Annales de l'Institut Henri Poincaré - Probabilités et Statistiques
2020, Vol. 56, No. 2, 1391–1418
https://doi.org/10.1214/19-AIHP1007

www.imstat.org/aihp

Parameter recovery in two-component contamination mixtures: The L^2 strategy

Sébastien Gadat[a], Jonas Kahn[b], Clément Marteau[c] and Cathy Maugis-Rabusseau[d]

[a]*Toulouse School of Economics, Université Toulouse 1 – Capitole, 21 allées de Brienne, 31000 Toulouse, France. E-mail: sebastien.gadat@tse-fr.eu*
[b]*Institut de Mathématiques de Toulouse; UMR5219, Université de Toulouse; CNRS, UPS, F-31062 Toulouse Cedex 9, France. E-mail: jonas.kahn@math.univ-toulouse.fr*
[c]*Univ Lyon, Université Claude Bernard Lyon 1, CNRS UMR 5208, Institut Camille Jordan, F-69622 Villeurbanne, France. E-mail: marteau@math.univ-lyon1.fr*
[d]*Institut de Mathématiques de Toulouse; UMR5219, Université de Toulouse; CNRS, INSA, F-31077 Toulouse Cedex 4, France. E-mail: cathy.maugis@insa-toulouse.fr*

Received 20 February 2018; revised 21 November 2018; accepted 27 May 2019

Abstract. In this paper, we consider a parametric density contamination model. We work with a sample of i.i.d. data with a common density, $f^\star = (1-\lambda^\star)\phi + \lambda^\star\phi(\cdot - \mu^\star)$, where the shape ϕ is assumed to be known. We establish the optimal rates of convergence for the estimation of the mixture parameters $(\lambda^\star, \mu^\star) \in (0,1) \times \mathbb{R}^d$. In particular, we prove that the classical parametric rate $1/\sqrt{n}$ cannot be reached when at least one of these parameters is allowed to tend to 0 with n.

Résumé. Dans cet article, nous étudions un modèle de contamination paramétrique. Nous considérons un échantillon i.i.d de densité $f^\star = (1-\lambda^\star)\phi + \lambda^\star\phi(\cdot - \mu^\star)$, où la fonction ϕ est supposée connue. Nous établissons des vitesses de convergence optimales pour l'estimation des paramètres de mélange $(\lambda^\star, \mu^\star) \in (0,1) \times \mathbb{R}^d$. En particulier, nous prouvons que la vitesse paramétrique usuelle $1/\sqrt{n}$ ne peut pas être atteinte quand au moins un de ces paramètres est amené à tendre vers 0 avec n.

MSC: Primary 62G05; 62F15; secondary 62G20

Keywords: $\mathbb{L}^2$ contrast; Parameter estimation; Rate of convergence; Two-component contamination mixture model

1. Introduction

Because of their wide range of flexibility, finite mixtures are a popular tool to model the unknown distribution of heterogeneous data. They are found in several domains and have been at the core of several mathematical investigations. For a complete introduction to mixtures, we refer the reader to [25] and [11]. In most cases of interest, a sample $\mathcal{S}_n := (X_1, \ldots, X_n)$ of i.i.d. data is at our disposal, and each entry admits the probability density $f^\star$ w.r.t. the Lebesgue measure. For a finite mixture model, the density $f^\star$ is assumed to have the following shape:

$$f^\star = \sum_{k=1}^{K} \lambda_k \phi_k. \tag{1.1}$$

With such a representation, the population of interest can in some sense be decomposed into K different groups where each group k has a proportion λ_k and is distributed according to the density ϕ_k. For practical purposes, parametric models are often considered. In such cases, the densities ϕ_k are assumed to be known, at least up to some finite parameters, and the parameter estimation problem is often addressed using an EM-type algorithm [10]. In contrast, with the impressive range of applications based on mixtures, theoretical issues related to mixture models are somewhat poorly understood.

Among the available theoretical results for mixtures, some of them are particularly linked to the density estimation problem. The works [13,14] and [20] develop a nonparametric Bayesian point of view, while exploiting both the approximation capacity of mixtures and their metric entropy size, first with Gaussian distributions and later with exponential

power distributions. A Gaussian mixture estimator based on a non asymptotic penalized likelihood criterion is proposed in [23] and the adaptive properties of this estimator are investigated in [24].

In the mixture models, the focus on the parameters themselves has received less theoretical attention because of their great mathematical difficulty despite their natural interest. It is indeed highly informative to obtain the estimation of the mixing distribution, and many applied works use this estimation for descriptive statistics. Among them, the unsupervised clustering with Bayesian interpretation is certainly one of the most widely used applications of mixtures (see, *e.g.*, [25]). Given a dictionary of densities, [5] propose an estimation procedure based on the minimization of an $\mathbb{L}^2$ empirical criterion with a sparsity constraint, providing an estimation of the parameters of interest when the location parameters $\mu_k^\star$ (here $\phi_k = \phi(\cdot - \mu_k^\star)$) are not too close to each other. [9] studied the estimation of the mixing distribution under a strong identifiability condition. As observed in the recent works of [17,26] and [15], the optimal rate depends on the knowledge of the number of components. [16] show that the parameter estimation rates are slower for some weakly identifiable mixtures. Other extensions are available in [17]. Identifiability (and estimation) issues are discussed in [19] under the assumption that the ϕ_k can be written as $\phi_k = \phi(\cdot - \mu_k)$ for some sequence $(\mu_k)_{k=1..K}$ and a symmetric probability density ϕ.

Finally, the EM algorithm (see, *e.g.*, [10]) is a popular alternative for the analysis of the latent structures involved in the mixture models, but the analysis of the convergence rate of the final estimator is somewhat intricate. A first positive result about the *convergence* of this method is given in [29] when the density is unimodal and certain smoothness conditions hold. However, when multimodality occurs, the behavior of the EM method remains mysterious and is suspected to fall into local traps of the log-likelihood. Some recent advances in the analysis of this famous method were brought by [2], where a general result is given for a convergence of the sample-based EM towards the population one, up to initialization, Lipschitz and concavity conditions.

