Sparse dictionary learning

Colloque Jeunes probabilistes et Statisticiens 2021

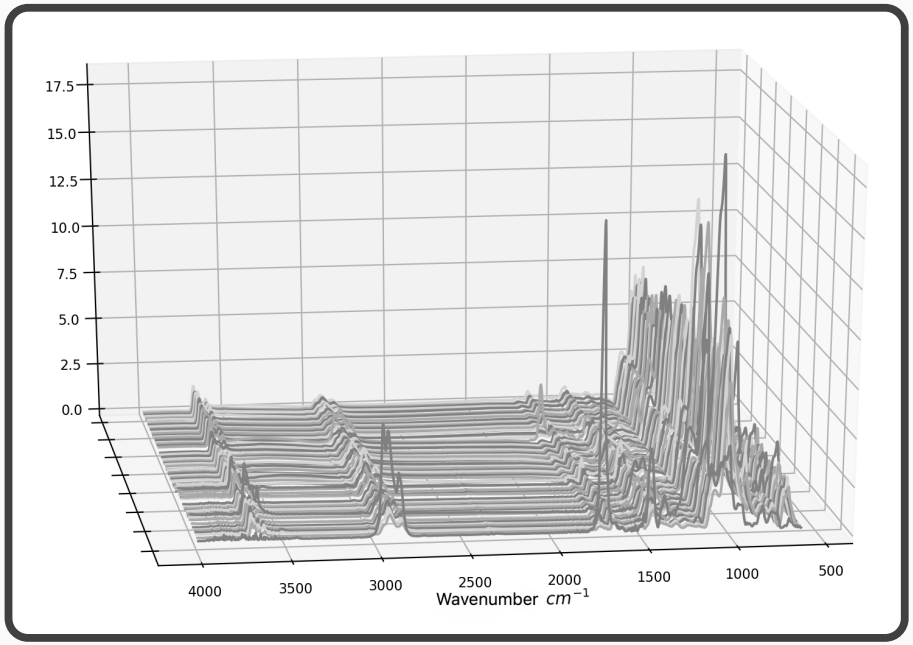
C. Butucea (ENSAE), J.-F. Delmas (Ecole des Ponts), A. Dutfoy (EDF R&D),
C. Hardy (EDF R&D, Ecole des Ponts)

Déconvolution de pics

La spectroscopie infrarouge

Wave numbers (cm-1)	Peak assignment
3690-3400-3364-3200-3014	-OH
2952-2920-2850	$\nu - CH_2, CH_3$ Aliphatic
1731	$\nu - C = O$
1647	$\nu - C = C$ de $HC = CH_2$
1540	$\nu - C = C$ de R-CR=CH-R, δ CH2 Aliphatic
1419	δCH_2, δ-CH Aliphatic
1160-1082	ν Si-O (SiO_2)
1009-909	ν Si-O (Si-OH)
825	C-Cl
664	CH Aromatic

Table des positions des pics et groupes chimiques associés pour des échantillons de néoprène ([Tchalla, 2017]).



