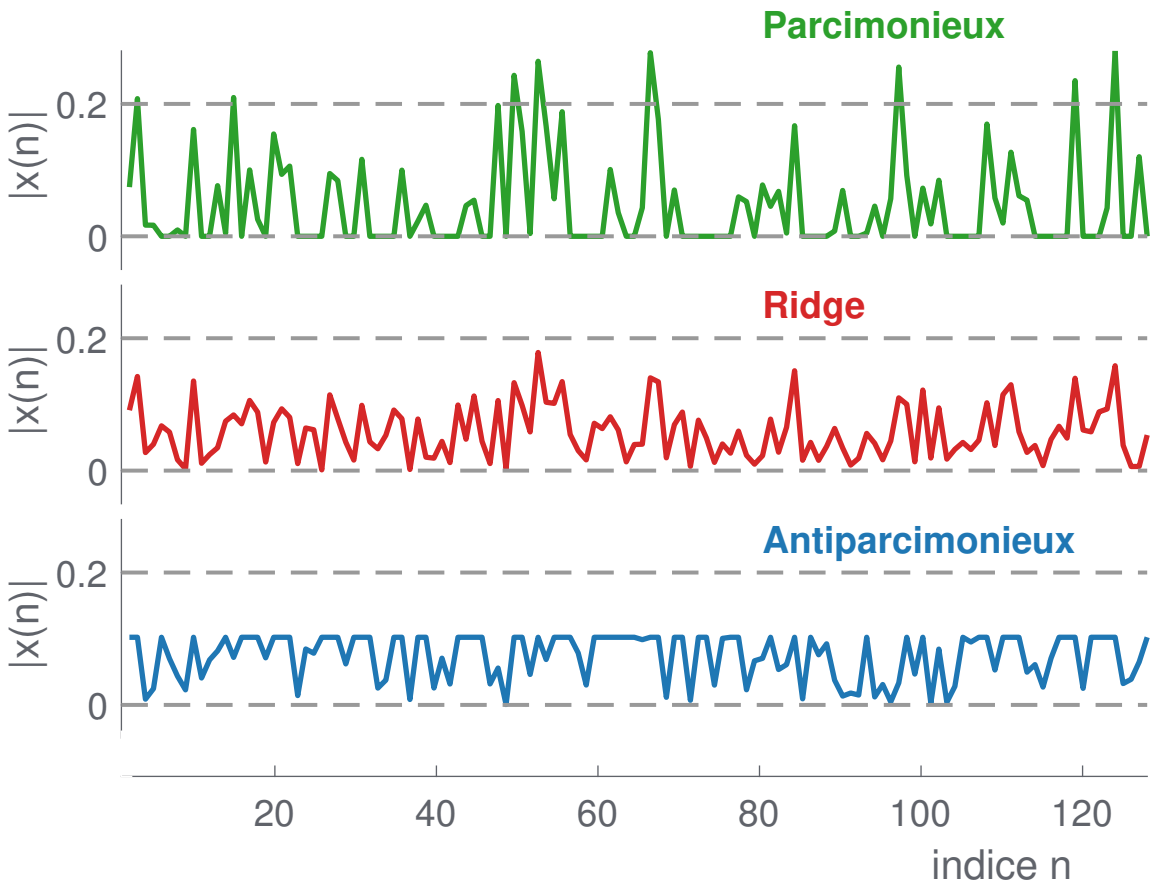
Partie I

codes antiparcimonieux
et loi démocratique

Codage linéaire et antiparcimonieux



Codage linéaire et antiparcimonieux



peu de coefficients

minimum d'énergie

$$x_k \simeq \pm \frac{\|\mathbf{y}\|_2}{\sqrt{K}}$$

Applications

- Recherche approximative de plus proches voisins

Jegou & al. (2012)

$x_k = \pm\alpha \iff$ binarisation + vie privée

distance binaire = XOR $\implies$ plus rapide

- Automatique : répartition des efforts moteurs *Cadzow (1971)*
- Réduction d'erreur de quantification *Lyubarskii & Vershynin (2010)*
- télecom : Réduction facteur de crête (PAPR)

Ilic & Strohmer (2009)

$$\forall \mathbf{x} \in \mathbb{R}^K, \quad \mathrm{PAPR}(\mathbf{x}) = \frac{K\|\mathbf{x}\|_\infty^2}{\|\mathbf{x}\|_2^2}$$

Applications

- Recherche approximative de plus proches voisins

Jegou & al. (2012)

$x_k = \pm\alpha \iff$ binarisation + vie privée

distance binaire = XOR $\implies$ plus rapide

- Automatique : répartition des efforts moteurs *Cadzow (1971)*
- Réduction d'erreur de quantification *Lyubarskii & Vershynin (2010)*
- télecom : Réduction facteur de crête (PAPR)

Ilic & Strohmer (2009)

$$\forall \mathbf{x} \in \mathbb{R}^K, \quad \mathrm{PAPR}(\mathbf{x}) = \frac{K\|\mathbf{x}\|_\infty^2}{\|\mathbf{x}\|_2^2}$$

Code antiparcimonieux & problème inverse

$$\mathbf{y} \simeq \mathbf{H}\mathbf{x} \qquad \text{et} \qquad |x_k| \simeq \frac{\|\mathbf{y}\|_2}{\sqrt{K}}$$

$$(\mathrm{P}_\infty^\varepsilon) \quad \min_{\mathbf{x}\in\mathbb{R}^K} \|\mathbf{x}\|_\infty \quad \text{s.c.} \quad \|\mathbf{y} - \mathbf{H}\mathbf{x}\|_2^2 \leq \varepsilon$$

$\exists\ C_\ell$ et C_u telles que *Lyubarskii & Vershynin (2010)*

$$\frac{C_\ell(\mathbf{H},\varepsilon)}{\sqrt{K}} \left(\|\mathbf{y}\|_2 - \varepsilon\right) \leq \|\widehat{\mathbf{x}}\|_\infty \leq \frac{C_u(\mathbf{H},\varepsilon)}{\sqrt{K}} \left(\|\mathbf{y}\|_2 - \varepsilon\right)$$

La loi démocratique

Résoudre

$$\hat{\mathbf{x}} = \underset{\mathbf{x}\in\mathbb{R}^K}{\operatorname{argmin}} \underbrace{\frac{1}{2}\|\mathbf{y} - \mathbf{H}\mathbf{x}\|_2^2}_{\text{attache aux données}} + \underbrace{\beta\|\mathbf{x}\|_\infty}_{\text{pénalité}} = \underset{\mathbf{x}\in\mathbb{R}^K}{\operatorname{argmin}} J(\mathbf{x}, \beta)$$

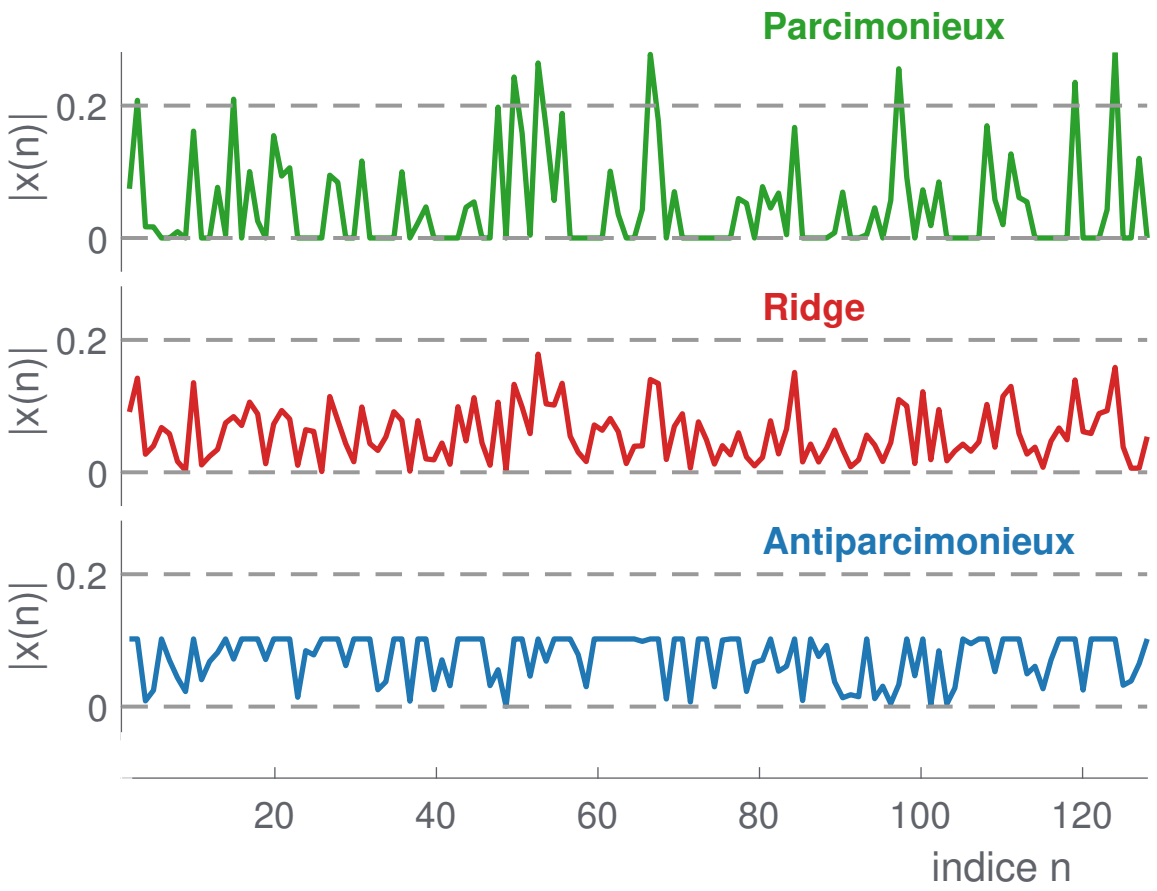
avec une méthode *forward-backward* FITRA *Studer & Larsson (2013)*

$$\Longleftrightarrow$$

MAP de vraisemblance gaussienne + prior démocratique

Densité de la loi démocratique

$$\forall \boldsymbol{x} \in \mathbb{R}^K \qquad \mathrm{p}(\boldsymbol{x}) = \frac{\lambda^K}{2^K K!} \exp\left(-\lambda\|\boldsymbol{x}\|_\infty\right)$$



$$\frac{1}{2}\|\mathbf{y}-\mathbf{H}\mathbf{x}\|_2^2+\beta\|\mathbf{x}\|_1$$
Laplace

$$\frac{1}{2}\|\mathbf{y}-\mathbf{H}\mathbf{x}\|_2^2+\beta\|\mathbf{x}\|_2^2$$
Gaussienne

$$\frac{1}{2}\|\mathbf{y}-\mathbf{H}\mathbf{x}\|_2^2+\beta\|\mathbf{x}\|_\infty$$
?

$\longrightarrow$ Démocratique

La loi démocratique

Résoudre

$$\hat{\mathbf{x}} = \underset{\mathbf{x}\in\mathbb{R}^K}{\operatorname{argmin}} \underbrace{\frac{1}{2}\|\mathbf{y} - \mathbf{H}\mathbf{x}\|_2^2}_{\text{attache aux données}} + \underbrace{\beta\|\mathbf{x}\|_\infty}_{\text{pénalité}} = \underset{\mathbf{x}\in\mathbb{R}^K}{\operatorname{argmin}} J(\mathbf{x}, \beta)$$

avec une méthode *forward-backward* FITRA *Studer & Larsson (2013)*

$$\Longleftrightarrow$$

MAP de vraisemblance gaussienne + prior démocratique

Densité de la loi démocratique

$$\forall \boldsymbol{x} \in \mathbb{R}^K \qquad \mathrm{p}(\boldsymbol{x}) = \frac{\lambda^K}{2^K K!} \exp\left(-\lambda\|\boldsymbol{x}\|_\infty\right)$$

Contributions & plan de la partie

- Formulation bayésienne du codage antiparcimonieux
- Étude de la loi démocratique

$$p(\mathbf{x}) \propto \exp\left(-\lambda \|\mathbf{x}\|_\infty\right)$$

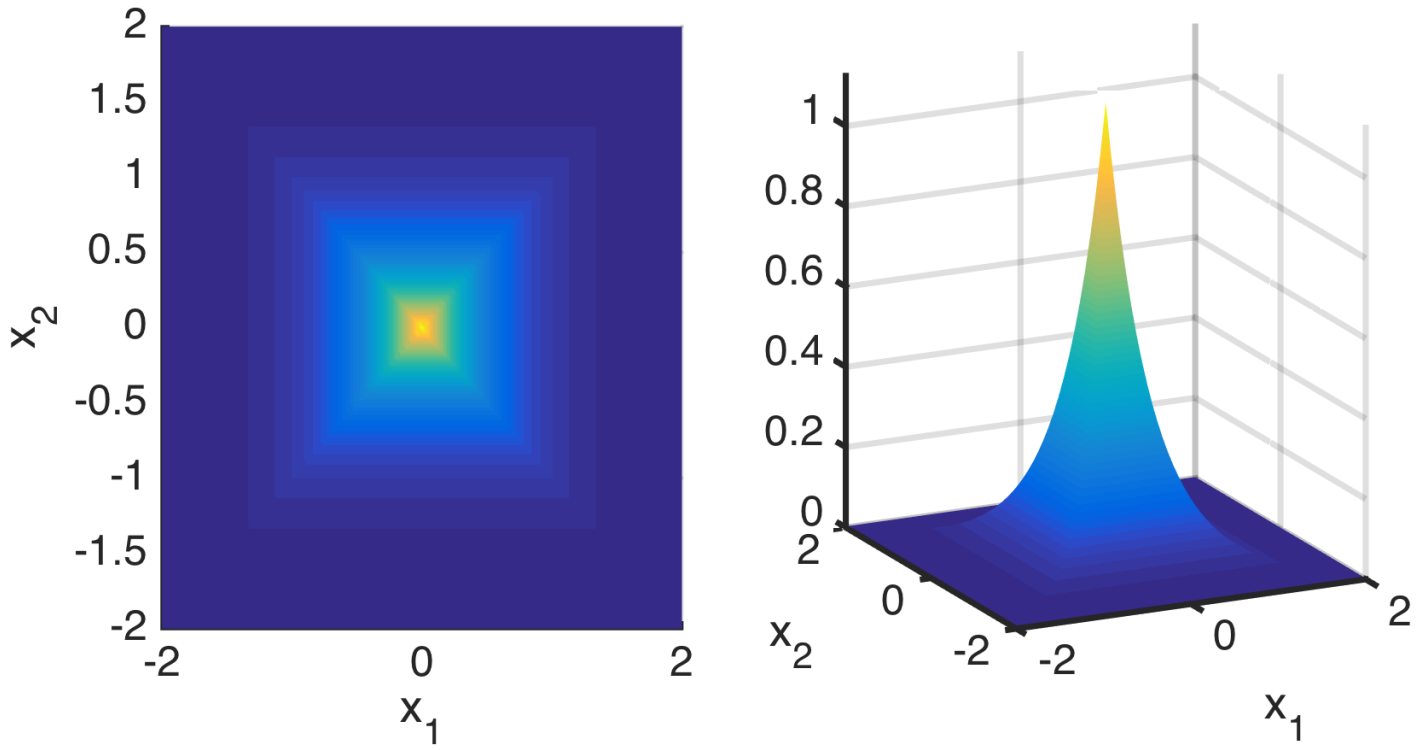
- 3 algorithmes de Monte Carlo : Gibbs, P-MALA et P-MYULA
- Applications numériques : réduction de PAPR

$$\forall \mathbf{x} \in \mathbb{R}^K, \quad \mathrm{PAPR}(\mathbf{x}) = \frac{K\|\mathbf{x}\|_\infty^2}{\|\mathbf{x}\|_2^2}$$

Valorisation : IEEE TSP, IEEE SSP & Gretsi

La loi démocratique

$$p(\boldsymbol{x}) = \frac{\lambda^K}{2^K K!} \exp\left(-\lambda \|\boldsymbol{x}\|_\infty\right)$$



Distribution marginale d'une composante x_k

$$\mathrm{p}(x_k) = \frac{\lambda}{2K}\left(\sum_{j=0}^{K-1}\frac{\lambda^j}{j!}|x_k|^j\right)\exp\left(-\lambda|x_k|\right)$$