In this paper, we focus on the multivariate parameter estimation problem when the density of interest is a two-component contamination mixture:

$$f^\star = \big(1 - \lambda^\star\big)\phi + \lambda^\star \phi\big(\cdot - \mu^\star\big),$$

where the density ϕ is *known* and the parameters $(\lambda^\star, \mu^\star) \in (0,1) \times \mathbb{R}^d$ are to be estimated. This model is a particular case of the Huber contamination model ([18]).

The estimation of the couple $(\lambda^\star, \mu^\star)$ has already been considered in the literature. In [4], a slightly different model is considered where $f^\star = (1-\lambda^\star)\phi(\cdot - \mu_1^\star) + \lambda^\star\phi(\cdot - \mu_2^\star)$ and ϕ is assumed to be symmetric and unknown. Using a recurrence procedure based on an inversion formula, they propose an estimator for $\theta^\star = (\lambda^\star, \mu_1^\star, \mu_2^\star)$ and the function ϕ. In particular, the parameter $\lambda^\star$ is estimated at the 'classical' parametric rate $1/\sqrt{n}$, while the rate $n^{-1/4}$ is obtained for location parameters $(\mu_1^\star, \mu_2^\star)$. A similar problem is addressed in [6] where the rate $1/\sqrt{n}$ is reached for the estimation of the whole parameter $\theta^\star$. The estimation procedure is based on a computation of an empirical Fourier transform. More recently, [27] considered the situation where the distribution of one of the component of the mixture is known. In such a case, they provide an estimator of both the mixing parameter and of the distribution of the second component. In the setting considered here (i.e., when $f^\star$ is a two-component contamination mixture), [8] proposes an iterative procedure based on the empirical distribution function. In the so-called *sparse* setting where[1] $\lambda^\star \ll 1/\sqrt{n}$ and $\mu^\star \sim \sqrt{2r\log(n)}$ for some $r \in (0,1)$ as $n \to +\infty$, the authors derive rates of convergence for the estimation of $\lambda^\star$. In particular, they prove that the classical parametric rate cannot be attained in such a setting.

In all the aforementioned contributions except [8], it is implicitly assumed that both location and proportion parameters are fixed with respect to n. The main aim of this paper is to fill this gap. We propose a procedure inspired by [5] and derive an estimator $(\hat\lambda_n, \hat\mu_n)$ for the couple $(\lambda^\star, \mu^\star)$. This estimator is based on the minimization of an $\mathbb{L}^2$ contrast instead of a usual maximum likelihood estimator of mixture parameters computed with an EM-type algorithm. Then, given a bound M s.t. $\max_{j=1\dots d} |\mu_j^\star| \le M$ and under mild assumptions on the shape ϕ, we prove that:

$$\sup_{(\lambda^\star,\mu^\star)\in(0,1)\times[-M,M]^d} \mathbb{E}_{\lambda^\star,\mu^\star}\Big[\big(\lambda^\star\big)^2 \big\|\mu^\star\big\|^2 \big\|\hat\mu_n - \mu^\star\big\|^2\Big] \lesssim \frac{\log^2 n}{n}, \tag{1.2}$$

and

$$\sup_{\substack{(\lambda^\star,\mu^\star)\in(0,1)\times[-M,M]^d \\ \lambda^\star\|\mu^\star\|^2 \gtrsim n^{-1/2}}} \mathbb{E}_{\lambda^\star,\mu^\star}\Big[\big\|\mu^\star\big\|^4 \big(\hat\lambda_n - \lambda^\star\big)^2\Big] \lesssim \frac{\log^2 n}{n}. \tag{1.3}$$

[1] All the notation used in this paper are made precise at the end of this section.

These results are completed by the corresponding lower bounds that ensure the optimality of (1.2) and (1.3), up to logarithmic factors. In particular, we can immediately observe that the parametric rate of $1/\sqrt{n}$ is attained when $\lambda^\star$ and $\mu^\star$ are fixed, but is deteriorated as soon as these parameters are allowed to tend to 0 with n.

Finally, we also obtain an interesting link between the $\mathbb{L}^2$ loss and the Wasserstein loss in our contamination mixture model:

$$\|f_{\lambda,\mu} - f_{\lambda',\mu'}\|_2 \geq c_\phi W_2^2(G_{\lambda,\mu}, G_{\lambda',\mu'}), \tag{1.4}$$

where the Wasserstein (L^p)-transportation distances between two probability measures m_1 and m_2 on Ω are defined by

$$W_p(m_1, m_2) := \left[\inf_{\pi\in\Pi(m_1,m_2)} \int d^p(x, y)\, d\pi(x, y)\right]^{\frac{1}{p}}, \tag{1.5}$$

$\Pi(m_1, m_2)$ being the set of probability measures on $\Omega \times \Omega$ such that their marginals are m_1 and m_2; and $G_{\lambda,\mu} = (1 - \lambda)\delta_0 + \lambda\delta_\mu$ is the mixing distribution associated to the density $f_{\lambda,\mu}$, where δ_θ is the Dirac peak at θ. This makes even more explicit the hardness of recovering the unknown parameters of the contamination mixture model.

The paper is organized as follows. First, a preliminary oracle inequality for $\mathbb{L}^2$ density estimation is established in Section 2. On the basis of this result, some rates of convergence for the estimation of $(\lambda^\star, \mu^\star)$ are deduced (see Section 3.2) under some assumptions on ϕ presented in Section 3.1. Some lower bounds are provided in Section 4, first in a strong contamination model ($\|\mu^\star\| > m$ with m independent of n; see Section 4.1); and second, in a weak contamination model ($\|\mu\|$ can tend to 0 when $n \to +\infty$; see Section 4.2). The main part of the paper ends with a discussion in Section 5 that reveals several insights between Wasserstein distances among mixing distributions and distances between the probability distributions. A few simulations are presented in Section 6. Proofs of the upper bounds (resp. lower bounds) are given in Section 7 (resp. Appendix B) while Section 8 provides the proof of the link between some Wasserstein transportation cost among mixing distributions and the $\mathbb{L}^2$ loss. Technical results are presented in Appendix A.