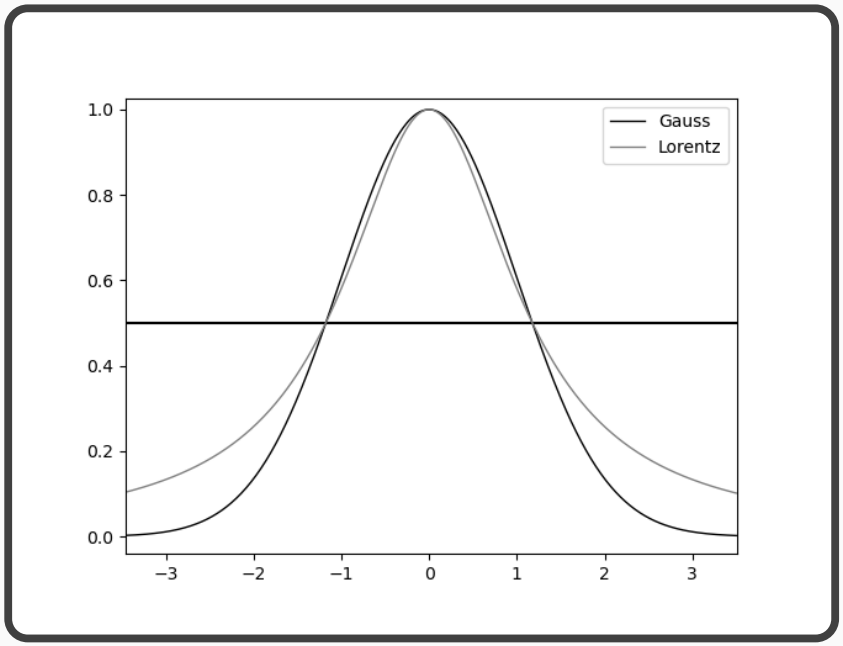
$$\mathbf{y(t)} = \sum_{k=1}^{s^\star} \beta_k^\star \, \phi(\theta_k^\star, t) + w(t), \quad (\phi(\theta, \cdot), \theta \in \Theta) \text{ dictionnaire continu.}$$

Déconvolution de pics

Choix du dictionnaire continu : $\phi(\theta, t) = \frac{\varphi(\theta,t)}{\|\varphi(\theta,\cdot)\|}$,

$$\varphi_{Gauss} : \Theta \times \mathbb{R} \to \mathbb{R}$$
$$((\mu, \nu), t) \mapsto e^{-\frac{(t-\mu)^2}{2\nu^2}},$$

$$\varphi_{Lorentz} : \Theta \times \mathbb{R} \to \mathbb{R}$$
$$((\mu, \nu), t) \mapsto \frac{1}{1 + \frac{(t-\mu)^2}{2\nu^2}}.$$



Déconvolution de pics

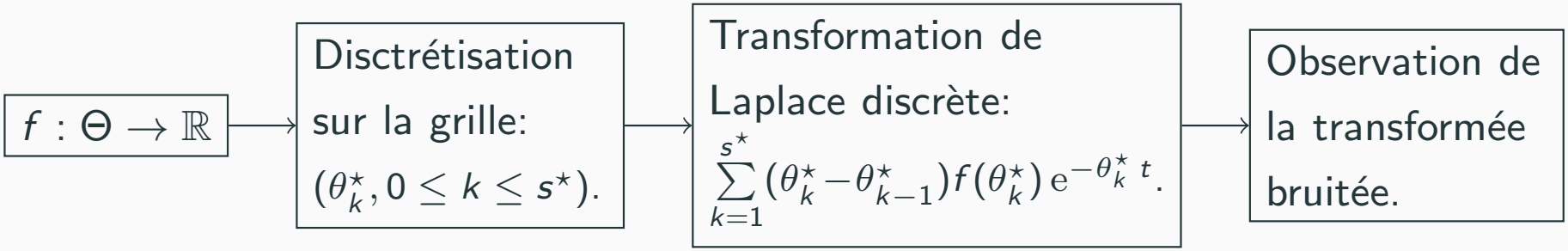
$$\mathbf{y(t)} = \sum_{k=1}^{s^\star} \beta_k^\star \, \phi(\theta_k^\star, t) + w(t).$$

$\rightarrow$ Retrouver le nombre $s^\star$ de fonctions paramétriques dans le mélange.

$\rightarrow$ Retrouver la position des pics $\theta_k^\star$ pour identifier les groupes chimiques.

$\rightarrow$ Retrouver les amplitudes des pics $\beta_k^\star$ pour déterminer les concentrations des espèces chimiques.

Inversion d'une transformée de Laplace



$$\mathbf{y(t)} = \sum_{k=1}^{s^\star} \beta_k^\star\, \phi(\theta_k^\star, t) + w(t).$$

$\phi(\theta, t) = \frac{\varphi(\theta,t)}{\|\varphi(\theta,\cdot)\|}$.