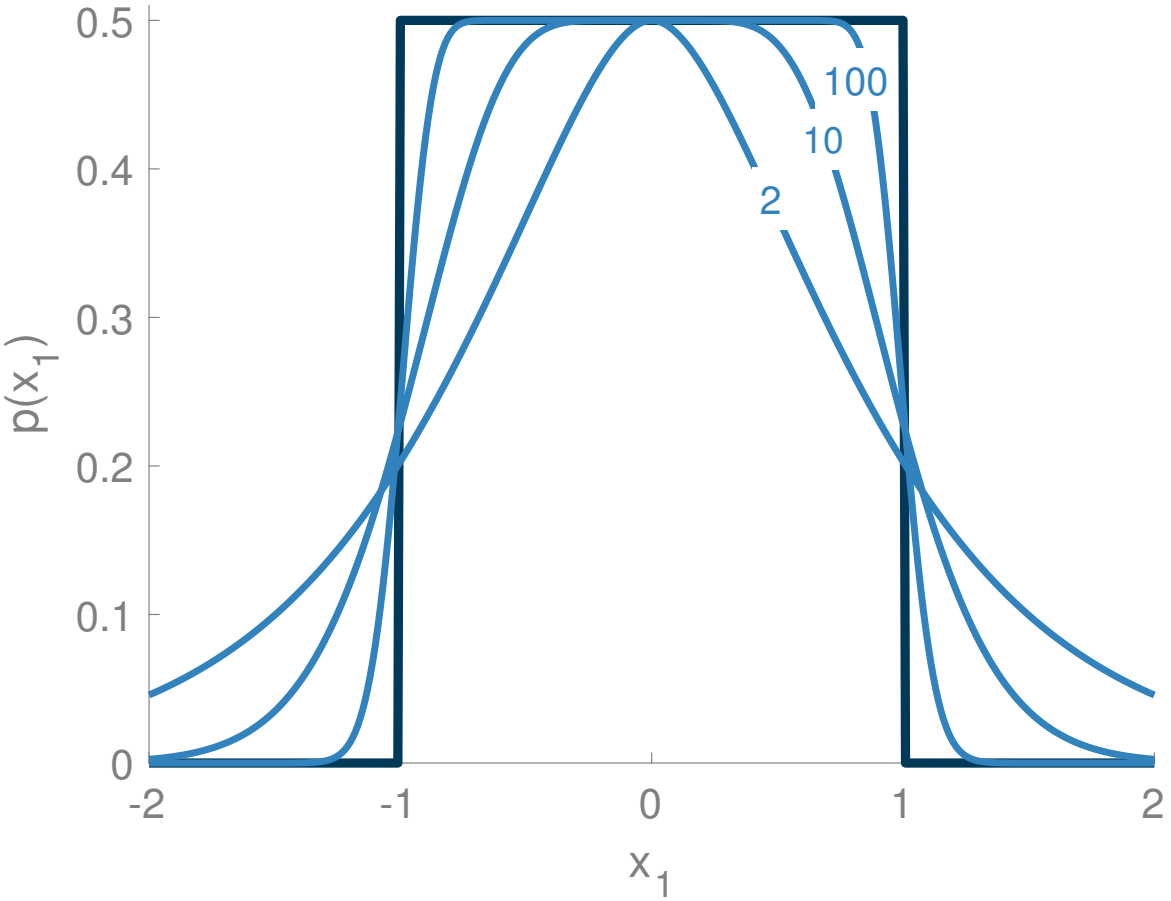
(série de Taylor tronquée de exp) $\times$ exp

Résultat

$$\frac{\lambda}{K}X_k \xrightarrow{\mathcal{L}} \mathcal{U}_{[-1,1]}$$

Distribution marginale d'une composante x_k

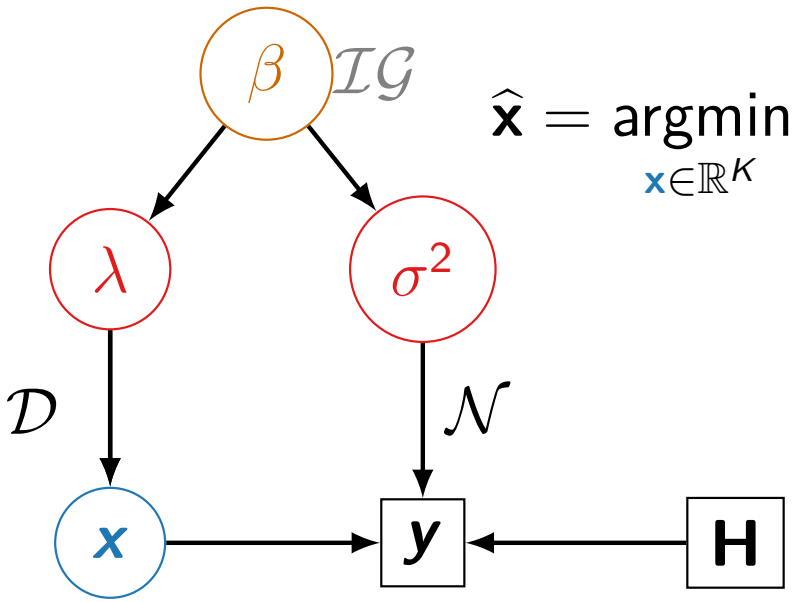
$$\frac{\lambda}{K} x_k \xrightarrow{\mathcal{L}} \mathcal{U}_{[-1,1]}$$



Distribution marginale renormalisée $\lambda = 3$ et $K = 2, 10, 100$

Formulation bayésienne du codage antiparcimonieux

$\boldsymbol{y} = \mathbf{H}\boldsymbol{x} + \boldsymbol{r}$ erreur résiduelle $\neq$ bruit



MAP :

$$\widehat{\mathbf{x}} = \underset{\mathbf{x}\in\mathbb{R}^K}{\operatorname{argmin}} \; \|\mathbf{y} - \mathbf{H}\mathbf{x}\|_2^2 + \underbrace{2\lambda\sigma^2}_{\beta} \|\mathbf{x}\|_\infty$$

connue, de taille $D \times K$

$D \leq K$

$$p(\boldsymbol{x}) \propto \exp(-\lambda \|\boldsymbol{x}\|_\infty)$$

Deux estimateurs

MAP marginalisé (mMAP) - mode *a posteriori*

$$\begin{aligned}\widehat{\mathbf{x}}_{\text{mMAP}} &= \underset{\mathbf{x}\in\mathbb{R}^K}{\text{argmax}} \int_{\mathbb{R}_+} f\left(\mathbf{x}, \beta|\mathbf{y}\right) \text{p}(\beta)\text{d}\beta \\ &\simeq \underset{\mathbf{x}^{(t)}\in\mathcal{X}}{\text{argmax}}\, f\left(\mathbf{x}^{(t)}|\mathbf{y}\right)\end{aligned}$$

MMSE - moyenne *a posteriori*

$$\begin{aligned}\widehat{\mathbf{x}}_{\text{MMSE}} &= \text{E}\left[\mathbf{x}|\mathbf{y}\right] \\ &\simeq \frac{1}{T_r} \sum_{t=T_{\text{bi}}+1}^{T_{\text{MC}}} \mathbf{x}^{(t)}\end{aligned}$$

Algorithme MCMC générique

pour $t \leftarrow 1$ **à** T_{MC} **faire**

Tirer $\beta^{(t)} \sim \mathcal{IG}$ - *conjuguée* ;

Tirer un code $\mathbf{x}^{(t)}$ selon

1. Composante x_k par composante : Gibbs

ou

2. Vecteur entier : P-MALA

ou

3. Vecteur entier : P-MYULA

fin

Sortie : une collection d'échantillons $\{\beta^{(t)}, \boldsymbol{x}^{(t)}\}$;

P-MYULA : principe

Durmus et al. (2017)

Processus de diffusion de Langevin discrétisé

$$\text{ULA} \quad : \quad \mathbf{x}^{(t_{n+1})} = \mathbf{x}^{(t_n)} + \delta \nabla \log \mathrm{p}(\mathbf{x}^{(t_n)}|\mathbf{y}, \beta) + \sqrt{2\delta}\mathbf{w}^{(t_{n+1})}$$

où

$$\mathrm{p}(\mathbf{x}|\mathbf{y}, \beta) \propto \exp\Big(- f_1(\boldsymbol{x}) - \lambda g_0(\boldsymbol{x}) \Big) \notin \mathcal{C}^1$$

avec $f_1(\boldsymbol{x}) = \dfrac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{H}\mathbf{x}\|_2^2$ et $g_0(\boldsymbol{x}) = \|\mathbf{x}\|_\infty$

diff. *non diff.*

Solution : g_0 est *régularisée* pour améliorer le mélange

$$\Longrightarrow g_\delta(\mathbf{x}) = \inf_{\mathbf{u}} \; g_0(\mathbf{u}) - \frac{1}{2\delta}\|\mathbf{x} - \mathbf{u}\|_2 \;=\; \mathrm{prox}_{\lambda\delta g_0}(\boldsymbol{x})$$

MY : enveloppe de Moreau-Yoshida de g_0

P-MYULA en pratique

1. Descente de gradient généralisée

$$\mathbf{x}^{(t+1/2)} = \left(1 - \frac{\gamma}{\delta}\right)\mathbf{x}^{(t)} - \gamma\nabla f_1\left(\mathbf{x}^{(t)}\right) + \frac{\gamma}{\delta}\mathrm{prox}_{\delta\lambda g_0}\left(\mathbf{x}^{(t)}\right)$$

où $\gamma = L_{f_1}^{-1}$ et $\delta = \frac{\gamma}{4}$ et

$$\mathrm{prox}_{\delta\lambda\|\cdot\|_\infty}(\mathbf{x}) = \mathbf{x} - \delta\lambda\Pi_{\mathbf{u},\|\mathbf{u}\|_1\leq 1}(\mathbf{x})$$

Condat (2015)

2. Faire un pas de marche aléatoire

$$\mathbf{x}^{(t+1)} \sim \mathcal{N}\left(\mathbf{x}^{(t+1/2)}, 2\gamma I_N\right)$$

Complexité d'une itération

Gibbs

- K gradients $\sim \mathcal{O}(DK^2)$
- K lois multinomiales
- K lois normales tronquées
- 1 loi uniforme sur les permutations de $1, \dots, K$

P-MYULA

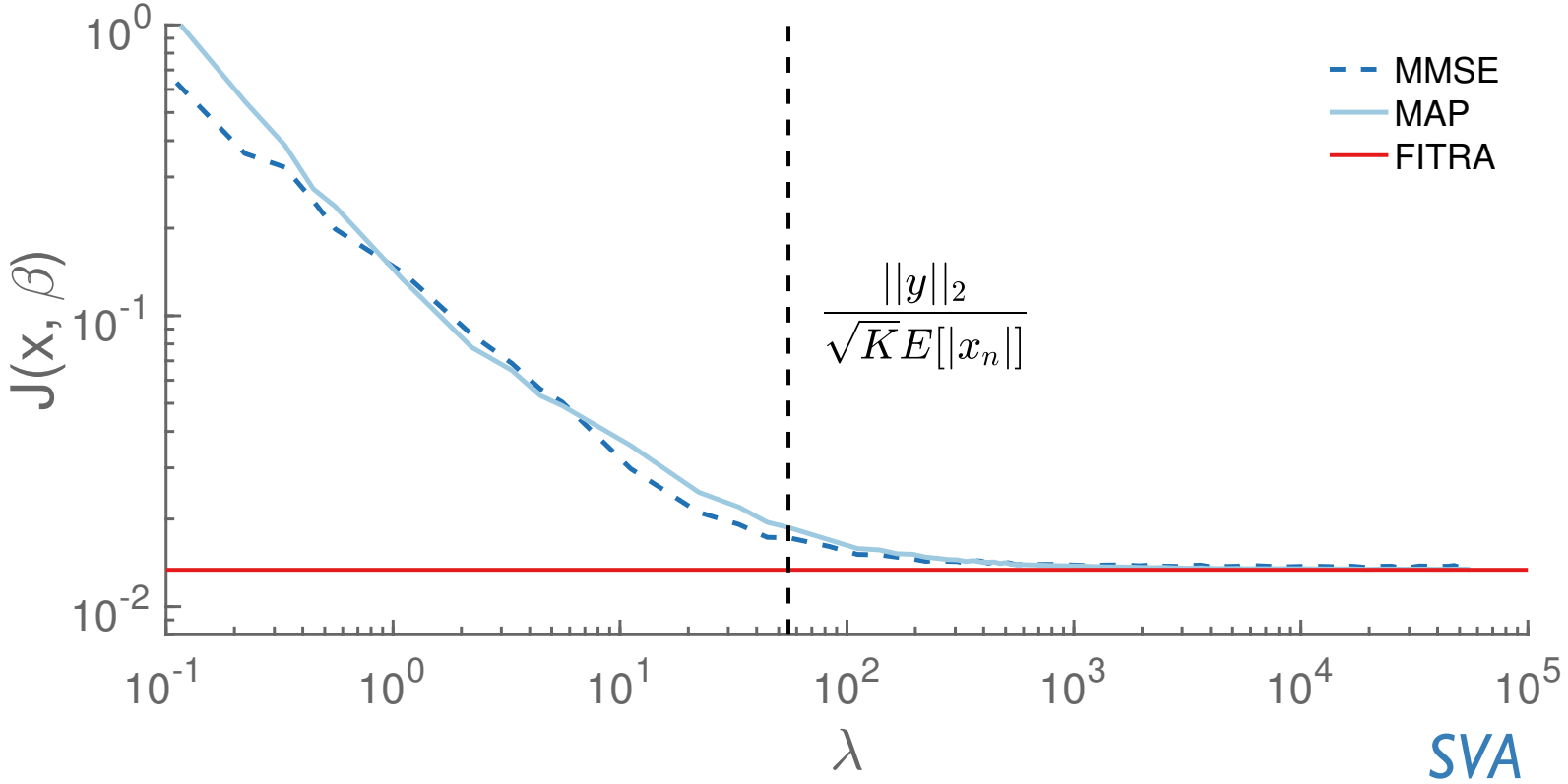
- 1 gradient $\sim \mathcal{O}(DK^2)$
- $\text{prox}_{\lambda\|\cdot\|_\infty} \sim \mathcal{O}(K)$
- K lois normales
- une loi uniforme sur $[0, 1]$

En pratique P-MYULA est $10\times$ plus rapide

Réglage des paramètres

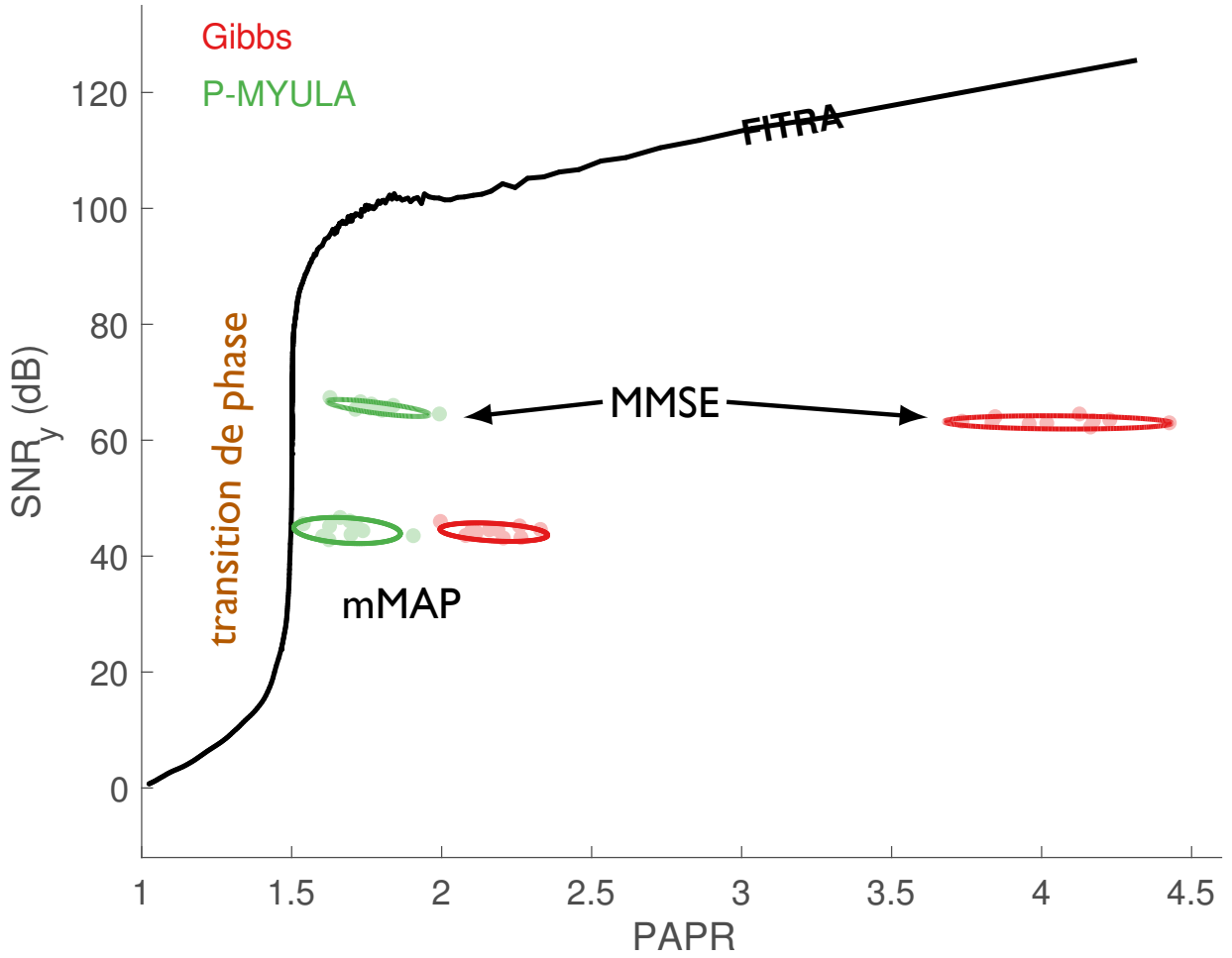
Le MAP dépend de : $J(\mathbf{x}, \beta) = \frac{1}{2} \|\mathbf{y} - \mathbf{H}\mathbf{x}\|_2^2 + \beta \|\mathbf{x}\|_\infty$

besoin de λ, σ^2 t.q. $2\lambda\sigma^2 = \beta \implies$ 1 degré de liberté



Réduction de PAPR $D = 30,\ K = 50$

$$\mathrm{SNR}_{\mathbf{y}}(\widehat{\mathbf{x}}) = \frac{\|\mathbf{y}\|_2^2}{\|\mathbf{y} - \mathbf{H}\widehat{\mathbf{x}}\|_2^2} \qquad \mathrm{PAPR}(\widehat{\mathbf{x}}) = \frac{K\|\widehat{\mathbf{x}}\|_\infty^2}{\|\widehat{\mathbf{x}}\|_2^2}$$



