Thesis

Mehdi Sebbar

2017

Contents

Chapter 1

Introduction

The broad goal of this thesis is to tackle a clustering problem in the scope of mixtures model framework. More precisely, we will study the clustering of points drawn from high-dimensional Gaussian mixtures distributions. Thus, in the first part of this section we present the Gaussian mixture model and the second part we describe the well know Expectation-Maximization algorithm (EM). We will also present the limitations of this algorithm in high-dimensional setting.

1.1 The Gaussian mixture model

The Gaussian mixture model is an important framework for clustering problems. It assumes that the observations are drawn from a mixture distribution the components of which are Gaussian with parameters $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$:

$$\varphi_{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k}(x) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_k|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k)\right) \qquad (1.1)$$

Let $\boldsymbol{\theta}$ be the list containing all the unknown parameters of a Gaussian mixture model: the family of means $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K) \in (\mathbb{R}^p)^K$, the family of covariance matrices $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_k) \in (\mathcal{S}_{++}^p)^K$ and the vector of cluster probabilities $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k) \in [0,1]^K$ such that $\mathbf{1}_p^\top \boldsymbol{\pi} = 1$. The density of one observation $\boldsymbol{X}_1$ is then given by:

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{k=1}^K \pi_k \varphi_{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k}(\boldsymbol{x}), \qquad \forall \boldsymbol{x} \in \mathbb{R}^p, \qquad (1.2)$$

where $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})$.

This model can be interpreted from a latent variable perspective. Let Z be a discrete random variable taking its values in the set $[K]$ and such

that $\mathbf{P}(Z = k) = \pi_k$ for every $k \in [K]$. The random variable Z indicates the cluster from which the observation $\boldsymbol{X}$ is drawn. Considering that all the conditional distributions $\boldsymbol{X}|Z = k$ are Gaussian, we get the following formula for the marginal density of X:

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{k=1}^{K} \mathbf{P}(Z = k) p_{\boldsymbol{\theta}}(\boldsymbol{x}|Z = k) = \sum_{k=1}^{K} \pi_k \varphi_{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k}(\boldsymbol{x}), \qquad \forall \boldsymbol{x} \in \mathbb{R}^p. \quad (1.3)$$

In the clustering problem, the goal is to assign X to a cluster or, equivalently, to predict the cluster Z of the vector $\boldsymbol{X}$. A prediction function in such a context is $g : \mathbb{R}^p \to [K]$ such that $g(\boldsymbol{X})$ is as close as possible to Z. If we measure the risk of a prediction function g in terms of misclassification error rate $R_{\boldsymbol{\theta}}(g) = \mathbf{P}_{\boldsymbol{\theta}}(g(\boldsymbol{X}) \neq Z)$, then it is well known that the optimal (Bayes) predictor $g_{\boldsymbol{\theta}}^* \in \arg\min_g R_{\boldsymbol{\theta}}(g)$ is provided by the rule

$$g_{\boldsymbol{\theta}}^*(\boldsymbol{x}) = \arg\max_{k \in [K]} \tau_k(\boldsymbol{x}, \boldsymbol{\theta}),$$

where $\tau_k(\boldsymbol{x}, \boldsymbol{\theta}) = p_{\boldsymbol{\theta}}(Z = k | \boldsymbol{X} = \boldsymbol{x})$ stands for the conditional probability of the latent variable Z given $\boldsymbol{X}$. In the Gaussian mixture model, Bayes's rule implies that

$$\tau_k(\boldsymbol{x}, \boldsymbol{\theta}) = \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}|Z = k) \mathbf{P}(Z = k)}{p_{\boldsymbol{\theta}}(\boldsymbol{x})} = \frac{\pi_k \varphi_{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k}(\boldsymbol{x})}{\sum_{k'=1}^{K} \pi_{k'} \varphi_{\boldsymbol{\mu}_{k'}, \boldsymbol{\Sigma}_{k'}}(\boldsymbol{x})} \qquad (1.4)$$

Since the true value of the parameter $\boldsymbol{\theta}$ is not available, formula (1.4) can not be directly used for solving the problem of clustering. Instead, a natural strategy is to estimate $\boldsymbol{\theta}$ by some vector $\widehat{\boldsymbol{\theta}}$, based on a sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ drawn from the density $p_{\boldsymbol{\theta}}$, and then to define the clustering rule by

$$\widehat{g}(\boldsymbol{x}) = g_{\widehat{\boldsymbol{\theta}}}^*(\boldsymbol{x}) = \arg\max_{k \in [K]} \tau_k(\boldsymbol{x}, \widehat{\boldsymbol{\theta}}) = \arg\max_{k \in [K]} \widehat{\pi}_k \varphi_{\widehat{\boldsymbol{\mu}}_k, \widehat{\boldsymbol{\Sigma}}_k}(\boldsymbol{x}). \qquad (1.5)$$

A common approach to estimating the parameter $\boldsymbol{\theta}$ is to rely on the likelihood maximization.

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ with $\boldsymbol{X}_i \in \mathbb{R}^p$ be a set of iid observations drawn from the density $p_{\boldsymbol{\theta}}$ given by (1.2). The following graphical model depicts the scheme of the observations:

The log-likelihood of the Gaussian mixture model is

$$\ell_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}_i) = \sum_{i=1}^{n} \log \left\{ \sum_{k=1}^{K} \pi_k \varphi_{(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}(\boldsymbol{x}_i) \right\}. \qquad (1.6)$$
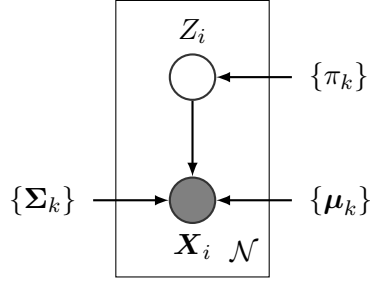
Because of the presence in this equation of the logarithm of a sum, the maximization of the log-likelihood is a difficult nonlinear and nonconvex problem. In particular, this is not a exponential family distribution yielding simple expressions. A commonly used approach for approximately maximizing (1.6) with respect to $\boldsymbol{\theta}$ is the Expectation-Maximization (EM) Algorithm [12] that we recall below.

Summarizing the content of this section, we can describe the following natural approach to solving the clustering problem under Gaussian mixture modeling assumption:

Input: data vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^p$ and the number of clusters K
Output: function $\widehat{g} : \mathbb{R}^p \to [K]$
1: Estimate $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ by maximizing the log-likelihood:

$$\widehat{\boldsymbol{\theta}} \in \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \ell(\boldsymbol{\theta}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \arg\max_{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}} \sum_{i=1}^{n} \log \left\{ \sum_{k=1}^{K} \pi_k \varphi_{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k}(\boldsymbol{x}_i) \right\}. \tag{1.7}$$

2: Output the clustering rule:

$$\widehat{g}(\cdot) = \arg\max_{k \in [K]} \widehat{\pi}_k \varphi_{\widehat{\boldsymbol{\mu}}_k, \widehat{\boldsymbol{\Sigma}}_k}(\cdot). \tag{1.8}$$

Figure 1.1: Clustering under Gaussian mixture modeling

1.2 EM Algorithm

The goal of the EM algorithm is to approximate a solution of the problem (1.7). Since this optimization problem contains a nonconvex cost function, it is impossible to design a polynomial time algorithm that provably converges to the global maximum point. Instead, the EM algorithm provides a

sequence $\{\widehat{\boldsymbol{\theta}}(t)\}_{t\in\mathbb{N}}$ of parameter values such that the cost function (*i.e.*, the log-likelihood) evaluated at these values forms an increasing sequence that converges to a local maximum.

The main idea underlying the EM algorithm is the following representation of the log-likelihood of one observation derived from the log-sum inequality:

$$\log\left\{\sum_{k=1}^{K}\pi_k\varphi_{\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k}(\boldsymbol{x}_i)\right\} = \max_{\boldsymbol{\tau}\in[0,1]^K\ \boldsymbol{\tau}^\top\mathbf{1}_K=1}\sum_{k=1}^{K}\left\{\tau_k\log\varphi_{\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k}(\boldsymbol{x}_i)+\tau_k\log(\pi_k/\tau_k)\right\}.$$
(1.9)

Let us denote by $\boldsymbol{\mathcal{T}} = (\tau_{i,k})$ a $n\times K$ matrix with nonnegative entries such that $\boldsymbol{\mathcal{T}}\mathbf{1}_K = \mathbf{1}_n$, that is each row of $\boldsymbol{\mathcal{T}}$ is a probability distribution on $[K]$. Combining (1.7) and (1.9), we get

$$\widehat{\boldsymbol{\theta}} \in \underset{\boldsymbol{\theta}=(\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma})}{\arg\max}\max_{\boldsymbol{\mathcal{T}}}\sum_{i=1}^{N}\sum_{k=1}^{K}\left\{\tau_{i,k}\log\varphi_{\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k}(\boldsymbol{x}_i) + \tau_{i,k}\log(\pi_k/\tau_{i,k})\right\}.$$
(1.10)

The great advantage of this new representation of the log-likelihood function is that the cost function in (1.10), considered as a function of $\boldsymbol{\theta}$ and $\boldsymbol{\mathcal{T}}$, is biconcave, *i.e.*, it is concave with respect to $\boldsymbol{\theta}$ for every fixed $\boldsymbol{\mathcal{T}}$ and concave with respect to $\boldsymbol{\mathcal{T}}$ for every fixed $\boldsymbol{\theta}$. In such a situation, one can apply the alternating maximization approach to sequentially improve on an initial point. In the present context, an additional attractive feature of the cost function in (1.10) is that the two optimization problems involved in the alternating maximization procedure admit explicit solutions.

Lemma 1. *Let us introduce the cost function*

$$F(\boldsymbol{\theta},\boldsymbol{\mathcal{T}}) = \sum_{i=1}^{n}\sum_{k=1}^{K}\left\{\tau_{i,k}\log\varphi_{\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k}(\boldsymbol{x}_i) + \tau_{i,k}\log(\pi_k/\tau_{i,k})\right\}.$$
(1.11)

Then, the following two optimization problems

$$\widehat{\boldsymbol{\theta}}(\boldsymbol{\mathcal{T}}) \in \underset{\boldsymbol{\theta}}{\arg\max}\, F(\boldsymbol{\theta},\boldsymbol{\mathcal{T}}), \qquad \widehat{\boldsymbol{\mathcal{T}}}(\boldsymbol{\theta}) \in \underset{\boldsymbol{\mathcal{T}}}{\arg\max}\, F(\boldsymbol{\theta},\boldsymbol{\mathcal{T}})$$
(1.12)

has explicit solutions given by

$$\widehat{\pi}_k = \frac{1}{n}\sum_{i=1}^{n}\tau_{i,k}, \qquad \widehat{\boldsymbol{\mu}}_k = \frac{1}{n\widehat{\pi}_k}\sum_{i=1}^{n}\tau_{i,k}\boldsymbol{x}_i, \qquad \forall k\in[K],$$
(1.13)

$$\widehat{\boldsymbol{\Sigma}}_k = \frac{1}{n\widehat{\pi}_k}\sum_{i=1}^{n}\tau_{i,k}(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_k)(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_k)^\top, \qquad \forall k\in[K],$$
(1.14)

$$\widehat{\tau}_{i,k} = \frac{\pi_k\varphi_{\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k}(\boldsymbol{x}_i)}{\sum_{k'\in[K]}\pi_{k'}\varphi_{\boldsymbol{\mu}_{k'},\boldsymbol{\Sigma}_{k'}}(\boldsymbol{x}_i)}, \qquad \forall k\in[K],\ \forall i\in[n].$$
(1.15)

Input: data vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^p$ and the number of clusters K
Output: parameter estimate $\widehat{\boldsymbol{\theta}} = \{\widehat{\boldsymbol{\mu}}_k, \widehat{\boldsymbol{\Sigma}}_k, \pi_k\}_{k \in [K]}$
1: Initialize $t = 0$, $\boldsymbol{\theta} = \boldsymbol{\theta}^0$.
2: **Repeat**
    3: Update the parameter $\mathcal{T}$:

$$\tau_{i,k}^t = \frac{\pi_k^t \varphi_{\boldsymbol{\mu}_k^t, \boldsymbol{\Sigma}_k^t}(\boldsymbol{x}_i)}{\sum_{k' \in [K]} \pi_{k'}^t \varphi_{\boldsymbol{\mu}_{k'}^t, \boldsymbol{\Sigma}_{k'}^t}(\boldsymbol{x}_i)}.$$

    4: Update the parameter $\boldsymbol{\theta}$:

$$\pi_k^{t+1} = \frac{1}{n} \sum_{i=1}^n \tau_{i,k}^t, \qquad \boldsymbol{\mu}_k^{t+1} = \frac{1}{n\pi_k^{t+1}} \sum_{i=1}^n \tau_{i,k}^t \boldsymbol{x}_i,$$

$$\boldsymbol{\Sigma}_k^{t+1} = \frac{1}{n\pi_k^{t+1}} \sum_{i=1}^n \tau_{i,k}^t (\boldsymbol{x}_i - \boldsymbol{\mu}_k^{t+1})(\boldsymbol{x}_i - \boldsymbol{\mu}_k^{t+1})^\top.$$

    5: increment t: $t = t + 1$.
6: **Until** stopping rule.
7: **Return** $\boldsymbol{\theta}^t$.

Figure 1.2: EM algorithm for Gaussian mixtures

Based on this result, the EM algorithm is defined as in Figure 1.2. The algorithm operates iteratively and needs a criterion to determine when the iterations should be stopped. There is no clear consensus on this point in the statistical literature, but it is a commonly used practice to stop when one of the following conditions is fulfilled:

i) The number of iterations t exceeds a pre-specified level $t_{\max}$.

ii) The increase of the log-likelihood over past t_0 iterations is not significantly different from zero: $\ell_n(\boldsymbol{\theta}^t) - \ell_n(\boldsymbol{\theta}^{t-t_0}) \leq \varepsilon$ for some pre-specified values $t_0 \in \mathbb{N}$ and $\varepsilon > 0$.

EM is conceptually easy and each iteration increases the log-likelihood:

$$\ell_n(\boldsymbol{\theta}^{t+1}) \geq \ell_n(\boldsymbol{\theta}^t), \qquad \forall t \in \mathbb{N}.$$

The complexity at each step of the EM algorithm is $O(Knp^2)$ and it usually requires many iterations to converge. In a high-dimensional setting when p is large, the quadratic dependence on p may result in prohibitively large running times. However, the computation of the elements of the covariance

matrices $\boldsymbol{\Sigma}_k^t$ and the mean vectors $\boldsymbol{\mu}_k^t$ can be parallelized which may lead to considerable savings in the running time.

Chapter 2

Graphical Lasso for Gaussian mixtures

The EM algorithm experiences severe performance degradation in high-dimensional setting. A technique to avoid this degradation is by regularizing the parameters of the model. The following algorithm is inspired by the Graphical lasso [14] [1] which penalizes the components of the precision matrix of a Gaussian graphical model.

We consider $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ a sample of n points drawn from a p-dimensional Gaussian mixture distribution. In this problem, we will consider the estimation of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ with $\theta_k = (\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Omega}_k)$ where $\boldsymbol{\Omega}_k$ is the precision matrix regarding the k component of the mixture. We denote $\varphi_{(\boldsymbol{\mu}_k, \boldsymbol{\Omega}_k)}$ the gaussian density of mean $\boldsymbol{\mu}_k$ and precision matrix $\boldsymbol{\Omega}_k$. The penalized log-likelihood is

$$\ell_n^{pen}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - pen(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \left\{ \sum_{k=1}^{K} \pi_k \varphi_{(\boldsymbol{\mu}_k, \boldsymbol{\Omega}_k)}(\boldsymbol{x}_i) \right\} - pen(\boldsymbol{\theta}).$$
(2.1)

In this problem, we suppose that within a cluster k, most pairs of features (X^i, X^j) are independent given the other features X^l with $i, j, l \in [p], l \notin \{i, j\}$. This property entails the sparsity of $\boldsymbol{\Omega}_k$. Therefore, we consider an l_1 regularization $pen(\theta_k) = \lambda_k ||\boldsymbol{\Omega}_k||_{1,1}$ with $\lambda_k > 0$.

The penalization of the log-likelihood concerns only the precision matrices $\boldsymbol{\Omega}_k$. Regarding the other parameters $(\pi_k, \boldsymbol{\mu}_k)$, our algorithm is the same as EM and we can use the same iteration technique as in lemma 1 to maximize

the following cost function

$$F^{pen}(\boldsymbol{\theta}, \boldsymbol{\mathcal{T}}) = \sum_{k=1}^{K} \left(\sum_{i=1}^{n} \left\{ \tau_{i,k} \log \varphi_{\boldsymbol{\mu}_k, \boldsymbol{\Omega}_k}(\boldsymbol{x}_i) + \tau_{i,k} \log(\pi_k/\tau_{i,k}) \right\} - \lambda_k ||\boldsymbol{\Omega}_k||_{1,1} \right)$$
(2.2)

Then, we have the two following optimization problems

$$\widehat{\boldsymbol{\theta}}(\boldsymbol{\mathcal{T}}) \in \arg\max_{\boldsymbol{\theta}} F^{pen}(\boldsymbol{\theta}, \boldsymbol{\mathcal{T}}), \qquad \widehat{\boldsymbol{\mathcal{T}}}(\boldsymbol{\theta}) \in \arg\max_{\boldsymbol{\mathcal{T}}} F^{pen}(\boldsymbol{\theta}, \boldsymbol{\mathcal{T}})$$
(2.3)

which has explicit solutions. For a $\widehat{\boldsymbol{\mathcal{T}}}$ given, estimates of $(\pi, \boldsymbol{\mu})$ obtained by the first optimization problem in 2.3 are

$$\widehat{\pi}_k = \frac{1}{n} \sum_{i=1}^{n} \widehat{\tau}_{i,k}, \qquad \widehat{\boldsymbol{\mu}}_k = \frac{1}{n\widehat{\pi}_k} \sum_{i=1}^{n} \widehat{\tau}_{i,k} \boldsymbol{x}_i, \qquad \forall k \in [K]$$
(2.4)

And for a $\widehat{\boldsymbol{\theta}}$ given, an estimate of $\boldsymbol{\mathcal{T}}$ obtained by the second optimization problem is

$$\widehat{\tau}_{i,k} = \frac{\widehat{\pi}_k \varphi_{\widehat{\boldsymbol{\mu}}_k, \widehat{\boldsymbol{\Omega}}_k}(\boldsymbol{x}_i)}{\sum_{k' \in [K]} \widehat{\pi}_{k'} \varphi_{\widehat{\boldsymbol{\mu}}_{k'}, \widehat{\boldsymbol{\Omega}}_{k'}}(\boldsymbol{x}_i)}, \qquad \forall k \in [K], \ \forall i \in [n].$$
(2.5)

However, due to the penality $\lambda_k ||\boldsymbol{\Omega}_k||_{1,1}$, the estimation of $\boldsymbol{\Omega}_k$ is not straightforward.

We introduce the weighted empirical covariance matrix

$$\boldsymbol{\Sigma}_{n,k} = \frac{1}{n} \frac{\sum_{i=1}^{n} \tau_{i,k} (\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_k)(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_k)^\top}{\sum_{i=1}^{n} \tau_{i,k}}$$
(2.6)

The Gaussian density in equation (2.2) can be expanded as follows

$$F^{pen}(\boldsymbol{\theta}, \boldsymbol{\mathcal{T}}) = \sum_{k=1}^{K} \left(\sum_{i=1}^{n} \left\{ \tau_{i,k} \left(-\frac{p}{2} \log(2\pi) + \frac{1}{2} \log |\boldsymbol{\Omega}_k| \right. \right. \right.$$
$$\left. -\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Omega}_k(\boldsymbol{x}_i - \boldsymbol{\mu}_k) \right) + \tau_{i,k} \log(\pi_k/\tau_{i,k}) \right\} - \lambda_k ||\boldsymbol{\Omega}_k||_{1,1} \right)$$
$$= -\frac{np}{2} \log(2\pi) + \sum_{k=1}^{K} \left(\frac{n\pi_k}{2} \log |\boldsymbol{\Omega}_k| \right.$$
$$\left. + \sum_{i=1}^{n} \left\{ -\frac{\tau_{i,k}}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Omega}_k(\boldsymbol{x}_i - \boldsymbol{\mu}_k) + \tau_{i,k} \log(\pi_k/\tau_{i,k}) \right\} - \lambda_k ||\boldsymbol{\Omega}_k||_{1,1} \right).$$

The opposite minimization problem regarding each $\boldsymbol{\Omega}_k$ is

$$\boldsymbol{\Omega}_k \in \underset{\boldsymbol{\Omega} \succeq 0}{\arg\min} \left\{ -\frac{n\pi_k}{2} \log|\boldsymbol{\Omega}| + \frac{1}{2} \sum_{i=1}^{n} \tau_{i,k} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Omega} (\boldsymbol{x}_i - \boldsymbol{\mu}_k) + \lambda_k ||\boldsymbol{\Omega}||_{1,1} \right\}$$

$$(2.7)$$

Using the well-known commutativity property of the trace operator and dividing by $n\pi_k$

$$\boldsymbol{\Omega}_k \in \underset{\boldsymbol{\Omega} \succeq 0}{\arg\min} \left\{ -\frac{1}{2} \log|\boldsymbol{\Omega}| + \frac{1}{2} tr(\boldsymbol{\Sigma}_{n,k} \boldsymbol{\Omega}) + \frac{\lambda_k}{n\pi_k} ||\boldsymbol{\Omega}||_{1,1} \right\} \qquad (2.8)$$

Our algorithm solves a graphical lasso problem within each cluster. We use a block coordinate ascent algorithm [25] to solve this convex problem as in the graphical lasso implementation in R, see `http://statweb.stanford.edu/~tibs/glasso/` The alternating maximization procedure is summarized in the following algorithm

Input: data vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^p$ and the number of clusters K

Output: parameter estimate $\widehat{\boldsymbol{\theta}} = \{\widehat{\boldsymbol{\mu}}_k, \widehat{\boldsymbol{\Omega}}_k, \widehat{\pi}_k\}_{k \in [K]}$

1: Initialize $t = 0$, $\boldsymbol{\theta} = \boldsymbol{\theta}^0$.

