Off-the-grid learning of mixtures

ETICS 2022

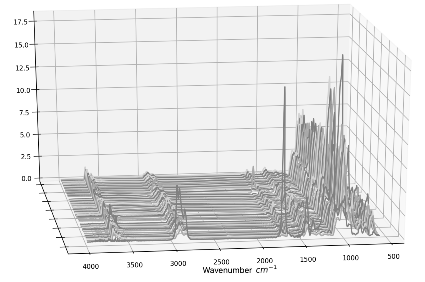
C. Butucea (ENSAE), J.-F. Delmas (Ecole des Ponts), A. Dutfoy (EDF R&D),
C. Hardy (EDF R&D, Ecole des Ponts)

Sparse spike deconvolution

Infrared spectroscopy

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 of the location of peaks and their corresponding bonds for polychloroprene samples ([Tchalla, 2017]).



$$\mathbf{y(t)} = \sum_{k=1}^{s} \beta_k \, \varphi(\theta_k, t) + w(t), \ (\varphi(\theta, \cdot), \theta \in \Theta) \text{ continuous dictionary.}$$

Some examples of dictionaries

- **Sparse spike deconvolution:** $$\begin{aligned}\varphi\colon \Theta \times \mathbb{R} &\to \mathbb{R}\\ (\theta, t) &\mapsto \mathrm{e}^{-\frac{(\theta-t)^2}{2\sigma^2}}\,.\end{aligned}$$

- **Scaling model:** $$\begin{aligned}\varphi\colon \Theta \times \mathbb{R}_+ &\to \mathbb{R}\\ (\theta, t) &\mapsto \mathrm{e}^{-\theta\, t}\,.\end{aligned}$$

- **Multiresolution approximation:** $$\begin{aligned}\varphi_j\colon \Theta \times \mathbb{R} &\to \mathbb{R}\\ (\theta, t) &\mapsto \mathrm{sinc}(2^j t - \theta).\end{aligned}$$

- **One hidden layer neural networks:** $$\begin{aligned}\varphi\colon \Theta \times \mathbb{R}^d &\to \mathbb{R}\\ (\theta, x) &\mapsto \xi(\langle x, \theta\rangle)\end{aligned}$$
  where ξ is the ReLU or the sigmoid function.

Model

We observe a random element y of the Hilbert space $(H_T, < \cdot, \cdot >_T)$, for $T \in \mathbb{N}$.

Continuous dictionary $\{\varphi_T(\theta), \theta \in \Theta\}$ of non-degenerate elements of H_T and the normalized functions

$$\phi_T(\theta) = \frac{\varphi_T(\theta)}{\|\varphi_T(\theta)\|_T}.$$

Model

We observe a random element y of the Hilbert space $(H_T, < \cdot, \cdot >_T)$, for $T \in \mathbb{N}$.

Continuous dictionary $\{\varphi_T(\theta), \theta \in \Theta\}$ of non-degenerate elements of H_T and the normalized functions

$$\phi_T(\theta) = \frac{\varphi_T(\theta)}{\|\varphi_T(\theta)\|_T}.$$

We assume

$$y = \sum_{k=1}^{K} \beta_k^\star \cdot \phi_T(\theta_k^\star) + w_T,$$

where

- w_T is a centered Gaussian element of H_T,
- $\beta^\star$ in $\mathbb{R}^K$, $s-$sparse,
- $\{\theta_k^\star\}_{k=1}^K$ included in Θ.

Model

$$\boxed{y = \beta^\star \Phi_T(\vartheta^\star) + w_T, \quad \text{in} \quad H_T.} \qquad \text{(model)}$$

For all $\vartheta = (\theta_1, \cdots, \theta_K) \in \Theta^K$,

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

is a multivariate function defined on Θ^K. (K is a bound on s that can be taken arbitrarily large.)

$$S^\star = \{k, \quad \beta_k^\star \neq 0\}, \; Card\, S^\star = s < K.$$

Examples

We observe a process y in $H_T = L^2(\lambda_T)$.

- **Discrete example:** Regular grid on $[0,1]$, $\lambda_T = \frac{1}{T}\sum_{j=1}^{T} \delta_{t_j}$ with $t_j = j/T$ and , $w_T(t_j) \underset{i.i.d}{\sim} \mathcal{N}(0,\sigma^2)$.