Conclusion partielle

- Formulation bayésienne du codage antiparcimonieux
- loi démocratique $p(\mathbf{x}) \propto \exp(-\lambda\|\mathbf{x}\|_\infty)$
- Lois conditionnelles conjuguées …

  & algorithmes de Monte Carlo proximaux rapides et efficaces
- Aucun réglage de paramètre $\longrightarrow$ non supervisé

« Bayesian anti-sparse coding » IEEE TSP 2017 + SSP'16 + Gretsi'17

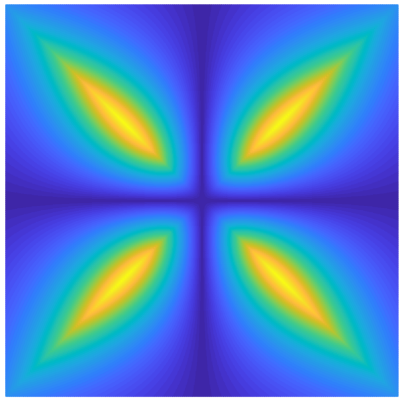
Toolbox Matlab / Mex disponible en ligne

`github.com/c-elvira/bayesian_antisparse_algorithm`

Perspectives

- Généralisation de la loi démocratique = nouvelle régularisation

$$f(\mathbf{x}) = \frac{\lambda^{K\gamma}\gamma^{K}}{2^{K}\Gamma(K\gamma+1)} \prod_{k=1}^{K} |x_k|^{\gamma} e^{-\lambda\|\mathbf{x}\|_\infty},$$



- Nouvel estimateur ? Les solutions de

$$(\mathrm{P}_\infty): \quad \min \ \|\mathbf{x}\|_\infty \quad \text{t.q.} \quad \mathbf{y} = \mathbf{H}\mathbf{x}$$

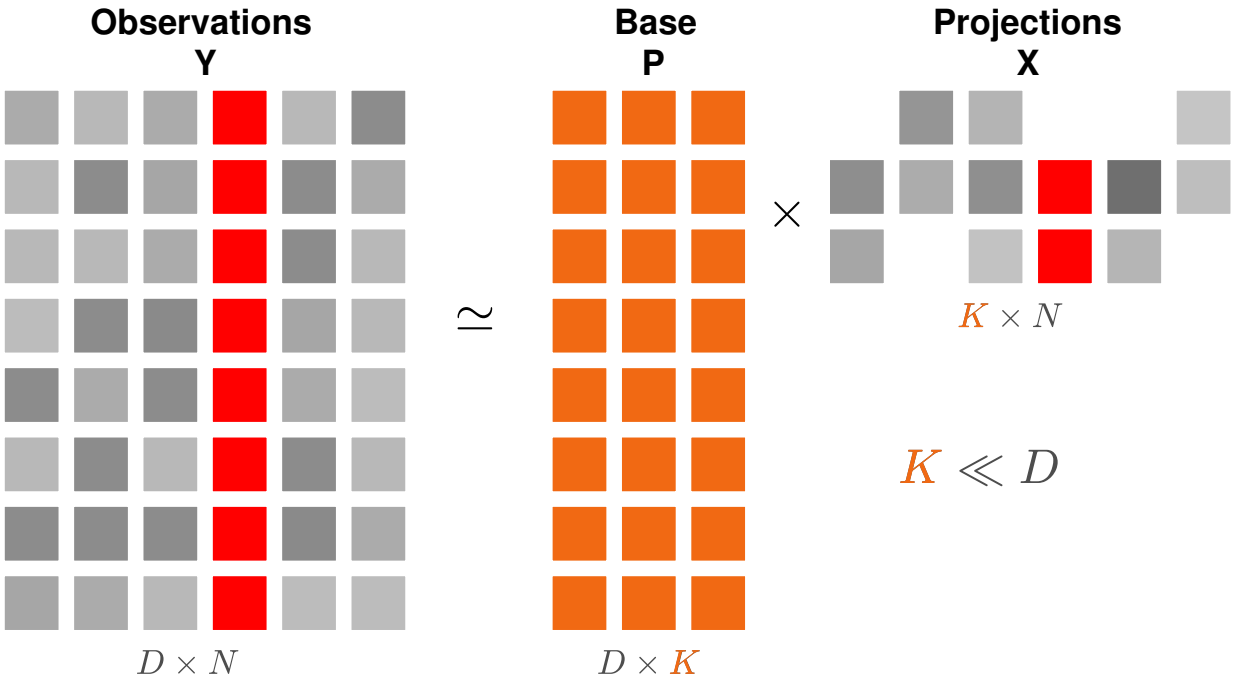
$\Rightarrow K - \mathrm{rang}(\mathbf{H}) + 1$ valeurs extrêmes *(Cadzow, 1971)*

signe($\mathbf{x}$)|$\mathbf{y}$ pour trouver la position des valeurs extrêmes ?

Partie II

Bayésien non paramétrique pour l'estimation de sous-espaces

Représentation en plus petite dimension



Valeur de K $\longrightarrow$ pertinence de la réduction de la dimension
$\longrightarrow$ impacte les performances
$\longrightarrow$ stockage
$\longrightarrow$ bruit capturé

Estimation de sous-espaces & dimension

- Incontournable : Analyse en Composantes Principales
- ACP probabiliste $\rightarrow$ facteurs latents *Tipping & Bishop (1999)*

  Parcimonie

  $\rightarrow$ ACP parcimonieuse *Zou & al. (2006)*

  $\rightarrow$ SVD régularisée *Dobigeon & Tourneret (2010)*

- Estimation de K

  $\rightarrow$ approximations analytiques *Minka (2000)*, *Šmídl & Quinn (2007)*

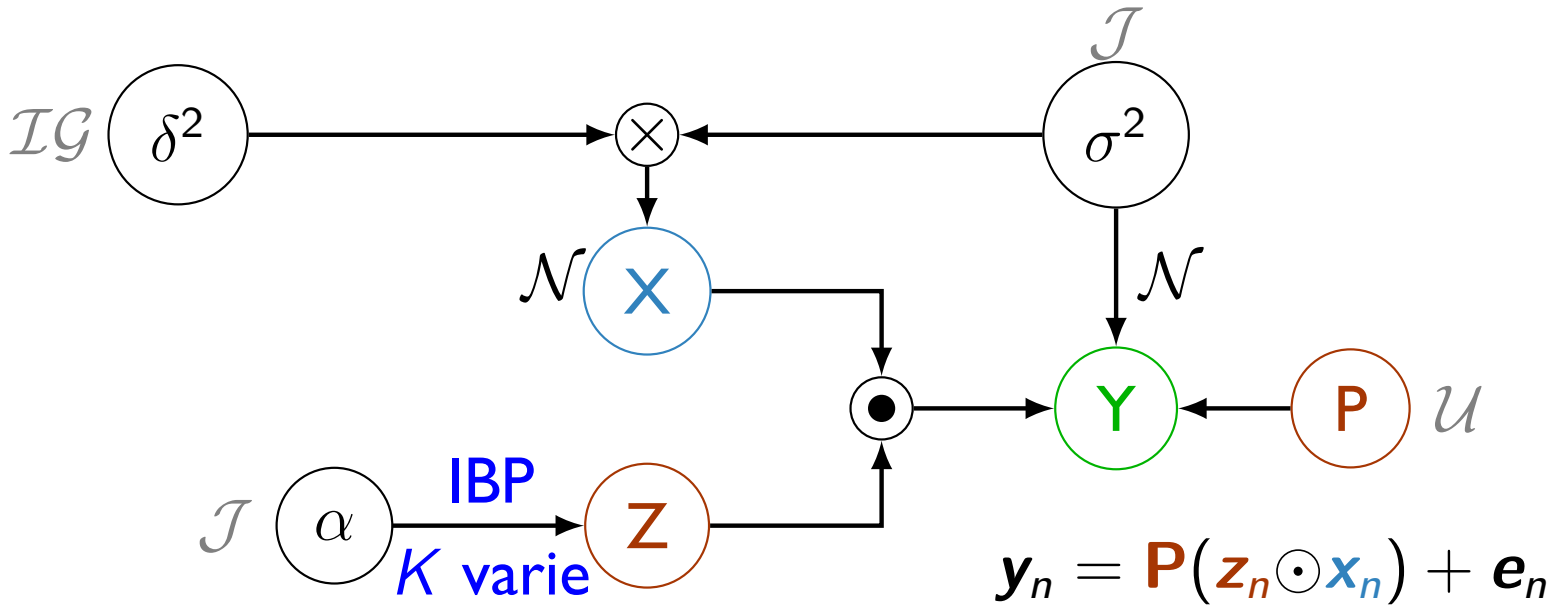
  $\rightarrow$ RJMCMC *Zhang & al. (2004)*

Contributions

- ACP non paramétrique parcimonieuse et flexible
  $\rightarrow$ matrice orthonormée + processus du buffet indien
- Méthode d'échantillonnage + estimateurs
- Étude théorique de la consistance de $\widehat{K} \,|\, \boldsymbol{Y}$
- Étude numérique
- Application
  $\longrightarrow$ Couplage avec un séparateur linéaire
  $\longrightarrow$ Hyperspectral

Article soumis + IEEE'ICASSP

Modèle BNP-PCA



$e_n \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_D)$, bruit additif gaussien

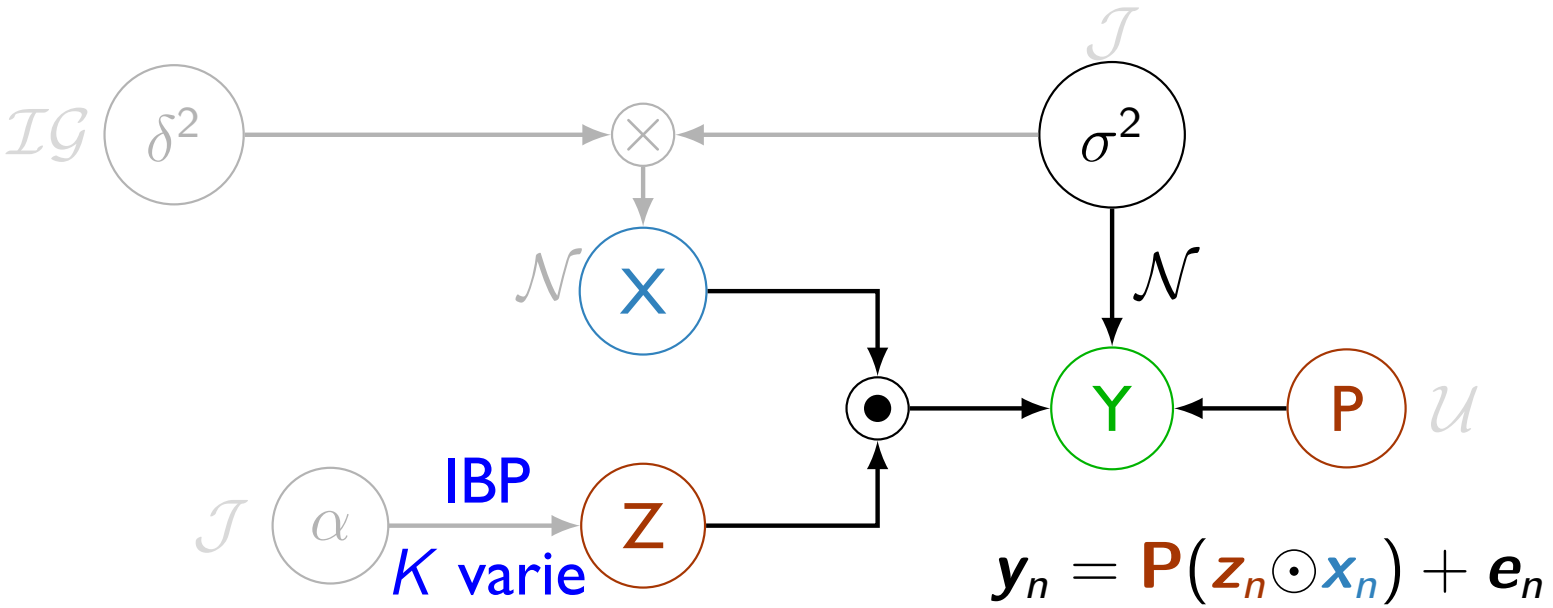
$\mathbf{P} = [\mathbf{p}_1 \dots \mathbf{p}_K, \mathbf{p}_{K+1} \dots \mathbf{p}_D]$, $\mathbf{P}^t\mathbf{P} = \mathbf{I}_D$, et $\mathbf{P} \sim \mathcal{U}_{\mathcal{O}_D}$

$Z = [z_1 \dots z_N] \sim \text{IBP}(\alpha)$ matrice binaire $\rightarrow K$

$x_n = [x_{n,1} \dots x_{n,K}]$ $\forall k,\ x_{n,k} \sim \mathcal{N}(0, \delta_k^2 \sigma^2)$

$\theta = \{\delta^2, \sigma^2, \alpha\}$ loi conjuguée vague

Modèle BNP-PCA



$\boldsymbol{e}_n \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_D)$, bruit additif gaussien

$\mathbf{P} = [\mathbf{p}_1 \dots \mathbf{p}_K, \mathbf{p}_{K+1} \dots \mathbf{p}_D]$, $\mathbf{P}^t\mathbf{P} = \mathbf{I}_D$, et $\mathbf{P} \sim \mathcal{U}_{\mathcal{O}_D}$

$\boldsymbol{Z} = [\boldsymbol{z}_1 \dots \boldsymbol{z}_N] \sim \text{IBP}(\alpha)$ matrice binaire $\rightarrow K$

$\boldsymbol{x}_n = [x_{n,1} \dots x_{n,K}]$ $\forall k,\ x_{n,k} \sim \mathcal{N}(0, \delta_k^2 \sigma^2)$

$\theta = \{\delta^2, \sigma^2, \alpha\}$ loi conjuguée vague

Le processus du Buffet Indien *Griffiths and Ghahramani (2006)*

- *A priori* sur les matrices binaires $\rightarrow$ parcimonie

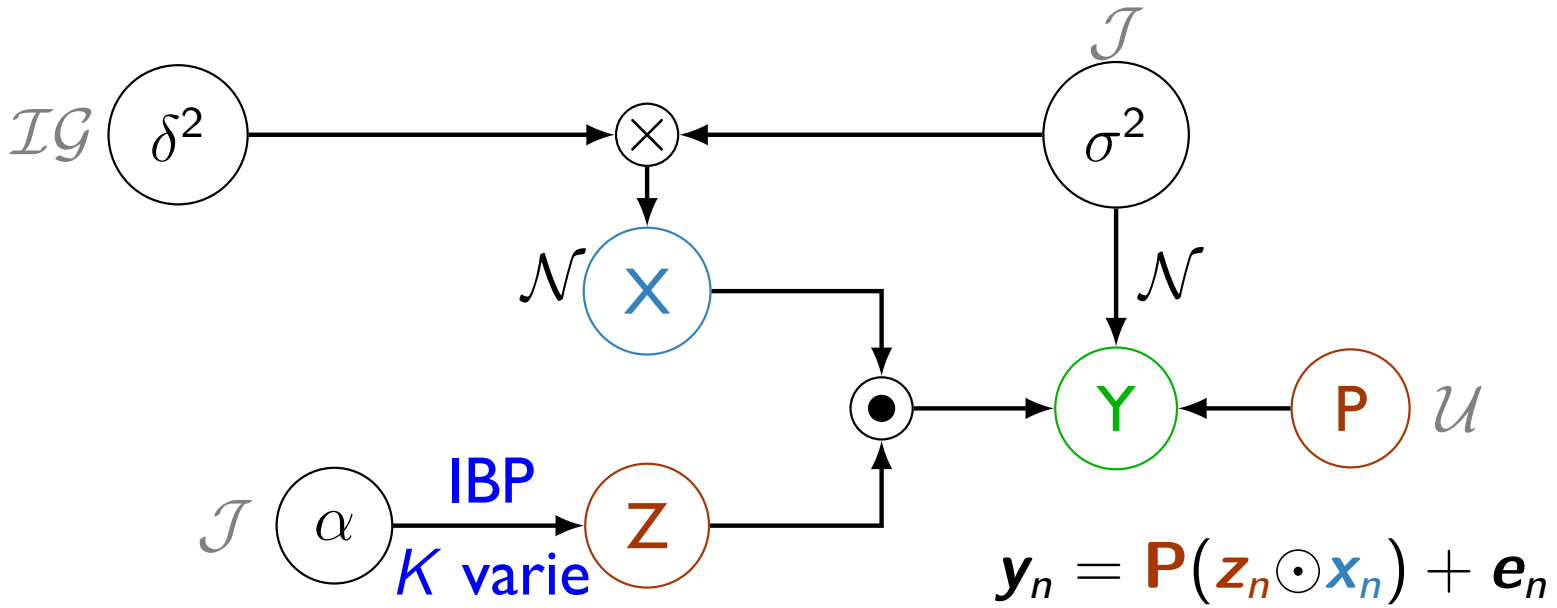
  cf. # coefs non nuls pour $\mathbf{y}_n$