2: **Repeat**

3:      Update the parameter $\mathcal{T}$:

$$\tau_{i,k}^t = \frac{\pi_k^t \varphi_{\boldsymbol{\mu}_k^t, \boldsymbol{\Omega}_k^t}(\boldsymbol{x}_i)}{\sum_{k' \in [K]} \pi_{k'}^t \varphi_{\boldsymbol{\mu}_{k'}^t, \boldsymbol{\Omega}_{k'}^t}(\boldsymbol{x}_i)}.$$

4:      Update the parameter $\boldsymbol{\theta}$:

$$\pi_k^{t+1} = \frac{1}{n} \sum_{i=1}^n \tau_{i,k}^t,$$

$$\boldsymbol{\mu}_k^{t+1} = \frac{1}{n\pi_k^{t+1}} \sum_{i=1}^n \tau_{i,k}^t \boldsymbol{x}_i$$

$$\boldsymbol{\Sigma}_{n,k} = \frac{1}{n^2 \pi_k^{t+1}} \sum_{i=1}^n \tau_{i,k}^{t+1} (\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_k^{t+1})(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_k^{t+1})^\top$$

$$\boldsymbol{\Omega}_k^{t+1} \in \underset{\boldsymbol{\Omega} \succeq 0}{\arg\min} \left\{ -\frac{1}{2} \log |\boldsymbol{\Omega}| + \frac{1}{2} tr(\boldsymbol{\Sigma}_{N,k} \boldsymbol{\Omega}) + \frac{\lambda_k}{n\pi_k^{t+1}} ||\boldsymbol{\Omega}||_{1,1} \right\}$$

5:          increment t: $t = t + 1$.

6: **Until** stopping rule.

7: **Return** $\boldsymbol{\theta}^t$.

Figure 2.1: Graphical lasso algorithm for Gaussian mixtures

Chapter 3

Estimating the number of clusters

The idea is to add a regularization term on the estimation of the $n \times K$ matrix $\mathcal{T}$, the estimate of the number of clusters K will be the number of non-empty columns of $\mathcal{T}$.

We consider a maximum number of clusters M, we note the convex set $A = \{\tau \in \mathbb{R}^M : \sum_{k=1}^{M} \tau_k = 1, \tau_k \geq 0 \quad \forall k \in [M]\}$ and the "indicator" function $\chi_A(.)$ defined by:

$$\chi_A(x) = \begin{cases} 0 & \text{if } x \in A, \\ \infty & \text{if } x = 0 \end{cases}$$

We note $\mathcal{T}_{.,k}$ the kth column and $\mathcal{T}_{i,.}$ the ith line of $\mathcal{T}$. We will estimate $\mathcal{T}$ using the same equation 2.2, 2.3 with a regularization term:

$$F^{pen}(\boldsymbol{\theta}, \mathcal{T}) = \sum_{k=1}^{K} \left(\sum_{i=1}^{n} \left\{ \tau_{i,k} \log \varphi_{\boldsymbol{\mu}_k, \boldsymbol{\Omega}_k}(\boldsymbol{x}_i) + \tau_{i,k} \log(\pi_k/\tau_{i,k}) \right\} - \lambda_k ||\boldsymbol{\Omega}_k||_{1,1} \right)$$
$$+ \sum_{k=1}^{K} ||\mathcal{T}_{.,k}||_2 + \sum_{i=1}^{n} \chi_A(\mathcal{T}_{i,.})$$

Removing the penalization on $\boldsymbol{\Omega}$:

$$F^{pen}(\boldsymbol{\theta}, \mathcal{T}) = \sum_{k=1}^{K} \left(\sum_{i=1}^{n} \left\{ \tau_{i,k} \log \varphi_{\boldsymbol{\mu}_k, \boldsymbol{\Omega}_k}(\boldsymbol{x}_i) + \tau_{i,k} \log(\pi_k/\tau_{i,k}) \right\} \right.$$
$$+ \sum_{k=1}^{K} ||\mathcal{T}_{.,k}||_2 + \sum_{i=1}^{n} \chi_A(\mathcal{T}_{i,.})$$

and the optimization problem:

$$\widehat{\mathcal{T}}(\boldsymbol{\theta}) \in \arg\max_{\mathcal{T}} F^{pen}(\boldsymbol{\theta}, \mathcal{T}) \tag{3.1}$$

Unfortunately, the regularization term prevents to derive explicit solution as in previous chapters. Furthermore, we cant separate the objective function since we optimize along columns and lines of $\mathcal{T}$. The objective function $F^{pen}(\boldsymbol{\theta}, \mathcal{T})$ rewritten $F^{pen}_{\boldsymbol{\theta}}(\mathcal{T})$ can be split into two terms:

$$F^{pen}_{\boldsymbol{\theta}}(\mathcal{T}) = f(\mathcal{T}) + g(\mathcal{T}) \tag{3.2}$$

with:

$$f(\mathcal{T}) = \sum_{k=1}^{K} \left(\sum_{i=1}^{n} \left\{ \tau_{i,k} \log \varphi_{\boldsymbol{\mu}_k, \boldsymbol{\Omega}_k}(\boldsymbol{x}_i) + \tau_{i,k} \log(\pi_k/\tau_{i,k}) \right\} + \sum_{k=1}^{K} ||\mathcal{T}_{.,k}||_2$$

$$g(\mathcal{T}) = \sum_{i=1}^{n} \chi_A(\mathcal{T}_{i,.})$$

f is convex and differentiable on its domain, g is also convex but not smooth. We will tackle this problem by using a proximal method:

$$\mathcal{T}^{k+1} = \mathbf{prox}_{\lambda g}(\mathcal{T}^k - \lambda \nabla f(\mathcal{T}^k)) \;\; = \qquad\qquad P_A(\mathcal{T}^k - \lambda \nabla f(\mathcal{T}^k))$$

$$= \arg\min_{\mathcal{T}: \forall K, \mathcal{T}^k \in A} \left(||\mathcal{T} - (\mathcal{T}^k - \lambda \nabla f(\mathcal{T}^k))||_2^2 \right)$$

The gradient of f on $\mathcal{T}$ is given by:

$$\left[\nabla_{\mathcal{T}} f(\mathcal{T}) \right]_{i,j} = \left[\frac{\partial f}{\partial \mathcal{T}_{ij}}(\mathcal{T}) \right]_{i,j}$$

$$= \log(\varphi_{\boldsymbol{\mu}_j, \boldsymbol{\Omega}_j}(x_i)) + \log(\frac{\pi_j}{\tau_{i,j}}) + \frac{\tau_{i,j}}{||\mathcal{T}_{.,j}||_2} - 1$$

We will use FISTA to accelerate the convergence

We use the algorithm of last chapter with the new estimation procedure of $\mathcal{T}$

Input:
Output: parameter estimate $\mathcal{T}$
1:   Initialize $t_1 = 1$ and $\boldsymbol{\xi}^0$ with

$$\xi^0_{i,k} = \frac{\pi^0_k \varphi_{\boldsymbol{\mu}^0_k, \boldsymbol{\Omega}^0_k}(\boldsymbol{x}_i)}{\sum_{k' \in [K]} \pi^0_{k'} \varphi_{\boldsymbol{\mu}^0_{k'}, \boldsymbol{\Omega}^0_{k'}}(\boldsymbol{x}_i)}$$

2:   **Repeat**

$$\mathcal{T}^k = \underset{\mathcal{T}: \forall K, \mathcal{T}^k \in A}{\arg\min} \left(||\mathcal{T} - (\boldsymbol{\xi}^k - \lambda \nabla f(\boldsymbol{\xi}^k))||_2^2 \right)$$

$$t^{k+1} = \frac{1 + \sqrt{1 + 4 * (t^k)^2}}{2}$$

$$\boldsymbol{\xi}^{k+1} = \mathcal{T}^k + \left(\frac{t^k - 1}{t^{k+1}} \right) (\mathcal{T}^k - \mathcal{T}^{k-1})$$

Figure 3.1: $\mathcal{T}$ estimation with FISTA

Input: data vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^p$ and the number of clusters K
Output: parameter estimate $\widehat{\boldsymbol{\theta}} = \{\widehat{\boldsymbol{\mu}}_k, \widehat{\boldsymbol{\Omega}}_k, \widehat{\pi}_k\}_{k \in [K]}$
1:   Initialize $t = 0$, $\boldsymbol{\theta} = \boldsymbol{\theta}^0$.
2:   **Repeat**
3:       Update the parameter $\mathcal{T}$ with previous algorithm
4:       Update the parameter $\boldsymbol{\theta}$:

$$\pi_k^{t+1} = \frac{1}{n} \sum_{i=1}^n \tau_{i,k}^t,$$

$$\boldsymbol{\mu}_k^{t+1} = \frac{1}{n\pi_k^{t+1}} \sum_{i=1}^n \tau_{i,k}^t \boldsymbol{x}_i$$

$$\boldsymbol{\Sigma}_{n,k} = \frac{1}{n^2 \pi_k^{t+1}} \sum_{i=1}^n \tau_{i,k}^{t+1} (\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_k^{t+1})(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_k^{t+1})^\top$$

$$\boldsymbol{\Omega}_k^{t+1} \in \underset{\boldsymbol{\Omega} \succeq 0}{\arg\min} \left\{ -\frac{1}{2} \log |\boldsymbol{\Omega}| + \frac{1}{2} tr(\boldsymbol{\Sigma}_{N,k} \boldsymbol{\Omega}) + \frac{\lambda_k}{n\pi_k^{t+1}} ||\boldsymbol{\Omega}||_{1,1} \right\}$$

5:       increment t: $t = t + 1$.
6:   **Until** stopping rule.
7:   **Return** $\boldsymbol{\theta}^t$.

Figure 3.2: Graphical lasso algorithm for Gaussian mixtures with cluster number discovery

Chapter 4

Algorithm 2

We consider the diagonal matrix $D_\lambda = diag(\lambda_1, \ldots, \lambda_K)$.

Input: data vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^p$, the number of clusters K and D_λ

Output: parameter estimate $\widehat{\boldsymbol{\theta}} = \{\widehat{\boldsymbol{\mu}}_k, \widehat{\boldsymbol{\Omega}}_k, \pi_k\}_{k \in [K]}$

1:   Initialize $t = 0$, $\boldsymbol{\theta} = \boldsymbol{\theta}^0$.

2:   **Repeat**

    3:   Update the parameter $\mathcal{T}$:

$$\tau_{i,k}^t = \frac{\pi_k^t \varphi_{\boldsymbol{\mu}_k^t, \boldsymbol{\Omega}_k^t}(\boldsymbol{x}_i)}{\sum_{k' \in [K]} \pi_{k'}^t \varphi_{\boldsymbol{\mu}_{k'}^t, \boldsymbol{\Omega}_{k'}^t}(\boldsymbol{x}_i)}.$$

    4:   Update the parameter $\boldsymbol{\theta}$:

$$(\mu^k, B^k) \in \underset{(\mu, B) \in \mathbb{R}^p \times \mathbb{R}^{p \times p}, B_{jj} = 1}{\arg\min} \left\{ \frac{1}{N} \sum_{n=1}^N \tau_n^k(t) ||(x_n - \mu)^T B||_2^2 + ||D_\lambda B||_{1,1} \right\}$$

$$\pi_k^{t+1} =$$
$$\boldsymbol{\mu}_k^{t+1} =$$
$$\boldsymbol{\Omega}_k^{t+1}$$

    5:   increment t: $t = t + 1$.

6:   **Until** stopping rule.

7:   **Return** $\boldsymbol{\theta}^t$.

Figure 4.1: Lasso for Gaussian mixtures

Chapter 5

structural analysis on Σ approach

We consider a multivariate Gaussian distribution with mean $\boldsymbol{\mu}^*$ and covariance $\boldsymbol{\Sigma}^*$ and $Y_1, \ldots, Y_N \in \mathbb{R}^p$ iid drawn from this distribution. We would like to estimate $\boldsymbol{\mu}^*$ and $\boldsymbol{\Sigma}^*$. We know that $\widehat{\boldsymbol{\mu}}_n = \bar{Y}_n$, then wlog we consider $\mu^* = 0$, the problem is to estimate $\boldsymbol{\Sigma}^*$. We will study the precision matrix and consider that Σ^{-1} is sparse. We note $\Sigma^{-1} = \Omega$, Y_n the n-th random variable and Y_n^i the i-th component of this vector. If $\Sigma_{ij}^{-1} = 0 \Rightarrow Y^i \perp\!\!\!\perp Y^j$ conditionally to $Y^{l \neq \{i,j\}}$. Thus, it makes sense to impose a L_1 penalty on Σ^{-1} to increase its sparsity.

5.1   Graphical Lasso

Let consider a multivariate normal distribution with parameters μ^*, Σ^* with density;

$$\mathcal{N}(x|\mu^*, \Sigma^*) = \frac{1}{(2\pi)^{d/2}|\Sigma^*|^{1/2}} \exp^{-\frac{1}{2}(x-\mu^*)^T \Sigma^{-1*}(x-\mu^*)} \tag{5.1}$$

We consider $\mu = 0$. Given N datapoints $X_1, \ldots, X_N$ and $X_i \in \mathbb{R}^d$, the log-likelihood is given by:

$$\begin{aligned} \mathcal{L}(\Sigma) &= \log\left(\prod_{n=1}^{N} \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp^{-\frac{1}{2}(x_n)^T \Sigma^{-1}(x_n)}\right) \\ &= -\frac{dN}{2}\log 2\pi - \frac{N}{2}\sum_{n=1}^{N}\log|\Sigma^*| - \frac{1}{2}\sum_{n=1}^{N} x_n^T \Sigma^{*,-1} x_n \end{aligned} \tag{5.2}$$

Note that $x_n^T \Sigma^{*,-1} x_n = tr(x_n^T \Sigma^{*,-1} x_n)$, and therefore:

$$\sum_{n=1}^{N} x_n^T \Sigma^{*,-1} x_n = tr\Big(\sum_{n=1}^{N} x_n^T \Sigma^{*,-1} x_n\Big) = tr\Big(\Big[\sum_{n=1}^{N} x_n^T x_n\Big] \Sigma^{*,-1}\Big) = tr(S_N \Sigma^*)$$
(5.3)

Where S_N is the empirical covariance matrix. We can replace that in the log-likelihood expression:

$$\mathcal{L}(\Sigma) = -\frac{dN}{2} \log 2\pi - \frac{N}{2} \sum_{n=1}^{N} \log |\Sigma^*| - \frac{1}{2} tr(S_N \Sigma^*)$$
(5.4)

Finally:

$$\mathcal{L}(\Sigma) = C + \frac{N}{2} \log |\Sigma^{-1}| - \frac{1}{2} tr(S_n \Sigma^{-1})$$
(5.5)

Where C is a constant (dependent on N). Thus, considering the sparsity of the precision matrix $\Omega = \Sigma^{-1}$, we impose a penalization to the maximum likelihood estimator of Ω

$$\widehat{\Omega} \in argmin\big\{ \log |\Omega| - tr(S_N \Omega) - \lambda ||\Omega||_1 \big\}$$
(5.6)

A reason to use the L_1 penalization instead of the ridge is that for an L_p penalization, the problem is convex for $p \geq 1$ and we have parsimonious property for $p \leq 1$. This is a convex optimization problem, however the complexity is $O(p^3)$ (Source, high dim & var select Buhlmann 2006 ? Wassermann)

5.2   Column-Wise Lasso

We consider a gaussian vector $Y \in \mathbb{R}^d$, $Y \sim \mathcal{N}(0, \Sigma)$. We can write $Y = (Y^1, Y^{2:d})$. With this decomposition we can write the covariance matrix as following:

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{pmatrix}$$
(5.7)

and according to theorem[?]: If Σ_{22} is inversible, then:

$$\begin{array}{rcl} \mathbf{E}[Y^1|Y^{2:d}] & = & \Sigma_{12}\Sigma_{22}^{-1}Y^{2:d} \\ Var[Y^1|Y^{2:d}] & = & \sigma_1^2 - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T \end{array}$$
(5.8)

We have the following identity:

$$\begin{pmatrix} \omega_{11} & \Omega_{12} \\ \Omega_{12}^T & \Omega_{22} \end{pmatrix} \begin{pmatrix} \sigma_1^2 & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & I_{p-1} \end{pmatrix}$$
(5.9)

Which gives the following equations:

$$\begin{cases} \omega_{11}\sigma_1^2 + \Omega_{12}\Sigma_{12}^T & = 1 \quad (*) \\ \omega_{11}\Sigma_{12} + \Omega_{12}\Sigma_{22} & = 0 \quad (**) \\ \Omega_{12}^T\Sigma_{12} + \Omega_{22}\Sigma_{22} & = I_{p-1} \quad (***) \end{cases} \tag{5.10}$$

With (**) we have $-\omega_{11}\Sigma_{12}\Sigma_{22}^{-1} = \Omega_{12}$ and injected to (*) we have:

$$\begin{cases} \mathbf{E}[Y^1|Y^{2:d}] & = -\frac{1}{\omega_{11}}\Omega_{12}Y^{2:d} \\ Var[Y^1|Y^{2:d}] & = \frac{1}{\omega_{11}} \end{cases} \tag{5.11}$$

Finally, $Y^1 - \mathbf{E}[Y^1|Y^{2:d}]$ is a gaussian vector of $\mathbb{R}^{d-1}$, centered, independent of $Y^{2:d}$ and of covariance matrix $\sigma_1^2 - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T$. If we denote $\xi^1 \sim \mathcal{N}(0,1)$ we have $Y^1 - \mathbf{E}[Y^1|Y^{2:d}] = \frac{1}{\sqrt{\omega_{11}}}\xi^1$.

Therefore, for $Y_1, \ldots, Y_n$ iid of law $\mathcal{N}(0, \Sigma^*)$ we have:

$$\begin{aligned} Y_i^1 & = -\frac{1}{\omega_{11}^*}\Omega_{12}Y_i^{2:d} + \frac{1}{\sqrt{\omega_{11}^*}}\xi_i^1 \\ & = -\sum_{j=2}^d \frac{w_{ij}^*}{\omega_{11}^*}Y_i^j + \frac{1}{\sqrt{\omega_{11}^*}}\xi_i^1 \end{aligned} \tag{5.12}$$

and

$$\beta_1^{*T}Y_i = \frac{1}{\sqrt{\omega_{11}^*}}\xi_i^1 \Rightarrow \beta_1^{*T}\mathbf{Y} = \frac{1}{\sqrt{\omega_{11}^*}}\boldsymbol{\xi}^1 \tag{5.13}$$

with

$$\beta_1^* = \frac{1}{\sqrt{\omega_{11}^*}}\left[w_{11}^* \ w_{12} \ \vdots \ w_{1d}\right] \in \mathbb{R}^d \quad \text{and} \quad \mathbf{Y} = \left[verifier\right] \tag{5.14}$$

5.2.1   The Square-Root Lasso

Chapter 6

Optimal Kullback-Leibler Aggregation in Mixture Density Estimation by Maximum Likelihood

6.1   Abstract

We study the maximum likelihood estimator of density of n independent observations, under the assumption that it is well approximated by a mixture with a large number of components. The main focus is on statistical properties with respect to the Kullback-Leibler loss. We establish risk bounds taking the form of sharp oracle inequalities both in deviation and in expectation. A simple consequence of these bounds is that the maximum likelihood estimator attains the optimal rate $((\log K)/n)^{1/2}$, up to a possible logarithmic correction, in the problem of convex aggregation when the number K of components is larger than $n^{1/2}$. More importantly, under the additional assumption that the Gram matrix of the components satisfies the compatibility condition, the obtained oracle inequalities yield the optimal rate in the sparsity scenario. That is, if the weight vector is (nearly) D-sparse, we get the rate $(D \log K)/n$. As a natural complement to our oracle inequalities, we introduce the notion of nearly-D-sparse aggregation and establish matching lower bounds for this type of aggregation.

6.2 Introduction

Assume that we observe n independent random vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \mathcal{X}$ drawn from a probability distribution P^* that admits a density function f^* with respect to some reference measure ν. The goal is to estimate the unknown density by a mixture density. More precisely, we assume that for a given family of mixture components $f_1, \ldots, f_K$, the unknown density of the observations f^* is well approximated by a convex combination $f_{\boldsymbol{\pi}}$ of these components, where

$$f_{\boldsymbol{\pi}}(\boldsymbol{x}) = \sum_{j=1}^{K} \pi_j f_j(\boldsymbol{x}), \quad \boldsymbol{\pi} \in \mathbb{B}_+^K = \Big\{ \boldsymbol{\pi} \in [0,1]^K : \sum_{j=1}^{K} \pi_j = 1 \Big\}. \qquad (6.1)$$

The assumption that the component densities $\mathcal{F} = \{f_j : j \in [K]\}$ are known essentially means that they are chosen from a dictionary obtained on the basis of previous experiments or expert knowledge.