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

Examples

We observe a process y in $H_T = L^2(\lambda_T)$.

- **Discrete example:** Regular grid on $[0,1]$, $\lambda_T = \frac{1}{T}\sum_{j=1}^{T}\delta_{t_j}$ with $t_j = j/T$ and , $w_T(t_j) \underset{i.i.d}{\sim} \mathcal{N}(0,\sigma^2)$.

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

- **Continuous example:** $\lambda_T = \mathit{Lebesgue}$ on $[0,1]$ and w_T is a Brownian motion: $w_T = \frac{\sigma}{\sqrt{T}} B$,

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

Examples

We observe a process y in $H_T = L^2(\lambda_T)$.

- **Discrete example:** Regular grid on $[0,1]$, $\lambda_T = \frac{1}{T}\sum_{j=1}^{T}\delta_{t_j}$ with

  $t_j = j/T$ and , $w_T(t_j) \underset{i.i.d}{\sim} \mathcal{N}(0,\sigma^2)$.

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

- **Continuous example:** $\lambda_T = \mathit{Lebesgue}$ on $[0,1]$ and w_T is a Brownian motion: $w_T = \frac{\sigma}{\sqrt{T}} B$,

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

In both cases: $\forall f \in L^2(\lambda_T),\ \mathit{Var}\,\langle f, w_T\rangle_T \leq \frac{\sigma^2}{T}\|f\|_T^2$.

Off-the-grid methods - BLasso

They can be stated and applied to:

-learning mixtures, compressed sensing, one hidden layer neural networks, super-resolution in signal processing...

Off-the-grid methods - BLasso

They can be stated and applied to:

-learning mixtures, compressed sensing, one hidden layer neural networks, super-resolution in signal processing...

Beurling-Lasso (BLasso) de Castro and Gamboa, 2012 - convex optimization problem over a set of Radon measures $\mathcal{M}(\mathcal{T})$ on the design space $\mathcal{T}$:

$$\min_{\mu \in \mathcal{M}(\mathcal{T})} \frac{1}{2}\|y - \Phi\mu\|_T^2 + \kappa|\mu|_{TV}, \qquad (\mathcal{P}(\kappa))$$

where $\Phi : \mathcal{M}(\mathcal{T}) \to H_T$ is the acquisition operator and $|\mu|_{TV}$ denotes the total variation of the measure μ.

Remark: $\Phi\mu = \int \phi d\mu$ is equal to $\sum_k \beta_k^\star \phi(\theta_k^\star)$ for $d\mu(t) = \sum_k \beta_k^\star \delta_{\theta_k^\star}(dt)$.

Optimization problem

Remark: -the solution to the problem $\mathcal{P}(\kappa)$ is not necessarily a discrete measure (typically when $dim(H_T) = +\infty$). Therefore, we proceed with a slightly different optimization problem so that we recover a discrete mixture as solution.

We build estimators by solving a regularized optimization problem with a tuning parameter $\kappa > 0$:

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

$$\Theta_T \subset \Theta, \text{ compact interval.}$$

We assume that for all $k \in S^\star$, $\theta_k^\star \in \Theta_T$.

Optimization problem

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

The algorithms used to solve numerically the problem (also the BLasso):

- Sliding Frank-Wolfe algorithm (Denoyel et al. 2019)
- conic particle gradient descent (Chizat, 2021)

Optimization problem

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

The algorithms used to solve numerically the problem (also the BLasso):

- Sliding Frank-Wolfe algorithm (Denoyel et al. 2019)
- conic particle gradient descent (Chizat, 2021)

We will give high-probability bounds for the prediction risk

$$\|\hat{\beta}\Phi_T(\hat{\theta}) - \beta^\star\Phi_T(\vartheta^\star)\|_T^2$$

and some estimation results.

Optimization problem

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

The algorithms used to solve numerically the problem (also the BLasso):

- Sliding Frank-Wolfe algorithm (Denoyel et al. 2019)
- conic particle gradient descent (Chizat, 2021)

We will give high-probability bounds for the prediction risk

$$\|\hat{\beta}\Phi_T(\hat{\theta}) - \beta^\star\Phi_T(\vartheta^\star)\|_T^2$$

and some estimation results.

Bibliography:

-For known $\vartheta^\star$, linear regression model! [Bühlmann and van de Geer, 2011].