- Taille : $\infty \times N$

- Effet régularisant : $\mathbb{E}[K] = \alpha \log(N)$

  cf. réduction de la dimension

Modèle BNP-PCA



$\boldsymbol{e}_n \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_D)$, bruit additif gaussien

$\mathbf{P} = [\mathbf{p}_1 \dots \mathbf{p}_K, \mathbf{p}_{K+1} \dots \mathbf{p}_D]$, $\mathbf{P}^t\mathbf{P} = \mathbf{I}_D$, et $\mathbf{P} \sim \mathcal{U}_{\mathcal{O}_D}$

$\boldsymbol{Z} = [\boldsymbol{z}_1 \dots \boldsymbol{z}_N] \sim \text{IBP}(\alpha)$ matrice binaire $\rightarrow K$

$\boldsymbol{x}_n = [x_{n,1} \dots x_{n,K}]$ $\forall k,\ x_{n,k} \sim \mathcal{N}(0, \delta_k^2 \sigma^2)$

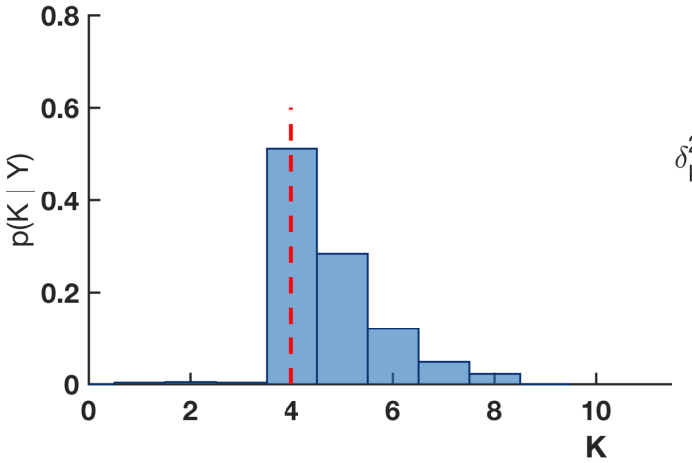
$\boldsymbol{\theta} = \{\boldsymbol{\delta^2}, \sigma^2, \alpha\}$ loi conjuguée vague

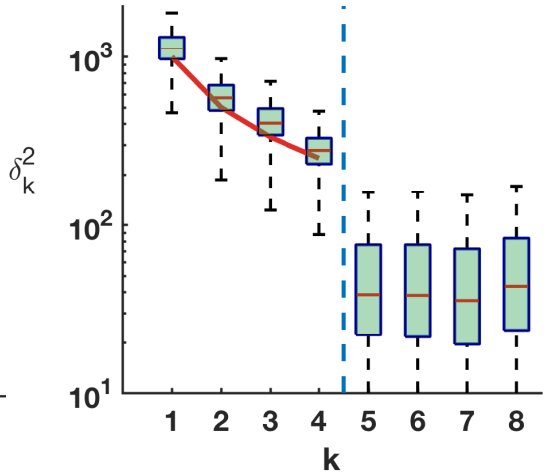
Exemple

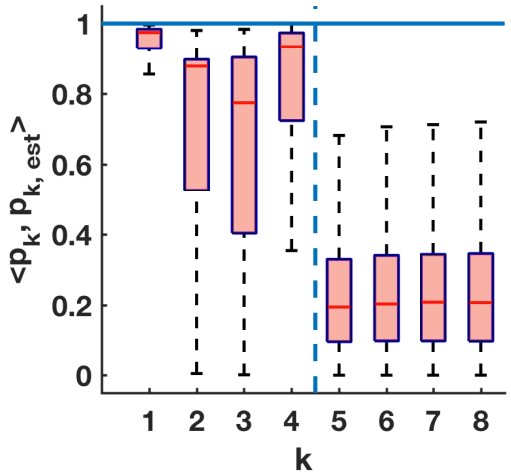
$N = 500$ observations, $D = 16$, $K_{\text{true}} = 4$ selon

$$\boldsymbol{y} = \mathbf{P}\boldsymbol{x} + \boldsymbol{e}$$

avec $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \Sigma)$, $\boldsymbol{e} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_D)$







Estimateur MAP conditionnel de $K|\mathbf{H}, \alpha$

Résultat théorique

$$\forall k,\ \limsup_{N\to+\infty}\ \mathrm{P}\Big[K_N = k \mid \boldsymbol{y}_1 \dots \boldsymbol{y}_N, \alpha\Big] < 1$$

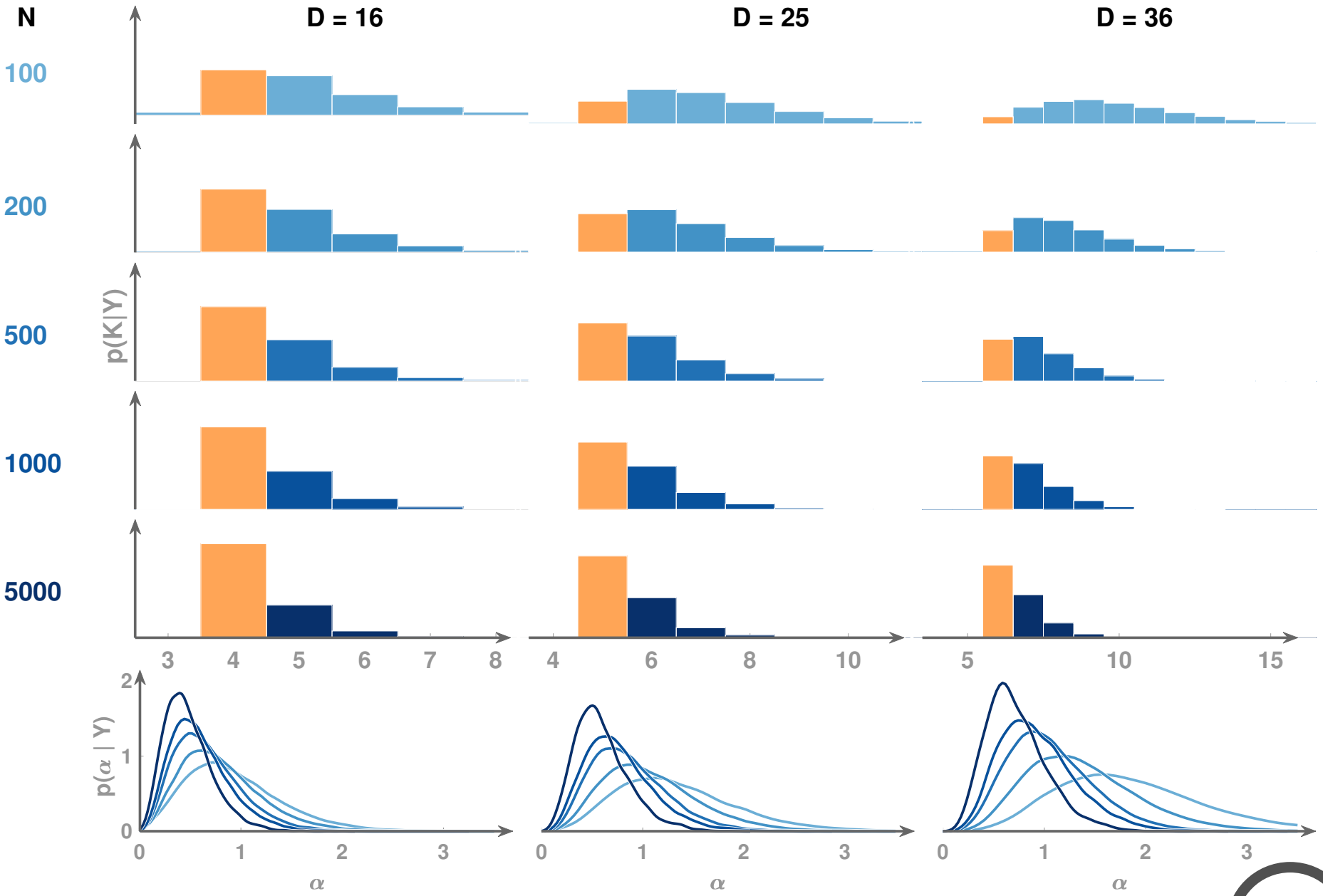
Si $\mathbf{y}_n$ = bruit blanc $\quad K_N \xrightarrow{\text{p.s.}} 0$

similaire à *Miller and Harrison (2014)* pour le clustering

$\Rightarrow$ L'estimateur MAP $\widehat{K}|\alpha$ n'est pas consistant

mauvaise nouvelle

Étude numérique de la consistance de $K|\mathbf{Y}$



Le processus du Buffet Indien *Griffiths and Ghahramani (2006)*

- *A priori* sur les matrices binaires $\rightarrow$ parcimonie

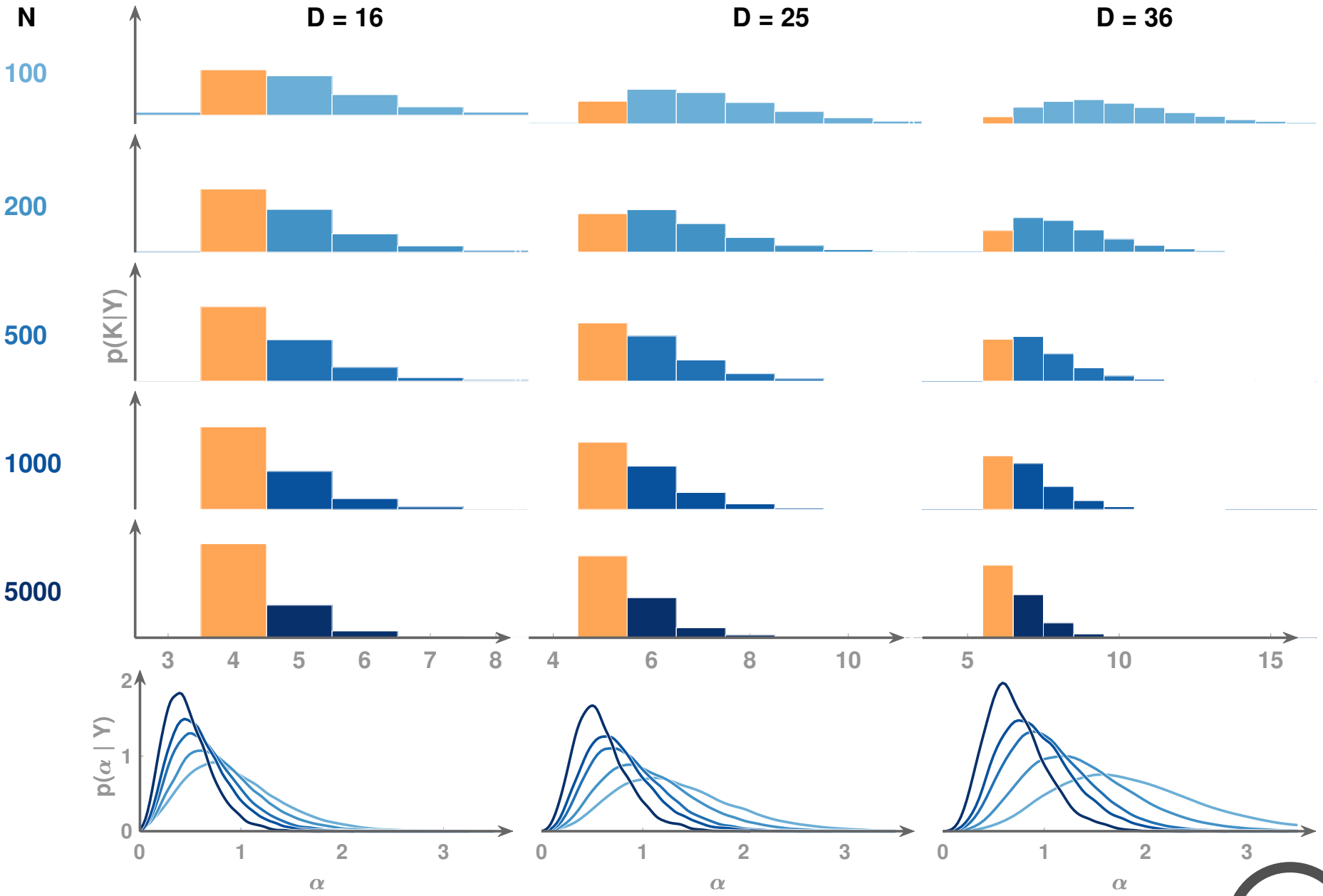
  cf. # coefs non nuls pour $\mathbf{y}_n$

- Taille : $\infty \times N$

- Effet régularisant : $\mathbb{E}[K] = \alpha \log(N)$

  cf. réduction de la dimension

Étude numérique de la consistance de $K|\mathbf{Y}$



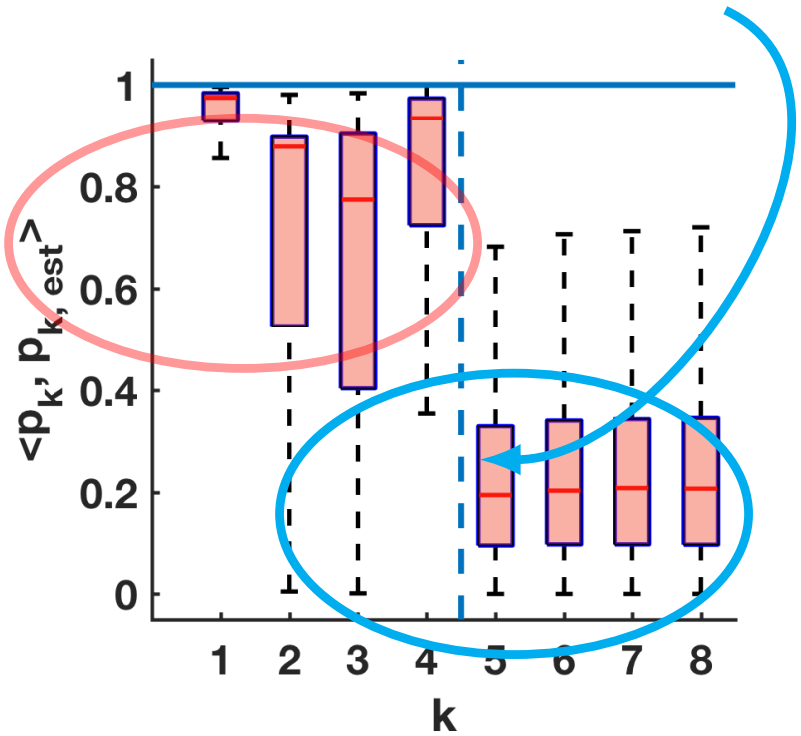
Nouvel estimateur de K consistant

- K directions pertinentes $\Rightarrow$ $D - K$ directions non pertinentes
- $\mathbf{p}_k$ non pertinent $\Rightarrow$ $\mathbf{p}_k|\mathbf{Y} \simeq$ uniforme

$\Rightarrow$ test d'uniformité

✗ $\mathbf{p}_k|\mathbf{Y}$: test vectoriel

✓ $\langle \mathbf{p}_k, \mathbf{u} \rangle|\mathbf{Y}$: test $1D$, support compact



$\Rightarrow$ distribution théorique

Nouvel estimateur de K

Soient $\|\boldsymbol{u}\|_2 = 1$ $\quad$ $\mathbf{p}$ uniforme sur $\mathcal{O}_{D-K}$ $\quad$ $W_{D-K} \triangleq |\langle \boldsymbol{u}, \mathbf{p} \rangle|$

$$\mathrm{p}\left(W_{D-K} \leq \lambda\right) \;=\; \text{expression connue}$$

$$\Rightarrow \text{test de Kolmogorov-Smirnov}$$

test statistique

$$\mathcal{H}_0^K : |\langle \boldsymbol{u}, \mathbf{p}_{K+1} \rangle|, \ldots, |\langle \boldsymbol{u}, \mathbf{p}_D \rangle| \sim W_{D-K}$$

