SuperMix: Sparse Regularization for Mixtures

Y. De Castro [1], S. Gadat [2], C. Marteau [3]
and C. Maugis-Rabusseau [4] .

Metting in Mathematical Statistics
CIRM 2020

1. Ecole Centrale de Lyon - Institut Camille Jordan
2. Université Toulouse I - Toulouse School of Economics
3. Université Lyon I - Institut Camille Jordan
4. INSA de Toulouse - Institut Mathématiques de Toulouse

Outline

Outline

The mixture model

We have at our disposal a sample $\mathcal{S} = (X_1, \dots, X_n)$ of i.i.d. random variables ($X_i \in \mathbb{R}^d$), having a common density $f^\star$.

In an unsupervised classification context, $f^\star$ can be considered of the form

$$f^\star = \sum_{k=1}^{K} a_k \varphi(. - t_k),$$

where φ is a **known** density, $a_k \in [0,1]$, $t_k \in \mathbb{R}^d$ and K are unknown parameters.

Classical statistical issues

- estimation of the sequences $(a_k)_{k=1 \dots K}$ and $(t_k)_{k=1 \dots K}$,
- estimation of the component number K (model selection task).

References

[1] F. Bunea, A. B. Tsybakov, M. H. Wegkamp, and A. Barbu. Spades and mixture models. The Annals of Statistics, 38(4) :2525–2558, 2010.

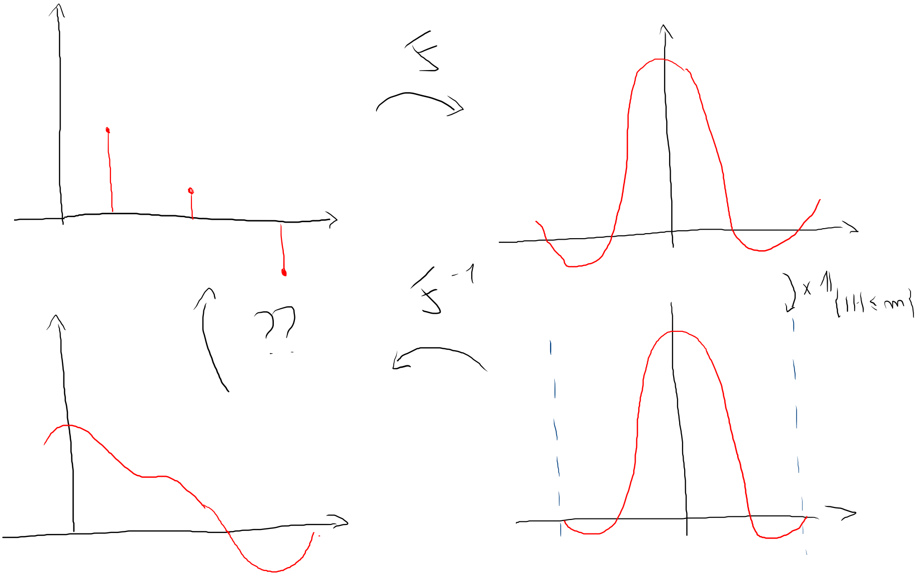
[2] C. Butucea and P. Vandekerkhove. Semiparametric mixtures of symmetric distributions. Scand. J. Stat., 41(1) :227–239, 2014.

[3] D. Donoho and J. Jin. Higher Criticism for detecting sparse heterogeneous mixtures, *Annals of Statistics*, **32**, (2004) 962-994.

[4] P. Heinrich and J. Kahn. Strong identifiability and optimal minimax rates for finite mixture estimation. The Annals of Statistics, 46 :2844–2870, 2018.

[5] C. Maugis-Rabusseau and B. Michel. A non asymptotic penalized criterion for Gaussian mixture model selection. ESAIM Probab. Stat., 15 :41–68, 2011.

The super resolution phenomenon



The super resolution phenomenon

Signal of interest : $x = \sum_{j=1}^{K} a_j^0 \delta_{t_j}$.

Observation : $y = \lambda * x$ with $\mathcal{F}[\lambda] = \mathbf{1}_{|.| \leq m}$.

Considered convex program :

$$\min_{\tilde{x}} \|\tilde{x}\|_{TV} \quad \text{s.t.} \quad y = \lambda * \tilde{x}.$$

$\longrightarrow$ perfect recovery provided $\Delta = \min_{i \neq j} |t_i - t_j| \geq 1/m$.