We focus on the problem of estimation of the density function $f_{\boldsymbol{\pi}}$ and the weight vector $\boldsymbol{\pi}$ from the simplex $\mathbb{B}_+^K$ under the sparsity scenario: the ambient dimension K can be large, possibly larger than the sample size n, but most entries of $\boldsymbol{\pi}$ are either equal to zero or very small.

Our goal is to investigate the statistical properties of the Maximum Likelihood Estimator (MLE), defined by

$$\widehat{\boldsymbol{\pi}} \in \underset{\boldsymbol{\pi} \in \Pi}{\arg\min} \Big\{ -\frac{1}{n} \sum_{i=1}^{n} \log f_{\boldsymbol{\pi}}(\boldsymbol{X}_i) \Big\}, \qquad (6.2)$$

where the minimum is computed over a suitably chosen subset Π of $\mathbb{B}_+^K$. In the present work, we will consider sets $\Pi = \Pi_n(\mu)$, depending on a parameter $\mu > 0$ and the sample $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$, defined by

$$\Pi_n(\mu) = \Big\{ \boldsymbol{\pi} \in \mathbb{B}_+^K : \min_{i \in [n]} \sum_{j=1}^{K} \pi_j f_j(\boldsymbol{X}_i) \geq \mu \Big\}. \qquad (6.3)$$

Note that the objective function in (6.2) is convex and the same is true for set (6.3). Therefore, the MLE $\widehat{\boldsymbol{\pi}}$ can be efficiently computed even for large K by solving a problem of convex programming. To ease notation, very often, we will omit the dependence of $\Pi_n(\mu)$ on μ and write Π_n instead of $\Pi_n(\mu)$.

The quality of an estimator $\widehat{\boldsymbol{\pi}}$ can be measured in various ways. For instance, one can consider the Kullback-Leibler divergence

$$\mathrm{KL}(f^*\|f_{\widehat{\boldsymbol{\pi}}}) = \begin{cases} \int_{\mathcal{X}} f^*(\boldsymbol{x}) \log \frac{f^*(\boldsymbol{x})}{f_{\widehat{\boldsymbol{\pi}}}(\boldsymbol{x})} \nu(d\boldsymbol{x}), & \text{if } P^*\big(f^*(\boldsymbol{X}) = 0 \text{ and } f_{\widehat{\boldsymbol{\pi}}}(\boldsymbol{X}) > 0\big) = 0, \\ +\infty, & \text{otherwise}, \end{cases}$$
$$(6.4)$$

which has the advantage of bypassing identifiability issues. One can also consider the (well-specified) setting where $f^* = f_{\boldsymbol{\beta}^*}$ for some $\boldsymbol{\beta}^* \in \mathbb{B}_+^K$ and measure the quality of estimation through a distance between the vectors $\widehat{\boldsymbol{\pi}}$ and $\boldsymbol{\pi}^*$ (such as the ℓ_1-norm $\|\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}^*\|_1$ or the Euclidean norm $\|\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}^*\|_2$).

The main contributions of the present work are the following:

(a) We demonstrate that in the mixture model there is no need to introduce sparsity favoring penalty in order to get optimal rates of estimation under the Kullback-Leibler loss in the sparsity scenario. In fact, the constraint that the weight vector belongs to the simplex acts as a sparsity inducing penalty. As a consequence, there is no need to tune a parameter accounting for the magnitude of the penalty.

(b) We show that the maximum likelihood estimator of the mixture density simultaneously attains the optimal rate of aggregation for the Kullback-Leibler loss for at least three types of aggregation: model-selection, convex and D-sparse aggregation.

(c) We introduce a new type of aggregation, termed *nearly D-sparse aggregation* that extends and unifies the notions of convex and D-sparse aggregation. We establish strong lower bounds for the nearly D-sparse aggregation and demonstrate that the maximum likelihood estimator attains this lower bound up to logarithmic factors.

6.2.1 Related work

The results developed in the present work aim to gain a better understanding (a) of the statistical properties of the maximum likelihood estimator over a high-dimensional simplex and (b) of the problem of aggregation of density estimators under the Kullback-Leibler loss. Various procedures of aggregation[1] for density estimation have been studied in the literature with respect to different loss functions. [9, 38, 15] investigated different variants of the progressive mixture rules, also known as mirror averaging [39, 11], with respect to the Kullback-Leibler loss and established model selection type oracle inequalities[2] in expectation. Same type of guarantees, but holding with high probability, were recently obtained in [2, 8] for the procedure termed Q-aggregation, introduced in other contexts by [10, 30].

---

[1]We refer the interested reader to [34] for an up to date introduction into aggregation of statistical procedures.

[2]This means that they prove that the expected loss of the aggregate is almost as small as the loss of the best element of the dictionary $\{f_1, \ldots, f_K\}$.

Aggregation of estimators of a probability density function under the L_2-loss was considered in [28], where it was shown that a suitably chosen unbiased risk estimate minimizer is optimal both for convex and linear aggregation. The goal in the present work is to go beyond the settings of the aforementioned papers in that we want simultaneously to do as well as the best element of the dictionary, the best convex combination of the dictionary elements but also the best sparse convex combination. Note that the latter task was coined D-aggregation in [24] (see also [6]). In the present work, we rename it in D-sparse aggregation, in order to make explicit its relation to sparsity.

Key differences between the latter work and ours are that we do not assume the sparsity index to be known and we are analyzing an aggregation strategy that is computationally tractable even for large K. This is also the case of [7, 4], which are perhaps the most relevant references to the present work. These papers deal with the L_2-loss and investigate the lasso and the Dantzig estimators, respectively, suitably adapted to the problem of density estimation. Their methods handle dictionary elements $\{f_j\}$ which are not necessarily probability density functions, but has the drawback of requiring the choice of a tuning parameter. This choice is a nontrivial problem in practice. Instead, we show here that the optimal rates of sparse aggregation with respect to the Kullback-Leibler loss can be attained by procedure which is tuning parameter free.

Risk bounds for the maximum likelihood and other related estimators in the mixture model have a long history [23, 22, 26]. For the sake of comparison we recall here two elegant results providing non-asymptotic guarantees for the Kullback-Leibler loss.

Theorem 6.2.1 (Theorem 5.1 in [22]). *Let $\mathcal{F}$ be a finite dictionary of cardinality K of density functions such that $\max_{f\in\mathcal{F}}\|f^*/f\|_\infty \leq V$. Then, the maximum likelihood estimator over $\mathcal{F}$, $\widehat{f}_{\mathcal{F}}^{\mathrm{ML}} \in \arg\max_{f\in\mathcal{F}}\sum_{i=1}^n \log f(\mathbf{X}_i)$, satisfies the inequality*

$$\mathbf{E}_{f^*}\big[\mathrm{KL}\big(f^*||\widehat{f}_{\mathcal{F}}^{\mathrm{ML}}\big)\big] \leq \big(2 + \log V\big)\left(\min_{f\in\mathcal{F}}\mathrm{KL}(f^*||f) + \frac{2\log K}{n}\right). \qquad (6.5)$$

Inequality (6.5) is an inexact oracle inequality in expectation that quantifies the ability of $\widehat{f}_{\mathcal{F}}^{\mathrm{ML}}$ to solve the problem of model-selection aggregation. The adjective inexact refers to the fact that the "bias term" $\min_{f\in\mathcal{F}}\mathrm{KL}(f^*||f)$ is multiplied by factor strictly larger than one. It is noteworthy that the remainder term $\frac{2\log K}{n}$ corresponds to the optimal rate of model-selection aggregation [16, 33]. In relation with Theorem 6.2.1, it is worth mentioning a result of [38] and [9], see also Theorem 5 in [18] and Corollary 5.4 in [15],

establishing a risk bound similar to (6.5) without the extra factor $2 + \log V$ for the so called mirror averaging aggregate.

Theorem 6.2.2 (page 226 in [26]). *Let $\mathcal{F}$ be a finite dictionary of cardinality K of density functions and let $\mathcal{C}_k = \{ f_{\boldsymbol{\pi}} : \|\boldsymbol{\pi}\|_0 \leq k \}$ be the set of all the mixtures of at most k elements of $\mathcal{F}$ ($k \in [K]$). Assume that f^* and the densities f_k from $\mathcal{F}$ are bounded from below and above by some positive constants m and M, respectively. Then, there is a constant C depending only on m and M such that, for any tolerance level $\delta \in (0,1)$, the maximum likelihood estimator over $\mathcal{C}_k$, $\widehat{f}_{\mathcal{C}_k}^{\mathrm{ML}} \in \arg\max_{f \in \mathcal{C}_k} \sum_{i=1}^n \log f(\boldsymbol{X}_i)$, satisfies the inequality*

$$\mathrm{KL}\big(f^* \| \widehat{f}_{\mathcal{C}_k}^{\mathrm{ML}}\big) \leq \min_{f \in \mathcal{C}_k} \mathrm{KL}(f^* \| f) + C \Big(\frac{\log(K/\delta)}{n} \Big)^{1/2} \tag{6.6}$$

with probability at least $1 - \delta$.

This result is remarkably elegant and can be seen as an exact oracle inequality in deviation for D-sparse aggregation (for $D = k$). Furthermore, if we choose $k = K$ in Theorem 6.2.2, then we get an exact oracle inequality for convex aggregation with a rate-optimal remainder term [33]. However, it fails to provide the optimal rate for D-sparse aggregation.

Closing this section, we would like to mention the recent work [37], where oracle inequalities for estimators of low rank density matrices are obtained. They share a common feature with those obtained in this work: the adaptation to the unknown sparsity or rank is achieved without any additional penalty term. The constraint that the unknown parameter belongs to the simplex acts as a sparsity inducing penalty.

6.2.2 Additional notation

In what follows, for any $i \in [n]$, we denote by $\boldsymbol{Z}_i$ the vector $[f_1(\boldsymbol{X}_i), \ldots, f_K(\boldsymbol{X}_i)]^\top$ and by $\mathbf{Z}$ the $n \times K$ matrix $[\boldsymbol{Z}_1^\top, \ldots, \boldsymbol{Z}_n^\top]^\top$. We also define $\ell(u) = -\log u$, $u \in (0, +\infty)$, so that the MLE $\widehat{\boldsymbol{\pi}}$ is the minimizer of the function

$$L_n(\boldsymbol{\pi}) = \frac{1}{n} \sum_{i=1}^n \ell\big(\boldsymbol{Z}_i^\top \boldsymbol{\pi}\big). \tag{6.7}$$

For any set of indices $J \subseteq [K]$ and any $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)^\top \in \mathbb{R}^K$, we define $\boldsymbol{\pi}_J$ as the K-dimensional vector whose j-th coordinate equals π_j if $j \in J$ and 0 otherwise. We denote the cardinality of any $J \subseteq [K]$ by $|J|$. For any

set $J \subset \{1, \ldots, K\}$ and any constant $c \geq 0$, we introduce the compatibility constants [35] of a $K \times K$ positive semidefinite matrix $\mathbf{A}$,

$$\kappa_{\mathbf{A}}(J, c) = \inf \left\{ \frac{c^2 |J| \|\mathbf{A}^{1/2} \boldsymbol{v}\|_2^2}{(c \|\boldsymbol{v}_J\|_1 - \|\boldsymbol{v}_{J^c}\|_1)^2} : \boldsymbol{v} \in \mathbb{R}^K, \|\boldsymbol{v}_{J^c}\| < c \|\boldsymbol{v}_J\|_1 \right\}, \quad (6.8)$$

$$\bar{\kappa}_{\mathbf{A}}(J, c) = \inf \left\{ \frac{|J| \|\mathbf{A}^{1/2} \boldsymbol{v}\|_2^2}{\|\boldsymbol{v}_J\|_1^2} : \boldsymbol{v} \in \mathbb{R}^K, \|\boldsymbol{v}_{J^c}\|_1 < c \|\boldsymbol{v}_J\|_1 \right\}. \quad (6.9)$$

The risk bounds established in the present work involve the factors $\kappa_{\mathbf{A}}(J, 3)$ and $\bar{\kappa}_{\mathbf{A}}(J, 1)$. One can easily check that $\bar{\kappa}_{\mathbf{A}}(J, 3) \leq \kappa_{\mathbf{A}}(J, 3) \leq \frac{9}{4} \bar{\kappa}_{\mathbf{A}}(J, 1)$. We also recall that the compatibility constants of a matrix $\mathbf{A}$ are bounded from below by the smallest eigenvalue of $\mathbf{A}$.

Let us fix a function $f_0 : \mathcal{X} \to \mathbb{R}$ and denote $\bar{f}_k = f_k - f_0$ and $\bar{\boldsymbol{Z}}_i = [\bar{f}_1(\boldsymbol{X}_i), \ldots, \bar{f}_K(\boldsymbol{X}_i)]^\top$ for $i \in [n]$. In the results of this work, the compatibility factors are used for the empirical and population Gram matrices of vectors $\bar{\boldsymbol{Z}}_k$, that is when $\mathbf{A} = \widehat{\boldsymbol{\Sigma}}_n$ and $\mathbf{A} = \boldsymbol{\Sigma}$ with

$$\widehat{\boldsymbol{\Sigma}}_n = \frac{1}{n} \sum_{i=1}^n \bar{\boldsymbol{Z}}_i \bar{\boldsymbol{Z}}_i^\top, \qquad \boldsymbol{\Sigma} = \mathbf{E}[\bar{\boldsymbol{Z}}_1 \bar{\boldsymbol{Z}}_1^\top]. \quad (6.10)$$

The general entries of these matrices are respectively $(\widehat{\boldsymbol{\Sigma}}_n)_{k,l} = \frac{1}{n} \sum_{i=1}^n \bar{f}_k(\boldsymbol{X}_i) \bar{f}_l(\boldsymbol{X}_i)$ and $(\boldsymbol{\Sigma})_{k,l} = \mathbf{E}[\bar{f}_k(\boldsymbol{X}_1) \bar{f}_l(\boldsymbol{X}_1)]$.

We assume that there exist positive constants m and M such that for all densities f_k with $k \in [K]$, we have

$$\forall x \in \mathcal{X}, \quad m \leq f_k(x) \leq M. \quad (6.11)$$

We use the notation $V = M/m$. It is worth mentioning that the set of dictionaries satisfying simultaneously this boundedness assumption and the aforementioned compatibility condition is not empty. For instance, one can consider the functions $f_k(x) = 1 + \frac{1}{2} \sin(2\pi k x)$ for $k \in [K]$. These functions are probability densities w.r.t. the Lebesgue measure on $\mathcal{X} = [0, 1]$. They are bounded from below and from above by $\frac{1}{2}$ and $\frac{3}{2}$, respectively. Taking $f_0(x) = 1$, the corresponding Gram matrix is $\boldsymbol{\Sigma} = \frac{1}{8} \mathbf{I}_K$, which has all eigenvalues equal to $\frac{1}{8}$.

6.2.3 Agenda

The rest of the paper is organized as follows. In Section 6.3, we state our main theoretical contributions and discuss their consequences. Possible relaxations of the conditions, as well as lower bounds showing the tightness

of the established risk bounds, are considered in Section 6.4. A brief summary of the paper and some future directions of research are presented in Section 6.5. The proofs of all theoretical results are postponed to Section 6.6 and Section 6.7.

6.3 Oracle inequalities in deviation and in expectation

In this work, we prove several non-asymptotic risk bounds that imply, in particular, that the maximum likelihood estimator is optimal in model-selection aggregation, convex aggregation and D-sparse aggregation (up to log-factors). In all the results of this section we assume the parameter μ in (6.3) to be equal to 0.

Theorem 6.3.1. *Let $\mathcal{F}$ be a set of $K \geq 4$ densities satisfying the boundedness condition (6.11). Denote by $f_{\widehat{\boldsymbol{\pi}}}$ the mixture density corresponding to the maximum likelihood estimator $\widehat{\boldsymbol{\pi}}$ over Π_n defined in (6.7). There are constants $c_1 \leq 32V^3$, $c_2 \leq 288M^2V^6$ and $c_3 \leq 128M^2V^6$ such that, for any $\delta \in (0, 1/2)$, the following inequalities hold*

$$\mathrm{KL}(f^*||f_{\widehat{\boldsymbol{\pi}}}) \leq \inf_{\substack{J \subset [K] \\ \boldsymbol{\pi} \in \mathbb{B}_+^K}} \left\{ \mathrm{KL}(f^*||f_{\boldsymbol{\pi}}) + c_1 \Big(\frac{\log(K/\delta)}{n} \Big)^{1/2} \|\boldsymbol{\pi}_{J^c}\|_1 + \frac{c_2|J|\log(K/\delta)}{n\kappa_{\widehat{\boldsymbol{\Sigma}}_n}(J, 3)} \right\},$$

(6.12)

$$\mathrm{KL}(f^*||f_{\widehat{\boldsymbol{\pi}}}) \leq \inf_{J \subset [K]} \inf_{\substack{\boldsymbol{\pi} \in \mathbb{B}_+^K \\ \boldsymbol{\pi}_{J^c}=0}} \left\{ \mathrm{KL}(f^*||f_{\boldsymbol{\pi}}) + \frac{c_3|J|\log(K/\delta)}{n\bar{\kappa}_{\widehat{\boldsymbol{\Sigma}}_n}(J, 1)} \right\}$$

(6.13)

with probability at least $1 - \delta$.

The proof of this and the subsequent results stated in this section are postponed to Section 6.6. Comparing the two inequalities of the above theorem, one can notice two differences. First, the term proportional to $\|\boldsymbol{\pi}_{J^c}\|_1$ is absent in the second risk bound, which means that the risk of the MLE is compared to that of the best mixture with a weight sequences supported by J. Hence, this risk bound is weaker than the first one provided by (6.12). Second, the compatibility factor $\bar{\kappa}_{\widehat{\boldsymbol{\Sigma}}_n}(J, 1)$ in (6.13) is larger that its counterpart $\kappa_{\widehat{\boldsymbol{\Sigma}}_n}(J, 3)$ in (6.12). This entails that in the cases where the oracle is expected to be sparse, the remainder term of the bound in (6.12) is slightly looser than that of (6.13).

A first and simple consequence of Theorem 6.2.1 is obtained by taking $J = \varnothing$ in the right hand side of the first inequality. Then, $\|\boldsymbol{\pi}_{J^c}\|_1 = \|\boldsymbol{\pi}\|_1 = 1$ and we get

$$\mathrm{KL}(f^*||f_{\widehat{\boldsymbol{\pi}}}) \leq \inf_{\boldsymbol{\pi} \in \mathbb{B}_+^K} \mathrm{KL}(f^*||f_{\boldsymbol{\pi}}) + c_1 \Big(\frac{\log(K/\delta)}{n}\Big)^{1/2}. \qquad (6.14)$$

This implies that for every dictionary $\mathcal{F}$, without any assumption on the smallness of the coherence between its elements, the maximum likelihood estimator achieves the optimal rate of convex aggregation, up to a possible[3] logarithmic correction, in the high-dimensional regime $K \geq n^{1/2}$. In the case of regression with random design, an analogous result has been proved by [20] and [19]. One can also remark that the upper bound in (6.14) is of the same form as the one of Theorem 6.2.2 stated in section 6.2.1 above.

The main compelling feature of our results is that they show that the MLE adaptively achieves the optimal rate of aggregation not only in the case of convex aggregation, but also for the model-selection aggregation and D-(convex) aggregation. For handling these two cases, it is more convenient to get rid of the presence of the compatibility factor of the empirical Gram matrix $\widehat{\boldsymbol{\Sigma}}_n$. The latter can be replaced by the compatibility factor of the population Gram matrix, as stated in the next result.

Theorem 6.3.2. *Let $\mathcal{F}$ be a set of K densities satisfying the boundedness condition (6.11). Denote by $f_{\widehat{\boldsymbol{\pi}}}$ the mixture density corresponding to the maximum likelihood estimator $\widehat{\boldsymbol{\pi}}$ over Π_n defined in (6.7). There are constants $c_4 \leq 32V^3 + 4$, $c_5 \leq 4.5M^2(8V^3 + 1)^2$ and $c_6 \leq 2M^2(8V^3 + 1)^2$ such that, for any $\delta \in (0, 1/2)$, the following inequalities hold*

$$\mathrm{KL}(f^*||f_{\widehat{\boldsymbol{\pi}}}) \leq \inf_{\substack{J \subset [K] \\ \boldsymbol{\pi} \in \mathbb{B}_+^K}} \Bigg\{\mathrm{KL}(f^*||f_{\boldsymbol{\pi}}) + c_4 \Big(\frac{\log(K/\delta)}{n}\Big)^{1/2}\|\boldsymbol{\pi}_{J^c}\|_1 + \frac{c_5|J|\log(K/\delta)}{n\kappa_{\boldsymbol{\Sigma}}(J,3)}\Bigg\},$$

$$(6.15)$$

$$\mathrm{KL}(f^*||f_{\widehat{\boldsymbol{\pi}}}) \leq \inf_{J \subset [K]} \inf_{\substack{\boldsymbol{\pi} \in \mathbb{B}_+^K \\ \boldsymbol{\pi}_{J^c}=0}} \Bigg\{\mathrm{KL}(f^*||f_{\boldsymbol{\pi}}) + \frac{c_6|J|\log(K/\delta)}{n\bar{\kappa}_{\boldsymbol{\Sigma}}(J,1)}\Bigg\} \qquad (6.16)$$

with probability at least $1 - 2\delta$.