$\varphi : \Theta \times \mathbb{R} \to \mathbb{R}$

$(\theta, t) \mapsto e^{-t\,\theta}$.

$\to$ Retrouver $s^\star$, $(\beta_k^\star, 1 \leq k \leq s^\star)$ et $(\theta_k^\star, 1 \leq k \leq s^\star)$ pour reconstruire f.

Sommaire

I) Le modèle

On observe un signal y sur le support d'une mesure λ_T.

- λ_T discrète (ex: $\lambda_T = \Delta_T \sum_{j=1}^{T} \delta_{t_j}$) $\rightarrow$ observations sur une grille.
- λ_T continu (ex: $\lambda_T = Lebesgue$) $\rightarrow$ observations continues.

I) Le modèle

On observe un signal y sur le support d'une mesure λ_T.

- λ_T discrète (ex: $\lambda_T = \Delta_T \sum_{j=1}^{T} \delta_{t_j}$) $\rightarrow$ observations sur une grille.
- λ_T continu (ex: $\lambda_T = Lebesgue$) $\rightarrow$ observations continues.

On observe un signal y bruité par un processus stochastique w_T.

- Pour tout $f \in L^2(\lambda_T)$, $Var \langle f, w_T \rangle_T \leq \sigma^2 \Delta_T \|f\|_T^2$.

On définit,

$$\langle f, g \rangle_T = \int_{\mathbb{R}} f(t) g(t) \lambda_T(dt) \quad \text{et} \quad \|f\|_T = \langle f, f \rangle_T^{1/2} \quad \text{pour} \quad f, g \in L^2(\lambda_T).$$

I) Le modèle

On observe un signal y sur le support d'une mesure λ_T.

- λ_T discrète (ex: $\lambda_T = \Delta_T \sum_{j=1}^{T} \delta_{t_j}$) $\rightarrow$ observations sur une grille.
- λ_T continu (ex: $\lambda_T = Lebesgue$) $\rightarrow$ observations continues.

On observe un signal y bruité par un processus stochastique w_T.

- Pour tout $f \in L^2(\lambda_T)$, $Var \langle f, w_T \rangle_T \leq \sigma^2 \Delta_T \|f\|_T^2$.

On définit,

$$\langle f, g \rangle_T = \int_{\mathbb{R}} f(t) g(t) \lambda_T(dt) \quad \text{et} \quad \|f\|_T = \langle f, f \rangle_T^{1/2} \quad \text{pour} \quad f, g \in L^2(\lambda_T).$$

Le paramètre T correspond à la qualité des observations: $\Delta_T \underset{T \to +\infty}{\rightarrow} 0$.

I) Le modèle

On observe le signal

$$y = \beta^\star \Phi_T(\vartheta^\star) + w_T, \quad \lambda_T - p.p. \qquad \text{(modèle)}$$

Pour tout $\vartheta = (\theta_1, \cdots, \theta_K) \in \Theta^K$,

$$\Phi_T(\vartheta) = \begin{pmatrix} \phi_T(\theta_1) \\ \vdots \\ \phi_T(\theta_K) \end{pmatrix}$$

est la fonction multivariée $\Phi_T(\vartheta)$ définie sur $\mathbb{R}$ à valeurs dans $\mathbb{R}^K$, $t \mapsto \Phi_T(\vartheta, t)$.

Le paramètre K est une borne arbitrairement grande pour $s^\star$.