E. J. Candès and C. Fernandez-Granda. Towards a Mathematical Theory of Super-resolution. Communications on Pure and Applied Mathematics, 67(6) :906–956, 2014.

Mixture as an inverse problem

The estimation of the mixture parameters turns to be a discret inverse (deconvolution) problem. Indeed,

$$X_i = U_i + \epsilon_i, \quad \forall i \in \{1, \dots, n\},$$

where $\epsilon_i \sim \varphi$ (error term) and U_i are associated to the discrete measure $\mu_0 = \sum_{k=1}^{K} a_k \delta_{t_k}$. Then,

$$f^\star = \varphi * \mu_0 \quad \text{and} \quad \mathcal{F}[f^\star] = \mathcal{F}[\varphi] \times \mathcal{F}[\mu_0].$$

In this context, the 'classical' deconvolution tools are not available.

Outline

Notation

- The empirical measure

$$\hat{f}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}.$$

- The total variation norm. For any $\mu \in \mathcal{M}(\mathbb{R}^d, \mathbb{R})$,

$$\begin{aligned} \|\mu\|_1 &= \sup\left\{\int_{\mathbb{R}^d} f\, d\mu : \ f \text{ is } \mu-\text{measurable and } |f| \leq 1\right\}, \\ &= \int_{\mathbb{R}^d} d|\mu|. \end{aligned}$$

- Convolution operator $Lf = \lambda * f$, where **in this talk**, the filter λ is such that $\mathcal{F}[\lambda](t) = \mathbf{1}_{\{|t| \leq m\}}$.

A Beurling-Lasso approach

We define $\hat{\mu}_n$ as

$$\hat{\mu}_n \in \arg \min_{\mu \in \mathcal{M}(\mathbb{R}^d, \mathbb{R})} \left\{ \| L\hat{f}_n - L\varphi * \mu \|_{\mathbb{L}}^2 + \kappa \|\mu\|_1 \right\},$$

for some regularization parameter κ.

Items not discussed in this talk

- Is $\hat{\mu}_n$ a discrete measure ? (yes if $d = 1$).
- Algorithms to compute $\hat{\mu}_n$.

L. Chizat. Sparse optimization on measures with over-parameterized gradient descent. arXiv :1907.10300, 2019.

Q. Denoyelle, V. Duval, G. Peyré, and E. Soubies. The sliding frank-wolfe algorithm and its application to super-resolution microscopy. Inverse Problems, 2019.

Outline

Theoretical bound

Using simple inequalities, we can prove that if $\kappa = \rho_n / \|c_{0,m}\|_{\mathbb{L}}^2$ then

$$\mathbb{E}[\mathcal{D}_{\mathcal{P}_m}(\hat{\mu}_n, \mu_0)] \lesssim \frac{\rho_n \|\mathcal{P}_m\|_2}{\sqrt{\inf_{|t|\leq m} |\mathcal{F}[\varphi](t)|^2}},$$

where

$$\mathcal{D}_{\mathcal{P}_m}(\hat{\mu}_n, \mu_0) := \|\hat{\mu}_n\|_1 - \|\mu_0\|_1 - \int_{\mathbb{R}^d} \mathcal{P}_m d(\hat{\mu}_n - \mu_0),$$

is the Bregman divergence between $\hat{\mu}_n$ and μ_0, $(\rho_n)_n$ is such that

$$\mathbb{E}[\|L\hat{f}_n - L\varphi * \mu^0\|^2] \leq \rho_n^2 \quad \forall n \in \mathbb{N},$$

and $\mathcal{P}_m$ is a **dual certificate** s.t. $\mathcal{P}_m = \varphi * c_{0,m}$ with $\mathcal{F}[c_{0,m}](t) = 0$ for any $|t| > m$.