The main advantage of the upper bounds provided by Theorem 6.3.2 as compared with those of Theorem 6.3.1 is that the former is deterministic,

---

[3]In fact, the optimal rate of convex aggregation when $K \geq n^{1/2}$ is of order $\big(\log(K/n^{1/2})/n\big)^{1/2}$. Therefore, even the $\log K$ term is optimal whenever $K \geq Cn^{1/2+\alpha}$ for some $\alpha > 0$.

whereas the latter involves the compatibility factor of the empirical Gram matrix which is random. The price to pay for getting rid of randomness in the risk bound is the increased values of the constants c_4, c_5 and c_6. Note, however, that this price is not too high, since obviously $1 \leq M \leq L$ and, therefore, $c_4 \leq 1.25c_1$, $c_5 \leq 1.56c_2$ and $c_6 \leq 1.56c_3$. In addition, the absence of randomness in the risk bound allows us to integrate it and to convert the bound in deviation into a bound in expectation.

Theorem 6.3.3 (Bound in Expectation). *Let $\mathcal{F}$ be a set of K densities satisfying the boundedness condition (6.11). Denote by $f_{\widehat{\boldsymbol{\pi}}}$ the mixture density corresponding to the maximum likelihood estimator $\widehat{\boldsymbol{\pi}}$ over Π_n defined in (6.7). There are constants $c_7 \leq 20V^3 + 8$, $c_8 \leq M^2(22V^3 + 3)^2$ and $c_9 \leq M^2(15V^3 + 2)^2$ such that*

$$\mathbf{E}[\mathrm{KL}(f^*||f_{\widehat{\boldsymbol{\pi}}})] \leq \inf_{\substack{J \subset [K] \\ \boldsymbol{\pi} \in \mathbb{B}_+^K}} \left\{ \mathrm{KL}(f^*||f_{\boldsymbol{\pi}}) + c_7\Big(\frac{\log K}{n}\Big)^{1/2}\|\boldsymbol{\pi}_{J^c}\|_1 + \frac{c_8|J|\log K}{n\kappa_{\boldsymbol{\Sigma}}(J,3)} \right\},$$
(6.17)

$$\mathbf{E}[\mathrm{KL}(f^*||f_{\widehat{\boldsymbol{\pi}}})] \leq \inf_{J \subset [K]} \inf_{\substack{\boldsymbol{\pi} \in \mathbb{B}_+^K \\ \boldsymbol{\pi}_{J^c}=0}} \left\{ \mathrm{KL}(f^*||f_{\boldsymbol{\pi}}) + \frac{c_9|J|\log K}{n\bar{\kappa}_{\boldsymbol{\Sigma}}(J,1)} \right\}.$$
(6.18)

In inequality (6.18), upper bounding the infimum over all sets J by the infimum over the singletons, we get

$$\mathbf{E}[\mathrm{KL}(f^*||f_{\widehat{\boldsymbol{\pi}}})] \leq \inf_{j \in [K]} \left\{ \mathrm{KL}(f^*||f_j) + \frac{c_9 \log K}{n\bar{\kappa}_{\boldsymbol{\Sigma}}(J,1)} \right\}.$$
(6.19)

This implies that the maximum likelihood estimator $f_{\widehat{\boldsymbol{\pi}}}$ achieves the rate $\frac{\log K}{n}$ in model-selection type aggregation. This rate is known to be optimal in the model of regression [30]. If we compare this result with Theorem 6.2.1 stated in Section 6.2.1, we see that the remainder terms of these two oracle inequalities are of the same order (provided that the compatibility factor is bounded away from zero), but inequality (6.19) has the advantage of being exact.

We can also apply (6.18) to the problem of convex aggregation with small dictionary, that is for K smaller than $n^{1/2}$. Upper bounding $|J|$ by $|K|$, we get

$$\mathbf{E}[\mathrm{KL}(f^*||f_{\widehat{\boldsymbol{\pi}}})] \leq \inf_{\boldsymbol{\pi} \in \mathbb{B}_+^K} \mathrm{KL}(f^*||f_{\boldsymbol{\pi}}) + \frac{c_9 K \log K}{n\bar{\kappa}_{\boldsymbol{\Sigma}}([K],1)}.$$
(6.20)

Assuming, for instance, the smallest eigenvalue of $\boldsymbol{\Sigma}$ bounded away from zero (which is a quite reasonable assumption in the context of low dimensionality),

the above upper bound provides a rate of convex aggregation of the order of $\frac{K \log K}{n}$. Up to a logarithmic term, this rate is known to be optimal for convex aggregation in the model of regression.

Finally, considering all the sets J of cardinal smaller than D (with $D \leq K$) and setting $\bar{\kappa}_{\mathbf{\Sigma}}(D, 1) = \inf_{J:|J| \leq D} \bar{\kappa}_{\mathbf{\Sigma}}(J, 1)$, we deduce from (6.18) that

$$\mathbf{E}[\mathrm{KL}(f^* || f_{\hat{\pi}})] \leq \inf_{\pi \in \mathbb{B}_+^K : \|\pi\|_0 \leq D} \mathrm{KL}(f^* || f_{\pi}) + \frac{c_9 D \log K}{n \bar{\kappa}_{\mathbf{\Sigma}}(D, 1)}. \tag{6.21}$$

According to [31, Theorem 5.3], in the regression model, the optimal rate of D-sparse aggregation is of order $(D/n) \log(K/D)$, whenever $D = o(n^{1/2})$. Inequality (6.21) shows that the maximum likelihood estimator over the simplex achieves this rate up to a logarithmic factor. Furthermore, this logarithmic inflation disappears when the sparsity D is such that, asymptotically, the ratio $\frac{\log D}{\log K}$ is bounded from above by a constant $\alpha < 1$. Indeed, in such a situation the optimal rate $\frac{D \log(K/D)}{n} = \frac{D \log K}{n}(1 - \frac{\log D}{\log K})$ is of the same order as the remainder term in (6.21), that is $\frac{D \log K}{n}$.

6.4 Discussion of the conditions and possible extensions

In this section, we start by announcing lower bounds for the Kullback-Leibler aggregation in the problem of density estimation. Then we discuss the implication of the risk bounds of the previous section to the case where the target is the weight vector π rather than the mixture density f_{π}. Finally, we present some extensions to the case where the boundedness assumption is violated.

6.4.1 Lower bounds for nearly-D-sparse aggregation

As mentioned in previous section, the literature is replete with lower bounds on the minimax risk for various types of aggregation. However most of them concern the regression setting either with random or with deterministic design. Lower bounds of aggregation for density estimation were first established by [29] for the L_2-loss. In the case of Kullback-Leibler aggregation in density estimation, the only lower bounds we are aware are those established by [18] for model-selection type aggregation. It is worth emphasizing here that the results of the aforementioned two papers provide weak lower bounds. Indeed, they establish the existence of a dictionary for which the minimax excess risk is lower bounded by the suitable quantity. In contrast with this, we

establish here strong lower bounds that hold for every dictionary satisfying the boundedness and the compatibility conditions.

Let $\mathcal{F} = \{f_1, \ldots, f_K\}$ be a dictionary of density functions on $\mathcal{X} = [0, 1]$. We say that the dictionary $\mathcal{F}$ satisfies the boundedness and the compatibility assumptions if for some positive constants m, M and κ, we have $m \leq f_j(x) \leq M$ for all $j \in [K]$, $x \in \mathcal{X}$. In addition, we assume in this subsection that all the eigenvalues of the Gram matrix Σ belong to the interval $[\varkappa_*, \varkappa^*]$, with $\varkappa_* > 0$ and $\varkappa^* < \infty$.

For every $\gamma \in (0, 1)$ and any $D \in [K]$, we define the set of nearly-D-sparse convex combinations of the dictionary elements $f_j \in \mathcal{F}$ by

$$\mathcal{H}_{\mathcal{F}}(\gamma, D) = \left\{ f_{\boldsymbol{\pi}} : \boldsymbol{\pi} \in \mathbb{B}_+^K \text{ such that } \exists\, J \subset [K] \text{ with } \|\boldsymbol{\pi}_{J^c}\|_1 \leq \gamma \text{ and } |J| \leq D \right\}. \tag{6.22}$$

In simple words, $f_{\boldsymbol{\pi}}$ belongs to $\mathcal{H}_{\mathcal{F}}(\gamma, D)$ if it admits a γ-approximately D-sparse representation in the dictionary $\mathcal{F}$. We are interested in bounding from below the minimax excess risk

$$\mathcal{R}\big(\mathcal{H}_{\mathcal{F}}(\gamma, D)\big) = \inf_{\widehat{f}} \sup_{f^*} \left\{ \mathbf{E}[\mathrm{KL}(f^* \| \widehat{f})] - \inf_{f_{\boldsymbol{\pi}} \in \mathcal{H}_{\mathcal{F}}(\gamma, D)} \mathrm{KL}(f^* \| f_{\boldsymbol{\pi}}) \right\}, \tag{6.23}$$

where the inf is over all possible estimators of f^* and the sup is over all density functions over $[0, 1]$. Note that the estimator $\widehat{f}$ is not necessarily a convex combination of the dictionary elements. Furthermore, it is allowed to depend on the parameters γ and D characterizing the class $\mathcal{H}_{\mathcal{F}}(\gamma, D)$. It follows from (6.17), that if the dictionary satisfies the boundedness and the compatibility condition, then

$$\mathcal{R}\big(\mathcal{H}_{\mathcal{F}}(\gamma, D)\big) \leq C\left\{ \left(\frac{\gamma^2 \log K}{n}\right)^{1/2} + \frac{D \log K}{n} \right\} \bigwedge \left(\frac{\log K}{n}\right)^{1/2}, \tag{6.24}$$

for some constant C depending only on m, M and $\varkappa_*$. Note that the last term accounts for the following phenomenon: If the sparsity index D is larger than a multiple of $\sqrt{n}$, then the sparsity bears no advantage as compared to the ℓ_1 constraint. The next result implies that this upper bound is optimal, at least up to logarithmic factors.

Theorem 6.4.1. *Assume that $\log(1 + eK) \leq n$. Let $\gamma \in (0, 1)$ and $D \in [K]$ be fixed. There exists a constant A depending only on m, M, $\varkappa_*$ and $\varkappa^*$ such that*

$$\mathcal{R}(\mathcal{H}_{\mathcal{F}}(\gamma, D)) \geq A\left\{ \left[\frac{\gamma^2}{n} \log\left(1 + \frac{K}{\gamma\sqrt{n}}\right)\right]^{1/2} + \frac{D \log(1 + K/D)}{n} \right\} \bigwedge \left[\frac{1}{n} \log\left(1 + \frac{K}{\sqrt{n}}\right)\right]^{1/2}. \tag{6.25}$$

This is the first result providing lower bounds on the minimax risk of aggregation over nearly-D-sparse aggregates. To the best of our knowledge, even in the Gaussian sequence model, such a result has not been established to date. It has the advantage of unifying the results on convex and D-sparse aggregation, as well as extending them to a more general class. Let us also stress that the condition $\log(1 + eK) \leq n$ is natural and unavoidable, since it ensures that the right hand side of (6.24) is smaller than the trivial bound $\log V$.

6.4.2 Weight vector estimation

The risk bounds carried out in the previous section for the problem of density estimation in the Kullback-Leibler loss imply risk bounds for the problem of weight vector estimation. Indeed, under the boundedness assumption (6.11), the Kullback-Leibler divergence between two mixture densities can be shown to be equivalent to the squared Mahalanobis distance between the weight vectors of these mixtures with respect to the Gram matrix. In order to go from the Mahalanobis distance to the Euclidean one, we make use of the restricted eigenvalue

$$\kappa_{\mathbf{\Sigma}}^{\mathrm{RE}}(s, c) = \inf \big\{ \|\mathbf{\Sigma}^{1/2}\boldsymbol{v}\|_2^2 \, : \, \exists\, J \subset [K] \text{ s.t. } |J| \leq s, \, \|\boldsymbol{v}_{J^c}\|_1 \leq c\|\boldsymbol{v}_J\|_1 \text{ and } \|\boldsymbol{v}_J\|_2 = 1 \big\}.$$
(6.26)

This strategy leads to the next result.

Proposition 1. *Let $\mathcal{F}$ be a set of $K \geq 4$ densities satisfying condition (6.11). Denote by $f_{\widehat{\boldsymbol{\pi}}}$ the mixture density corresponding to the maximum likelihood estimator $\widehat{\boldsymbol{\pi}}$ over Π_n defined in (6.7). Let $\boldsymbol{\pi}^*$ the weight-vector of the best mixture density: $\boldsymbol{\pi}^* \in \arg\min_{\boldsymbol{\pi}} \mathrm{KL}(f^*\|f_{\boldsymbol{\pi}})$, and let J^* be the support of $\boldsymbol{\pi}^*$. There are constants $c_{10} \leq M^2(64V^3 + 8)$ and $c_{11} \leq 4M^2(8V^3 + 1)$ such that, for any $\delta \in (0, 1/2)$, the following inequalities hold*

$$\|\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}^*\|_1 \leq \frac{c_{10}|J^*|}{\bar{\kappa}_{\mathbf{\Sigma}}(J^*, 1)} \left(\frac{\log(K/\delta)}{n} \right)^{1/2},$$
(6.27)

$$\|\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}^*\|_2 \leq \frac{c_{11}}{\kappa_{\mathbf{\Sigma}}^{\mathrm{RE}}(|J^*|, 1)} \left(\frac{2|J^*| \log(K/\delta)}{n} \right)^{1/2},$$
(6.28)

$$\|\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}^*\|_2^2 \leq \frac{c_{11}}{\kappa_{\mathbf{\Sigma}}^{\mathrm{RE}}(|J^*|, 1)} \left(\frac{2 \log(K/\delta)}{n} \right)^{1/2}$$
(6.29)

with probability at least $1 - 2\delta$.

In simple words, this result tells us that the wight estimator $\widehat{\boldsymbol{\pi}}$ attains the minimax rate of estimation $|J^*|(\frac{\log(K)}{n})^{1/2}$ over the intersection of the ℓ_1

and ℓ_0 balls, when the error is measured by the ℓ_1-norm, provided that the compatibility factor of the dictionary $\mathcal{F}$ is bounded away from zero. The optimality of this rate—up to logarithmic factors—follows from the fact that the error of estimation of each nonzero coefficients of $\boldsymbol{\pi}^*$ is at least $cn^{-1/2}$ (for some $c > 0$), leading to a sum of the absolute values of the errors at least of the order $|J^*|n^{-1/2}$. The logarithmic inflation of the rate is the price to pay for not knowing the support J^*. It is clear that this reasoning is valid only when the sparsity $|J^*|$ is of smaller order than $n^{1/2}$. Indeed, in the case $|J^*| \geq cn^{1/2}$, the trivial bound $\|\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}^*\|_1 \leq 2$ is tighter than the one in (6.27).

Concerning the risk measured by the Euclidean norm, we underline that there are two regimes characterized by the order between upper bounds in (6.28) and (6.29). Roughly speaking, when the signal is highly sparse in the sense that $|J^*|$ is smaller than $(n/\log K)^{1/2}$, then the smallest bound is given by (6.28) and is of the order $\frac{|J^*|\log(K)}{n}$. This rate is can be compared to the rate $\frac{|J^*|\log(K/|J^*|)}{n}$, known to be optimal in the Gaussian sequence model. In the second regime corresponding to mild sparsity, $|J^*| > (n/\log K)^{1/2}$, the smallest bound is the one in (6.29). The latter is of order $(\frac{\log(K)}{n})^{1/2}$, which is known to be optimal in the Gaussian sequence model. For various results providing lower bounds in regression framework we refer the interested reader to [27, 31, 36].

6.4.3 Extensions to the case of vanishing components

In the previous sections we have deliberately avoided any discussion of the role of the parameter μ, present in the search space $\Pi_n(\mu)$ of the problem (6.2)-(6.3). In fact, when all the dictionary elements are separated from zero by a constant m, a condition assumed throughout previous sections, choosing any value of $\mu \leq m$ is equivalent to choosing $\mu = 0$. Therefore, the choice of this parameter does not impact the quality of estimation. However, this parameter might have strong influence in practice both on statistical and computational complexity of the maximum likelihood estimator. A first step in understanding the influence of μ on the statistical complexity is made in the next paragraphs.

Let us consider the case where the condition $\min_x \min_j f_j(x) \geq m > 0$ fails, but the upper-boundedness condition $\max_x \max_j f_j(x) \leq M$ holds true. In such a situation, we replace the definition $V = M/m$ by $V = M/\mu$. We also define the set $\Pi^*(\mu) = \{\boldsymbol{\pi} \in \mathbb{B}_+^K : P^*(f_{\boldsymbol{\pi}}(\boldsymbol{X}) \geq \mu) = 1\}$. In order to keep mathematical formulae simple, we will only state the equivalent of (6.13) in the case of $m = 0$. All the other results of the previous section can be extended in a similar way.

Proposition 2. *Let $\mathcal{F}$ be a set of $K \geq 2$ densities satisfying the boundedness condition $\sup_{\boldsymbol{x} \in \mathcal{X}} f_j(\boldsymbol{x}) \leq M$. Denote by $f_{\widehat{\boldsymbol{\pi}}}$ the mixture density corresponding to the maximum likelihood estimator $\widehat{\boldsymbol{\pi}}$ over $\Pi_n(\mu)$ defined in (6.7). There is a constant $\bar{c} \leq 128M^2 V^4$ such that, for any $\delta \in (0, 1/2)$,*

$$\mathrm{KL}(f^* \| f_{\widehat{\boldsymbol{\pi}}}) \leq \inf_{J \subset [K]} \inf_{\substack{\boldsymbol{\pi} \in \Pi^*(\mu) \\ \boldsymbol{\pi}_{J^c} = 0}} \left\{ \mathrm{KL}(f^* \| f_{\boldsymbol{\pi}}) + \frac{\bar{c}|J| \log(K/\delta)}{n \bar{\kappa}_{\widehat{\boldsymbol{\Sigma}}_n}(J, 1)} \right\} + \int_{\mathcal{X}} (\log \mu - \log f_{\widehat{\boldsymbol{\pi}}})_+ f^* d\nu$$

$$(6.30)$$

on an event of probability at least $1 - \delta$. Furthermore, if $\inf_{\boldsymbol{x} \in \mathcal{X}} f^*(\boldsymbol{x}) \geq \mu$, then, on the same event, we have

$$\|f^* - f_{\widehat{\boldsymbol{\pi}}}\|^2_{L^2(P^*)} \leq 2M^2 \inf_{J \subset [K]} \inf_{\substack{\boldsymbol{\pi} \in \Pi^*(\mu) \\ \boldsymbol{\pi}_{J^c} = 0}} \left\{ \mathrm{KL}(f^* \| f_{\boldsymbol{\pi}}) + \frac{\bar{c}|J| \log(K/\delta)}{n \bar{\kappa}_{\widehat{\boldsymbol{\Sigma}}_n}(J, 1)} \right\}. \quad (6.31)$$

The last term present in the first upper bound, $\int_{\mathcal{X}} (\log \mu - \log f_{\widehat{\boldsymbol{\pi}}})_+ f^* d\nu$ is the price we pay for considering densities that are not lower bounded by a given constant. A simple, non-random upper bound on this term is $\int_{\mathcal{X}} \max_{k \in [K]} (\log \mu - \log f_k)_+ f^* d\nu$. Providing a tight upper bound on this kind or remainder terms is an important problem which lies beyond the scope of the present work.

6.5 Conclusion

In this paper, we have established exact oracle inequalities for the maximum likelihood estimator of a mixture density. This oracle inequality clearly highlights the interplay of three sources of error: misspecification of the model of mixture, departure from D-sparsity and stochastic error of estimating D nonzero coefficients. We have also proved a lower bound that show that the remainder terms of our upper bounds are optimal, up to logarithmic terms. This lower bound is valid not only for the maximum likelihood estimator, but for any estimator of the density function. As a consequence, the maximum likelihood estimator has a nearly optimal excess risk in the minimax sense.

In all the results of the present paper, we have assumed that the components of the mixture model are deterministic. From a practical point of view, it might be reasonable to choose these components in a data driven way, using, for instance, a hold-out sample. This question, as well as the problem of tuning the parameter μ, constitute interesting and challenging avenues for future research.

6.6 Proofs of results stated in previous sections

This section collects the proofs of the theorems and claims stated in previous sections.

6.6.1 Proof of Theorem 6.3.1

The main technical ingredients of the proof are a strong convexity argument and a control of the maximum of an empirical process. The corresponding results are stated in Lemma 6.6.2 and Proposition 6.6.1, respectively, deferred to Section 6.6.6. We denote by $\bar{\mathbf{Z}}$ the $n \times K$ matrix $[\bar{\boldsymbol{Z}}_1, \ldots, \bar{\boldsymbol{Z}}_K]$.