Notation. Above and below, we use in this paper some specific notation. For any real sequences $(u_n)_{n\in\mathbb{N}}$ and $(v_n)_{n\in\mathbb{N}}$, $u_n \ll v_n$ means that $u_n/v_n \to 0$ as $n \to +\infty$. Similarly, $u_n \sim v_n$ (resp. $u_n \lesssim v_n$ and $u_n \gtrsim v_n$) means that there exists $a, b \in \mathbb{R}^+$ such that $av_n \leq u_n \leq bv_n$ (resp. $u_n \leq bv_n$ and $av_n \leq u_n$) for any $n \in \mathbb{N}$. For any $x \in \mathbb{R}^d$, $\|x\|$ will denote the classical euclidian norm (namely $\|x\|^2 = \sum_{j=1}^d x_j^2$) while $\|f\|_2$ will denote the $\mathbb{L}^2$ norm of any $f \in \mathbb{L}^2(\mathbb{R}^d)$ associated to the corresponding scalar product $\langle\cdot,\cdot\rangle$. Finally, $\mathbb{P}_\theta$ will alternatively (the meaning will be clear following the context) correspond to the measure of a single observation X_i or of the whole sample $(X_1, \ldots, X_n)$ associated to any mixture parameter $\theta = (\lambda, \mu)$. The associated expectation will be alternatively denoted by $\mathbb{E}_\theta$, $\mathbb{E}_{\lambda,\mu}$ or $\mathbb{E}$, according to the context.

2. A preliminary result on $\mathbb{L}^2$ density estimation

2.1. *Statistical setting and identifiability*

We recall that we have at our disposal an i.i.d. sample of size n denoted $\mathcal{S}_n := (X_1, \ldots, X_n)$, where the distribution of each X_i is associated with a two-component contamination mixture model. More precisely, we assume that each X_i admits an unknown density $f^\star$ with respect to the Lebesgue measure on $\mathbb{R}^d$, which is given by:

$$f^\star = (1 - \lambda^\star)\phi + \lambda^\star\phi(\cdot - \mu^\star). \tag{2.1}$$

In the following text, $\theta^\star = (\lambda^\star, \mu^\star) \in (0, 1) \times \mathbb{R}^d$ refers to the *parameters* of the two-component contamination mixture model. We assume that the density ϕ is a *known* function and that a real contamination of this baseline density ϕ occurs ($\lambda^\star > 0$). Finally, we assume that the unknown contamination shift $\mu^\star$ belongs to a bounded interval $[-M, M]^d$ where $M > 0$ is known.

Here and below, for any $\theta = (\lambda, \mu) \in (0, 1) \times \mathbb{R}^d$, we write:

$$f_\theta = f_{\lambda,\mu} = (1 - \lambda)\phi + \lambda\phi_\mu,$$

where ϕ_μ is defined according to the standard notation in location models:

$$\forall \mu \in \mathbb{R}^d \quad \phi_\mu : x \longmapsto \phi(x - \mu).$$

In particular, as a slight abuse of notation, we write $f^\star = f_{\theta^\star} = f_{\lambda^\star,\mu^\star}$ and (when the meaning is clear following the context) $\hat{f} = f_{\hat{\theta}} = f_{\hat{\lambda},\hat{\mu}}$ for any estimator $\hat{\theta}$ of $\theta^\star$.

We aim to recover the unknown parameter $\theta^\star$ from the sample $\mathcal{S}_n$. This might be possible according to the next identifiability result, whose proof is given in Appendix A.

Proposition 2.1. *Any two-component contamination mixture model is identifiable*: $f_{\theta_1} = f_{\theta_2}$ *if and only if* $\theta_1 = \theta_2$.

Such an identifiability result is well known in some more general cases up to additional assumptions on the baseline density ϕ (see, *e.g.*, [19] or Theorem 2.1 of [4] where the symmetry of ϕ is added to ensure the identifiability of the general mixture model without contamination). Here, the fact that one of the components of the mixture is constrained to be centered makes it possible to get rid of any additional assumption on ϕ. In particular, Proposition 2.1 holds as soon as ϕ is non-negative with $\int_{\mathbb{R}^d} \phi = 1$.

2.2. *Estimation strategy and oracle inequality on the* $\mathbb{L}^2$ *norms*

Our estimator will be built according to an optimal $\mathbb{L}^2$ density estimation constrained to the contamination models. For this purpose, we first define a grid over the possible values of λ and μ through:

$$\mathcal{M}_{\Lambda,\mathfrak{M}} := \big\{(\lambda,\mu) : \lambda \in \Lambda = \{\lambda_1,\dots,\lambda_p\} \text{ and } \mu \in \mathfrak{M} = \{\mu_1,\dots,\mu_q\}\big\},$$

where Λ, $\mathfrak{M}$ will depend on n to obtain good properties both from the statistical and approximation point of view. To obtain a good estimation of $f^\star$ and $\theta^\star$, we adopt a SURE approach (see, *e.g.*, [28]) and choose an estimator that minimizes $\|f^\star - f_{\lambda,\mu}\|_2^2$ over the grid $\mathcal{M}_{\Lambda,\mathfrak{M}}$. Observing that:

$$\big\|f^\star - f_{\lambda,\mu}\big\|_2^2 - \big\|f^\star\big\|_2^2 = -2\big\langle f^\star, f_{\lambda,\mu}\big\rangle + \|f_{\lambda,\mu}\|_2^2,$$

and since $\|f^\star\|_2^2$ does not depend on (λ,μ), it is natural to introduce the following contrast function:

$$\forall(\lambda,\mu) \in \mathcal{M}_{\Lambda,\mathfrak{M}} \quad \gamma_n(\lambda,\mu) := -\frac{2}{n}\sum_{i=1}^n f_{\lambda,\mu}(X_i) + \|f_{\lambda,\mu}\|_2^2,$$

leading to the estimator:

$$(\hat\lambda_n,\hat\mu_n) = \arg\min_{(\lambda,\mu)\in\mathcal{M}_{\Lambda,\mathfrak{M}}} \gamma_n(\lambda,\mu). \tag{2.2}$$

Our first main result, stated below, quantifies the performances of $\hat f_n := f_{\hat\lambda_n,\hat\mu_n}$.