I) Le modèle

- **Exemple discret:** Grille régulière sur $[0,1]$, $\lambda_T = \Delta_T \sum_{j=1}^{T} \delta_{t_j}$ avec $t_j = j/T$ et $\Delta_T = 1/T$, $w_T(t_j) \underset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$.

$$y\left(\frac{j}{T}\right) = \beta^\star \Phi_T\left(\vartheta^\star, \frac{j}{T}\right) + w_j, \quad w_j \underset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2), \quad j = 1, \cdots, T.$$

I) Le modèle

- **Exemple discret:** Grille régulière sur $[0,1]$, $\lambda_T = \Delta_T \sum_{j=1}^{T} \delta_{t_j}$ avec $t_j = j/T$ et $\Delta_T = 1/T$, $w_T(t_j) \underset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$.

$$y\left(\frac{j}{T}\right) = \beta^\star \Phi_T\left(\vartheta^\star, \frac{j}{T}\right) + w_j, \quad w_j \underset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2), \quad j = 1, \cdots, T.$$

- **Exemple continu:** $\lambda_T = \textit{Lebesgue}$ sur $[0,1]$ et w_T est un Brownien: $w_T = \sigma\sqrt{\Delta_T}\, B$, $\Delta_T = 1/T$,

$$y = \beta^\star \Phi(\vartheta^\star) + \frac{\sigma}{\sqrt{T}} B, \quad \textit{Lebesgue}\text{-p.p.}$$

I) Le modèle

- **Exemple discret:** Grille régulière sur $[0,1]$, $\lambda_T = \Delta_T \sum_{j=1}^{T} \delta_{t_j}$ avec $t_j = j/T$ et $\Delta_T = 1/T$, $w_T(t_j) \underset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$.

$$y\left(\frac{j}{T}\right) = \beta^\star \Phi_T\left(\vartheta^\star, \frac{j}{T}\right) + w_j, \quad w_j \underset{i.i.d}{\sim} \mathcal{N}(0,\sigma^2), \quad j = 1, \cdots, T.$$

- **Exemple continu:** $\lambda_T = Lebesgue$ sur $[0,1]$ et w_T est un Brownien: $w_T = \sigma\sqrt{\Delta_T}\, B$, $\Delta_T = 1/T$,

$$y = \beta^\star \Phi(\vartheta^\star) + \frac{\sigma}{\sqrt{T}} B, \quad Lebesgue\text{-p.p.}$$

Dans les deux cas : $\forall f \in L^2(\lambda_T),\ Var\,\langle f, w_T\rangle_T \leq \sigma^2\, \Delta_T\, \|f\|_T^2$.

I) Le modèle

$$y = \beta^\star \Phi_T(\vartheta^\star) + w_T, \quad \lambda_T - p.p. \qquad \text{(modèle)}$$

On considère une extension non linéaire du modèle de régression parcimonieuse en grande dimension : $S^\star = \{k, \quad \beta_k^\star \neq 0\}$, $Card\, S^\star = s^\star < K$.

I) Le modèle

$$y = \beta^\star \Phi_T(\vartheta^\star) + w_T, \quad \lambda_T - p.p. \qquad \text{(modèle)}$$

On considère une extension non linéaire du modèle de régression parcimonieuse en grande dimension : $S^\star = \{k, \quad \beta_k^\star \neq 0\}$, $Card\, S^\star = s^\star < K$.

$\rightarrow$ Estimateurs pour $\beta^\star$ et $\vartheta^\star_{S^\star}$?

$\rightarrow$ Risque de prévision ?

(à une permutation jointe sur les composantes de $\beta^\star$ et $\vartheta^\star$ près)

II) Problème d'optimisation

Formulation d'un problème d'optimisation avec une pénalisation Lasso pondérée par $\kappa > 0$:

$$(\hat{\beta}, \hat{\vartheta}) \in \underset{\beta \in \mathbb{R}^K, \vartheta \in \Theta_T^K}{\operatorname{argmin}} \quad \frac{1}{2}||y - \beta\Phi(\vartheta)||_T^2 + \kappa||\beta||_{\ell_1}$$

$\Theta_T \subset \Theta$, intervalle compact.

On suppose que pour tout $k \in S^\star$, $\theta_k^\star \in \Theta_T$.