Since $\widehat{\boldsymbol{\pi}}$ is a minimizer of $L_n(\cdot)$, see (6.2) and (6.7), we know that $L_n(\widehat{\boldsymbol{\pi}}) \leq L_n(\boldsymbol{\pi})$ for every π. However, this inequality can be made sharper using the (local) strong convexity of the function $\ell(u) = -\log(u)$. Indeed, Lemma 6.6.2 below shows that

$$\frac{1}{n}\sum_{i=1}^{n}\ell(f_{\widehat{\boldsymbol{\pi}}}(\boldsymbol{X}_i)) \leq \frac{1}{n}\sum_{i=1}^{n}\ell(f_{\boldsymbol{\pi}}(\boldsymbol{X}_i)) - \frac{1}{2M^2n}\|\bar{\mathbf{Z}}(\widehat{\boldsymbol{\pi}}-\boldsymbol{\pi})\|_2^2. \tag{6.32}$$

On the other hand, if we set $\varphi(\pi, \boldsymbol{x}) = \int(\log f_{\boldsymbol{\pi}})f^*d\nu - \log f_{\boldsymbol{\pi}}(\boldsymbol{x})$, we have $\mathbf{E}_{f^*}[\varphi(\boldsymbol{\pi}, \boldsymbol{X}_i)] = 0$ and

$$\ell(f_{\boldsymbol{\pi}}(\boldsymbol{X}_i)) = \mathrm{KL}(f^*||f_{\boldsymbol{\pi}}) - \int_{\mathcal{X}} f^* \log f^* d\nu + \varphi(\pi, \boldsymbol{X}_i). \tag{6.33}$$

Combining inequalities (6.32) and (6.33), we get

$$\mathrm{KL}(f^*||f_{\widehat{\boldsymbol{\pi}}}) \leq \mathrm{KL}(f^*||f_{\boldsymbol{\pi}}) - \frac{1}{2M^2n}\|\bar{\mathbf{Z}}(\widehat{\boldsymbol{\pi}}-\boldsymbol{\pi})\|_2^2 + \frac{1}{n}\sum_{i=1}^{n}\big(\varphi(\boldsymbol{\pi}, \boldsymbol{X}_i) - \varphi(\widehat{\boldsymbol{\pi}}, \boldsymbol{X}_i)\big). \tag{6.34}$$

The next step of the proof consists in establishing a suitable upper bound on the noise term $\Phi_n(\boldsymbol{\pi}) - \Phi_n(\widehat{\boldsymbol{\pi}})$ where

$$\Phi_n(\boldsymbol{\pi}) = \frac{1}{n}\sum_{i=1}^{n}\varphi(\boldsymbol{\pi}, \boldsymbol{X}_i). \tag{6.35}$$

According to the mean value theorem, setting $\zeta_n := \sup_{\bar{\boldsymbol{\pi}}\in\boldsymbol{\Pi}_n}\big\|\nabla\Phi_n(\bar{\boldsymbol{\pi}})\big\|_\infty$, for every vector $\boldsymbol{\pi} \in \boldsymbol{\Pi}_n$, it holds that

$$|\Phi_n(\widehat{\boldsymbol{\pi}}) - \Phi_n(\boldsymbol{\pi})| \leq \sup_{\bar{\boldsymbol{\pi}}\in\boldsymbol{\Pi}_n}\big\|\nabla\Phi_n(\bar{\boldsymbol{\pi}})\big\|_\infty\|\widehat{\boldsymbol{\pi}}-\boldsymbol{\pi}\|_1 = \zeta_n\|\widehat{\boldsymbol{\pi}}-\boldsymbol{\pi}\|_1. \tag{6.36}$$

This inequality, combined with (6.34), yields

$$\mathrm{KL}(f^*||f_{\widehat{\boldsymbol{\pi}}}) \le \mathrm{KL}(f^*||f_{\boldsymbol{\pi}}) - \frac{1}{2M^2 n}\|\bar{\mathbf{Z}}(\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi})\|_2^2 + \zeta_n\|\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_1. \quad (6.37)$$

Using the Gram matrix $\widehat{\boldsymbol{\Sigma}}_n = {}^1\!/\!_n \bar{\mathbf{Z}}^\top \bar{\mathbf{Z}}$, the quantity $\|\bar{\mathbf{Z}}(\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi})\|_2$ can be rewritten as

$$\|\bar{\mathbf{Z}}(\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi})\|_2^2 = n\|\widehat{\boldsymbol{\Sigma}}_n^{1/2}(\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi})\|_2^2. \quad (6.38)$$

We proceed with applying the following result [3, Lemma 2].

Lemma 6.6.1 ([3], Lemma 2). *For any pair of vectors $\boldsymbol{\pi}, \boldsymbol{\pi}' \in \mathbb{R}^K$, for any pair of scalars $\mu > 0$ and $\gamma > 1$, for any $K \times K$ symmetric matrix $\mathbf{A}$ and for any set $J \subset [p]$, the following inequality is true*

$$2\mu\gamma^{-1}(\|\boldsymbol{\pi}-\widehat{\boldsymbol{\pi}}\|_1 + \gamma\|\boldsymbol{\pi}\|_1 - \gamma\|\widehat{\boldsymbol{\pi}}\|_1) - \|\mathbf{A}(\boldsymbol{\pi}-\widehat{\boldsymbol{\pi}})\|_2^2 \le 4\mu\|\boldsymbol{\pi}_{J^c}\|_1 + \frac{(\gamma+1)^2\mu^2|J|}{\gamma^2\kappa_{\mathbf{A}^2}(J, c_\gamma)},$$
$$(6.39)$$

where $c_\gamma = (\gamma + 1)/(\gamma - 1)$.

Choosing $\mathbf{A} = \widehat{\boldsymbol{\Sigma}}_n^{1/2}/(\sqrt{2}\,M)$, $\mu = \zeta_n$ and $\gamma = 2$ (thus $c_\gamma = 3$) we get the inequality

$$\zeta_n\|\boldsymbol{\pi} - \widehat{\boldsymbol{\pi}}\|_1 - \|\mathbf{A}(\boldsymbol{\pi} - \widehat{\boldsymbol{\pi}})\|_2^2 \le 4\zeta_n\|\boldsymbol{\pi}_{J^c}\|_1 + \frac{9\zeta_n^2|J|}{4\kappa_{\mathbf{A}^2}(J, 3)}, \qquad \forall J \in \{1, \ldots, p\}.$$
$$(6.40)$$

One can check that $\kappa_{\mathbf{A}^2}(J, 3) = \kappa_{\widehat{\boldsymbol{\Sigma}}_n}(J, 3)/(2M^2)$. Combining the last inequality with (6.37), we arrive at

$$\mathrm{KL}(f^*||f_{\widehat{\boldsymbol{\pi}}}) \le \mathrm{KL}(f^*||f_{\boldsymbol{\pi}}) + 4\zeta_n\|\boldsymbol{\pi}_{J^c}\|_1 + \frac{9M^2\zeta_n^2|J|}{2\kappa_{\widehat{\boldsymbol{\Sigma}}_n}(J, 3)}. \quad (6.41)$$

Since the last inequality holds for every $\boldsymbol{\pi}$, we can insert an $\inf_{\boldsymbol{\pi}}$ in the right hand side. Furthermore, in view of Proposition 6.6.1 below, with probability larger than $1-\delta$, ζ_n is bounded from above by $8V^3(\frac{\log(K/\delta)}{n})^{1/2}$. This completes the proof of (6.12).

To prove (6.13), we follow the same steps as above up to inequality (6.37). Then, we remark that for every $\boldsymbol{\pi}$ in the simplex satisfying $\boldsymbol{\pi}_{J^c} = 0$, it holds

$$\|(\widehat{\boldsymbol{\pi}}-\boldsymbol{\pi})_{J^c}\|_1 = \|\widehat{\boldsymbol{\pi}}_{J^c}\|_1 = 1 - \|\widehat{\boldsymbol{\pi}}_J\|_1 = \|\boldsymbol{\pi}_J\|_1 - \|\widehat{\boldsymbol{\pi}}_J\|_1 \le \|(\widehat{\boldsymbol{\pi}}-\boldsymbol{\pi})_J\|_1. \quad (6.42)$$

Therefore, $\|\widehat{\boldsymbol{\Sigma}}_n^{1/2}(\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi})\|_2^2 \geq$ we have with probability at least $1 - \delta$

$$\zeta_n\|\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_1 - \frac{1}{2M^2n}\|\mathbf{Z}(\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi})\|_2^2 \leq 2\zeta_n\|(\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi})_J\|_1 - \frac{1}{2M^2}\|\widehat{\boldsymbol{\Sigma}}_n^{1/2}(\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi})\|_2^2 \tag{6.43}$$

$$\leq 2\zeta_n\|(\boldsymbol{\pi} - \widehat{\boldsymbol{\pi}})_J\|_1 - \frac{\bar{\kappa}_{\widehat{\boldsymbol{\Sigma}}_n}(J, 1)\|(\boldsymbol{\pi} - \widehat{\boldsymbol{\pi}})_J\|_1^2}{2M^2|J|} \tag{6.44}$$

$$\leq \frac{2\zeta_n^2 M^2|J|}{\bar{\kappa}_{\widehat{\boldsymbol{\Sigma}}_n}(J, 1)}. \tag{6.45}$$

Replacing the right hand term in (6.37) and taking the infimum, we get the claim of the corollary. Since, in view of Proposition 6.6.1 below, with probability larger than $1 - \delta$, ζ_n is bounded from above by $8V^3(\frac{\log(K/\delta)}{n})^{1/2}$, we get the claim of (6.13).

6.6.2   Proof of Theorem 6.3.2

Let us denote $\boldsymbol{v} = \widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}$. According to (6.37) and (6.38), we have

$$\mathrm{KL}(f^*||f_{\widehat{\boldsymbol{\pi}}}) \leq \mathrm{KL}(f^*||f_{\boldsymbol{\pi}}) + \zeta_n\|\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_1 - \frac{1}{2M^2}\|\widehat{\boldsymbol{\Sigma}}_n^{1/2}(\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi})\|_2^2 \tag{6.46}$$

$$\leq \mathrm{KL}(f^*||f_{\boldsymbol{\pi}}) + \zeta_n\|\boldsymbol{v}\|_1 - \frac{1}{2M^2}\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{v}\|_2^2 + \frac{1}{2M^2}\boldsymbol{v}^\top(\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_n)\boldsymbol{v}. \tag{6.47}$$

As $\boldsymbol{v}$ is the difference of two vectors lying on the simplex, we have $\|\boldsymbol{v}\|_1 \leq 2$. Let $\|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_n\|_\infty = \max_{j,j'}|(\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_n)_{j,j'}|$ stand for the largest (in absolute values) element of the matrix $\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_n$. We have

$$\boldsymbol{v}^\top(\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_n)\boldsymbol{v} \leq \|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_n\|_\infty\|\boldsymbol{v}\|_1^2 \leq 2\|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_n\|_\infty\|\boldsymbol{v}\|_1. \tag{6.48}$$

Setting $\bar{\zeta}_n = \zeta_n + M^{-2}\|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}_n\|_\infty$, we get

$$\mathrm{KL}(f^*||f_{\widehat{\boldsymbol{\pi}}}) \leq \mathrm{KL}(f^*||f_{\boldsymbol{\pi}}) + \bar{\zeta}_n\|\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_1 - \frac{1}{2M^2}\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi})\|_2^2. \tag{6.49}$$

Following the same steps as those used for obtaining (6.41), we arrive at

$$\mathrm{KL}(f^*||f_{\widehat{\boldsymbol{\pi}}}) \leq \mathrm{KL}(f^*||f_{\boldsymbol{\pi}}) + 4\bar{\zeta}_n\|\boldsymbol{\pi}_{J^c}\|_1 + \frac{9\bar{\zeta}_n^2 M^2|J|}{2\kappa_{\boldsymbol{\Sigma}}(J, 3)}. \tag{6.50}$$

The last step consists in evaluating the quantiles of the random variable $\bar{\zeta}_n$. To this end, one checks that the Hoeffding inequality combined with the union bound yields

$$\mathbf{P}\Big\{\|\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}_n\|_\infty > t\Big\} \leq K(K-1)\exp(-2nt^2/M^4), \qquad \forall t > 0. \quad (6.51)$$

In other terms, for every $\delta \in (0,1)$, we have

$$\mathbf{P}\Big\{\|\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}_n\|_\infty \leq M^2\Big(\frac{\log(K^2/\delta)}{2n}\Big)^{1/2}\Big\} \geq 1 - \delta. \quad (6.52)$$

Note that for $\delta \leq 1$, we have $\log(K^2/\delta) \leq 2\log(K/\delta)$. Combining with Proposition 6.6.1, this implies that $\bar{\zeta}_n \leq (8V^3+1)\big(\frac{\log(K/\delta)}{n}\big)^{1/2}$ with probability larger than $1 - 2\delta$. This completes the proof of (6.15). The proof of (6.16) is omitted since it repeats the same arguments as those used for proving (6.13).

6.6.3   Proof of Theorem 6.3.3

According to (6.50), for any $\boldsymbol{\pi} \in \Pi$ and any $J \subset \{1,\ldots,K\}$, we have

$$\mathbf{E}[\mathrm{KL}(f^*\|f_{\widehat{\boldsymbol{\pi}}})] \leq \mathrm{KL}(f^*\|f_{\boldsymbol{\pi}}) + 4\|\boldsymbol{\pi}_{J^c}\|_1\mathbf{E}[\bar{\zeta}_n] + \frac{9M^2|J|}{2\kappa_{\mathbf{\Sigma}}(J,3)}\,\mathbf{E}[\bar{\zeta}_n^2]. \quad (6.53)$$

Recall now that $\bar{\zeta}_n = \zeta_n + M^{-2}\|\widehat{\mathbf{\Sigma}}_n - \mathbf{\Sigma}\|_\infty$ and, according to Proposition 6.6.1, we have

$$\mathbf{E}[\zeta_n] \leq 4V^3\Big(\frac{2\log(2K^2)}{n}\Big)^{1/2} \qquad \text{and} \qquad \mathbf{Var}[\zeta_n] \leq \frac{V^2}{2n}. \quad (6.54)$$

Using Theorem 6.7.2, one easily checks that

$$\mathbf{E}[\|\widehat{\mathbf{\Sigma}}_n - \mathbf{\Sigma}\|_\infty] \leq M^2\Big(\frac{\log(2K^2)}{2n}\Big)^{1/2}. \quad (6.55)$$

This implies that

$$\mathbf{E}[\bar{\zeta}_n] \leq (8V^3 + 1)\Big(\frac{\log(2K^2)}{2n}\Big)^{1/2}. \quad (6.56)$$

Similarly, in view of the Efron-Stein inequality, we have $\mathbf{Var}[\|\widehat{\mathbf{\Sigma}}_n - \mathbf{\Sigma}\|_\infty] \leq \frac{M^4}{2n}$. This implies that

$$\mathbf{E}[\bar{\zeta}_n^2] \leq (\mathbf{E}[\bar{\zeta}_n])^2 + \big\{(\mathbf{Var}[\zeta_n])^{1/2} + M^{-2}(\mathbf{Var}[\|\widehat{\mathbf{\Sigma}}_n - \mathbf{\Sigma}\|_\infty])^{1/2}\big\}^2 \quad (6.57)$$

$$\leq (8V^3 + 1)^2\frac{\log(2K^2)}{2n} + \frac{(V+1)^2}{2n} \quad (6.58)$$

$$\leq 1.615(8V^3 + 1)^2\frac{\log K}{n}. \quad (6.59)$$

Combining (6.56), (6.59) and (6.53), we get the desired result.

6.6.4 Proof of Proposition 1

Using the strong convexity of the function $u \mapsto \log u$ over the interval $[m, M]$ and the fact that $\boldsymbol{\pi}^*$ minimizes the convex function $\boldsymbol{\pi} \mapsto \mathrm{KL}(f^* \| f_{\boldsymbol{\pi}})$, we get

$$\mathrm{KL}(f^* \| f_{\widehat{\boldsymbol{\pi}}}) \geq \mathrm{KL}(f^* \| f_{\boldsymbol{\pi}^*}) + \frac{1}{2M^2} \|\widehat{\boldsymbol{\Sigma}}_n^{1/2}(\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}^*)\|_2^2. \tag{6.60}$$

Combining with (6.49), in which we replace $\boldsymbol{\pi}$ by $\boldsymbol{\pi}^*$, we get

$$\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}^*)\|_2^2 \leq 2M^2 \bar{\zeta}_n \|\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}^*\|_1. \tag{6.61}$$

Let us set $\boldsymbol{v} = \widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}^*$. If $\boldsymbol{v} = 0$, then the claims are trivial. In the rest of this proof, we assume $\|\boldsymbol{v}\|_1 > 0$. In view of (6.42), we have $\|\boldsymbol{v}\|_1 \leq 2\|\boldsymbol{v}_{J^*}\|_1$. Therefore, using the definition of the compatibility factor, we get

$$\|\boldsymbol{v}\|_1^2 \leq 4\|\boldsymbol{v}_{J^*}\|_1^2 \leq \frac{4|J^*| \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{v}\|_2^2}{\bar{\kappa}(J^*, 1)} \leq \frac{8|J^*| M^2 \bar{\zeta}_n \|\boldsymbol{v}\|_1}{\bar{\kappa}(J^*, 1)}. \tag{6.62}$$

We have already checked that $\bar{\zeta}_n \leq (8V^3 + 1)\big(\frac{\log(K/\delta)}{n}\big)^{1/2}$ with probability larger than $1 - 2\delta$. Dividing both sides of inequality (6.62) by $\|\boldsymbol{v}\|_1$ and using the aforementioned upper bound on $\bar{\zeta}_n$, we get the desired bound on $\|\boldsymbol{v}\|_1 = \|\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}^*\|_1$.

In order to bound the error $\boldsymbol{v} = \widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}^*$ in the Euclidean norm, we denote by $\widehat{J}$ the set of $D = |J^*|$ indices corresponding to D largest entries of the vector $(|v_1|, \ldots, |v_K|)$. Since $\|\boldsymbol{v}\|_1 \leq 2\|\boldsymbol{v}_{J^*}\|_1$, we clearly have $\|\boldsymbol{v}\|_1 \leq 2\|\boldsymbol{v}_{\widehat{J}}\|_1$. Therefore,

$$\|\boldsymbol{v}\|_2^2 = \|\boldsymbol{v}_{\widehat{J}}\|_2^2 + \|\boldsymbol{v}_{\widehat{J}^c}\|_2^2 \tag{6.63}$$

$$\leq \|\boldsymbol{v}_{\widehat{J}}\|_2^2 + \|\boldsymbol{v}_{\widehat{J}^c}\|_\infty \|\boldsymbol{v}_{\widehat{J}^c}\|_1 \tag{6.64}$$

$$\leq \|\boldsymbol{v}_{\widehat{J}}\|_2^2 + \frac{\|\boldsymbol{v}_{\widehat{J}}\|_1}{D} \|\boldsymbol{v}_{\widehat{J}^c}\|_1 \tag{6.65}$$

$$\leq \|\boldsymbol{v}_{\widehat{J}}\|_2^2 + \frac{1}{D} \|\boldsymbol{v}_{\widehat{J}}\|_1^2 \leq 2\|\boldsymbol{v}_{\widehat{J}}\|_2^2. \tag{6.66}$$

Combining this inequality with the definition of the restricted eigenvalue and inequality (6.61) above, we arrive at

$$\|\boldsymbol{v}_{\widehat{J}}\|_2^2 \leq \frac{\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{v}\|_2^2}{\kappa^{\mathrm{RE}}(D, 1)} \leq \frac{2M^2 \bar{\zeta}_n \|\boldsymbol{v}\|_1}{\kappa^{\mathrm{RE}}(D, 1)} \leq \frac{4M^2 \bar{\zeta}_n(\|\boldsymbol{v}_{\widehat{J}}\|_1 \wedge 1)}{\kappa^{\mathrm{RE}}(D, 1)} \leq \frac{4M^2 \bar{\zeta}_n(\sqrt{D}\|\boldsymbol{v}_{\widehat{J}}\|_2 \wedge 1)}{\kappa^{\mathrm{RE}}(D, 1)}. \tag{6.67}$$

Dividing both sides by $\|\boldsymbol{v}_{\widehat{J}}\|_2$, taking the square and using (6.66), we get

$$\|\boldsymbol{v}\|_2 \leq \sqrt{2} \|\boldsymbol{v}_{\widehat{J}}\|_2 \leq \frac{4\sqrt{2}M^2 |J^*|^{1/2} \bar{\zeta}_n}{\kappa^{\mathrm{RE}}(|J^*|, 1)} \bigwedge \frac{2\sqrt{2}M \bar{\zeta}_n^{1/2}}{\kappa^{\mathrm{RE}}(|J^*|, 1)^{1/2}}. \tag{6.68}$$
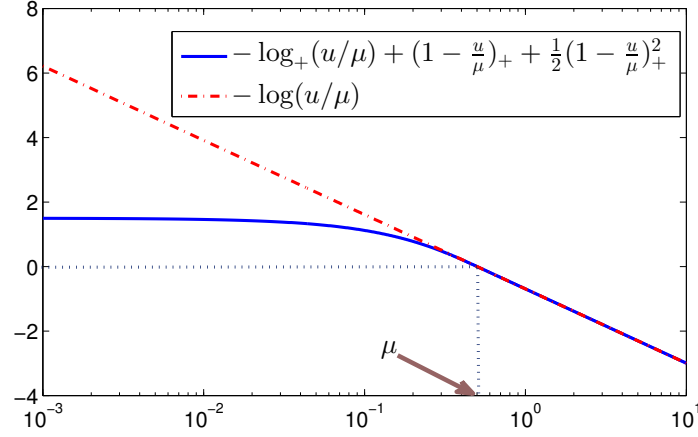
Figure 6.1: The plot of the function $u \mapsto \bar{\ell}(u)$, used in the proof of Proposition 2, superposed on the plot of the function $u \mapsto \ell(u) = -\log u$. We see that the former is a strongly convex surrogate of the latter.