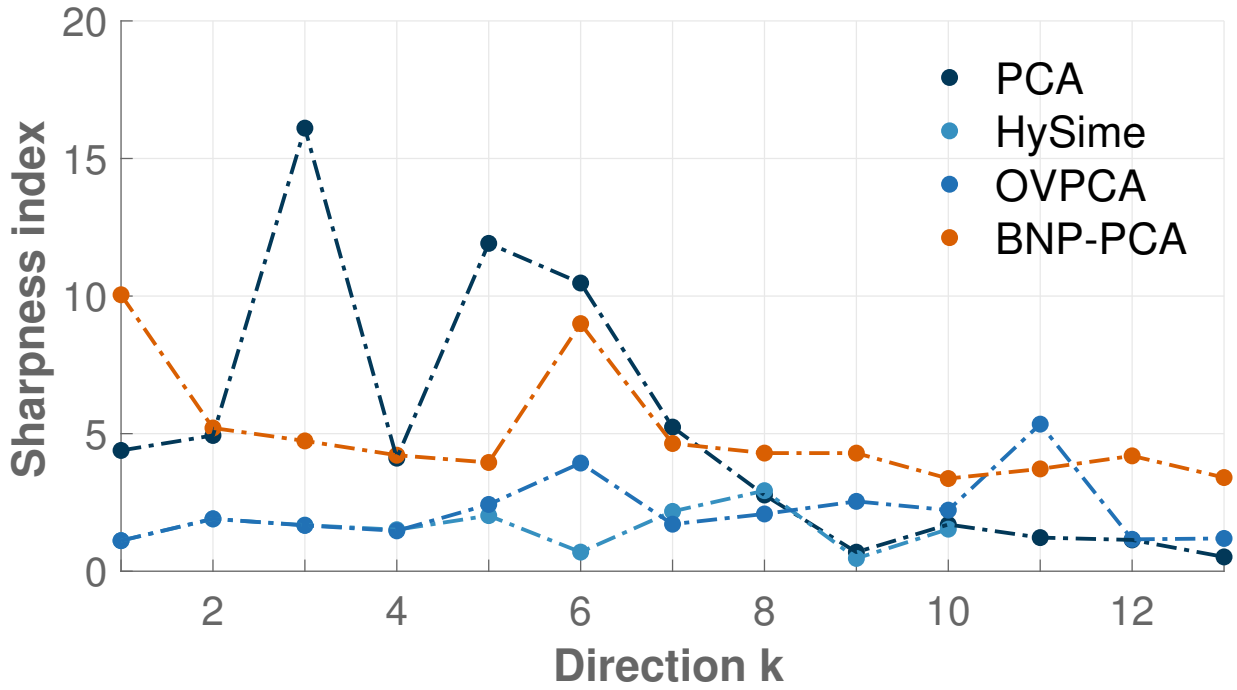
nouvel estimateur

$$\widehat{K}_{\mathrm{KS}} = \min_K \left\{ K \;:\; \mathcal{H}_0^K \text{ est vraie} \right\}$$

Application : sous-espace d'un cube hyperspectral

$\boldsymbol{y} \in \mathbb{R}^{\#\text{long. onde}}$ vérité terrain : $K \simeq 10$

Algorithme	$\widehat{K}$
L-PCA	25
OVPCA	23
HySime	10
$\widehat{K}_{\text{KS}}$	13
$\widehat{K}_{\text{MAPm}}$	25



HySime *Bioucas-Dias & Nascimento (2008)*

Estimer **P** *Besson et al (2009)*

Sharpness index *Leclaire & Moisan (2015)*

Conclusion partielle

- BNP-PCA : IBP + variété de Stiefel *aucun réglage*
- Étude théorique du MAP conditionnel - α *fixé*
  $\longrightarrow$ inconsistant
- Étude expérimentale du MAP - α *échantillonné*
  $\longrightarrow$ empiriquement consistant
- Nouvel estimateur : test d'uniformité
  $\longrightarrow$ théoriquement consistant

Preprint "Towards Bayesian non parametric PCA" + IEEE'ICASSP 2017

`https://arxiv.org/pdf/1709.05667.pdf`

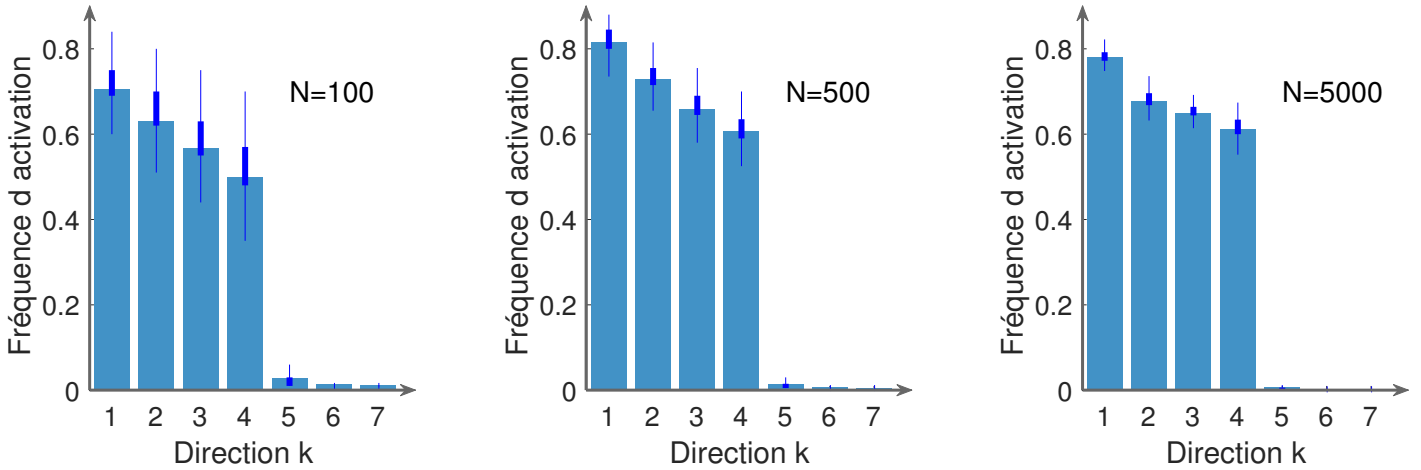
codes et mwe Matlab disponibles en ligne

Perspectives

- Complexité numérique / implémentation
  & robustesse des estimateurs aux modèles génératifs

- Généraliser les résultats d'inconsistance

- Fréquences d'activation $\rightarrow$ nouvel estimateur ?



$\Rightarrow$

ACP fonctionnelle ?

- $p(\mathbf{P}, \boldsymbol{Z}, \theta | \boldsymbol{Y}) \rightarrow$ dépend d'un produit scalaire

Conclusion

Codage antiparcimonieux

- Bayésien $=$ meilleure compréhension du problème
- Régularisation $\rightarrow$ nouvelle loi de probabilité
$$p(\mathbf{x}) \propto \exp(-\lambda \|\mathbf{x}\|_\infty)$$
- Échantillonnage efficace : Proximal Monte Carlo

Estimation de sous-espaces

- BNP-PCA
- Étude théorique et numérique du comportement des estimateurs de la dimension
- Applications numériques

Publications

Revues

[A1] "Bayesian Antisparse Coding", dans *IEEE Transactions on Signal Processing*, 2017

[A2] "Bayesian nonparametric principal component analysis", soumis en juillet 2017

Conférences internationales

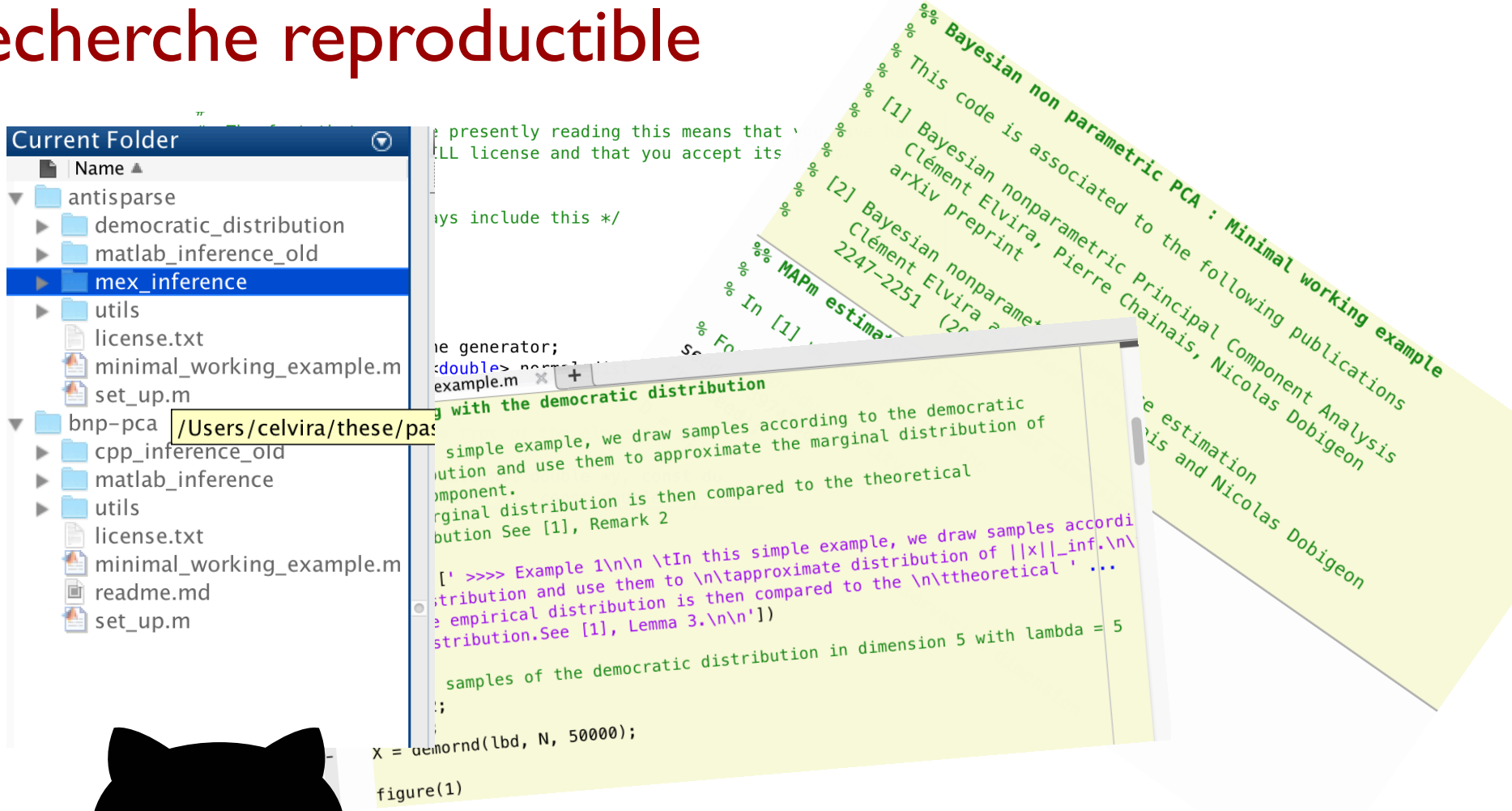
[C1] "Democratic prior for antisparse coding", dans *Proc. IEEE Statistical Signal Processing Workshop (SSP)*, Juin 2016

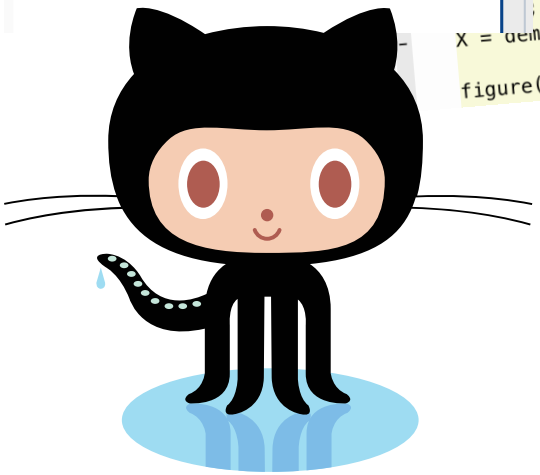
[C2] "Bayesian nonparametric subspace estimation", dans *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing (ICASSP)*, mars 2017
student paper contest finalist

Conférence nationale

[N1] "Une formulation bayésienne du codage antiparcimonieux", dans *Actes du Colloque GRETSI*, septembre 2017

Recherche reproductible





https://github.com/c-elvira

Modèles bayésiens pour l'identification de représentations antiparcimonieuses et l'analyse en composantes principales non paramétrique

Clément Elvira

Directeurs de thèse :
Pierre Chainais et Nicolas Dobigeon

10 novembre 2017

Annexes

Codage antiparcimonieux

- Construction de **H**
- échantillonnage prior
- Gibbs
- P-MALA
- Fom versus *K*
- Vers un processus

Estimation de sous-espaces

- Échantillonnage
- Complexité
- Preuve
- Consistance
- BNP 3 paramètres
- Manip bruit blanc
- Manip clustering

Construction des matrices H

retour

Condition sur le noyau — *Cadzow (1971)*

les solutions de $\mathrm{P}_{\infty}^{\varepsilon}$ possèdent au moins $K - \mathrm{rang}(\mathbf{H}) + 1$ valeurs extrêmes

Principe d'incertitude matricielle

H vérifie le PIM ssi $\exists \nu, \gamma < 1$ t.q.

$$\|\boldsymbol{x}\|_0 \leq \nu K \quad \Longrightarrow \quad \|\mathbf{H}\boldsymbol{x}\|_2 \leq \gamma \|\boldsymbol{x}\|_2$$

$$\text{PIM} \quad \Longrightarrow \quad \forall \boldsymbol{y},\ \exists \boldsymbol{x} \ \text{ t.q. } \ \boldsymbol{y} = \mathbf{H}\boldsymbol{x} \quad \text{et} \quad \|\boldsymbol{x}\|_\infty \leq \frac{1}{(1-\nu\sqrt{\gamma})} \frac{\|\boldsymbol{x}\|_2}{K}$$

Lyubarskii et Vershynin (2010)

PIM = version affaiblie de la condition RIP

$\longrightarrow$ pas d'unicité

Matrices vérifiant la condition PIM

retour

$\longrightarrow$ recourt à l'aléatoire

Lyubarskii et Vershynin (2010)

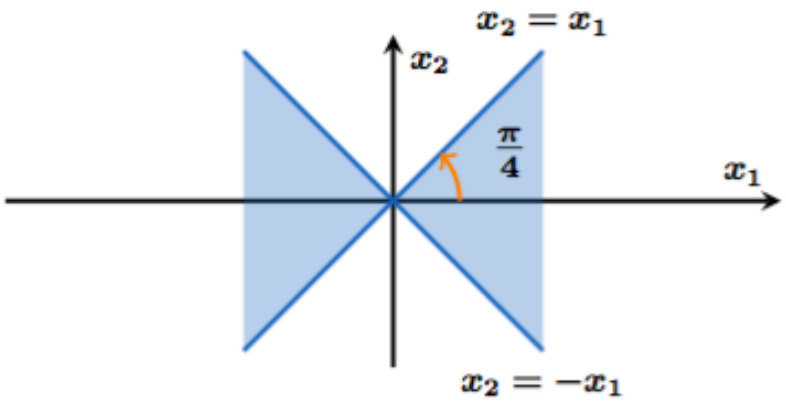
- Sous-échantillonnage de matrices orthogonales
- Sous-échantillonnage de matrices de coefficients de Fourrier
- Matrice dont les coefficients sont des v.a. sous-gaussiennes

Loi conditionnelle

retour

$\{\mathcal{C}_k\}$ ensemble de cônes

$$\mathcal{C}_k \triangleq \left\{ \mathbf{x} = [x_1, \dots, x_K]^T \in \mathbb{R}^K \ : \ \forall j \neq k, \ |x_k| > |x_j| \right\}$$



- $\mathrm{P}[\boldsymbol{x} \in \mathcal{C}_k] = \frac{1}{K}$
- $x_k \mid \boldsymbol{x} \in \mathcal{C}_k$ *(di-gamma)*
- $\boldsymbol{x}_{\setminus k} \mid x_k; \boldsymbol{x} \in \mathcal{C}_n$ *(uniforme)*
- $\boldsymbol{x}_{\setminus k} \mid \boldsymbol{x} \in \mathcal{C}_k$ *(democratique...)*

Échantillonneur exacte

retour

$$\mathrm{p}(\mathbf{x}) = \sum_{k=1}^{K} \left[\prod_{j \neq n} \mathrm{p}\left(x_j | x_k, \mathbf{x} \in \mathcal{C}_k\right) \right] \mathrm{p}\left(x_k | \mathbf{x} \in \mathcal{C}_k\right) \mathrm{P}\left[\mathbf{x} \in \mathcal{C}_k\right]$$