Problème non-convexe mais des méthodes numériques performantes existent (sliding Franck-Wolfe algorithm [Denoyelle et al, 2019]).

II) Problème d'optimisation

On peut réécrire le modèle

$$y = \beta^\star \Phi_T(\vartheta^\star) + w_T, \quad \lambda_T - p.p$$

sous la forme

$$y = \int \phi_T(\theta)\mu^\star(d\theta) + w_T, \quad \lambda_T - p.p$$

avec $\mu^\star = \sum_{k=1}^{s^\star} \beta_k^\star \delta_{\theta_k^\star}$.

II) Problème d'optimisation

Formulation d'un problème sur un espace de mesures (Beurling Lasso [De Castro & Gamboa 2012]),

$$\tilde{\mu} \in \underset{\mu \in \mathcal{M}(\Theta)}{\operatorname{argmin}} \quad \frac{1}{2}||y - \int \phi_T(\theta)\mu(d\theta)||_T^2 + \kappa||\mu||_{TV}.$$

Problème convexe !

Remarque: pour $\mu = \sum_{k=1}^{s} \beta_k \delta_{\theta_k}$ on a $\|\mu\|_{TV} = \|\beta\|_{\ell_1}$.

II) Problème d'optimisation

$$\tilde{\mu} \in \underset{\mu \in \mathcal{M}(\Theta)}{\operatorname{argmin}} \quad \frac{1}{2}||y - \int \phi_T(\theta)\mu(d\theta)||_T^2 + \kappa||\mu||_{TV}.$$

Obtient-on une mesure $\tilde{\mu}$ discrète ?

II) Problème d'optimisation

$$\tilde{\mu} \in \underset{\mu \in \mathcal{M}(\Theta)}{\operatorname{argmin}} \quad \frac{1}{2}||y - \int \phi_T(\theta)\mu(d\theta)||_T^2 + \kappa||\mu||_{TV}.$$

Obtient-on une mesure $\tilde{\mu}$ discrète ?

$\rightarrow$ Lorsque λ_T est discrète, si l'ensemble des solutions est non vide, il existe une solution discrète [Boyer et al, 2019].

$\rightarrow$ Sous des hypothèses sur φ et $\mu^\star$ et lorsque κ et $\frac{\|w_T\|_T}{\kappa}$ sont suffisamment petits les solutions sont discrètes et composées de $s^\star$ Diracs [Duval & Peyré 2015].

II) Problème d'optimisation

$$\tilde{\mu} \in \underset{\mu \in \mathcal{M}(\Theta)}{\operatorname{argmin}} \quad \frac{1}{2}||y - \int \phi_T(\theta)\mu(d\theta)||_T^2 + \kappa||\mu||_{TV}.$$

Borne sur le risque de prévision $\left\|\tilde{\beta}\Phi_T(\tilde{\vartheta}) - \beta^\star\Phi_T(\vartheta^\star)\right\|_T$?

$$\tilde{\mu} = \sum_{k=1}^{\tilde{s}} \tilde{\beta}_k \delta_{\tilde{\theta}_k}.$$

II) Problème d'optimisation

$$\tilde{\mu} \in \underset{\mu \in \mathcal{M}(\Theta)}{\operatorname{argmin}} \quad \frac{1}{2}||y - \int \phi_T(\theta)\mu(d\theta)||_T^2 + \kappa||\mu||_{TV}.$$

Borne sur le risque de prévision $\left\|\tilde{\beta}\Phi_T(\tilde{\vartheta}) - \beta^\star\Phi_T(\vartheta^\star)\right\|_T$?

$$\tilde{\mu} = \sum_{k=1}^{\tilde{s}} \tilde{\beta}_k \delta_{\tilde{\theta}_k}.$$