This inequality, in conjunction with the upper bound on $\bar{\zeta}_n$ used above, completes the proof of the second claim.

6.6.5 Proof of Proposition 2

We repeat the proof of Theorem 6.3.1 with some small modifications. First of all, we replace the function $\ell(u) = -\log(u)$ by the function

$$
\bar{\ell}(u) = \begin{cases} -\log(u/\mu), & \text{if } u \geq \mu, \\ (1 - \frac{u}{\mu}) + \frac{1}{2}(1 - \frac{u}{\mu})^2, & \text{if } u \in (0, \mu). \end{cases} \tag{6.69}
$$

One easily checks that this function is twice continuously differentiable with a second derivative satisfying $M^{-2} \leq \bar{\ell}''(u) \leq \mu^{-2}$ for every $u \in (0, M)$. Furthermore, since $\bar{\ell}(u) = \ell(u/\mu)$ for every $u \geq \mu$, we have $\bar{L}_n(\widehat{\boldsymbol{\pi}}) = L_n(\widehat{\boldsymbol{\pi}})$, where we have used the notation $\bar{L}_n(\boldsymbol{\pi}) = \frac{1}{n} \sum_{i=1}^n \bar{\ell}(f_{\boldsymbol{\pi}}(\boldsymbol{X}_i))$. Therefore, similarly to (6.32), we get

$$
\frac{1}{n} \sum_{i=1}^n \bar{\ell}(f_{\widehat{\boldsymbol{\pi}}}(\boldsymbol{X}_i)) \leq \frac{1}{n} \sum_{i=1}^n \bar{\ell}(f_{\boldsymbol{\pi}}(\boldsymbol{X}_i)) - \frac{1}{2M^2n} \|\bar{\mathbf{Z}}(\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi})\|_2^2, \tag{6.70}
$$

for every $\boldsymbol{\pi} \in \Pi^*(\mu)$. Let us define $\bar{\varphi}(\boldsymbol{\pi}, \boldsymbol{x}) = \bar{\ell}(f_{\boldsymbol{\pi}}(\boldsymbol{x})) - \int \bar{\ell}(f_{\boldsymbol{\pi}}) f^* d\nu$ and $\bar{\Phi}_n(\boldsymbol{\pi}) = \frac{1}{n} \sum_{i=1}^n \bar{\varphi}(\boldsymbol{\pi}, \boldsymbol{X}_i)$. We have

$$\int \bar{\ell}(f_{\widehat{\boldsymbol{\pi}}}) \, f^* d\nu \leq \int \bar{\ell}(f_{\boldsymbol{\pi}}) \, f^* d\nu - \frac{1}{2M^2 n} \|\bar{\mathbf{Z}}(\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi})\|_2^2 + \frac{1}{n} \sum_{i=1}^n \left(\varphi(\boldsymbol{\pi}, \boldsymbol{X}_i) - \varphi(\widehat{\boldsymbol{\pi}}, \boldsymbol{X}_i) \right)$$

$$(6.71)$$

$$\leq \int \bar{\ell}(f_{\boldsymbol{\pi}}) \, f^* d\nu - \frac{1}{2M^2 n} \|\bar{\mathbf{Z}}(\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi})\|_2^2 + \underbrace{\sup_{\boldsymbol{\pi} \in \Pi_n(0)} \|\nabla \bar{\Phi}_n(\boldsymbol{\pi})\|_\infty \|\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_1}_{:=\xi_n}.$$

$$(6.72)$$

Notice that $\boldsymbol{\pi} \in \Pi^*(\mu)$ implies that $\bar{\ell}(f_{\boldsymbol{\pi}}) = \log \mu - \log f_{\boldsymbol{\pi}}$ and that $\bar{\ell}(f_{\widehat{\boldsymbol{\pi}}}) \geq \log \mu - \log f_{\widehat{\boldsymbol{\pi}}} - (\log \mu - \log f_{\widehat{\boldsymbol{\pi}}})_+$. Therefore, along the lines of the proof of (6.13) (see, namely, (6.45)), we get

$$\mathrm{KL}(f^* || f_{\widehat{\boldsymbol{\pi}}}) \leq \mathrm{KL}(f^* || f_{\boldsymbol{\pi}}) + \frac{2\xi_n^2 M^2 |J|}{\bar{\kappa}_{\widehat{\boldsymbol{\Sigma}}_n}(J, 1)} + \int_{\mathcal{X}} (\log \mu - \log f_{\widehat{\boldsymbol{\pi}}})_+ f^* d\nu. \quad (6.73)$$

We can repeat now the arguments of Proposition 6.6.1 with some minor modifications. We first rewrite ξ_n as $\xi_n = \max_{l=1,\dots,K} \xi_{l,n}$ with $\xi_{l,n} = \sup_{\boldsymbol{\pi} \in \Pi_n(0)} |\partial_l \bar{\Phi}_n(\boldsymbol{\pi})|$. One checks that the bounded difference inequality and the Efron-Stein inequality can be applied with an additional factor 2, since for $F_l(\mathbf{X}) = \sup_{\boldsymbol{\pi} \in \Pi_n(0)} |\partial_l \bar{\Phi}_n(\boldsymbol{\pi})|$, we have

$$|F_l(\mathbf{X}) - F_l(\mathbf{X}')| \leq \frac{2M}{n\mu} = \frac{2V}{n}. \quad (6.74)$$

Therefore, for every $l \in [K]$, with probability larger than $1 - (\delta/K)$, we have $\xi_{l,n} \leq \mathbf{E}[\xi_{l,n}] + V(\frac{2\log(K/\delta)}{n})^{1/2}$ and $\mathbf{Var}[\xi_n] \leq (2V)^2/n$. By the union bound, we obtain that with probability larger than $1 - \delta$, $\xi_n \leq \max_l \mathbf{E}[\xi_{l,n}] + V(\frac{2\log(K/\delta)}{n})^{1/2}$. Thus, to upper bound $\mathbf{E}[\xi_{l,n}]$, we use the symmetrization argument:

$$\mathbf{E}[\xi_{l,n}] \leq 2\mathbf{E}\left[\sup_{\boldsymbol{\pi} \in \Pi_n(0)} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i \bar{\ell}'(f_{\boldsymbol{\pi}}(\boldsymbol{X}_i)) f_l(\boldsymbol{X}_i) \right| \right] \quad (6.75)$$

$$\leq 2M \mathbf{E}\left[\sup_{\boldsymbol{\pi} \in \Pi_n(0)} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i \bar{\ell}'(f_{\boldsymbol{\pi}}(\boldsymbol{X}_i)) \right| \right] \quad [5, \text{Th. } 11.5] \quad (6.76)$$

$$\leq \frac{2M}{\mu} \mathbf{E}\left[\left| \frac{1}{n} \sum_{i=1}^n \epsilon_i \right| \right] + 2M \mathbf{E}\left[\sup_{\boldsymbol{\pi} \in \Pi_n(0)} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i [\bar{\ell}'(f_{\boldsymbol{\pi}}(\boldsymbol{X}_i)) - \bar{\ell}'(0)] \right| \right].$$

$$(6.77)$$

Note that the function $\bar{\ell}'$, the derivative of $\bar{\ell}$ defined in (6.69), is by construction Lipschitz with constant $1/\mu^2$. Therefore, in view of the contraction principle,

$$\mathbf{E}[\xi_{l,n}] \leq \frac{2M}{\mu}\mathbf{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\right)^2\right]^{1/2} + \frac{4M}{\mu^2}\mathbf{E}\left[\sup_{\boldsymbol{\pi}\in\Pi_n(0)}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f_{\boldsymbol{\pi}}(\boldsymbol{X}_i)\right] \quad (6.78)$$

$$\leq \frac{2M}{\mu\sqrt{n}} + \frac{4M}{\mu^2}\mathbf{E}\left[\sup_{k\in[K]}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f_k(\boldsymbol{X}_i)\right] \quad (6.79)$$

$$\leq \frac{2M}{\mu\sqrt{n}} + \frac{8M^2}{\mu^2}\left(\frac{\log K}{2n}\right)^{1/2} \leq \frac{2V^2(1 + 2\sqrt{2\log K})}{\sqrt{n}}. \quad (6.80)$$

As a consequence, we proved that with probability larger than $1 - \delta$, we have $\xi_n \leq 8V^2(\frac{\log K}{n})^{1/2}$. This completes the proof of the first inequality. In order to prove the second one, we simply change the way we have evaluated the term $\int \bar{\ell}(f_{\widehat{\boldsymbol{\pi}}})f^*$ in the left hand side of (6.71). Since $\bar{\ell}$ is strongly convex with a second order derivative bounded from below by $1/M^2$, we have $\bar{\ell}(f_{\widehat{\boldsymbol{\pi}}}) \geq \bar{\ell}(f^*) + \bar{\ell}'(f^*)(f_{\widehat{\boldsymbol{\pi}}} - f^*) + \frac{1}{2M^2}(f_{\widehat{\boldsymbol{\pi}}} - f^*)^2$. Since f^* is always larger than μ, the derivative $\bar{\ell}'(f^*)$ equals $1/f^*$. Integrating over $\mathcal{X}$, we get the second inequality of the proposition.

6.6.6   Auxiliary results

We start by a general convex result based on the strong convexity of the $-log$ function to derive a bound on the estimated log-likelihood.

Lemma 6.6.2. *Let us assume that $M = \max_{j\in[K]}\|f_j\|_\infty < \infty$. Then, for any $\boldsymbol{\pi} \in \mathbb{B}_+^K$, it holds that*

$$L_n(\widehat{\boldsymbol{\pi}}) \leq L_n(\boldsymbol{\pi}) - \frac{1}{2M^2n}\|\bar{\mathbf{Z}}(\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi})\|_2^2. \quad (6.81)$$

Proof. Recall that $\widehat{\boldsymbol{\pi}}$ minimizes the function L_n defined in (6.7) over Π_n. Furthermore, the function $u \mapsto \ell(u)$ is clearly strongly convex with a second order derivative bounded from below by $1/M^2$ over the set $u \in (0, M]$. Therefore, for every $\widehat{u} \in (0, M]$, the function $\widetilde{\ell}$ given by:

$$\widetilde{\ell}(u) = \ell(u) - \frac{1}{2M^2}(\widehat{u} - u)^2, \quad u \in (0, M], \quad (6.82)$$

is convex. This implies that the mapping

$$\boldsymbol{\pi} \mapsto \widetilde{L}_n(\boldsymbol{\pi}) = L_n(\boldsymbol{\pi}) - \frac{1}{2M^2n}\|\mathbf{Z}(\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi})\|_2^2 \quad (6.83)$$

is convex over the set $\boldsymbol{\pi} \in \mathbb{B}_+^K$. This yields[4]

$$\widetilde{L}_n(\boldsymbol{\pi}) - \widetilde{L}_n(\widehat{\boldsymbol{\pi}}) \geq \sup_{\boldsymbol{v} \in \partial \widetilde{L}_n(\widehat{\boldsymbol{\pi}})} \boldsymbol{v}^\top (\boldsymbol{\pi} - \widehat{\boldsymbol{\pi}}), \qquad \forall \boldsymbol{\pi} \in \mathbb{B}_+^K. \qquad (6.84)$$

Using the Karush-Kuhn-Tucker conditions and the fact that $\widehat{\boldsymbol{\pi}}$ minimizes L_n, we get $\mathbf{0}_K \in \partial L_n(\widehat{\boldsymbol{\pi}}) = \partial \widetilde{L}_n(\widehat{\boldsymbol{\pi}})$. This readily gives $\widetilde{L}_n(\boldsymbol{\pi}) - \widetilde{L}_n(\widehat{\boldsymbol{\pi}}) \geq 0$, for any $\boldsymbol{\pi} \in \mathbb{B}_+^K$. The last step is to remark that $\mathbf{Z}(\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}) = \bar{\mathbf{Z}}(\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi})$, since both $\widehat{\boldsymbol{\pi}}$ and $\boldsymbol{\pi}$ have entries summing to one. $\qquad \square$

The core of our results lies in the following proposition which bound the deviations of the empirical process part.

Proposition 6.6.1 (Supremum of Empirical Process). *For any* $\boldsymbol{\pi} \in \mathbb{B}_+^K$ *and* $\boldsymbol{x} \in \mathcal{X}$*, define* $\varphi(\pi, \boldsymbol{x}) = \int (\log f_{\boldsymbol{\pi}}) f^* - \log f_{\boldsymbol{\pi}}(\boldsymbol{x})$ *and consider* $\Phi_n(\boldsymbol{\pi}) = \frac{1}{n} \sum_{i=1}^n \varphi(\boldsymbol{\pi}, \boldsymbol{X}_i)$*. If* $K \geq 2$*, then for any* $\delta \in (0,1)$*, with probability at least* $1 - \delta$*, we have*

$$\zeta_n = \sup_{\boldsymbol{\pi} \in \Pi_n} \left\| \nabla \Phi_n(\boldsymbol{\pi}) \right\|_\infty \leq 8V^3 \left(\frac{\log(K/\delta)}{n} \right)^{1/2}. \qquad (6.85)$$

Furthermore, we have $\mathbf{E}[\zeta_n] \leq 4V^3 \left(\frac{2\log(2K^2)}{n} \right)^{1/2}$ *and* $\mathbf{Var}[\zeta_n] \leq V^2/(2n)$*.*

Proof. To ease notation, let us denote $g_{\boldsymbol{\pi},l}(x) = \frac{f_l(x)}{f_{\boldsymbol{\pi}}(x)} - \mathbf{E}\left[\frac{f_l(\boldsymbol{X})}{f_{\boldsymbol{\pi}}(\boldsymbol{X})} \right]$ and

$$F(\mathbf{X}) = \sup_{\boldsymbol{\pi} \in \Pi_n} \left\| \nabla \Phi_n(\boldsymbol{\pi}) \right\|_\infty = \sup_{(\boldsymbol{\pi}, l) \in \Pi_n \times [K]} \left| \frac{1}{n} \sum_{i=1}^n g_{\boldsymbol{\pi},l}(\boldsymbol{X}_i) \right|, \qquad (6.86)$$

where $\mathbf{X} = (\boldsymbol{X}_1, \dots, \boldsymbol{X}_n)$. To derive a bound on F, we will use the McDiarmid concentration inequality that requires the bounded difference condition to hold for F. For some $i_0 \in [n]$, let $\mathbf{X}' = (\boldsymbol{X}_1, \dots, \boldsymbol{X}'_{i_0}, \dots, \boldsymbol{X}_n)$ be a new sample obtained from $\mathbf{X}$ by modifying the i_0-th element $\boldsymbol{X}_i$ and by leaving all the others unchanged. Then, we have

$$F(\mathbf{X}) - F(\mathbf{X}') = \sup_{(\boldsymbol{\pi}, l) \in \Pi_n \times [K]} \left| \frac{1}{n} \sum_{i=1}^n g_{\boldsymbol{\pi},l}(\boldsymbol{X}_i) \right| - \sup_{(\boldsymbol{\pi}, l) \in \Pi \times [K]} \left| \frac{1}{n} \sum_{i=1}^n g_{\boldsymbol{\pi},l}(\boldsymbol{X}'_i) \right|$$
$$(6.87)$$

$$\leq \sup_{(\boldsymbol{\pi}, l) \in \Pi_n \times [K]} \left| \frac{1}{n} \sum_{i=1}^n g_{\boldsymbol{\pi},l}(\boldsymbol{X}_i) - \frac{1}{n} \sum_{i=1}^n g_{\boldsymbol{\pi},l}(\boldsymbol{X}'_i) \right| \qquad (6.88)$$

$$= \sup_{(\boldsymbol{\pi}, l) \in \Pi_n \times [K]} \left| \frac{1}{n} \left(g_{\boldsymbol{\pi},l}(\boldsymbol{X}_{i_0}) - g_{\boldsymbol{\pi},l}(\boldsymbol{X}'_{i_0}) \right) \right| \leq \frac{V}{n}, \qquad (6.89)$$

---

[4] We denote by ∂g the sub-differential of a convex function g.

where the last inequality is a direct consequence of assumption (6.11). Therefore, using the McDiarmid concentration inequality recalled in Theorem 6.7.3 below, we check that the inequality

$$F(\mathbf{X}) \leq \mathbf{E}(F(\mathbf{X})) + V\sqrt{\frac{\log(1/\delta)}{2n}} \tag{6.90}$$

holds with probability at least $1 - \delta$. Furthermore, in view of the Efron-Stein inequality, we have

$$\mathbf{Var}[\zeta_n] = \mathbf{Var}[F(\mathbf{X})] \leq \frac{V^2}{2n}. \tag{6.91}$$

Let us denote $\mathcal{G} := \{(f_l/f_{\boldsymbol{\pi}}) - 1, (\boldsymbol{\pi}, l) \in \Pi_n \times [K]\}$ and $\mathfrak{R}_{n,q}(\mathcal{G})$ the Rademacher complexity of $\mathcal{G}$ given by

$$\mathfrak{R}_n(\mathcal{G}) = \mathbf{E}_{\epsilon}\left[\sup_{(\boldsymbol{\pi},l)\in\Pi_n\times[K]}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i\left(\frac{f_l(\boldsymbol{X}_i)}{f_{\boldsymbol{\pi}}(\boldsymbol{X}_i)} - 1\right)\right|\right], \tag{6.92}$$

with $\epsilon_1, \ldots, \epsilon_n$ independent and identically distributed Rademacher random variables independent of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. Using the symmetrization inequality (see, for instance, Theorem 2.1 in [17]) we have

$$\mathbf{E}[F(\mathbf{X})] = \mathbf{E}[\zeta_n] \leq 2\mathbf{E}[\mathfrak{R}_n(\mathcal{G})]. \tag{6.93}$$

Lemma 6.6.3. *The Rademacher complexity defined in* (6.92) *satisfies*

$$\mathfrak{R}_n(\mathcal{G}) \leq 4V^3\sqrt{\frac{\log K}{n}}. \tag{6.94}$$

Proof. The proof relies on the contraction principle of [21] that we recall in Section 6.7.3 for the convenience. We apply this principle to the random variables $X_{i,(\boldsymbol{\pi},l)} = f_{\boldsymbol{\pi}}(\boldsymbol{X}_i)/f_l(\boldsymbol{X}_i) - 1$ and to the function $\psi(x) = (1+x)^{-1} - 1$. Clearly ψ is Lipschitz on $[\frac{1}{V} - 1, V - 1]$ with the Lipschitz constant equal to V^2 and $\psi(0) = 0$. Therefore

$$\mathfrak{R}_n(\mathcal{G}) \leq \mathbf{E}_{\epsilon}\left[\sup_{(\boldsymbol{\pi},l)}\frac{1}{n}\sum_{i=1}^n \epsilon_i\psi(\boldsymbol{X}_{i,(\boldsymbol{\pi},l)})\right] + \mathbf{E}_{\epsilon}\left[\sup_{(\boldsymbol{\pi},l)}\frac{1}{n}\sum_{i=1}^n \epsilon_i(-\psi)(\boldsymbol{X}_{i,(\boldsymbol{\pi},l)})\right] \tag{6.95}$$

$$\leq 2V^2\mathbf{E}_{\epsilon}\left[\sup_{(\boldsymbol{\pi},l)\in\Pi_n\times[K]}\frac{1}{n}\sum_{i=1}^n \epsilon_i\boldsymbol{X}_{i,(\boldsymbol{\pi},l)}\right] \tag{6.96}$$

$$= 2V^2\mathbf{E}_{\epsilon}\left[\sup_{(\boldsymbol{\pi},l)\in\Pi_n\times[K]}\frac{1}{n}\sum_{i=1}^n \epsilon_i\left(\frac{f_{\boldsymbol{\pi}}(\boldsymbol{X}_i)}{f_l(\boldsymbol{X}_i)} - 1\right)\right]. \tag{6.97}$$

Expanding $f_{\boldsymbol{\pi}}(\boldsymbol{X}_i)$ we obtain

$$\mathbf{E}_\epsilon\left[\sup_{(\boldsymbol{\pi},l)} \frac{1}{n}\sum_{i=1}^n \epsilon_i\left(\frac{f_{\boldsymbol{\pi}}(\boldsymbol{X}_i)}{f_l(\boldsymbol{X}_i)}-1\right)\right] = \mathbf{E}_\epsilon\left[\sup_{(\boldsymbol{\pi},l)} \sum_{k=1}^K \frac{\pi_k}{n}\sum_{i=1}^n \epsilon_i\left(\frac{f_k(\boldsymbol{X}_i)}{f_l(\boldsymbol{X}_i)}-1\right)\right]$$

(6.98)

$$= \mathbf{E}_\epsilon\left[\max_{k,l\in[K]} \frac{1}{n}\sum_{i=1}^n \epsilon_i\left(\frac{f_k(\boldsymbol{X}_i)}{f_l(\boldsymbol{X}_i)}-1\right)\right].$$

(6.99)

We apply now Theorem 6.7.2 with $s=(k,l)$, $N=K^2$, $a=-V$, $b=V$ and $Y_{i,s}=\epsilon_i\left(\frac{f_k(\boldsymbol{X}_i)}{f_l(\boldsymbol{X}_i)}-1\right)$. This yields

$$\mathbf{E}_\epsilon\left[\max_{k,l\in[K]} \frac{1}{n}\sum_{i=1}^n \epsilon_i\left(\frac{f_k(\boldsymbol{X}_i)}{f_l(\boldsymbol{X}_i)}-1\right)\right] \le 2V\left(\frac{\log K^2}{2n}\right)^{1/2}.$$

(6.100)

This completes the proof of the lemma. □

   Combining inequalities (6.90,6.93) and Lemma 6.6.3, we get that the inequality

$$F(\mathbf{X}) \le 8V^3\left(\frac{\log K}{n}\right)^{1/2} + V\left(\frac{\log(1/\delta)}{2n}\right)^{1/2}$$

(6.101)

holds with probability at least $1-\delta$. Noticing that $V\ge 1$ and, for $K\ge 2$, $\delta\in(0,K^{-1/31})$ we have $8\sqrt{\log K}+\sqrt{(1/2)\log(1/\delta)} \le 8\sqrt{\log(K/\delta)}$, we get the first claim of the proposition. The second claim is a direct consequence of Lemma 6.6.3 and (6.93). □

6.7 Proof of the lower bound for nearly-D-sparse aggregation

We prove the minimax lower bound for estimation in Kullback-Leibler risk using the following slightly adapted version of Theorem 2.5 from [32]. Throughout this section, we denote by $\lambda_{\min,\boldsymbol{\Sigma}}(k)$ and $\lambda_{\max,\boldsymbol{\Sigma}}(k)$, respectively, the smallest and the largest eigenvalue of all $k\times k$ principal minors of the matrix $\boldsymbol{\Sigma}$.