Entrées : λ, K

% Indice de la composante dominante

1. Tirer k_{dom} uniformément sur $\{1 \dots K\}$;

% Valeur de la composante dominante

2. Tirer $x_{k_{\text{dom}}}$ selon une loi Gamma symétrisée ;

% Valeurs des composantes non dominantens

pour $j \leftarrow 1$ **à** K $(j \neq k_{\text{dom}})$ **faire**

    3. Tirer x_j uniformément sur $[-x_{k_{\text{dom}}}, +x_{k_{\text{dom}}}]$;

fin

Output : $\mathbf{x} = [x_1, \dots, x_K]^T \sim \mathcal{D}_K(\lambda)$

Composante par composante : Gibbs

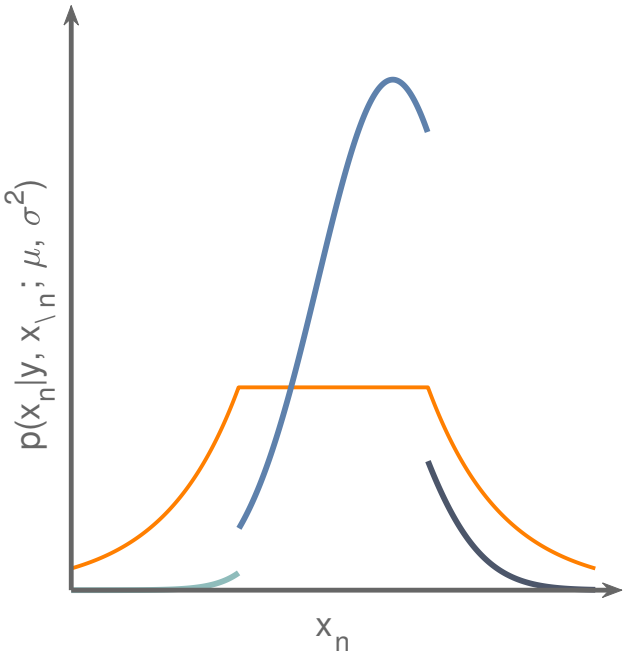
retour

Rappel : $$p\left(x_k \mid \mathbf{x}_{\setminus k}\right) \propto c_1 \mathbf{1}_{\mathcal{I}_{2k}}(x_k) + c_2 e^{-\lambda\left(|x_k| - \left\|\mathbf{x}_{\setminus k}\right\|_\infty\right)} \mathbf{1}_{\mathbb{R}\setminus\mathcal{I}_{2k}}(x_k)$$

Posterior distribution in a linear inverse model

$$x_k | \mathbf{x}_{\setminus k}, \mu, \sigma^2, \mathbf{y} \sim \sum_{i=1}^{3} \omega_{ik} \mathcal{N}_{\mathcal{I}_{ik}}\left(\mu_{ik}, s_k^2\right)$$

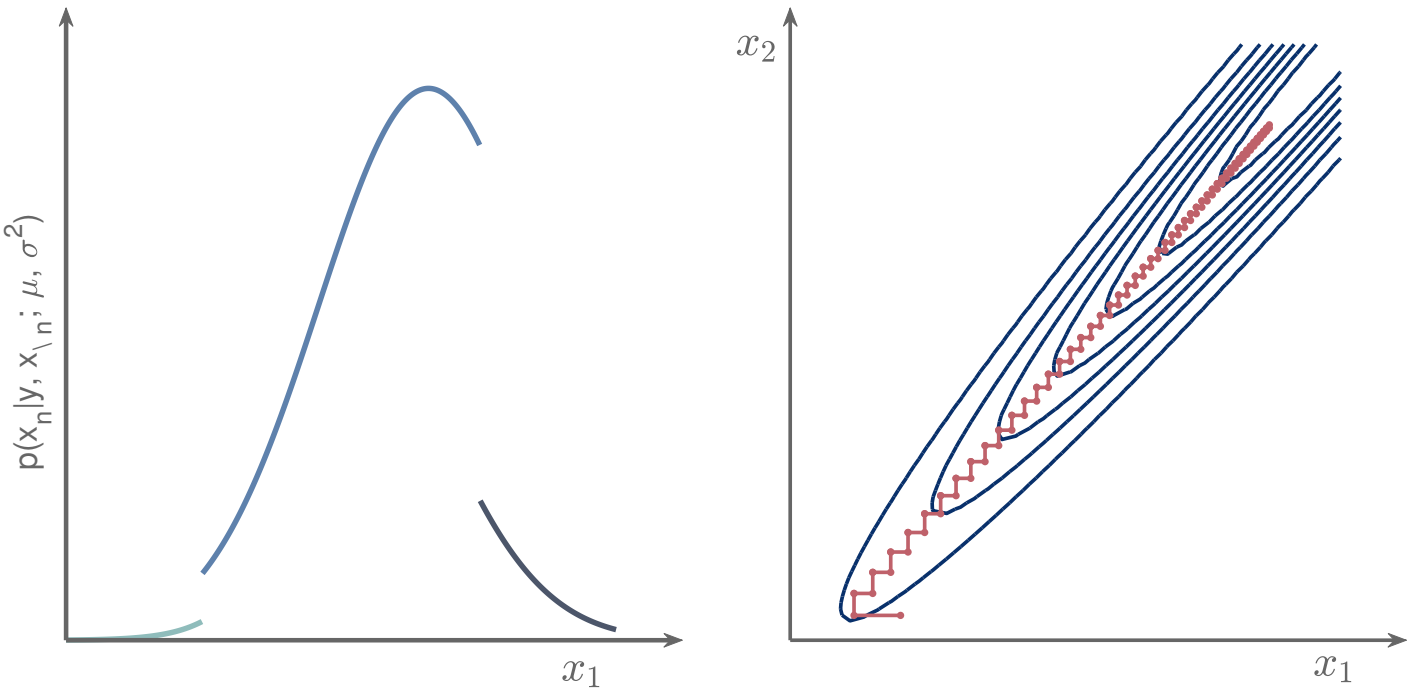
$$\mathcal{I}_{1n} = \left(-\infty, -\left\|\mathbf{x}_{\setminus k}\right\|_\infty\right)$$
$$\mathcal{I}_{2n} = \left(-\left\|\mathbf{x}_{\setminus k}\right\|_\infty, \left\|\mathbf{x}_{\setminus k}\right\|_\infty\right)$$
$$\mathcal{I}_{3n} = \left(\left\|\mathbf{x}_{\setminus k}\right\|_\infty, +\infty\right)$$



Nécessité d'une méthode plus efficace

retour

Gibbs $\Rightarrow$ mélange de la chaîne lent



- Combinatoire $\rightarrow 3^K$
- Échantillonnage
- Multimodale
- Modes étroits

P-MALA en pratique

retour

1. Descente de gradient généralisée

$$\mathbf{x}^{(t+1/2)} = \text{prox}_{\delta\lambda g_0/2}\left(\mathbf{x}^{(t)} + \delta\nabla f_1(\boldsymbol{x}^{(t)})\right)$$

où δ est ajusté pour atteindre $40-60\%$ d'acceptation

$$\text{prox}_{\delta\lambda\|\cdot\|_\infty}(\mathbf{x}) = \mathbf{x} - \delta\lambda\Pi_{\mathbf{u},\|\mathbf{u}\|_1\leq 1}(\mathbf{x})$$

Calcule de $\Pi_{\mathbf{u},\|\mathbf{u}\|_1\leq 1}$: *Condat (2015)*

2. Faire un pas de marche aléatoire

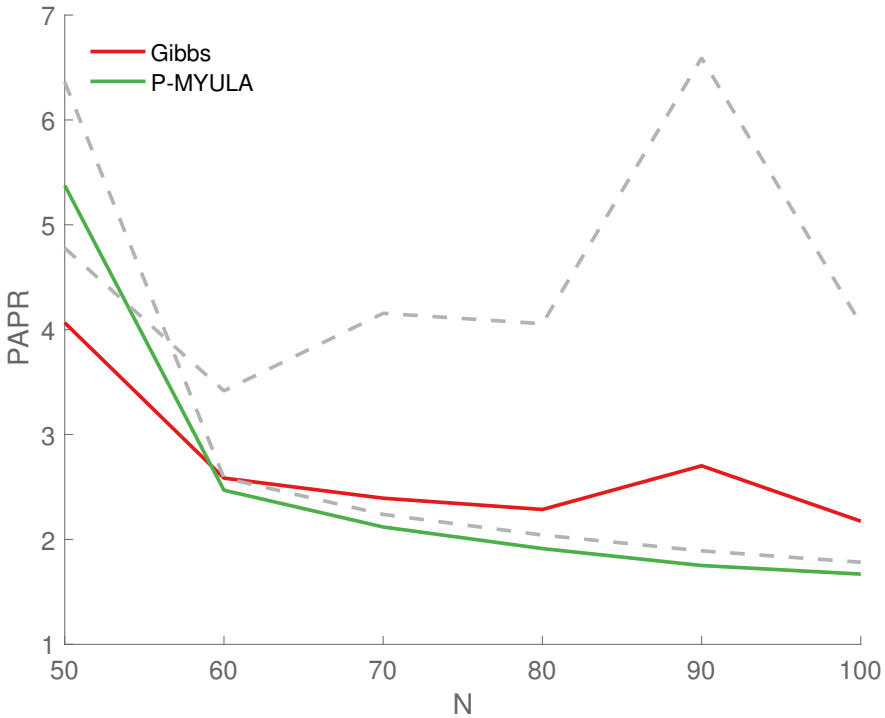
$$\mathbf{x}^* \sim \mathcal{N}\left(\mathbf{x}^{(t+1/2)}, \delta I_N\right)$$