$\rightarrow$ Pour $(\varphi(\theta)\colon t \mapsto \mathrm{e}^{i2\pi t\theta}, \theta \in \mathbb{T})$ et $\Delta_T = 1/T$, on peut choisir κ de telle sorte que

$$\left\|\tilde{\beta}\Phi_T(\tilde{\vartheta}) - \beta^\star\Phi_T(\vartheta^\star)\right\|_T = \mathcal{O}_{\mathbb{P}}\left(\sigma\sqrt{\frac{s^\star \log T}{T}}\right) \text{ [Tang et al 2014].}$$

III) Borne pour le risque de prédiction

On définit sur $\Theta \times \Theta$, $\mathcal{K}_T(\theta, \theta') = \langle \phi_T(\theta), \phi_T(\theta') \rangle_T$

Theorem ($\Theta \subset \mathbb{R}$)

Hypothèses :

- w_T *Gaussien et pour tout* $f \in L^2(\lambda_T)$, *Var* $\langle f, w_T \rangle_T \leq \sigma^2 \, \Delta_T \, \|f\|_T^2$.
- *La fonction* φ *est suffisamment régulière.*
- $\forall \, 1 \leq k \neq \ell \leq s^\star, \quad \mathfrak{d}_{FR}(\theta_k^\star, \theta_\ell^\star) > \delta$
- $\mathcal{K}_T$ *est suffisamment proche de* $\mathcal{K}_\infty$.

Alors pour $\kappa \geq \mathcal{C}_0 \sigma \sqrt{\tau \Delta_T \log T}$ *et* $\tau > 0$ *on a*

$$\left\| \hat{\beta} \Phi_T(\hat{\vartheta}) - \beta^\star \Phi_T(\vartheta^\star) \right\|_T \leq \mathcal{C}_1 \sqrt{s^\star} \, \kappa,$$

avec probabilité au moins $1 - \mathcal{C}_2 \left(\frac{|\Theta_T|}{T^\tau \log T} \vee \frac{1}{T^\tau} \right)$.

III) Borne pour le risque de prédiction

Le théorème montre que pour $\Delta_T = 1/T$ et en prenant $\kappa = \mathcal{C}_0 \sigma \sqrt{\tau \Delta_T \log T}$,

$$\left\| \hat{\beta} \Phi_T(\hat{\vartheta}) - \beta^\star \Phi_T(\vartheta^\star) \right\|_T \lesssim \sigma \sqrt{\frac{s^\star \log T}{T}},$$

en grande probabilité.

Lorsque $\vartheta^\star$ est connu $\hat{\beta}$ est l'estimateur Lasso. Sous des hypothèses de cohérence sur le dictionnaire, on peut choisir κ tel que :

$$\left\| (\hat{\beta} - \beta^\star) \Phi_T(\vartheta^\star) \right\|_T \lesssim \sigma \sqrt{\frac{s^\star \log K}{T}},$$

en grande probabilité ([Bickel et al, 2009]). Cette vitesse est minimax à un facteur logarithmique près ([Candès & Davenport, 2013]) .

III) Borne pour le risque de prédiction

Le théorème montre que pour $\Delta_T = 1/T$ et en prenant $\kappa = \mathcal{C}_0 \sigma \sqrt{\tau \Delta_T \log T}$,

$$\left\| \hat{\beta} \Phi_T(\hat{\vartheta}) - \beta^\star \Phi_T(\vartheta^\star) \right\|_T \lesssim \sigma \sqrt{\frac{s^\star \log T}{T}},$$

en grande probabilité.

Lorsque $\vartheta^\star$ est connu $\hat{\beta}$ est l'estimateur Lasso. Sous des hypothèses de cohérence sur le dictionnaire, on peut choisir κ tel que :

$$\left\| (\hat{\beta} - \beta^\star) \Phi_T(\vartheta^\star) \right\|_T \lesssim \sigma \sqrt{\frac{s^\star \log K}{T}},$$

en grande probabilité ([Bickel et al, 2009]). Cette vitesse est minimax à un facteur logarithmique près ([Candès & Davenport, 2013]) .

L'estimation des paramètres non linéaires dégrade les vitesses d'un facteur logarithmique.