Theorem 2.1. *Let* $(\lambda^\star,\mu^\star) \in (0,1)\times\mathbb{R}^d$. *Let* $(\hat\lambda_n,\hat\mu_n)$ *be the estimator defined in* (2.2). *Then, a positive constant* $\mathcal{C}$ *exists such that for all* $0<\alpha<1$:

$$\mathbb{E}\big[\big\|\hat f_n - f^\star\big\|_2^2\big] \le \left(\frac{1+\alpha}{1-\alpha}\right)\inf_{(\lambda,\mu)\in\mathcal{M}_{\Lambda,\mathfrak{M}}}\big\|f_{\lambda,\mu} - f^\star\big\|_2^2 + \frac{\mathcal{C}}{2\alpha}\frac{\log^2(|\mathcal{M}_{\Lambda,\mathfrak{M}}|)}{n}, \tag{2.3}$$

where $|\mathcal{M}_{\Lambda,\mathfrak{M}}|$ *corresponds to the cardinality of the grid* $\mathcal{M}_{\Lambda,\mathfrak{M}}$.

It is worth mentioning that the result above is almost assumption-free on the two-component contamination mixture model. Nevertheless, this result implicitly requires that the approximation term $\inf_{(\lambda,\mu)\in\mathcal{M}_{\Lambda,\mathfrak{M}}}\|f_{\lambda,\mu} - f^\star\|_2^2$ is comparable to the residual. In practice, this cannot be achieved unless we have an upper bound on the range for possible values of μ at our disposal. The proof of Theorem 2.1 is given in Section 7.1.

We stress that Theorem 2.1 is not the main interest of our work. It is a minimal requirement to further extend our analysis on the parameter estimation of the mixture models themselves. In particular, the following question now arises: *does the fact that* $\hat f_n$ *is a "good"* $\mathbb{L}^2$ *estimator of* $f^\star$ *imply that the corresponding* $\hat\theta_n$ *provides a satisfying estimator of* $\theta^\star$*?* The positive answer to this question is the main contribution of our work and is described in the next section. In order to establish this result, some mild restrictions on the class of possible densities ϕ are required.

3. Estimation of the parameter $\theta^\star$

3.1. *Baseline assumptions*

We now introduce mild and sufficient assumptions for an optimal recovery of $\theta^\star$ from the oracle inequality (2.3) (in terms of convergence rates). In the following, we denote by $\mathcal{C}^k(\mathbb{R}^d)$ the set of continuous functions that admits k continuous derivatives.

Assumption ($\mathbf{H_S}$). The density ϕ belongs to $\mathcal{C}^3(\mathbb{R}^d) \cap \mathbb{L}^2(\mathbb{R}^d)$.

The set of admissible densities considered in Assumption ($\mathbf{H_S}$) is very large, and contains many possible distributions (Gaussian, Cauchy, Gamma to name a few). Note that it is also possible to relax the smoothness assumption and handle piecewise differentiable densities with an additional symmetry assumption (see Appendix A). Note that since the density ϕ is continuous and in $\mathbb{L}^2(\mathbb{R}^d)$, this density is necessarily bounded on $\mathbb{R}^d$.

Our second important assumption is concerned with a tight link that may exist between $\phi - \phi_\mu$ and μ itself. It requires a type of Lipschitz upper bound in the translation model.

Assumption ($\mathbf{H_{Lip}}$). The density ϕ satisfies:

$$\exists g \in \mathbb{L}^2\big(\mathbb{R}^d\big)\ \forall x \in \mathbb{R}^d\ \forall \mu \in [-M, M]^d \quad \big|\phi(x) - \phi_\mu(x)\big| \leq \|\mu\| g(x), \tag{3.1}$$

and g satisfies the integrability condition:

$$\mathcal{J} := \int_{\mathbb{R}^d} g^2(x)\phi^{-1}(x)\,dx < +\infty.$$

This assumption will be of primary importance to obtain estimation results on the parameters of the mixture themselves. In particular, it will make it possible to derive a relationship between the $\mathbb{L}^2$ norm of $\phi - \phi_\mu$ and the size of $\|\mu\|$. Hence, under Assumption ($\mathbf{H_{Lip}}$), a good estimation of the density $f^\star$ for the $\mathbb{L}^2$ norm is assumed to yield a good estimation of the mixture parameters.

Remark 3.1. Instead of listing all the possible densities that both meet Assumptions ($\mathbf{H_S}$), ($\mathbf{H_{Lip}}$) (and later ($\mathbf{H_D}$) introduced in Section 4.2 for our lower bound results), we will show that *any log-concave* distribution ϕ written as:

$$\phi(\cdot) = e^{-u(\cdot)} \quad \text{with } u \text{ convex such that } \|\nabla u\| + \big\|D^2 u\big\| = o_\infty(u),$$

satisfies these three conditions.[2] The relationships between ($\mathbf{H_S}$), ($\mathbf{H_{Lip}}$), ($\mathbf{H_D}$) and the log-concave distributions are given in Appendix A.3.

Remark 3.2. An easy consequence of Remark 3.1 (see also Proposition A.2) is that the log-concave Gaussian distributions satisfy assumptions ($\mathbf{H_S}$) and ($\mathbf{H_{Lip}}$) so that all the results displayed below apply to these situations. It may be shown as well that our results apply for the Laplace distribution since the smoothness assumption ($\mathbf{H_S}$) may be replaced by a symmetry property (see Appendix A).

In the 1-dimensional Cauchy distribution case, we can compute $\phi - \phi_\mu$:

$$\big|\phi(x) - \phi_\mu(x)\big| = |\mu| \frac{|2x - \mu|}{\pi[1 + (x-\mu)^2][1 + x^2]} \leq C\phi(x)|\mu|,$$

for a large enough constant C. Hence, the assumptions ($\mathbf{H_S}$) and ($\mathbf{H_{Lip}}$) are satisfied with $g = C\phi$ for the Cauchy distribution.