Theorem 6.7.1. *For some integer $L\ge 4$ assume that $\mathcal{H}_\mathcal{F}(\gamma,D)$ contains L elements $f_{\boldsymbol{\pi}^{(1)}},\ldots,f_{\boldsymbol{\pi}^{(L)}}$ satisfying the following two conditions.*
   (i) $\mathrm{KL}(f_{\boldsymbol{\pi}^{(j)}}||f_{\boldsymbol{\pi}^{(k)}}) \ge 2s > 0$, for all pairs (j,k) such that $1\le j < k \le L$.

(ii) *For product densities f_ℓ^n defined on $\mathcal{X}^n$ by $f_\ell^n(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = f_{\boldsymbol{\pi}^{(\ell)}}(\boldsymbol{x}_1) \times \ldots \times f_{\boldsymbol{\pi}^{(\ell)}}(\boldsymbol{x}_n)$ it holds*

$$\max_{\ell \in [L]} \mathrm{KL}(f_\ell^n || f_1^n) \leq \frac{\log L}{16}. \tag{6.102}$$

Then

$$\inf_{\widehat{f}} \sup_{f \in \mathcal{H}_\mathcal{F}(\gamma, D)} \mathbf{P}_f\big(\mathrm{KL}(f || \widehat{f}) \geq s\big) \geq 0.17. \tag{6.103}$$

To establish the bound claimed in Theorem 6.4.1, we will split the problem into two parts, corresponding to the following two subsets of $\mathcal{H}_\mathcal{F}(\gamma, D)$

$$\begin{aligned}\mathcal{H}_\mathcal{F}(0, D) &= \big\{f_{\boldsymbol{\pi}} : \boldsymbol{\pi} \in \mathbb{B}_+^K \text{ such that } \exists J \subset [K] \text{ with } \|\boldsymbol{\pi}_{J^c}\|_1 = 0 \text{ and } |J| \leq D\big\}, \\ \mathcal{H}_\mathcal{F}(\gamma, 1) &= \big\{f_{\boldsymbol{\pi}} : \boldsymbol{\pi} \in \mathbb{B}_+^K \text{ such that } \pi_1 = 1 - \gamma \text{ and } \sum_{j=2}^K \pi_j = \gamma\big\}.\end{aligned} \tag{6.104}$$

We will show that over $\mathcal{H}_\mathcal{F}(0, D)$, we have a lower bound of order $\log(1 + K/D)/n$ while over $\mathcal{H}_\mathcal{F}(\gamma, 1)$, a lower bound of order $\big[\frac{\gamma^2}{n} \log\big(1 + K/(\gamma\sqrt{n})\big)\big]^{1/2}$ holds true. Therefore, the lower bound over $\mathcal{H}_\mathcal{F}(\gamma, D)$ is larger than the average of these bounds.

For any $M \geq 1$ and $k \in [M - 1]$, let Ω_k^M be the subset of $\{0, 1\}^M$ defined by

$$\Omega_k^M := \Big\{\boldsymbol{\omega} \in \{0, 1\}^M : \|\boldsymbol{\omega}\|_1 = k\Big\}. \tag{6.105}$$

Before starting, we remind here a version of the Varshamov-Gilbert lemma (see, for instance, [31, Lemma 8.3]) which will be helpful for deriving our lower bounds.

Lemma 6.7.1. *Let $M \geq 4$ and $k \in [M/2]$ be two integers. Then there exist a subset $\Omega \subset \Omega_k^M$ and an absolute constant C_1 such that*

$$\|\boldsymbol{\omega} - \boldsymbol{\omega}'\|_1 \geq \frac{k + 1}{4} \quad \forall \boldsymbol{\omega}, \boldsymbol{\omega}' \in \Omega \text{ s.t. } \boldsymbol{\omega} \neq \boldsymbol{\omega}' \tag{6.106}$$

and $L = |\Omega|$ satisfies $L \geq 4$ and

$$\log L \geq C_1 k \log\Big(1 + \frac{eM}{k}\Big). \tag{6.107}$$

We will also use the following lemma that allows us to relate the KL-divergence $\mathrm{KL}(f_{\boldsymbol{\pi}} || f_{\boldsymbol{\pi}'})$ to the Euclidean distance between the weight vectors $\boldsymbol{\pi}$ and $\boldsymbol{\pi}'$.

Lemma 6.7.2. *If the dictionary $\mathcal{F}$ satisfies the boundedness assumption (6.11), then for any $f_{\boldsymbol{\pi}}, f_{\boldsymbol{\pi}'} \in \mathcal{H}_\mathcal{F}(\gamma, D)$ we have*

$$\frac{1}{2V^2 M} \|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\pi}' - \boldsymbol{\pi})\|_2^2 \leq \mathrm{KL}(f_{\boldsymbol{\pi}} || f_{\boldsymbol{\pi}'}) \leq \frac{V^2}{2m} \|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\pi}' - \boldsymbol{\pi})\|_2^2. \tag{6.108}$$

Proof. Using the Taylor expansion, one can check that for any $u \in [1/L, L]$, we have $(1 - u) + \frac{1}{2V^2}(u - 1)^2 \leq -\log u \leq (1 - u) + \frac{V^2}{2}(u - 1)^2$. Therefore,

$$\frac{1}{2V^2} \int_{\mathcal{X}} \left(\frac{f_{\pi'}}{f_\pi} - 1\right)^2 f_\pi \, d\nu \leq \mathrm{KL}(f_\pi \| f_{\pi'}) \leq \frac{V^2}{2} \int_{\mathcal{X}} \left(\frac{f_{\pi'}}{f_\pi} - 1\right)^2 f_\pi \, d\nu. \quad (6.109)$$

Since $\mathcal{F}$ satisfies the boundedness assumption, we get

$$\frac{1}{2MV^2} \int_{\mathcal{X}} \left(f_{\pi'} - f_\pi\right)^2 d\nu \leq \mathrm{KL}(f_\pi \| f_{\pi'}) \leq \frac{V^2}{2m} \int_{\mathcal{X}} \left(f_{\pi'} - f_\pi\right)^2 d\nu. \quad (6.110)$$

The claim of the lemma follows from these inequalities and the fact that $\int_{\mathcal{X}} \left(f_{\pi'} - f_\pi\right)^2 d\nu = \|\mathbf{\Sigma}^{1/2}(\pi' - \pi)\|_2^2$. □

6.7.1 Lower bound on $\mathcal{H}_{\mathcal{F}}(0, D)$

We show here that the lower bound $(D/n) \log(1 + eK/D) \wedge \left((1/n)\log(1 + K/\sqrt{n})\right)^{1/2}$ holds when we consider the worst case error for f^* belonging to the set $\mathcal{H}_{\mathcal{F}}(0, D)$.

Proposition 3. *If $\log(1 + eK) \leq n$ then, for the constant*

$$C_2 = \frac{C_1 m \bar{\kappa}_{\mathbf{\Sigma}}(2D, 0)}{2^9 V^2 M (C_1 m \vee 4V^2 \lambda_{\max, \mathbf{\Sigma}}(2D))} \geq \frac{C_1 m \varkappa_*}{2^9 V^2 M (C_1 m \vee 4V^2 \varkappa^*)}, \quad (6.111)$$

we have

$$\inf_{\widehat{f}} \sup_{f \in \mathcal{H}_{\mathcal{F}}(0,D)} \mathbf{P}_f \left(\mathrm{KL}(f \| \widehat{f}) \geq C_2 \frac{D \log(1 + K/D)}{n} \wedge \left(\frac{\log(1 + K/\sqrt{n})}{n}\right)^{1/2}\right) \geq 0.17.$$
$$(6.112)$$

Proof. We assume that $D \leq K/2$. The case $D > K/2$ can be reduced to the case $D = K/2$ by using the inclusion $\mathcal{H}_{\mathcal{F}}(0, K/2) \subset \mathcal{H}_{\mathcal{F}}(0, D)$. Let us set $A_1 = 4 \vee 16V^2 \lambda_{\max, \mathbf{\Sigma}}(2D)/(C_1 m)$ and denote by d the largest integer such that

$$d \leq D \quad \text{and} \quad d^2 \log\left(1 + \frac{eK}{d}\right) \leq A_1 n. \quad (6.113)$$

According to Lemma 6.7.1, there exists a subset $\Omega = \{\boldsymbol{\omega}^{(\ell)} : \ell \in [L]\}$ of Ω_d^K of cardinality $L \geq 4$ satisfying $\log L \geq C_1 d \log(1 + eK/d)$ such that for any pair of distinct elements $\boldsymbol{\omega}^{(\ell)}, \boldsymbol{\omega}^{(\ell')} \in \Omega$ we have $\|\boldsymbol{\omega}^{(\ell)} - \boldsymbol{\omega}^{(\ell')}\|_1 \geq d/4$. Using these binary vectors $\boldsymbol{\omega}^{(\ell)}$, we define the set $\mathcal{D} = \{\boldsymbol{\pi}^{(1)}, \ldots, \boldsymbol{\pi}^{(L)}\} \subset \mathbb{B}_+^K$ as follows:

$$\boldsymbol{\pi}^{(1)} = \boldsymbol{\omega}^{(1)}/d, \quad \boldsymbol{\pi}^{(\ell)} = (1 - \varepsilon)\boldsymbol{\pi}^{(1)} + \varepsilon \boldsymbol{\omega}^{(\ell)}/d, \quad \ell = 2, \ldots, L. \quad (6.114)$$

Clearly, for every $\varepsilon \in [0,1]$, the vectors $\boldsymbol{\pi}^{(\ell)}$ belong to $\mathbb{B}_+^K$. Furthermore, for any pair of distinct values $\ell, \ell' \in [L]$, we have $\|\boldsymbol{\pi}^{(\ell)} - \boldsymbol{\pi}^{(\ell')}\|_q^q = (\varepsilon/d)^q \|\boldsymbol{\omega}^{(\ell)} - \boldsymbol{\omega}^{(\ell')}\|_1 \geq (\varepsilon/d)^q d/4$. In view of Lemma 6.7.2, this yields

$$\mathrm{KL}(f_{\boldsymbol{\pi}^{(\ell)}} \| f_{\boldsymbol{\pi}^{(\ell')}}) \geq \frac{\bar{\kappa}_{\boldsymbol{\Sigma}}(2d, 0)}{4V^2 M d} \|\boldsymbol{\pi}^{(\ell)} - \boldsymbol{\pi}^{(\ell')}\|_1^2 \geq \frac{\bar{\kappa}_{\boldsymbol{\Sigma}}(2D, 0)}{64 V^2 M} \times \frac{\varepsilon^2}{d}. \quad (6.115)$$

Let us choose

$$\varepsilon^2 = \frac{d^2 \log(1 + eK/d)}{n A_1}. \quad (6.116)$$

It follows from (6.113) that $\varepsilon \leq 1$. Inserting this value of ε in (6.115), we get

$$\mathrm{KL}(f_{\boldsymbol{\pi}^{(\ell)}} \| f_{\boldsymbol{\pi}^{(\ell')}}) \geq 2 C_2 \frac{d \log(1 + eK/d)}{n}. \quad (6.117)$$

This shows that condition (i) of Theorem 6.7.1 is satisfied with $s = C_2 (d/n) \log(1 + eK/d)$. For the second condition of the same theorem, we have

$$\max_{\ell \in [L]} \mathrm{KL}(f_\ell^n \| f_1^n) = n \max_\ell \mathrm{KL}(f_{\boldsymbol{\pi}^{(\ell)}} \| f_{\boldsymbol{\pi}^{(1)}}) \quad (6.118)$$

$$\leq \frac{n V^2 \lambda_{\max, \boldsymbol{\Sigma}}(2d)}{2m} \max_\ell \|\boldsymbol{\pi}^{(\ell)} - \boldsymbol{\pi}^{(1)}\|_2^2 \quad (6.119)$$

$$\leq \frac{n V^2 \lambda_{\max, \boldsymbol{\Sigma}}(2D)}{m} \times \frac{\varepsilon^2}{d}, \quad (6.120)$$

since one can check that $\|\boldsymbol{\pi}^{(\ell)} - \boldsymbol{\pi}^{(1)}\|_2^2 \leq (\varepsilon/d)^2 \|\boldsymbol{\omega}^{(\ell)} - \boldsymbol{\omega}^{(1)}\|_1 \leq 2\varepsilon^2/d$. Therefore, using the definition of ε, we get

$$\max_{\ell \in [L]} \mathrm{KL}(f_\ell^n \| f_1^n) \leq \frac{n V^2 \lambda_{\max, \boldsymbol{\Sigma}}(2D)}{m} \times \frac{C_1 d m \log(1 + eK/d)}{16 n V^2 \lambda_{\max, \boldsymbol{\Sigma}}(2D)} \quad (6.121)$$

$$= \frac{C_1 d \log(1 + eK/d)}{16} \leq \frac{\log L}{16}. \quad (6.122)$$

Theorem 6.7.1 implies that

$$\inf_{\widehat{f}} \sup_{f \in \mathcal{H}_{\mathcal{F}}(0, D)} \mathbf{P}_f \left(\mathrm{KL}(f \| \widehat{f}) \geq C_2 \frac{d \log(1 + eK/d)}{n} \right) \geq 0.17. \quad (6.123)$$

We use the fact that d is the largest integer satisfying (6.113). Therefore, either $d + 1 > D$ or

$$(d + 1)^2 \log \left(1 + \frac{eK}{d+1} \right) \leq A_1 n. \quad (6.124)$$

If $d \geq D$, then the claim of the proposition follows from (6.123), since $d \log(1 + eK/d) \geq D \log(1 + eK/D)$. On the other hand, if (6.124) is true, then

$$d \log(1 + eK/d) \geq \frac{1}{2}(d+1) \log(1 + eK/(d+1)) \geq \frac{1}{2}\left(A_1 n \log(1 + eK/(d+1))\right)^{1/2}.$$
(6.125)

In addition, $d^2 \log(1 + eK/d) \leq A_1 n$ implies that $(d+1)^2 \leq A_1 n$. Combining the last two inequalities, we get the inequality $d \log(1 + eK/d) \geq \frac{1}{2}\left(A_1 n \log(1 + eK/\sqrt{A_1 n})\right)^{1/2} \geq \left(n \log(1 + eK/\sqrt{n})\right)^{1/2}$. Therefore, in view of (6.123), we get the claim of the proposition. $\square$

6.7.2 Lower bound on $\mathcal{H}_{\mathcal{F}}(\gamma, 1)$

Next result shows that the lower bound $\frac{\gamma^2}{n} \log\left(1 + \frac{K}{\gamma\sqrt{n}}\right)$ holds for the worst case error when f^* belongs to the set $\mathcal{H}_{\mathcal{F}}(\gamma, 1)$.

Proposition 4. *Assume that*

$$\left(\frac{\log(1 + eK)}{n}\right)^{1/2} \leq 2\gamma.$$
(6.126)

Then, for the constant $C_3 = \frac{C_1 m \bar{\kappa}_{\boldsymbol{\Sigma}}(2D,0)}{2^{12} V^4 M \lambda_{\max, \boldsymbol{\Sigma}}(2D)}$, *it holds that*

$$\inf_{\widehat{f}} \sup_{f \in \mathcal{H}_{\mathcal{F}}(\gamma,1)} \mathbf{P}_f\left(\mathrm{KL}(f||\widehat{f}) \geq C_3 \left\{\frac{\gamma^2}{n} \log\left(1 + \frac{K}{\gamma\sqrt{n}}\right)\right\}^{1/2}\right) \geq 0.17.$$
(6.127)

Proof. Let $C > 2$ be a constant the precise value of which will be specified later. Denote by d the largest integer satisfying

$$d\sqrt{\log(1 + eK/d)} \leq C\gamma\sqrt{n}.$$
(6.128)

Note that $d \geq 1$ in view of the condition $(\frac{\log(1+eK)}{n})^{1/2} \leq 2\gamma$ of the proposition. This readily implies that $d \leq C\gamma\sqrt{n}$ and, therefore,

$$\frac{\gamma}{d} \geq C^{-1}\left\{\frac{1}{n} \log\left(1 + \frac{eK}{C\gamma\sqrt{n}}\right)\right\}^{1/2} \geq 2C^{-2}\left\{\frac{1}{n} \log\left(1 + \frac{K}{\gamma\sqrt{n}}\right)\right\}^{1/2}.$$
(6.129)

Let us first consider the case $d \leq (K-1)/2$. According to Lemma 6.7.1, there exists a subset $\Omega \subset \Omega_d^{K-1}$ of cardinality L satisfying $\log L \geq C_1 \log\left(1 + \frac{e(K-1)}{d}\right)$ and $\|\boldsymbol{\omega}^{(\ell)} - \boldsymbol{\omega}^{(\ell')}\|_1 \geq d/4$ for any pair of distinct elements $\boldsymbol{\omega}, \boldsymbol{\omega}'$ taken

from Ω. With these binary vectors in hand, we define the set $\mathcal{D} \subset \mathbb{B}_+^K$ of cardinality L as follows:

$$\mathcal{D} = \left\{ \boldsymbol{\pi} = (1 - \gamma, \gamma\boldsymbol{\omega}/d) : \quad \omega \in \Omega \right\}. \tag{6.130}$$

It is clear that all the vectors of $\mathcal{D}$ belong to $\mathcal{H}_{\mathcal{F}}(\gamma, 1)$. Let us fix now an element of $\mathcal{D}$ and denote it by $\boldsymbol{\pi}^1$, the corresponding element of Ω being denoted by $\boldsymbol{\omega}^1$. We have

$$\max_{\boldsymbol{\pi} \in \mathcal{D}} \mathrm{KL}(f_{\boldsymbol{\pi}}^n || f_{\boldsymbol{\pi}^1}^n) \leq \frac{nV^2}{2m} \max_{\boldsymbol{\pi} \in \mathcal{D}} \| \boldsymbol{\Sigma}^{1/2}(\boldsymbol{\pi} - \boldsymbol{\pi}^1) \|_2^2 \tag{6.131}$$

$$\leq \frac{nV^2 \lambda_{\max, \boldsymbol{\Sigma}}(2d)\gamma^2}{2md^2} \max_{\boldsymbol{\omega} \in \Omega} \| \boldsymbol{\omega} - \boldsymbol{\omega}^1 \|_2^2 \tag{6.132}$$

$$\leq \frac{nV^2 \lambda_{\max, \boldsymbol{\Sigma}}(2d)\gamma^2}{md}. \tag{6.133}$$

The definition of d yields $(d+1)\sqrt{\log(1 + eK/(d+1))} > C\gamma\sqrt{n}$, which implies that

$$\frac{\gamma^2}{d} \leq 2(d+1)\frac{\gamma^2}{(d+1)^2} \leq 2(d+1)\frac{\log\left(1 + eK/(d+1)\right)}{nC^2} \leq \frac{4d\log\left(1 + e(K-1)/d\right)}{nC^2}. \tag{6.134}$$

Combined with eq. (6.133), this implies that

$$\max_{\boldsymbol{\pi} \in \mathcal{D}} \mathrm{KL}(f_{\boldsymbol{\pi}}^n || f_{\boldsymbol{\pi}^1}^n) \leq \frac{nV^2 \lambda_{\max, \boldsymbol{\Sigma}}(2d)}{m} \times \frac{4d\log\left(1 + e(K-1)/d\right)}{nC^2} \tag{6.135}$$

$$= \frac{4V^2 \lambda_{\max, \boldsymbol{\Sigma}}(2d)}{mC^2} \times d\log\left(1 + e(K-1)/d\right). \tag{6.136}$$

Choosing

$$C^2 = 2 \vee \frac{64V^2 \lambda_{\max, \boldsymbol{\Sigma}}(2d)}{C_1 m}$$

we get that $\max_{\boldsymbol{\pi} \in \mathcal{D}} \mathrm{KL}(f_{\boldsymbol{\pi}}^n || f_{\boldsymbol{\pi}^1}^n) \leq \frac{1}{16}C_1 d\log\left(1 + e(K-1)/d\right) \leq \frac{\log L}{16}$.