Theoretical bound

Assume that $\mathcal{P}_m$ is such that

$$\mathcal{P}_m(t_k) = 1 \quad \forall k \in \{1, \ldots, K\}.$$

Then,

$$\int_{\mathbb{R}^d} \mathcal{P}_m d\mu_0 = \sum_{k=1}^{K} a_k = \|\mu_0\|_1,$$

and

$$\begin{aligned}
\mathcal{D}_{\mathcal{P}_m}(\hat{\mu}_n, \mu_0) &= \|\hat{\mu}_n\|_1 - \|\mu_0\|_1 - \int_{\mathbb{R}^d} \mathcal{P}_m d(\hat{\mu}_n - \mu_0), \\
&= \|\hat{\mu}_n\|_1 - \int_{\mathbb{R}^d} \mathcal{P}_m d\hat{\mu}_n, \\
&= \int_{\mathbb{R}^d} (1 - \mathcal{P}_m) d\hat{\mu}_n^+ + \int_{\mathbb{R}^d} (1 + \mathcal{P}_m) d\hat{\mu}_n^- .
\end{aligned}$$

Theoretical bound

Hence

$$\mathbb{E}\left[\int_{\mathbb{R}^d}(1-\mathcal{P}_m)d\hat{\mu}_n^+ + \int_{\mathbb{R}^d}(1+\mathcal{P}_m)d\hat{\mu}_n^-\right] \lesssim \frac{\rho_n\|\mathcal{P}_m\|_2}{\sqrt{\inf_{|t|\leq m}|\mathcal{F}[\varphi](t)|^2}}.$$

In particular, if

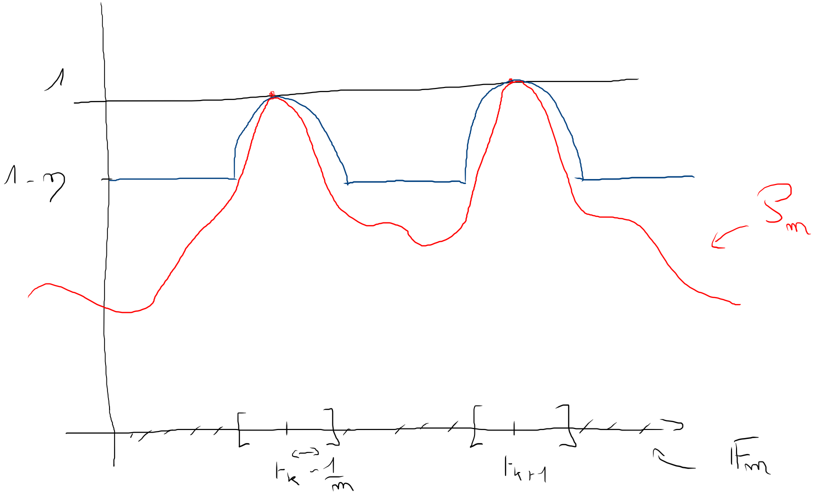
$$\mathcal{P}_m(x) \geq 0 \quad \forall x \in \mathbb{R}^d,$$

then

$$\mathbb{E}[\hat{\mu}_n^-(\mathbb{R}^d)] \leq \mathbb{E}\left[\int_{\mathbb{R}^d}(1+\mathcal{P}_m)d\hat{\mu}_n^-\right] \lesssim \frac{\rho_n\|\mathcal{P}_m\|_2}{\sqrt{\inf_{|t|\leq m}|\mathcal{F}[\varphi](t)|^2}}.$$

$\Rightarrow$ Control of the mass handled by the negative part of $\hat{\mu}_n$.

The dual certificate



Far region

For any $m \in \mathbb{R}^+$, define

$$\mathbb{F}_m = \bigcap_{k=1}^{K} \{t \in \mathbb{R}^d, \ \|t - t_k\| \gtrsim \frac{1}{m}\}.$$

Assume that, for some constant $\eta > 0$,

$$0 \leq \mathcal{P}_m(t) \leq 1 - \eta \quad \forall t \in \mathbb{F}_m.$$

Then

$$\eta \, \mathbb{E}[\hat{\mu}_n^+(\mathbb{F}_m)] \leq \mathbb{E}\left[\int (1 - \mathcal{P}_m) d\hat{\mu}_n^+\right] \lesssim \frac{\rho_n \|\mathcal{P}_m\|_2}{\sqrt{\inf_{|t| \leq m} |\mathcal{F}[\varphi](t)|^2}}.$$

Near region (spike detection)

Set $\mathbb{N}_m = \mathbb{F}_m^c$ and assume that

$$0 \leq \mathcal{P}_m(t) \leq 1 - Cm^2\|t - t_k\|^2 \quad \forall t \ s.t. \ \|t - t_k\| < \frac{1}{m}.$$