The skew Gaussian density[3] ϕ satisfies:

$$\big|\phi(x) - \phi_\mu(x)\big| \leq 2\psi(x)\big|\Psi(\alpha x) - \Psi\big(\alpha(x-\mu)\big)\big| + 2\Psi\big(\alpha(x-\mu)\big)\big|\psi(x) - \psi(x-\mu)\big|.$$

[2] Hereafter $o_\infty(u)$ denotes a quantity negligible compared to $u(x)$ as $\|x\| \to +\infty$

[3] It is defined as $\phi(\cdot) = 2\psi(\cdot)\Psi(\alpha\cdot)$ where ψ and Ψ denote respectively the density and cumulative function of a standard Gaussian distribution, and α an asymmetry parameter.

If we define g as $g(x) := 4\sup_{[x-M;x+M]}\psi(t) \times \sup_{[x-M;x+M]}\Psi(\alpha t)$, we can check that $(\mathbf{H}_{\mathcal{S}})$ and $(\mathbf{H}_{\mathbf{Lip}})$ are satisfied. In particular, the integrability condition $(\mathbf{H}_{\mathbf{Lip}})$ is satisfied for large x because $\Psi(\alpha x) \longrightarrow 1$ when $x \longrightarrow +\infty$. Conversely, if $x \longrightarrow -\infty$, we have:

$$\begin{aligned} g^2(x)\phi^{-1}(x) &\lesssim \big[\psi^{-1}(x)\Psi^{-1}(\alpha x)\big] \sup_{[x-M;x+M]} \psi^2(t) \times \sup_{[x-M;x+M]} \Psi^2(\alpha t) \\ &\lesssim \big[\alpha x e^{x^2/2} e^{\alpha^2 x^2/2}\big] e^{-(x-M)^2} \times e^{-\alpha^2(x-M)^2} \big[\alpha(x-M)\big]^{-2} \\ &\lesssim e^{-(x-2M)^2/4} e^{-\alpha^2(x-2M)^2/4}, \end{aligned}$$

which leads to the integrability condition around $-\infty$.

In the following text, we maintain a formalism that uses the two assumptions of Section 3.1 for the sake of generality.

3.2. *Consistency rates on the parameters* $(\lambda^\star, \mu^\star)$

We now use our assumptions on ϕ to deduce some rates of convergence for the estimation of the couple $(\lambda^\star, \mu^\star)$ from the oracle inequality of Theorem 2.1. According to the assumption $\mu^\star \in [-M, M]^d$ for some given $M > 0$, we define the grid $\mathcal{M}_n = \mathcal{M}_{\Lambda,\mathfrak{M}}$ as:

$$\begin{aligned} \mathcal{M}_n = \{(\lambda, \mu) : \lambda = \frac{i}{\sqrt{n}}, \mu = \big(\mu^{(1)}, \dots, \mu^{(d)}\big) \text{ with } \mu^{(j)} = \pm\frac{k_j}{\sqrt{n}} \\ \text{where } i \in \{1, \dots, \sqrt{n}\}, j \in \{1, \dots, d\}, k_j \in \{1, \dots, M\sqrt{n}\}\}, \end{aligned} \tag{3.2}$$

so that the approximation term $\inf_{(\lambda,\mu)\in\mathcal{M}_n} \|f_{\lambda,\mu} - f^\star\|_2^2$ in Equation (2.3) can be made lower than n^{-1}, while keeping the size of $\log(|\mathcal{M}_n|)$ reasonable and of order $d\log(n)$. The next result, whose proof is given in Section 7.2, explicitly gives a non-asymptotic consistency rate of the estimation of $\mu^\star$ in terms of the sample size n, of the amount of contamination $\mu^\star$, and of the probability $\lambda^\star$ of this contamination itself.

Theorem 3.1. *Let $(\hat\lambda_n, \hat\mu_n)$ be the estimator defined in (2.2) with $\mathcal{M}_n$ given in (3.2). If ϕ satisfies Assumptions $(\mathbf{H}_{\mathcal{S}})$ and $(\mathbf{H}_{\mathbf{Lip}})$, a positive constant C_1 exists such that*:

$$\forall n \in \mathbb{N} \quad \sup_{(\lambda^\star,\mu^\star)\in(0,1)\times[-M,M]^d} \mathbb{E}_{\lambda^\star,\mu^\star}\Big[\big(\lambda^\star\|\mu^\star\|\big)^2 \|\hat\mu_n - \mu^\star\|^2\Big] \le \frac{C_1\log^2 n}{n}.$$

In the 1-dimensional case ($d = 1$), an immediate consequence of Theorem 3.1 is that for a *fixed* couple $(\lambda^\star, \mu^\star) \in]0, 1[\times \mathbb{R} \setminus \{0\}$:

$$\mathbb{E}_{\lambda^\star,\mu^\star}\left[\left(\frac{\hat\mu_n}{\mu^\star} - 1\right)^2\right] \le \frac{C_1\log^2 n}{n\{\lambda^\star\}^2\{\mu^\star\}^4}.$$

In particular, since $\mu^\star$ is allowed to tend to 0 with n, the estimator $\hat\mu_n$ will be consistent as soon as $\sqrt{n}\lambda^\star\{\mu^\star\}^2 \to +\infty$ as $n \to +\infty$. In a detection context, a two-component mixture distribution can be distinguished from that of a single component as soon as $\sqrt{n}\lambda^\star|\mu^\star| > \mathcal{C}$ for some positive constant $\mathcal{C}$ (see, e.g., [7] or [21]). Naturally, detection is "easier" than estimation in the sense that the first task requires weaker conditions on the parameters of interest than the second. Since the contamination level $\mu^\star$ is assumed to be upper bounded, it is worth observing that we implicitly require that $\lambda^\star \gg 1/\sqrt{n}$ as $n \to +\infty$.

Before checking the optimality of this result (see Section 4), we investigate the estimation of the contamination proportion $\lambda^\star$. According to the previous discussion, we will assume that $\lambda^\star\|\mu^\star\|^2$ is significantly larger than $n^{-1/2}\log^2 n$. This ensures that the contamination level $\mu^\star$ is consistently estimated. For this purpose, we introduce the set $\Theta_n(M, (\ell_n)_n, \overline{\lambda})$ indexed by a sequence $(\ell_n)_n$:

$$\Theta_n\big(M, (\ell_n)_n, \overline{\lambda}\big) := \left\{\theta = (\lambda, \mu) : \frac{\ell_n}{\|\mu\|^2\sqrt{n}} \le \lambda \le \overline{\lambda}, \|\mu\|_\infty \le M\right\},$$

for some $\bar\lambda \in (0, 1)$.