Furthermore, for any $\boldsymbol{\pi}, \boldsymbol{\pi}' \in \mathcal{D}$, in view of Lemma 6.7.2 and (6.129), we have

$$\mathrm{KL}(f_{\boldsymbol{\pi}} || f_{\boldsymbol{\pi}'}) \geq \frac{\bar{\kappa}_{\boldsymbol{\Sigma}}(2d, 0)}{4V^2 Md} \| \boldsymbol{\pi} - \boldsymbol{\pi}' \|_1^2 = \frac{\bar{\kappa}_{\boldsymbol{\Sigma}}(2d, 0)\gamma^2}{4V^2 Md^3} \| \boldsymbol{\omega} - \boldsymbol{\omega}' \|_1^2 \tag{6.137}$$

$$\geq \frac{\bar{\kappa}_{\boldsymbol{\Sigma}}(2d, 0)}{64V^2 M} \times \frac{\gamma^2}{d} \tag{6.138}$$

$$\geq \frac{\bar{\kappa}_{\boldsymbol{\Sigma}}(2d, 0)}{32V^2 MC^2} \times \left\{ \frac{\gamma^2}{n} \log\left(1 + \frac{K}{\gamma\sqrt{n}}\right) \right\}^{1/2}. \tag{6.139}$$

Since $\frac{\bar{\kappa}_{\boldsymbol{\Sigma}}(2d,0)}{32V^2MC^2} = 2C_3$, this implies that Theorem 6.7.1 can be applied, which leads to the inequality

$$\inf_{\widehat{f}} \sup_{f \in \mathcal{H}_{\mathcal{F}}(\gamma,1)} \mathbf{P}_f\left(\mathrm{KL}(f||\widehat{f}) \geq C_3\left\{\frac{\gamma^2}{n}\log\left(1 + \frac{K}{\gamma\sqrt{n}}\right)\right\}^{1/2}\right) \geq 0.17. \quad (6.140)$$

To complete the proof of the proposition, we have to consider the case $d > (K-1)/2$. In this case, we can repeat all the previous arguments for $d = K/2$ and get the desired inequality. $\qquad\square$

6.7.3 Lower bound holding for all densities

Now that we have lower bounds in probability for $\mathcal{H}_{\mathcal{F}}(0, D)$ and $\mathcal{H}_{\mathcal{F}}(\gamma, 1)$, we can derive a lower bound in expectation for $\mathcal{H}_{\mathcal{F}}(\gamma, D)$. In particular, to prove Theorem 6.4.1, we will use the inequality

$$\mathcal{R}\big(\mathcal{H}_{\mathcal{F}}(\gamma, D)\big) \geq \inf_{\widehat{f}} \sup_{f^* \in \mathcal{H}_{\mathcal{F}}(0,D) \cup \mathcal{H}_{\mathcal{F}}(\gamma,1)} \mathbf{E}[\mathrm{KL}(f^*||\widehat{f})]. \quad (6.141)$$

Proof of Theorem 6.4.1. To ease notation, let us define

$$r(n, K, \gamma, D) = \left[\frac{\gamma^2}{n}\log\left(1 + \frac{K}{\gamma\sqrt{n}}\right)\right]^{1/2} + \frac{D\log(1 + K/D)}{n} \bigwedge \left(\frac{\log(1 + K/\sqrt{n})}{n}\right)^{1/2}. \quad (6.142)$$

We first consider the case where the dominating term is the first one, that is

$$\left[\frac{\gamma^2}{n}\log\left(1 + \frac{K}{\gamma\sqrt{n}}\right)\right]^{1/2} \geq \frac{3D\log(1 + K/D)}{n}. \quad (6.143)$$

On the one hand, since $D \geq 1$, we have

$$\frac{3D\log(1 + K/D)}{n} \geq \frac{\log(1 + eK)}{n}. \quad (6.144)$$

On the other hand, using the inequality $\log(1 + x) \leq x$, we get

$$\left[\frac{\gamma^2}{n}\log\left(1 + \frac{K}{\gamma\sqrt{n}}\right)\right]^{1/2} \leq \frac{\gamma}{\sqrt{n}}\left[\log(1 + eK) + \log\left(1 + \frac{1}{e^2\gamma^2 n}\right)\right]^{1/2} \quad (6.145)$$

$$\leq \gamma\left[\frac{\log(1 + eK)}{n}\right]^{1/2} + \frac{\gamma}{\sqrt{n}}\left[\frac{1}{e^2\gamma^2 n}\right]^{1/2} \quad (6.146)$$

$$\leq \gamma\left[\frac{\log(1 + eK)}{n}\right]^{1/2} + \frac{\log(1 + eK)}{2n}. \quad (6.147)$$

Combining (6.143), (6.144) and (6.147), we get

$$\left(\frac{\log(1+eK)}{n}\right)^{1/2} \leq 2\gamma. \tag{6.148}$$

This implies that we can apply Proposition 4, which yields

$$\inf_{\widehat{f}} \sup_{f \in \mathcal{H}_{\mathcal{F}}(\gamma,D)} \mathbf{P}_f\left(\mathrm{KL}(f||\widehat{f}) \geq C_3 \left\{\frac{\gamma^2}{n}\log\left(1+\frac{K}{\gamma\sqrt{n}}\right)\right\}^{1/2}\right) \geq 0.17. \tag{6.149}$$

In view of (6.143), this implies that

$$\inf_{\widehat{f}} \sup_{f \in \mathcal{H}_{\mathcal{F}}(\gamma,D)} \mathbf{P}_f\left(\mathrm{KL}(f||\widehat{f}) \geq \frac{3}{4}C_3\, r(n,K,\gamma,D)\right) \geq 0.17. \tag{6.150}$$

We now consider the second case, where the dominating term in the rate is the second one, that is

$$\left[\frac{\gamma^2}{n}\log\left(1+\frac{K}{\gamma\sqrt{n}}\right)\right]^{1/2} \leq \frac{3D\log(1+K/D)}{n} \bigwedge \left(\frac{\log(1+K/\sqrt{n})}{n}\right)^{1/2}. \tag{6.151}$$

In view of Proposition 3, we have

$$\inf_{\widehat{f}} \sup_{f \in \mathcal{H}_{\mathcal{F}}(\gamma,D)} \mathbf{P}_f\left(\mathrm{KL}(f||\widehat{f}) \geq C_2\frac{D\log(1+K/D)}{n} \bigwedge \left(\frac{\log(1+K/\sqrt{n})}{n}\right)^{1/2}\right) \geq 0.17. \tag{6.152}$$

In view of (6.151), we get

$$\inf_{\widehat{f}} \sup_{f \in \mathcal{H}_{\mathcal{F}}(\gamma,D)} \mathbf{P}_f\left(\mathrm{KL}(f||\widehat{f}) \geq \frac{1}{4}C_2\, r(n,K,\gamma,D)\right) \geq 0.17. \tag{6.153}$$

Thus, we have proved that $\log(1+eK) \leq n$ implies that $\inf_{\widehat{f}} \sup_{f \in \mathcal{H}_{\mathcal{F}}(\gamma,D)} \mathbf{P}_f\big(\mathrm{KL}(f||\widehat{f}) \geq C_4\, r(n,K,\gamma,D)\big) \geq 0.17$ for some constant $C_4 > 0$, whatever the relation between γ and D. The desired lower bound follows now from the Tchebychev inequality $\mathbf{E}\big[\mathrm{KL}(f||\widehat{f})\big] \geq C_4\, r(n,K,\gamma,D)\mathbf{P}_f\big(\mathrm{KL}(f||\widehat{f}) \geq C_4\, r(n,K,\gamma,D)\big)$. $\square$

Appendix A: Concentration inequalities

This section contains some well-known results, which are recalled here for the sake of the self-containedness of the paper.

Theorem 6.7.2. *For each $s = 1, \ldots, N$, let $Y_{1,s}, \ldots, Y_{n,s}$ be n independent and zero mean random variables such that for some real numbers a, b we have $\mathbf{P}(Y_{i,s} \in [a, b]) = 1$ for all $i \in [n]$ and $s \in [N]$. Then, we have*

$$\mathbf{E}\Big[\max_{s \in [N]} \frac{1}{n} \sum_{i=1}^{n} Y_{i,s}\Big] \leq (b-a)\Big(\frac{\log N}{2n}\Big)^{1/2}, \quad \mathbf{E}\Big[\max_{s \in [N]} \Big|\frac{1}{n} \sum_{i=1}^{n} Y_{i,s}\Big|\Big] \leq (b-a)\Big(\frac{\log(2N)}{2n}\Big)^{1/2}.$$
(6.154)

Proof. We denote $Z_s = \frac{1}{n} \sum_{i=1}^{n} Y_{i,s}$ for $s = 1, \ldots, N$ and $Z_s = -\frac{1}{n} \sum_{i=1}^{n} Y_{i,s}$ for $s = N+1, \ldots, 2N$. For every $s \in [2N]$, the logarithmic moment generating function $\psi_s(\lambda) = \log \mathbf{E}[e^{\lambda Z_s}]$ satisfies

$$\psi_s(\lambda) = \log\Big(\prod_i \mathbf{E}[e^{\lambda Y_{i,s}/n}]\Big) = \sum_{i=1}^{n} \log \mathbf{E}[e^{\lambda Y_{i,s}/n}] \leq \frac{\lambda^2 (b-a)^2}{8n}, \quad (6.155)$$

where the last inequality is a consequence of the Hoeffding lemma (see, for instance, Lemma 2.2 in [5]). This means that Z_s is sub-Gaussian with variance-factor $\nu = (b-a)^2/4n$. Therefore, Theorem 2.5 from [5] yields $\mathbf{E}[\max_s Z_s] \leq \sqrt{2\nu \log(2N)}$, which completes the proof. $\square$

We group and state together the bounded differences and the Efron-Stein inequalities ([5], Theorems 6.2 and 3.1, respectively).

Theorem 6.7.3. *Assume that a function f satisfies the bounded difference condition: there exist constants c_i, $i = 1, \ldots, n$ such that for all $i = 1, \ldots, n$, all $X = (X_1, \ldots, X_i, \ldots, X_n)$ and $X' = (X_1, \ldots, X_i', \ldots, X_n)$ where only the ith vector is changed*

$$|f(X) - f(X')| \leq c_i. \quad (6.156)$$

Denote

$$\nu = \sum_{i=1}^{n} c_i^2. \quad (6.157)$$

Let $Z = f(X_1, \ldots, X_n)$ where X_i are independent. Then, for every $\delta \in (0, 1)$,

$$\mathbf{P}\Big\{Z \leq \mathbf{E}Z + \Big(\frac{\nu \log(1/\delta)}{2}\Big)^{1/2}\Big\} \geq 1 - \delta, \quad and \quad \mathbf{Var}[Z] \leq \frac{\nu}{2}. \quad (6.158)$$

Next we state the contraction principle of [21]; a proof can be found in ([5], Theorem 11.6).

Theorem 6.7.4. *Let $x_1, \ldots, x_n$ be vectors whose real-valued components are indexed by $\mathcal{T}$, that is, $x_i = (x_{i,s})_{s \in \mathcal{T}}$. For each $i = 1, \ldots, n$ let $\varphi_i : \mathbb{R} \to \mathbb{R}$ be a 1-Lipschitz function such that $\varphi_i(0) = 0$. Let $\epsilon_1, \ldots, \epsilon_n$ be independent*

Rademacher random variables, and let $\Psi : [0, \infty) \to \mathbb{R}$ be a non-decreasing convex function. Then

$$\mathbf{E}\left[\Psi\left(\frac{1}{2}\sup_{s\in\mathcal{T}}\left|\sum_{i=1}^{n}\epsilon_i\varphi_i(x_{i,s})\right|\right)\right] \leq \mathbf{E}\left[\Psi\left(\sup_{s\in\mathcal{T}}\left|\sum_{i=1}^{n}\epsilon_i x_{i,s}\right|\right)\right] \qquad (6.159)$$

$$\mathbf{E}\left[\Psi\left(\sup_{s\in\mathcal{T}}\sum_{i=1}^{n}\epsilon_i\varphi_i(x_{i,s})\right)\right] \leq \mathbf{E}\left[\Psi\left(\sup_{s\in\mathcal{T}}\sum_{i=1}^{n}\epsilon_i x_{i,s}\right)\right]. \qquad (6.160)$$

Chapter 7

Numerical Experiments

In this section we describe the implementation of the algorithm. We also compare its performance with different alternative methods.

7.1 Implementation details

In this section we decribe the implementation of the algorithm described in the previous chapter. But before anything else, we remind the reader the problem setting and the estimator considered.

7.1.1 Problem considered

We observe n independent random vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \mathcal{X}$ drawn from a probability distribution P^* that admits a density function f^*. Given a family of mixture components $f_1, \ldots, f_K$, we assumed that this unknown density is well approximated by a convex combination $f_{\boldsymbol{\pi}}$ of these components

$$f_{\boldsymbol{\pi}}(\boldsymbol{x}) = \sum_{j=1}^{K} \pi_j f_j(\boldsymbol{x}), \quad \boldsymbol{\pi} \in \mathbb{B}_{+}^{K} = \Big\{ \boldsymbol{\pi} \in [0,1]^K : \sum_{j=1}^{K} \pi_j = 1 \Big\}. \qquad (7.1)$$

The component densities $\mathcal{F} = \{f_j : j \in [K]\}$ are assumed to be given by previous experiments or expert knowledge. The problem of construction of this family is an open problem that we try to address in section ?. Our objective is to estimate the weight vector $\boldsymbol{\pi}$ from the simplex $\mathbb{B}_{+}^{K}$ under the sparsity scenario and investigate the statistical properties of the Maximum Likelihood Estimator (MLE), defined by

$$\widehat{\boldsymbol{\pi}} \in \underset{\boldsymbol{\pi} \in \Pi}{\arg\min} \Big\{ -\frac{1}{n} \sum_{i=1}^{n} \log f_{\boldsymbol{\pi}}(\boldsymbol{X}_i) \Big\}, \qquad (7.2)$$

where the minimum is computed over a suitably chosen subset Π of $\mathbb{B}_+^K$.

7.1.2   Implementation

The computation of eq. (7.2) is fairly simple as the objective function and
the set Π are convex and can be easily solved by convex programming. We
wrote the algorithm in Python and used the modeling language CVXPY[13].
We give a dictionnary of densities in input.
Give pseudo code

```python
from sklearn.base import BaseEstimator
import numpy as np
from cvxpy import *


class WeightEstimator(BaseEstimator):

    def __init__(self, densities_dict, select_threshold=10e-3):
        self.densities_dict  = densities_dict
        self.K = len(self.densities_dict)
        self.select_threshold = select_threshold

    def fit(self, X):
        self.F = np.array([self.densities_dict[i].pdf(X) for i in range(sel
        F = np.array([self.densities_dict[i].pdf(X) for i in range(self.K)]
        self.pi = Variable(self.K)
        constraints = [sum_entries(self.pi) == 1, self.pi >= 0, self.pi <=
        objective = Minimize(-sum_entries(log(F * self.pi)))
        prob = Problem(objective, constraints)
        prob.solve()
```

7.2   Alternative methods

7.2.1   SPADES

From [7]

7.2.2   Adaptive Dantzig density estimation

http://www.cmap.polytechnique.fr/ lepennec/l1en.php
    From [4]

7.2.3 Kernel density estimation

The kernel density estimator is a well established non-parametric way of estimating the probability density function of a random variable. We will recall in this section some material about KDE.

Let $X_1, \ldots, X_n$ be i.i.d random variables drawn from an unknown probability density f with respect to the Lebesgue measure on $\mathbb{R}$. The kernel density estimator $\widehat{f}_h$ is given by

$$\widehat{f}_h(x) \triangleq \frac{1}{n} \sum_{i=1}^{n} K\Big(\frac{X_i - x}{h}\Big) \tag{7.3}$$

where $K : \mathbb{R} \to \mathbb{R}$ and $\int K(u)du = 1$ is called a Kernel and h is the bandwidth. We used gaussian kernels and the Shaeter Jones method for the bandwidth selection procedure

7.3 Experimental Evaluations

7.4 Real use case

The code is available on github

Bibliography

[1] O. Banerjee, L. El Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine Learning Research*, 2008.

[2] P. C. Bellec. Optimal exponential bounds for aggregation of density estimators. Technical report, arXiv:1405.3907, May 2014.

[3] Pierre C. Bellec, Arnak S. Dalalyan, Edwin Grappin, and Quentin Paris. On the prediction loss of the lasso in the partially labeled setting. Technical report, arXiv:1606.06179, June 2016.

[4] K. Bertin, E. Le Pennec, and V. Rivoirard. Adaptive Dantzig density estimation. *Ann. Inst. Henri Poincaré Probab. Stat.*, 47(1):43–74, 2011.

[5] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence.* OUP Oxford, 2013.

[6] Florentina Bunea, Alexandre B. Tsybakov, and Marten H. Wegkamp. Aggregation for gaussian regression. *Ann. Statist.*, 35(4):1674–1697, 08 2007.

[7] Florentina Bunea, Alexandre B. Tsybakov, Marten H. Wegkamp, and Adrian Barbu. Spades and mixture models. *Ann. Statist.*, 38(4):2525–2558, 2010.

[8] C. Butucea, J.-F. Delmas, A. Dutfoy, and R. Fischer. Optimal exponential bounds for aggregation of estimators for the Kullback-Leibler loss. Technical report, arXiv:1601.05686, January 2016.

[9] O Catoni. The mixture approach to universal model selection. Technical report, 1997.

[10] Dong Dai, Philippe Rigollet, and Tong Zhang. Deviation optimal learning using greedy Q-aggregation. *Ann. Statist.*, 40(3):1878–1905, 2012.

[11] Arnak S. Dalalyan and Alexandre B. Tsybakov. Mirror averaging with sparsity priors. *Bernoulli*, 18(3):914–944, 2012.

[12] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39, No. 1:1–38, 1977.

[13] Steven Diamond and Stephen Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

[14] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 2007.

[15] A. Juditsky, P. Rigollet, and A. B. Tsybakov. Learning by mirror averaging. *Ann. Statist.*, 36(5):2183–2206, 2008.

[16] Anatoli Juditsky and Arkadii Nemirovski. Functional aggregation for nonparametric regression. *The Annals of Statistics*, 28(3):681–712, 2000.

[17] Vladimir Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems*, volume 2033 of *Lecture Notes in Mathematics*. Springer, 2011. Lectures from the 38th Probability Summer School held in Saint-Flour, 2008.

[18] Guillaume Lecué. Lower bounds and aggregation in density estimation. *J. Mach. Learn. Res.*, 7:971–981, 2006.

[19] Guillaume Lecué. Empirical risk minimization is optimal for the convex aggregation problem. *Bernoulli*, 19(5B):2153–2166, 2013.

[20] Guillaume Lecué and Shahar Mendelson. On the optimality of the empirical risk minimization procedure for the convex aggregation problem. *Ann. Inst. Henri Poincaré Probab. Stat.*, 49(1):288–306, 2013.

[21] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer, Berlin, 1991.

[22] Jonathan Q. Li. *Estimation of Mixture Models*. Phd thesis, Yale University, 1999.

[23] Jonathan Q. Li and Andrew R. Barron. Mixture density estimation. In *Advances in Neural Information Processing Systems 12*, pages 279–285, 1999.

[24] K. Lounici. Generalized mirror averaging and *D*-convex aggregation. *Math. Methods Statist.*, 16(3):246–259, 2007.

[25] R. Mazumder. Topics in sparse multivariate statistics (thesis). 2012.

[26] Alexander Rakhlin, Dmitry Panchenko, and Sayan Mukherjee. Risk bounds for mixture density estimation. *ESAIM Probab. Stat.*, 9:220–229, 2005.

[27] Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Trans. Inform. Theory*, 57(10):6976–6994, 2011.

[28] Ph. Rigollet and A. B. Tsybakov. Linear and convex aggregation of density estimators. *Math. Methods Statist.*, 16(3):260–280, 2007.

[29] Philippe Rigollet. *Oracle inequalities, aggregation and adaptation*. Phd thesis, Université Pierre et Marie Curie - Paris VI, November 2006.

[30] Philippe Rigollet. Kullback-Leibler aggregation and misspecified generalized linear models. *Ann. Statist.*, 40(2):639–665, 2012.

[31] Philippe Rigollet and Alexandre Tsybakov. Exponential screening and optimal rates of sparse estimation. *Ann. Statist.*, 39(2):731–771, 2011.

[32] A.B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, 2009.

[33] Alexandre B. Tsybakov. Optimal rates of aggregation. In *Computational Learning Theory and Kernel Machines, COLT/Kernel, Proceedings*, pages 303–313, 2003.

[34] Alexandre B. Tsybakov. Aggregation and minimax optimality in high-dimensional estimation. In *Proceedings of the International Congress of Mathematicians (Seoul, August 2014)*, volume 3, pages 225–246, 2014.

[35] Sara van de Geer and Peter Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.*, 3:1360–1392, 2009.

[36] Zhan Wang, Sandra Paterlini, Fuchang Gao, and Yuhong Yang. Adaptive minimax regression estimation over sparse ℓ_q-hulls. *J. Mach. Learn. Res.*, 15:1675–1711, 2014.

[37] Dong Xia and Vladimir Koltchinskii. Estimation of low rank density matrices: Bounds in Schatten norms and other distances. *Electron. J. Stat.*, 10(2):2717–2745, 2016.

[38] Yuhong Yang. Mixing strategies for density estimation. *Ann. Statist.*, 28(1):75–87, 2000.

[39] A. B. Yuditskiĭ, A. V. Nazin, A. B. Tsybakov, and N. Vayatis. Recursive aggregation of estimators by the mirror descent method with averaging. *Problemy Peredachi Informatsii*, 41(4):78–96, 2005.