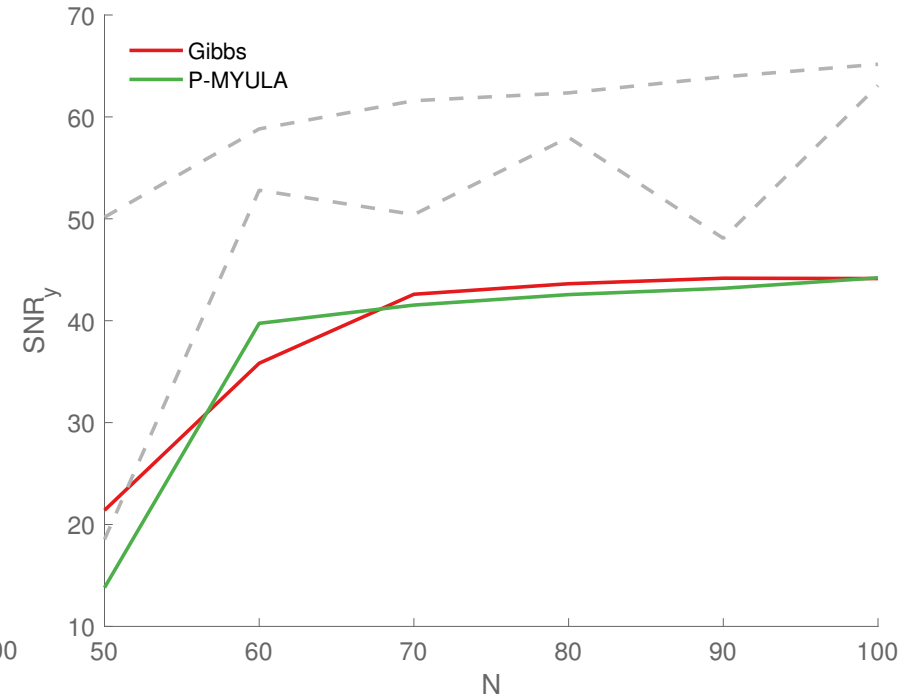
3. Metropolis Hastings

$\mathbf{x}(t+1) = \mathbf{x}^*$ avec proba

PAPR reduction

retour

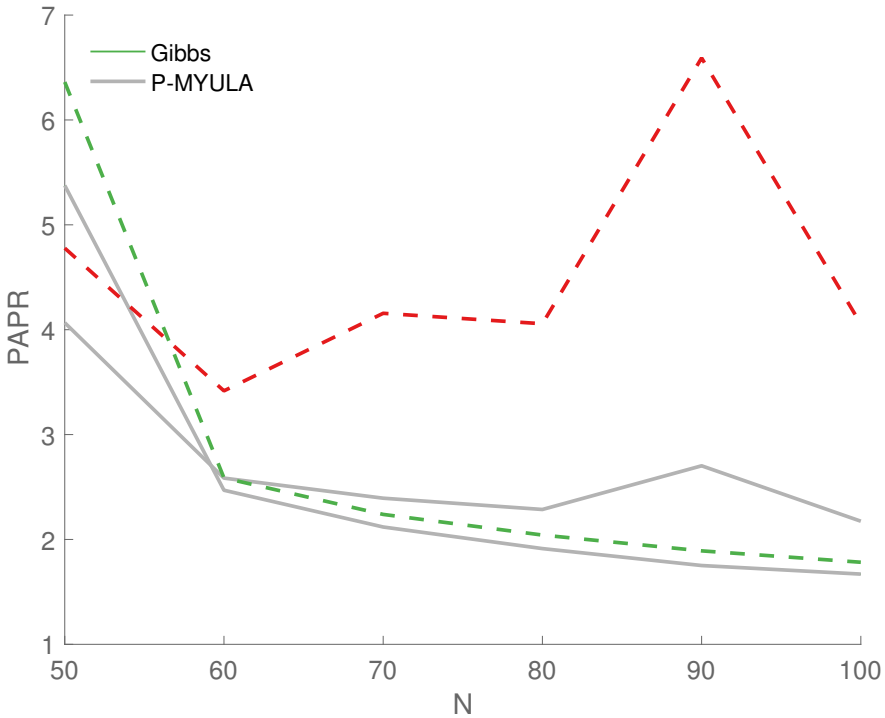


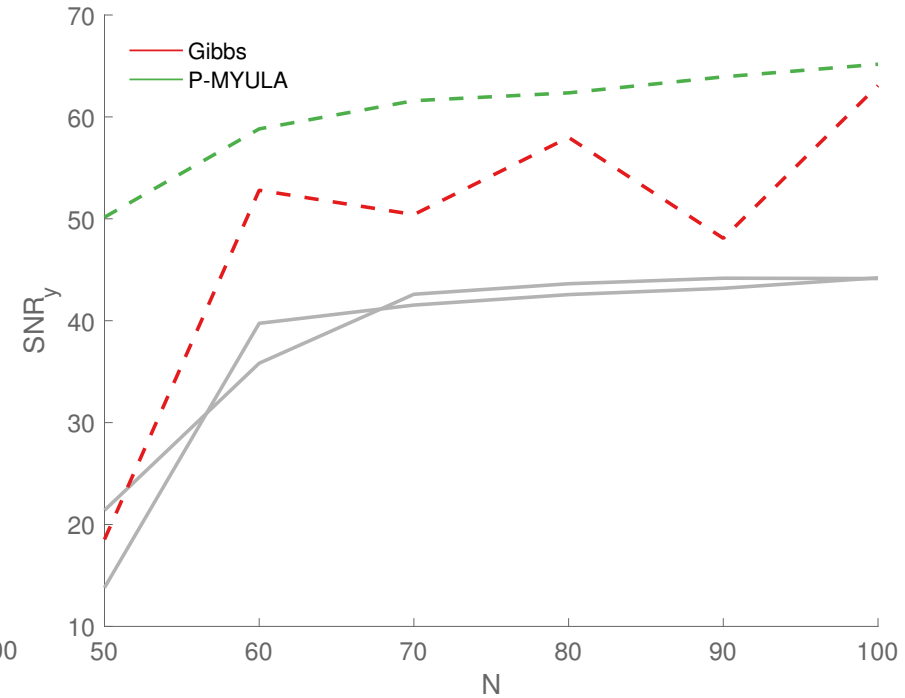


$$D = 30,\ K = 30, \ldots, 100$$

PAPR reduction

retour

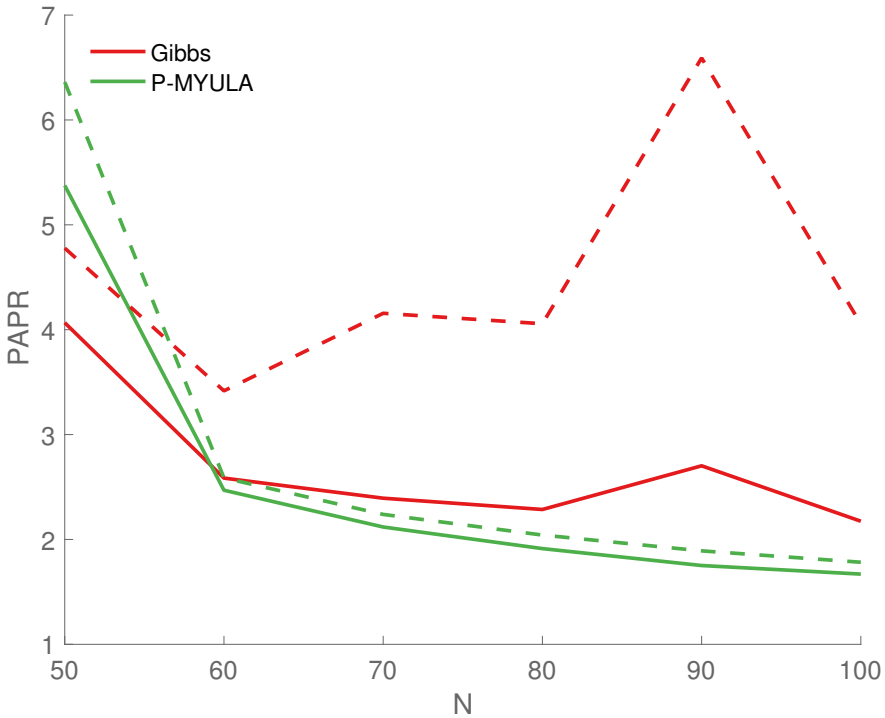


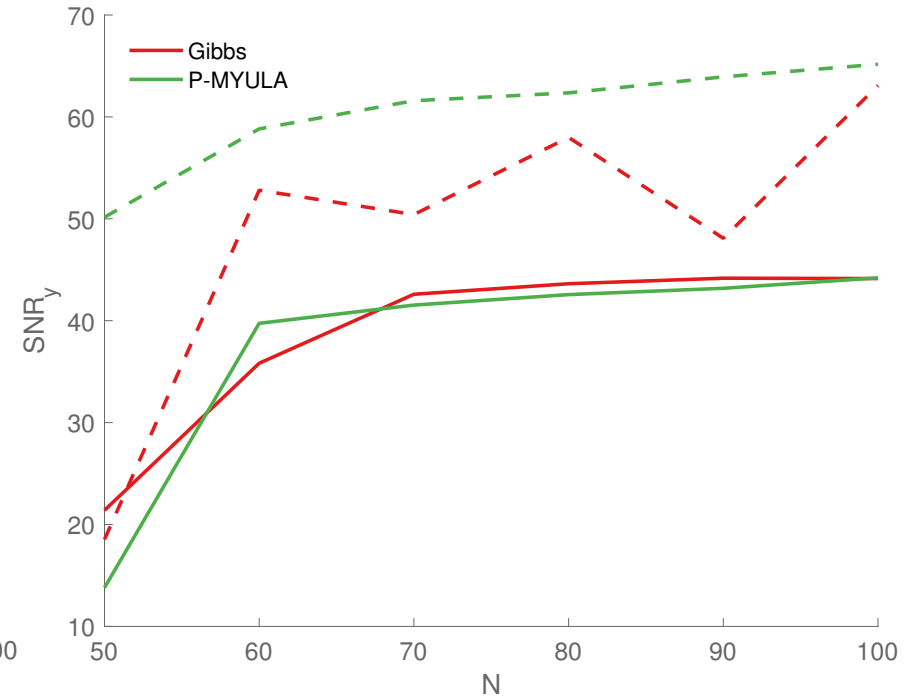


$$D = 30,\ K = 30, \ldots, 100$$

PAPR reduction

retour





$$D = 30,\ K = 30, \ldots, 100$$

Vers un processus stochastique

retour

- Processus démocratique ?

$$\forall \boldsymbol{y} \in \mathbb{R}_+^K, \quad \mathrm{p}\left(\boldsymbol{y}\right) = \frac{\gamma^{K-1}}{K \left\|\boldsymbol{y}\right\|_\infty^{K\gamma}} \prod_{\ell=1}^{K} y_\ell^{\gamma-1} \ (\boldsymbol{y}) \, \mathbb{1}_\Delta(\boldsymbol{y})$$

mais pas de propriété d'additivité

$$\boldsymbol{x}_{\setminus k} | \boldsymbol{x} \in \mathcal{C}_k \ \sim \ \mathcal{D}(\lambda)$$

Processus bêta démocratique - Bernoulli ?

Algorithme

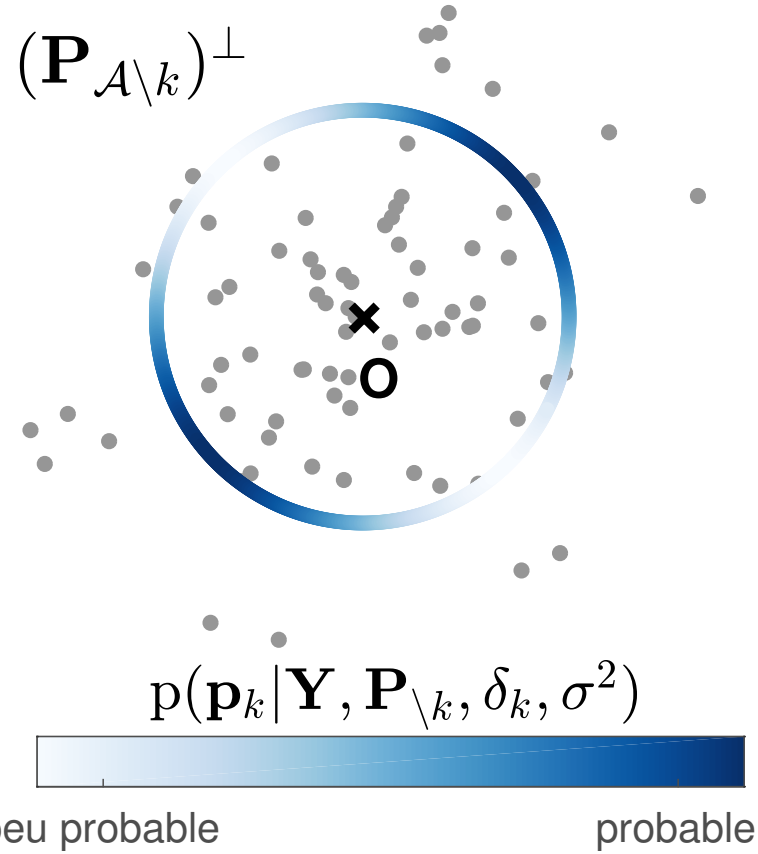
retour

pour chaque *itération* t **faire**
  pour $n \leftarrow 1$ **à** N **faire**
    // K est fixé
      Échantillonner directions $(z_{k,n})_k \sim$ Bernoulli ;
    // K varie
      ajouter / supprimer directions $\sim$ von Mises Fischer ;
  fin
  // K est fixé
  pour chaque *direction active* k **faire**
      Échantillonner le facteur d'échelle $\delta_k \sim$ lois Inverse Gamma translatée ;
      $\mathbf{p}_k | \mathbf{P}_{\setminus k} \sim$ Bingham ;
  fin
    Échantillonner $\sigma^2, \alpha \sim$ lois conjuguées ;
fin

Mise à jour des directions pertinentes

retour

$$\mathrm{p}\left(\mathbf{v}_k|\mathbf{P}_{\setminus k}\right) \propto \exp\left(\mathbf{v}_k^T \mathbf{A} \mathbf{v}_k\right) \quad \sim \text{ Bingham}$$



Échantillonnage : *Hoff (2009)*

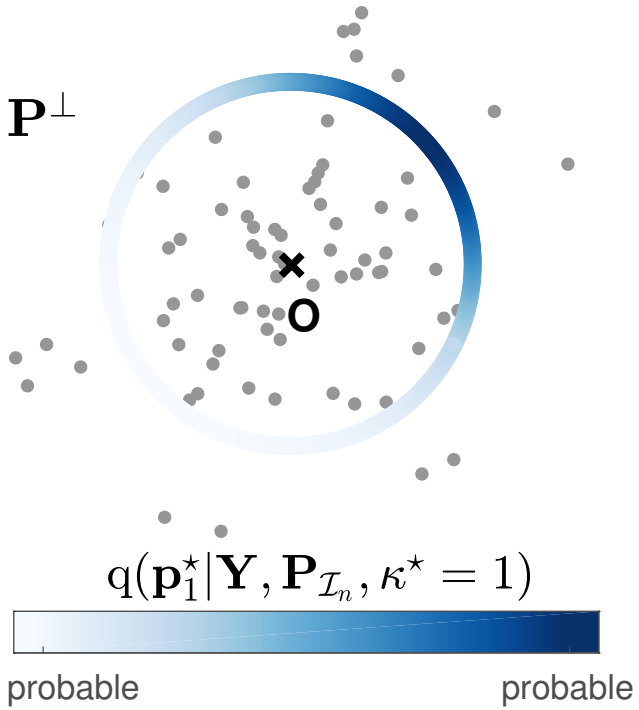
Mise à jour de K & exploration de l'espace

retour

1. Nouveau nombre de singletons $\kappa^* \sim \mathcal{P}(\alpha)$
2. Nouvelles directions $\mathbf{P}_{\text{new}} | \kappa^* \overset{d}{\sim}$ q = von Mises Fischer

$$\mathrm{q}\left(\mathbf{P}_{\text{new}}; \mathbf{A}\right) \propto \exp\left(-\sum_{k=1}^{\kappa^*} \mathbf{p}_{\text{new},k}^T \mathbf{a}_k\right)$$

3. Metropolis Hastings



Inconsistance avec un modèle génératif

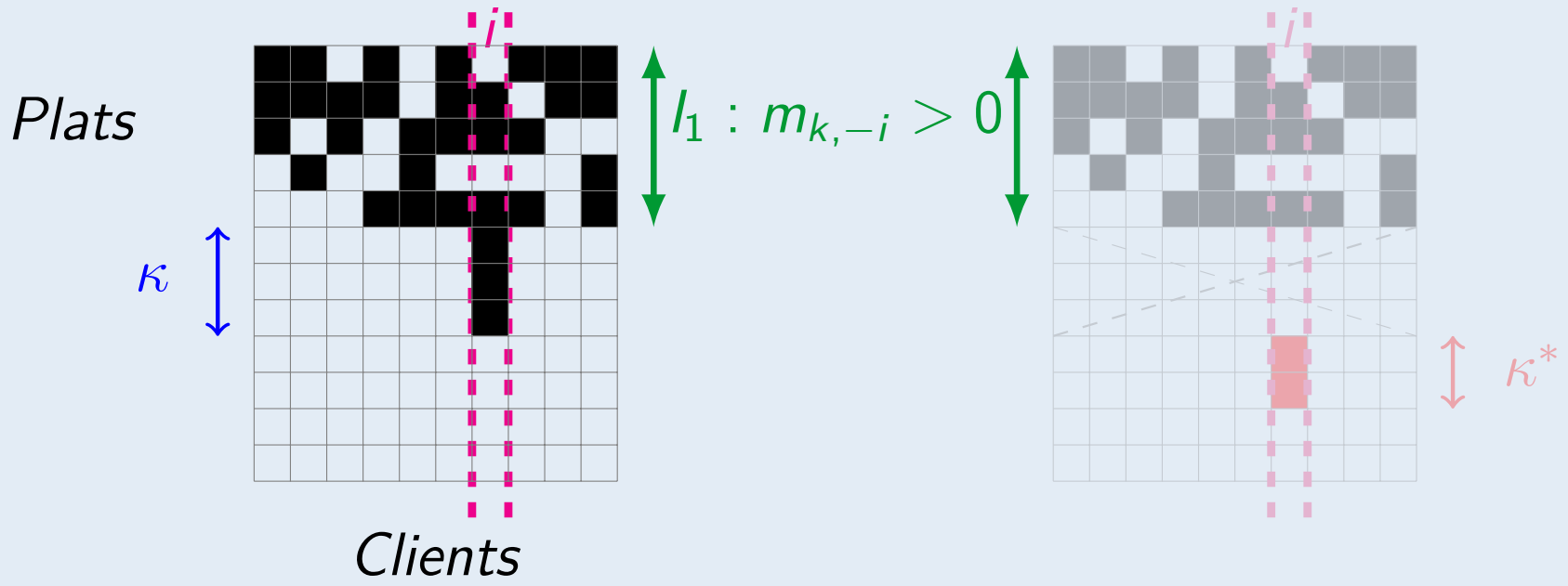
retour

$$\text{Si } \boldsymbol{y} \sim \mathcal{N}(0, \sigma^2 \mathrm{I}) \quad \mathrm{p}\left[K_N = 0 | \boldsymbol{y}_1 \dots \boldsymbol{y}_N, \alpha, \sigma^2\right] \xrightarrow[N \to +\infty]{p.s.} 0$$

Échantillonnage de Z

retour

Ajouter de nouveaux atomes

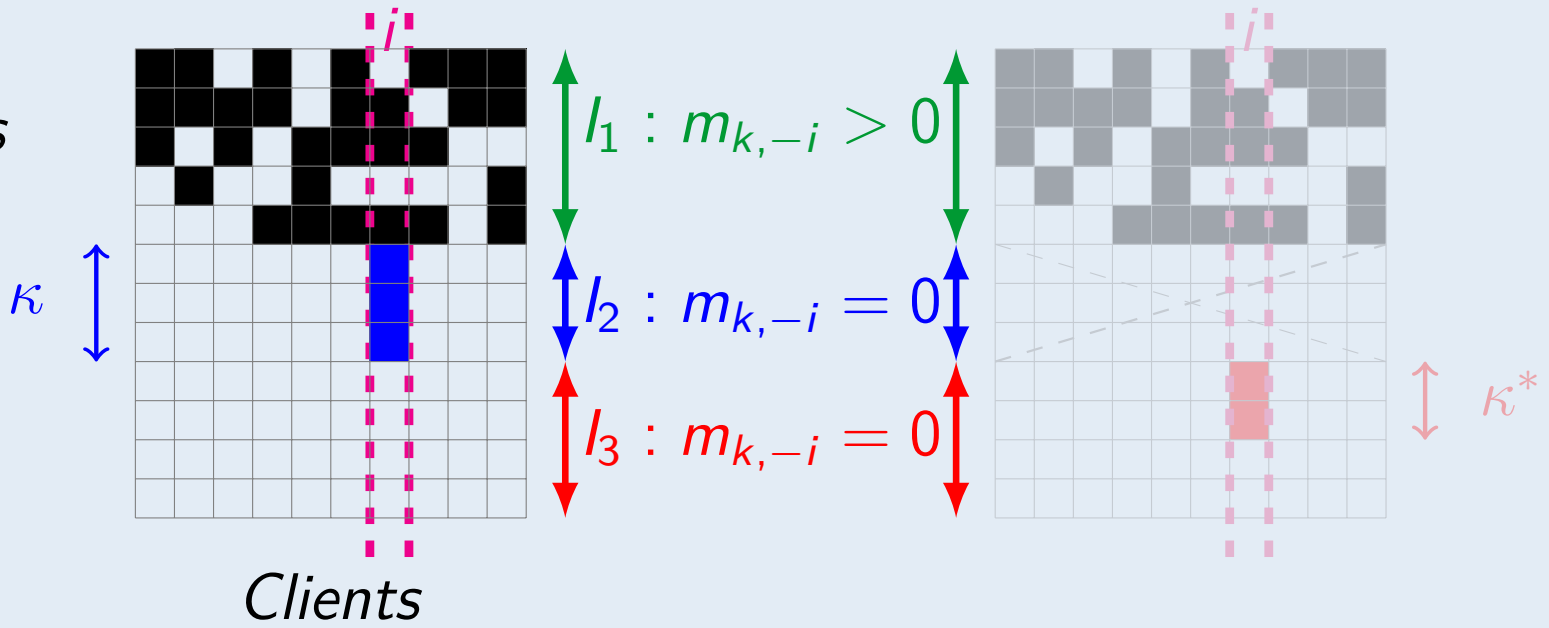


$$IBP(\alpha) \begin{cases} P(Z_{k,i} = 1) & = m_{k,-i}/N \\ k_{inew} & \sim \mathcal{P}(\alpha/N) \end{cases}$$

Échantillonnage de Z

retour

Ajouter de nouveaux atomes



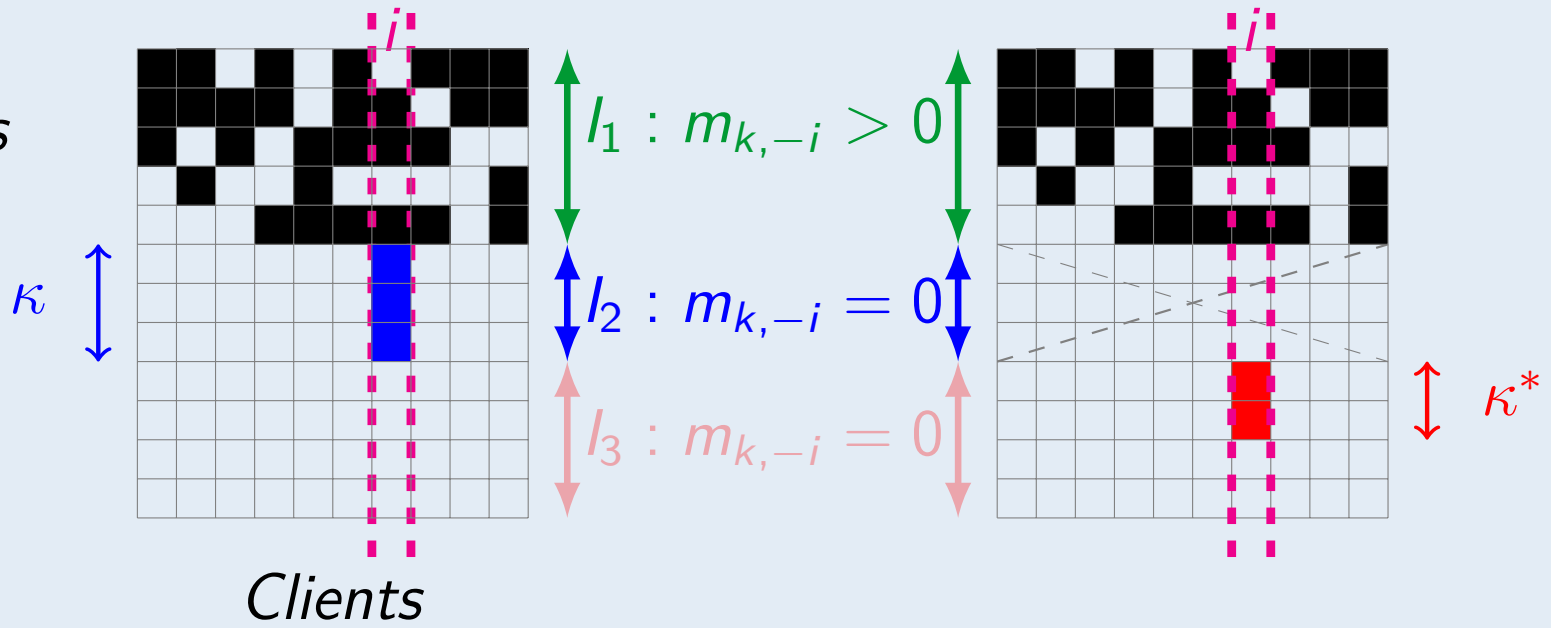
Metropolis-Hastings pour les singletons

Knowles & Ghahramani (2011)]