Theorem 3.2. *If ϕ satisfies Assumptions* $(\mathbf{H}_{\mathcal{S}})$ *and* $(\mathbf{H}_{\mathbf{Lip}})$ *and the sequence $(\ell_n)_n$ is such that $\lim_{n\to+\infty} \frac{\ell_n}{\log n} = +\infty$, then a positive constant C_2 exists such that*:

$$\sup_{(\lambda^\star,\mu^\star)\in\Theta_n(M,(\ell_n)_n,\overline{\lambda})} \mathbb{E}_{\lambda^\star,\mu^\star}\left[\left\|\mu^\star\right\|^4\left(\hat{\lambda}_n-\lambda^\star\right)^2\right] \leq \frac{C_2\log^2 n}{n}.$$

The proof is given in Section 7.3. Once again, we can immediately deduce from this bound that:

$$\mathbb{E}_{\lambda^\star,\mu^\star}\left[\left(\frac{\hat{\lambda}_n}{\lambda^\star}-1\right)^2\right] \leq \frac{C_2\log^2 n}{n\{\lambda^\star\}^2\|\mu^\star\|^4},$$

which only makes sense when $\sqrt{n}\lambda^\star\|\mu^\star\|^2 \to +\infty$ as $n \to +\infty$. We stress that in the particular case of fixed $\lambda^\star$ and $\mu^\star$ (w.r.t. n), these quantities can be estimated at the classical parametric rate of $1/\sqrt{n}$ (up to a logarithmic term).

Remark 3.3. The upper bounds displayed in Theorems 3.1 and 3.2 both involve a $(\log(n))^2$ term. This logarithmic term comes from the oracle inequality in Theorem 2.1 and is related to the complexity of the set, namely $\mathcal{M}_{\Lambda,\mathfrak{M}}$, over which our contrast is minimized. As we will see in the next section, such a term is missing from our lower bound. Up to our knowledge, a logarithmic gap between lower and upper bounds is a classical outcome when dealing with contrast minimization estimators.

4. Lower bounds

We now derive some lower bounds on the estimation of $\lambda^\star$ and $\mu^\star$ and show that our previous results are *minimax optimal* with respect to the values of n, $\lambda^\star$ and $\mu^\star$ up to some $\log^2 n$ terms.

4.1. *Strong contamination model*

For this purpose, we split our study into two cases and first consider the "standard" situation of a strong contamination, meaning that $\|\mu^\star\|$ is bounded from below by a constant independent on n: it translates the fact that the contamination is not negligible when $n \longrightarrow +\infty$. Let m and c be two positive constants, and:

$$\Theta_n(m,c) := \left\{\theta=(\lambda,\mu) : \frac{c}{\|\mu\|^2\sqrt{n}} \leq \lambda, m \leq \|\mu\|\right\}.$$

Note that this still allows a weak effect of contamination since $\lambda^\star$ can be on the order of $n^{-1/2}$. In this case, we obtain the lower bounds that matches (up to a log term) the upper bounds obtained in Theorems 3.1 and 3.2.

Theorem 4.1. *Consider two positive constants m and c such that $0 < \frac{c}{m^2\sqrt{n}} < 1$ so that $\Theta_n(m,c)$ is non empty. A density ϕ that satisfies* $(\mathbf{H}_{\mathcal{S}})$ *and* $(\mathbf{H}_{\mathbf{Lip}})$ *exists such that*:

(i) *a positive constant C_1 exists such that*:

$$\inf_{(\hat{\lambda},\hat{\mu})}\sup_{(\lambda,\mu)\in\Theta_n(m,c)} \mathbb{E}\left[\lambda^2\|\hat{\mu}-\mu\|^2\right] \geq \frac{C_1}{n}, \tag{4.1}$$

(ii) *a positive constant C_2 exists such that*:

$$\inf_{(\hat{\lambda},\hat{\mu})}\sup_{(\lambda,\mu)\in\Theta_n(m,c)} \mathbb{E}\left[(\hat{\lambda}-\lambda)^2\right] \geq \frac{C_2}{n}, \tag{4.2}$$

where the infimum is taken over all estimators $\hat{\theta}=(\hat{\lambda},\hat{\mu})$ in (4.1) and (4.2). The constants C_1 and C_2 depend on c, m and $\mathcal{J}$ (defined in $(\mathbf{H}_{\mathbf{Lip}})$).

Even though the proof relies on a Le Cam argument and leads to a n^{-1} rate, it clearly deserves a careful study for at least two reasons: the loss is asymmetric in (λ,μ) in i) and the balance between λ, μ and n is unclear. We give the proof of this result in Appendix B.2.

4.2. *Weak contamination model*

We now study the situation when the contamination $\|\mu\|$ is not yet bounded from below and can therefore tend to 0 as $n \longrightarrow +\infty$. Let $c > 0$, and:

$$\Theta_n(c) := \left\{\theta = (\lambda, \mu) : \frac{c}{\|\mu\|^2\sqrt{n}} \leq \lambda\right\}.$$

We introduce a sub-class of densities ϕ that satisfy the following assumption:

Assumption ($\mathbf{H_D}$). The density ϕ satisfies:

$$\mathcal{I}_\phi := \sup_{1\leq j\leq d} \int \left\{d_{j,j}\phi(x)\right\}^2 \phi^{-1}(x)\, dx < +\infty, \tag{4.3}$$

where $d_{j,j}$ refers to the second derivative of ϕ with respect to the variable j. Note that Assumption ($\mathbf{H_D}$) is needed for our lower bound results but is not necessary to obtain good estimation properties. However, this assumption is very mild and is again satisfied for many probability distributions as pointed out in Remark 3.1. Moreover, from the minimax paradigm, it is enough to obtain our lower bound results with a restricted subset of densities ϕ.