Then it is possible to prove that, $\forall A \subset \mathbb{R}^d$,

$$\mathbb{E}[\hat{\mu}_n^+(A)] \gtrsim \frac{\rho_n \|\mathcal{P}_m\|_2}{\sqrt{\inf_{|t|\leq m} |\mathcal{F}[\varphi](t)|^2}} \quad \Rightarrow \quad \min_{k\in[K]} \min_{t\in A} \|t - t_k\|_2^2 \lesssim \frac{1}{m^2}.$$

In some sense, m can be seen as a precision index.

Outline

Theoretical bounds

There exists $\mathcal{P}_m$ satisfying all the constraints mentioned above provided

$$m \geq \sqrt{K} d^{3/2} \Delta^{-1}.$$

Then

$$i) \quad \mathbb{E}[\hat{\mu}_n^-(\mathbb{R}^d)] \leq \frac{\rho_n \|\mathcal{P}_m\|_2}{\sqrt{\inf_{|t|\leq m} |\mathcal{F}[\varphi](t)|^2}}.$$

$$ii) \quad \mathbb{E}[\hat{\mu}_n(\mathbb{F}_m)] \lesssim \frac{\rho_n \|\mathcal{P}_m\|_2}{\sqrt{\inf_{|t|\leq m} |\mathcal{F}[\varphi](t)|^2}}.$$

$$iii) \quad \mathbb{E}[\hat{\mu}_n^+(A)] \gtrsim \frac{\rho_n \|\mathcal{P}_m\|_2}{\sqrt{\inf_{|t|\leq m} |\mathcal{F}[\varphi](t)|^2}} \Rightarrow \min_{k\in[K]} \min_{t\in A} \|t - t_k\|_2^2 \lesssim \frac{1}{m^2}$$

Behavior of these quantities for some specific cases ?

The Gaussian case

We consider the specific example of Gaussian mixtures ($d = 1$) :

$$\varphi(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \quad \text{and} \quad \mathcal{F}[\varphi](t) = e^{-\frac{t^2}{2}} \quad \forall t \in \mathbb{R}.$$

Then

- $\rho_n = \mathbb{E}\| L\hat{f}_n - L\varphi * \mu_0 \|_{\mathbb{L}}^2 \lesssim \dfrac{m}{n}.$
- $\inf_{|t| \leq m} \mathcal{F}[\varphi](t) = e^{-\frac{m^2}{2}}.$
- $\|\mathcal{P}_m\|_2^2 \leq \dfrac{1}{m}.$

The Gaussian case

Hence,

$$\frac{\rho_n \|\mathcal{P}_m\|_2}{\sqrt{\inf_{|t|\leq m} |\mathcal{F}[\varphi](t)|^2}} \lesssim \frac{1}{\sqrt{m}} \times e^{\frac{m^2}{2}} \times \sqrt{\frac{m}{n}} \sim \frac{e^{m^2/2}}{\sqrt{n}}.$$

Two possible scenarii

- m is constant (parametric rate but poor precision)

$$\max(\hat{\mu}_n^+(\mathbb{F}_m)) \lesssim \frac{1}{\sqrt{n}} \qquad \hat{\mu}_n(A) \gtrsim \frac{1}{\sqrt{n}} \Rightarrow \min_{k\in[K]} \min_{t\in A} \|t - t_k\|_2 \lesssim \frac{1}{m}.$$

- $m \sim \sqrt{r \log(n)}$ with $r < 1$. Then $\max(\hat{\mu}_n^+(\mathbb{F}_m)) \lesssim n^{\frac{r-1}{2}}$ and

$$\hat{\mu}_n(A) \gtrsim n^{\frac{r-1}{2}} \Rightarrow \min_{k\in[K]} \min_{t\in A} \|t - t_k\|_2^2 \lesssim \frac{1}{\log(n)}.$$

Outline

Conclusion

Possible outcomes

- Optimality (and improvement) of these results.
- Algorithms
- Considering heterogeneous mixtures.

Y. de Castro, S. Gadat, C. Marteau and C. Maugis-Rabusseau. SuperMix : Sparse regularization for mixtures. To appear in Annals of Statistics.