Bibliography

BLasso : [de Castro and Gamboa, 2012];

Super-resolution and **compressed sensing**: [Candès and Fernandez-Granda, 2013, 2014];[Tang et al, 2013]; ...

Off-the-grid methods

- Existence of atomic solutions when $dim(H_T) < +\infty$, [Boyer et al, 2019].
- Exact support recovery results in a small noise regime, [Duval & Peyré, 2015].
- Density mixture model, [De Castro et al, 2020].
- Prediction error bounds for the Fourier basis functions, [Tang et al 2014], [Boyer et al, 2017].

-Non translation invariant models: [Poon, Keriven, Peyré, 2021] describes the natural geometric framework of the BLasso.

Kernel and Riemannian metric

Assume $\Theta \subseteq \mathbb{R}$. We define the kernel $\mathcal{K}_T$ on Θ^2 by:

$$\mathcal{K}_T(\theta, \theta') = \langle \phi_T(\theta), \phi_T(\theta') \rangle_T = \frac{\langle \varphi_T(\theta), \varphi_T(\theta') \rangle_T}{\|\varphi_T(\theta)\|_T \|\varphi_T(\theta')\|_T}.$$

Kernel and Riemannian metric

Assume $\Theta \subseteq \mathbb{R}$. We define the kernel $\mathcal{K}_T$ on Θ^2 by:

$$\mathcal{K}_T(\theta, \theta') = \langle \phi_T(\theta), \phi_T(\theta') \rangle_T = \frac{\langle \varphi_T(\theta), \varphi_T(\theta') \rangle_T}{\|\varphi_T(\theta)\|_T \|\varphi_T(\theta')\|_T}.$$

We have

$$g_T(\theta) = \partial^2_{xy} \mathcal{K}_T(\theta, \theta),$$

defining an intrinsic **Riemannian metric** on Θ^2:

$$\mathfrak{d}_T(\theta, \theta') = |G_T(\theta) - G_T(\theta')|,$$

where G_T is a primitive of $\sqrt{g_T}$.

Results - Prediction and estimation

Assume we observe the random element y of H_T under the regression model with $\beta^\star$ a s-sparse vector and $\vartheta^\star = (\theta_1^\star, \cdots, \theta_K^\star)$ a vector with entries in Θ_T, a compact interval of $\mathbb{R}$, such that:

Assumption

- *w_T is Gaussian and there exists a noise level $\sigma > 0$ and a decay rate for the noise variance $\Delta_T > 0$ such that for all $f \in H_T$,*

$$Var \langle f, w_T \rangle_T \leq \sigma^2 \Delta_T \|f\|_T^2.$$

- *Smoothness conditions on φ.*
- *Local concavity and boundedness of $\mathcal{K}_\infty$.*
- *For all $1 \leq k \neq \ell \leq s, \quad \mathfrak{d}_T(\theta_k^\star, \theta_\ell^\star) > 2\,\delta(s)$.*
- *$\mathcal{K}_T$ is close enough from $\mathcal{K}_\infty$.*

Results - Prediction and estimation

Theorem (Butucea, Delmas, Dutfoy, H., 22)

For $\tau > 1$ *and* $\kappa \geq \mathcal{C}_1 \sigma \sqrt{\Delta_T \log \tau}$, *we have*

$$\left\| \hat{\beta} \Phi_T(\hat{\vartheta}) - \beta^\star \Phi_T(\vartheta^\star) \right\|_T \leq \mathcal{C}_0 \, \sigma \, \sqrt{s} \, \kappa$$

with probability at least $1 - \mathcal{C}_2 \left(\frac{|\Theta_T|_{\mathfrak{d}_T}}{\tau \log \tau} \vee \frac{1}{\tau} \right)$.

Results - Prediction and estimation

We define the following sets for $r > 0$:

- $\hat{S} = \left\{ \ell : \hat{\beta}_\ell \neq 0 \right\}$ the support of $\hat{\beta}$;

- $\tilde{S}_k(r) = \left\{ \ell \in \hat{S} : \mathfrak{d}_T(\hat{\theta}_\ell, \theta_k^\star) < r \right\}$ the set of indices ℓ in the support of $\hat{\beta}$ associated to a parameter $\hat{\theta}_\ell$ that is close to $\theta_k^\star$, for k in $S^\star$;

- $\tilde{S}(r) = \bigcup_{k \in S^\star} \tilde{S}_k(r)$.