Theorem 4.2. *An integer $N > 0$ and a function ϕ that satisfies ($\mathbf{H_S}$) and ($\mathbf{H_D}$) exists such that, for all $n > N$:*

(i) *a positive constant C_1 exists such that*:

$$\inf_{(\hat\lambda,\hat\mu)} \sup_{(\lambda,\mu)\in\Theta_n(c)} \mathbb{E}\left[\|\mu\|^4(\lambda - \hat\lambda)^2\right] \geq \frac{C_1}{n}, \tag{4.4}$$

(ii) *a positive constant C_2 exists such that*:

$$\inf_{(\hat\lambda,\hat\mu)} \sup_{(\lambda,\mu)\in\Theta_n(c)} \mathbb{E}\left[\lambda^2\|\mu\|^2\|\mu - \hat\mu\|^2\right] \geq \frac{C_2}{n}, \tag{4.5}$$

where the infimum is taken over all estimators $\hat\theta = (\hat\lambda, \hat\mu)$ in (4.4) and (4.5). The constant C_1 and C_2 depend on c and $\mathcal{I}_\phi$ (defined in ($\mathbf{H_D}$)).

Finally, we should also remark that estimating μ when λ becomes negligible comparing to $n^{-1/2}$ appears to be impossible as pointed out in (ii) of Theorem 4.2.

5. Discussion

5.1. *Related works on distances inequalities and mixture models*

In this paragraph, we provide some additional remarks on the links between several metrics used to describe mixture models in the particular situation of our two-component contamination model. As pointed out in [17] and [15], relating distances between probability distributions on the observations, and Wasserstein distances (defined in (1.5)) on the space of mixture measures is a popular subject of investigation. Of course, it makes sense when we handle some strong-identifiable models as remarked in the cited previous works. We will rely the rates for estimating contamination mixtures to rates for general mixtures. The latter are usually stated in terms of transportation distance between the mixing distributions G. For a contamination mixture, it reads:

$$G_{\lambda,\mu} = (1 - \lambda)\delta_0 + \lambda\delta_\mu, \tag{5.1}$$

where δ_θ is the Dirac peak at θ.

In [17], it is shown that the Total Variation distance denoted $V(f_{\lambda,\mu}, f_{\lambda^\star,\mu^\star})$ between the probability distributions dominates the Wasserstein distance $W_1(G_{\lambda,\mu}, G_{\lambda^\star,\mu^\star})$ when the number of components is known. When it is unknown, but we are only interested in the distance of the estimator to the true distribution, the rate deteriorates to $V(f_{\lambda,\mu}, f_{\lambda^\star,\mu^\star}) \gtrsim W_2^2(G_{\lambda,\mu}, G_{\lambda^\star,\mu^\star})$, under appropriate identifiability conditions.

When we are interested in local minimax rates of convergences, the situation worsens, as proved in [15]. It is shown that the supremum norm between the probability distributions $\|\cdot\|_\infty$ dominates the Wasserstein distance W_{2m-1}^{2m-1} where essentially $2m-1$ is the number of unknown positions to be estimated in the mixture model (the m possible locations and the $m-1$ dimensional weights distribution):

$$\|f_{\lambda,\mu} - f_{\lambda^\star,\mu^\star}\|_\infty \gtrsim W_{2m-1}^{2m-1}(G_{\lambda,\mu}, G_{\lambda^\star,\mu^\star}).$$

The Dvoretzky–Kiefer–Wolfowitz inequality then allows [15] to deduce a $n^{-1/(4m-2)}$ rate of convergence on the parameters.

Notice that for two components, the above speed is in $n^{-1/6}$, whereas our speeds here are in $n^{-1/4}$. This is because the bound by [15] is for generic mixture models, while in this work, we deal with a specific two-component contaminated model. Specifically, in typical cases, the minimax speed for estimating the parameters of mixture models is $n^{-1/2d}$ where d is the number of parameters. The generic two-component model has three parameters, whereas our contamination model has only two.

5.2. *Comparing W_2 and $\|\cdot\|_2$ in a two-component contamination model*

In this work, we have chosen to handle the $\mathbb{L}^2$ distance on probability distributions, instead of V or $\|\cdot\|_\infty$, nevertheless a relationship between $\|\cdot\|_2$ and W_p should exist. The next result essentially states this dependency.

Theorem 5.1. *For any density ϕ that satisfies* $(\mathbf{H}_\mathcal{S})$ *and* $(\mathbf{H_{Lip}})$, *a constant $c_\phi > 0$ exists such that*:

$$\forall(\lambda,\lambda') \in (0,1)^2\ \forall(\mu,\mu') \in [-M,M]^d \quad \|f_{\lambda,\mu} - f_{\lambda',\mu'}\|_2 \geq c_\phi W_2^2(G_{\lambda,\mu}, G_{\lambda',\mu'}).$$

Hence, $\hat{f}_n := f_{\hat{\lambda}_n,\hat{\mu}_n}$ defined by (2.2) satisfies

$$\mathbb{E}_{\lambda^\star,\mu^\star}\big[W_2^4(G_{\hat{\lambda}_n,\hat{\mu}_n}, G_{\lambda^\star,\mu^\star})\big] \lesssim \mathbb{E}\big[\|\hat{f}_n - f_{\lambda^\star,\mu^\star}\|_2^2\big] \lesssim \frac{(\log n)^2}{n}.$$

In other words, the $\mathbb{L}^2$ strategy investigated in this paper allows in fact to control the Wasserstein distance between the estimated mixture distribution $G_{\hat{\lambda}_n,\hat{\mu}_n}$ and the target $G_{\lambda^\star,\mu^\star}$. On the other hand, a lower bound on the minimax rate of convergence in term of the Wasserstein distance may not be directly deduced from our results displayed in Theorems 4.1 or 4.2 because of the lack of symmetry in (λ,μ) with respect to $(\hat{\lambda},\hat{\mu})$.

6. Simulation study

Distributions

In this section, we assess the performance of the $\mathbb{L}^2$-estimator given in (2.2) on four particular cases ($d=1$) of baseline density ϕ. We study the following features:

- Standard Gaussian case with $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$.
- Non-smooth distribution with the Laplace density $\phi(x) = \frac{1}{2}e^{-|x|}$.
- Heavy tailed distribution with the Cauchy density: $\phi(x) = \frac{1}{\pi(1+x^2)}$.
- Asymmetry with the skew Gaussian density: $\phi(x) = 2\psi(x)\Psi(\alpha x)$, where ψ and Ψ, respectively, denote the density and the cumulative function of the standard Gaussian distribution and where α is the asymmetry parameter different from 0 (in the simulations, we fix $\alpha = 10$). This example of asymmetric distributions has been introduced by [1].