Échantillonnage de Z

retour

Ajouter de nouveaux atomes



Metropolis-Hastings pour les singletons

Knowles & Ghahramani (2011)]

Complexité

retour

Par itération, sans optimisation

- σ^2 KND
- δ_k^2 $K \times ND$
- α $\mathcal{O}(1)$
- $\boldsymbol{P}$ $K \times D^3 + NK^2D^2$ (1 SVD + mise à jour)
- $\boldsymbol{Z}$ $\underbrace{NKD}_{\text{actif}} + \underbrace{NK^3 + N(D-K)D^2}_{\text{singletons}}$ (1 SVD + mise à jour)

$\Longrightarrow$ Le + cher : chercher l'orthogonal des composantes pertinentes

Algorithme

retour

pour chaque *itération* t **faire**
  pour $n \leftarrow 1$ **à** N **faire**
    // K est fixé
      Échantillonner directions $(z_{k,n})_k \sim$ Bernoulli ;
    // K varie
      ajouter / supprimer directions $\sim$ von Mises Fischer ;
  fin
  // K est fixé
  pour chaque *direction active* k **faire**
      Échantillonner le facteur d'échelle $\delta_k \sim$ lois Inverse Gamma translatée ;
      $\mathbf{p}_k | \mathbf{P}_{\setminus k} \sim$ Bingham ;
  fin
    Échantillonner $\sigma^2, \alpha \sim$ lois conjuguées ;
fin

Résultat d'inconsistance : commentaires

retour

$$\text{Preuve :} \begin{cases} \lim\limits_{N\to+\infty} \frac{1}{N} \max\limits_{\mathbf{Z}} \max\limits_{\mathbf{Z}'\simeq\mathbf{Z}} \frac{\mathrm{P}[\mathbf{Z}|\alpha]}{\mathrm{P}[\mathbf{Z}'|\alpha]} < +\infty & (1) \\ \forall(\mathbf{Z},\mathbf{Z}')\ \mathbf{Z}'\simeq\mathbf{Z}, \quad \mathrm{p}(\mathbf{Y}|\mathbf{Z}) \leq \kappa \mathrm{p}(\mathbf{Y}|\mathbf{Z}') & (2) \end{cases}$$

où $\mathbf{Z} \simeq \mathbf{Z}' \implies \mathbf{Z}' = \mathbf{Z}$ + colonne avec un seul 1

(1) $\longrightarrow$ propriété de l'IBP à 2 paramètres

(2) $\longrightarrow$ propriété du modèle

$\implies$ résultat plus général

Constance de l'estimateur $\widehat{K}_{KS}$

retour

A posteriori

$$p(\widetilde{\mathbf{P}}|\mathbf{H}, \dots) \propto \exp\left(\widetilde{\mathbf{P}}^T \mathbf{N}^T (\text{parcimonie} \times \mathbf{Y}\mathbf{Y}^T)\mathbf{N}\widetilde{\mathbf{P}}\right)$$

et

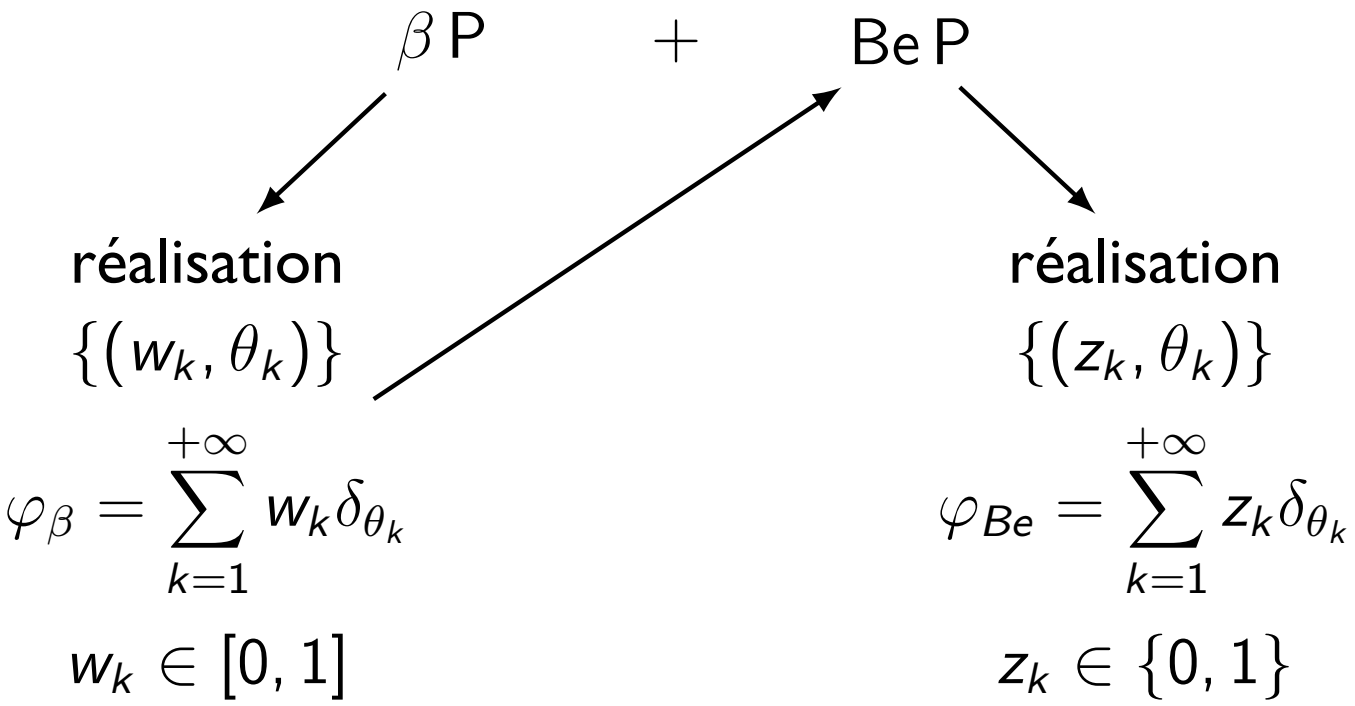
$$\frac{1}{N}\mathbf{N}^T\mathbf{Y}\mathbf{Y}^T\mathbf{N} \underset{N\to+\infty}{\longrightarrow} \sigma^2 \mathbb{I}_{D-K}$$

$$\Longrightarrow \text{loi uniforme}$$

Construction de l'IBP

retour

P.P.P à valeurs dans masse $\times$ paramètres



Construction de l'IBP

retour

Mesure intensité du processus Bêta

$$\mu_\beta(\mathrm{d}w, \mathrm{d}\theta) = \alpha w^{-1}(1-\omega^{-1})\mathrm{d}w\ G_0(\mathrm{d}\theta)$$

$\Longrightarrow$ calcul d'intégrales

$$\#\text{points dans } W \times O = \int_{W\times O} \mu_\beta(\mathrm{d}w, \mathrm{d}\theta)$$

En particulier

$$\#\text{ total points} = \mu_\beta([0,1], \Theta) = +\infty$$

$\Longrightarrow$ L'IBP favorise une infinité de paramètre cf. $\alpha \log(K)$

IBP à trois paramètres

retour

Mesure intensité

$$\widetilde{\mu}_\beta(\mathrm{d}w, \mathrm{d}\theta) = \alpha \frac{\Gamma(1+c)}{\Gamma(1-\sigma)\Gamma(c+\sigma)} w^{-\sigma-1}(1-\omega^{c+\sigma-1})\mathrm{d}w \; G_0(\mathrm{d}\theta)$$

- $\sigma > 0$ loi de puissance *Teh and Gorur, 2009*
- $\sigma < 0$ non encore étudié et

$$\widetilde{\mu}_\beta\left([0,1], \Theta\right) < +\infty$$

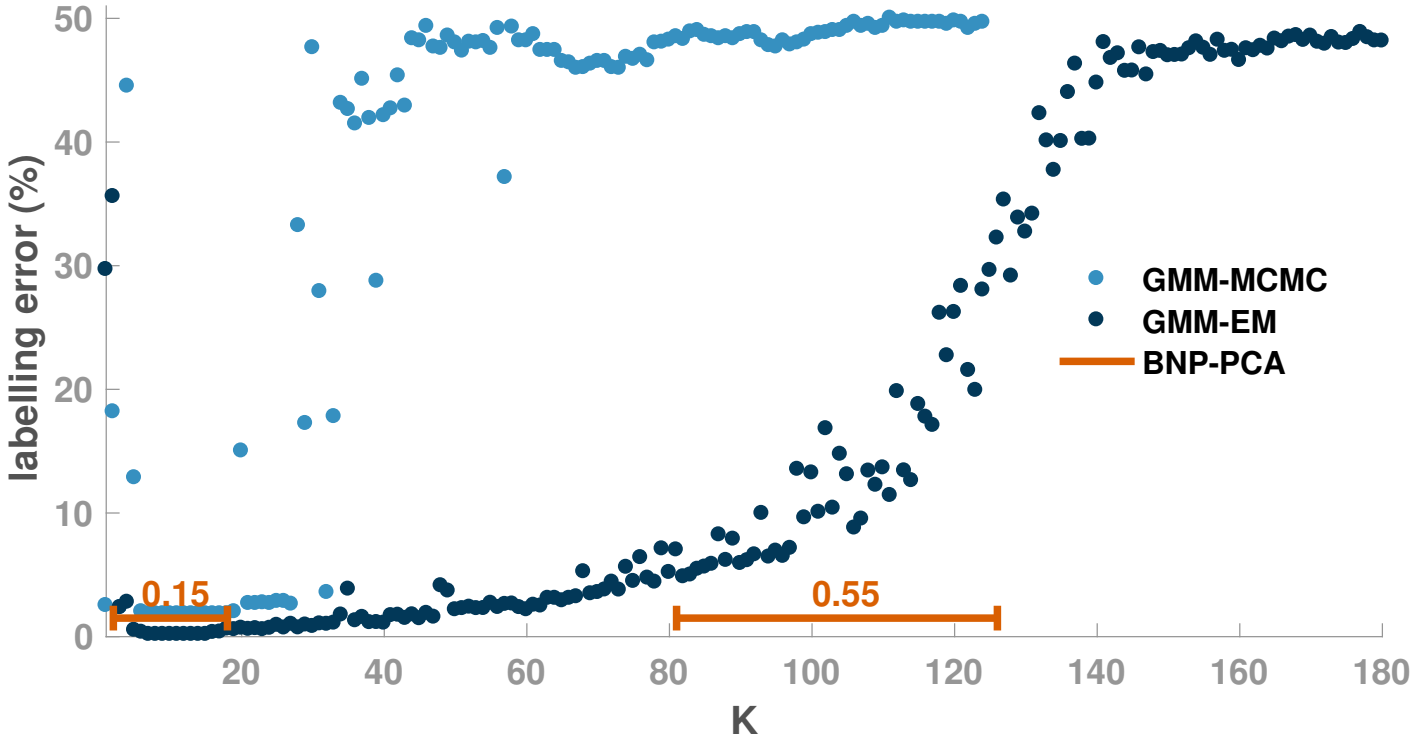
Good news ?

Application 2 : Couplage avec un séparateur linéaire

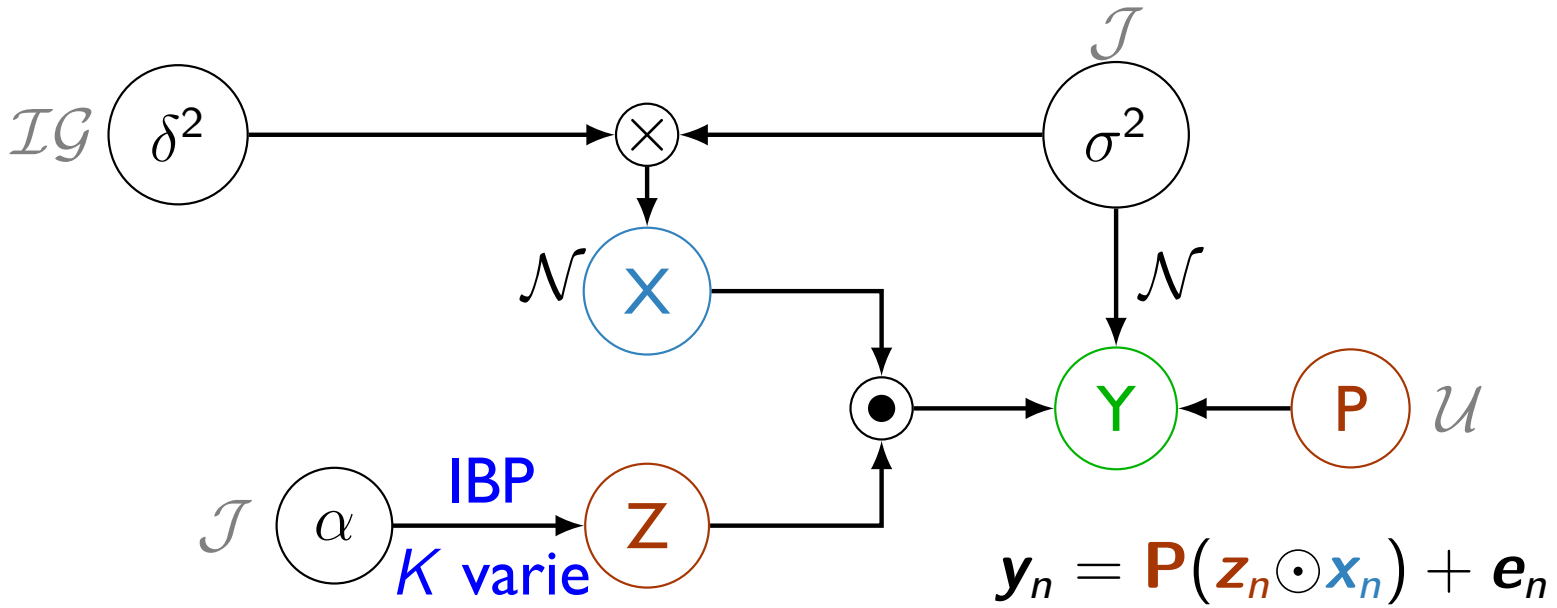
retour

$$x_n \sim \pi_1 \mathcal{N}(\mu_1, \delta_1^2 \sigma^2) + (1 - \pi_1)\mathcal{N}(\mu_2, \delta_2^2 \sigma^2)$$

2 chiffres (6 et 7) de MNIST



Modèle BNP-PCA



$\boldsymbol{e}_n \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_D)$, bruit additif gaussien

$\mathbf{P} = [\mathbf{p}_1 \dots \mathbf{p}_K, \mathbf{p}_{K+1} \dots \mathbf{p}_D]$, $\mathbf{P}^t\mathbf{P} = \mathbf{I}_D$, et $\mathbf{P} \sim \mathcal{U}_{\mathcal{O}_D}$

$\boldsymbol{Z} = [\boldsymbol{z}_1 \dots \boldsymbol{z}_N] \sim \text{IBP}(\alpha)$ matrice binaire $\rightarrow K$

$\boldsymbol{x}_n = [x_{n,1} \dots x_{n,K}]$ $\forall k,\ x_{n,k} \sim \mathcal{N}(0, \delta_k^2 \sigma^2)$

$\boldsymbol{\theta} = \{\boldsymbol{\delta^2}, \sigma^2, \alpha\}$ loi conjuguée vague

Application 2 : Couplage avec un séparateur linéaire

retour

$$x_n \sim \pi_1 \mathcal{N}(\mu_1, \delta_1^2 \sigma^2) + (1 - \pi_1)\mathcal{N}(\mu_2, \delta_2^2 \sigma^2)$$

2 chiffres (6 et 7) de MNIST