Results - Prediction and estimation

Theorem (Butucea, Delmas, Dutfoy, H., 22)

There exists $r > 0$ so that the sets $\tilde{S}_k(r)$ are disjoint and for $\tau > 1$ and $\kappa \geq \mathcal{C}_1 \sigma \sqrt{\Delta_T \log \tau}$, we have

$$\sum_{k \in S^\star} \left| |\beta_k^\star| - \sum_{\ell \in \tilde{S}_k(r)} |\hat{\beta}_\ell| \right| \lesssim \sigma s \kappa$$

$$\sum_{k \in S^\star} \left| \beta_k^\star - \sum_{\ell \in \tilde{S}_k(r)} \hat{\beta}_\ell \right| \lesssim \sigma s \kappa$$

$$\left\| \hat{\beta}_{\tilde{S}(r)^c} \right\|_{\ell_1} \lesssim \sigma s \kappa$$

with probability greater than $1 - \mathcal{C}_2 \left(\frac{|\Theta_T|_{\mathfrak{d}_T}}{\tau \sqrt{\log \tau}} \vee \frac{1}{\tau} \right)$.

Discussion

We consider a general framework including discrete and continuous models with Gaussian, possibly correlated, noise and various dictionaries of smooth functions.

Discussion

We consider a general framework including discrete and continuous models with Gaussian, possibly correlated, noise and various dictionaries of smooth functions.

The upper bound on the prediction risk:

- is nearly the same as for the linear regression in the discrete model (*i.e* $\vartheta^\star$ is known and $H_T = \mathbb{R}^T$),
- extends results obtained for a Fourier basis functions [Tang et al 2014], [Boyer et al, 2017].
- holds under strong separation conditions on the non-linear parameters (of order s in theory, can be reduced to constant for models of spike deconvolution)!
- is free of K
- involves controls of tails of sup of linear functionals of a Gaussian process (Azaïs and Wschebor, 2009)

In preparation

- Simultaneous learning of a continuum of signals

$$y(z) = \sum_{k=1}^{s} \beta(z)\Phi_T(\vartheta) + w_T(z), \qquad z \in \mathcal{Z}.$$

- Goodness-of-fit testing
- Testing if the features involved in the mixture belong to a known finite set of features.

Bonus

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

- *w_T is Gaussian and there exists a noise level $\sigma > 0$ and a decay rate for the noise variance $\Delta_T > 0$ such that for all $f \in H_T$,*

$$Var \langle f, w_T \rangle_T \leq \sigma^2 \Delta_T \|f\|_T^2.$$

- *Smoothness conditions on φ.*
- *Local concavity and boundedness of $\mathcal{K}_\infty$.*
- *$\delta(u, s) < +\infty$, where u is a computable constant.*
- *For all $1 \leq k \neq \ell \leq s$, $\quad \mathfrak{d}_T(\theta_k^\star, \theta_\ell^\star) > 2\,\delta(u, s)$.*
- *$\mathcal{K}_T$ is close enough from $\mathcal{K}_\infty$.*

$\delta(u,s) = \inf \left\{ \delta > 0\colon\ \max_{1 \leq \ell \leq s} \sum_{k=1, k \neq \ell}^{s} |\mathcal{K}_\infty^{[i,j]}(\theta_\ell, \theta_k)| < u \text{ for all } (i,j) \in \{0,1\} \times \{0,1,2\} \text{ and for all } (\theta_1, \cdots, \theta_s) \in \Theta^s \text{ s.t } \mathfrak{d}_\infty(\theta_k, \theta_\ell) > \delta \right\}.$

Bonus

Boundedness and local concavity on the diagonal of the kernel
Define:

$$\varepsilon_T(r) = 1 - \sup \{ |\mathcal{K}_T(\theta, \theta')|; \quad \theta, \theta' \in \Theta_T \text{ such that } \mathfrak{d}_T(\theta', \theta) \geq r \} ,$$

$$\nu_T(r) = -\sup \left\{ \mathcal{K}_T^{[0,2]}(\theta, \theta'); \quad \theta, \theta' \in \Theta_T \text{ such that } \mathfrak{d}_T(\theta', \theta) \leq r \right\} .$$

We shall require $\varepsilon_T(r)$ and $\nu_T(r)$ for some $r > 0$.