Our estimator requires the calculation of the contrast γ_n and, in particular, the value of the $\mathbb{L}^2$ norm:

$$\|f_{\lambda,\mu}\|_2^2 = \big[\lambda^2 + (1-\lambda)^2\big]\|\phi\|_2^2 + 2\lambda(1-\lambda)\langle\phi,\phi_\mu\rangle,$$

that involves the value of inner product $\langle\phi,\phi_\mu\rangle$ for any value of the location parameter $\mu \in [-M,M]$. In the first three examples of distributions, a closed formula exists:

- Gaussian density: $\langle\phi,\phi_\mu\rangle = (4\pi)^{-\frac{1}{2}} \exp[-\frac{1}{4}\mu^2]$
- Laplace density: $\langle\phi,\phi_\mu\rangle = \frac{1}{4}e^{-|\mu|}(1+|\mu|)$

- Cauchy density: $\langle \phi, \phi_\mu \rangle = \frac{2}{\pi(4+\mu^2)}$

Unfortunately, such a formula is not available (to our knowledge) for the skew Gaussian density: there is no analytical expression of $\langle \phi, \phi_\mu \rangle$. Instead, we used a Monte-Carlo procedure to evaluate this quantity for each value of μ in our grid $\mathcal{M}_n$ given in (3.2). To obtain a sufficient approximation of these inner products, we used a number of Monte-Carlo iterations T_{MC} each time of the order $T_{\mathrm{MC}} \propto n^2$ (where n will be the sample size used for our estimation problem).

Statistical setting

We have worked in 1-D with a fixed value of $\lambda^\star = \frac{1}{4}$ while $\mu^\star$ is allowed to vary with n. Below, we used the following relationship between $\mu^\star$ and n:

$$\mu^\star = \sqrt{\frac{1}{\lambda^\star n^\nu}} \quad \text{with } \nu = \frac{\alpha}{24}, \alpha \in \{1, \dots, 24\}.$$

For each value of the parameter $\mu^\star$, we used 10^3 Monte-Carlo simulations to obtain reliable results, while the grid size is determined by fixing the maximal value of the unknown $|\mu^\star|$ as $M = 10$. Finally, we sampled a set of $n = 5000$ observations each time.

In Figure 1, for each case of the mixture model, we represent the evolution of the mean square error for the estimation of $\lambda^\star$ and of $\mu^\star$ when ν varies between $1/24$ and 1:

$$\nu \longmapsto \mathrm{MSE}(\lambda) = \frac{1}{10^3} \sum_{j=1}^{10^3} \big(\hat{\lambda}_j - \lambda^\star\big)^2 \quad \text{and} \quad \nu \longmapsto \mathrm{MSE}(\mu) = \frac{1}{10^3} \sum_{j=1}^{10^3} \big(\hat{\mu}_j - \mu^\star\big)^2.$$

As pointed out in Figure 1, the estimation of $\lambda^\star$ and $\mu^\star$ performs quite well as soon as ν is lower than $1/2$ but becomes completely inconsistent when $\nu > 1/2$, even if we use a sample size of 5000 observations.

We also represent the violin plot of these estimations indicating the same behavior in each particular case (Gaussian and Laplace in Figure 2; Cauchy and skew Gaussian in Figure 3).

Again, a similar conclusion holds: the estimators derived from (2.2) exhibit a low bias and variance when ν is chosen small enough (lower than $1/2$, which corresponds to values greater than 12 in the horizontal axes of Figures 2–3). In contrast, the estimation is seriously damaged for values of ν greater than $1/2$ (which corresponds to values lower than 11 in the horizontal axes of Figures 2–3). Finally, it should be noted that the shape of the density ϕ does not seem to have a big influence on the estimation ability, even though the Cauchy distribution settings may be seen as the most difficult problem (as represented by the green MSE in Figure 1).

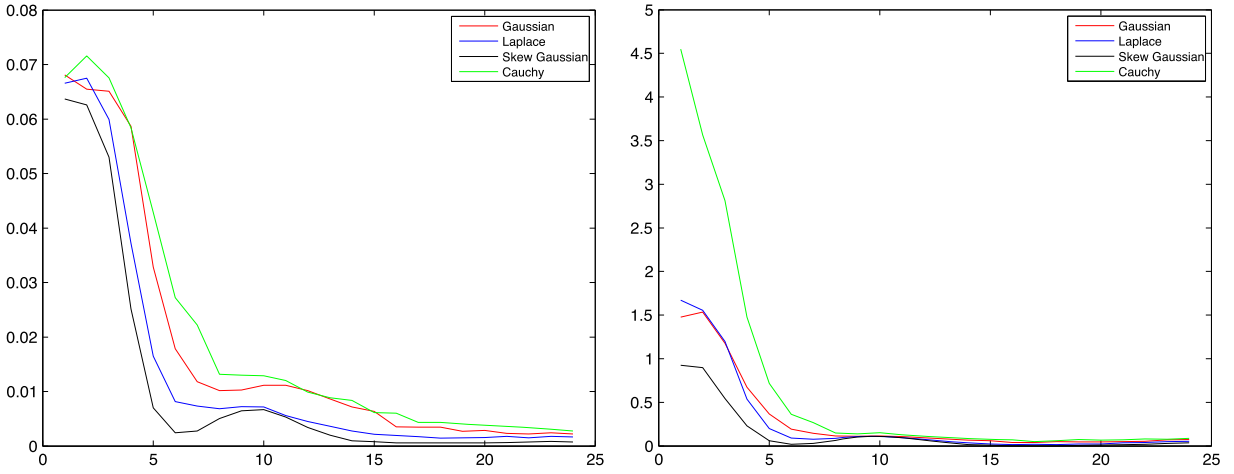

Fig. 1. Mean square error of estimating $\lambda^\star$ (left) and $\mu^\star$ (right) for the 24 values of ν in descending order.

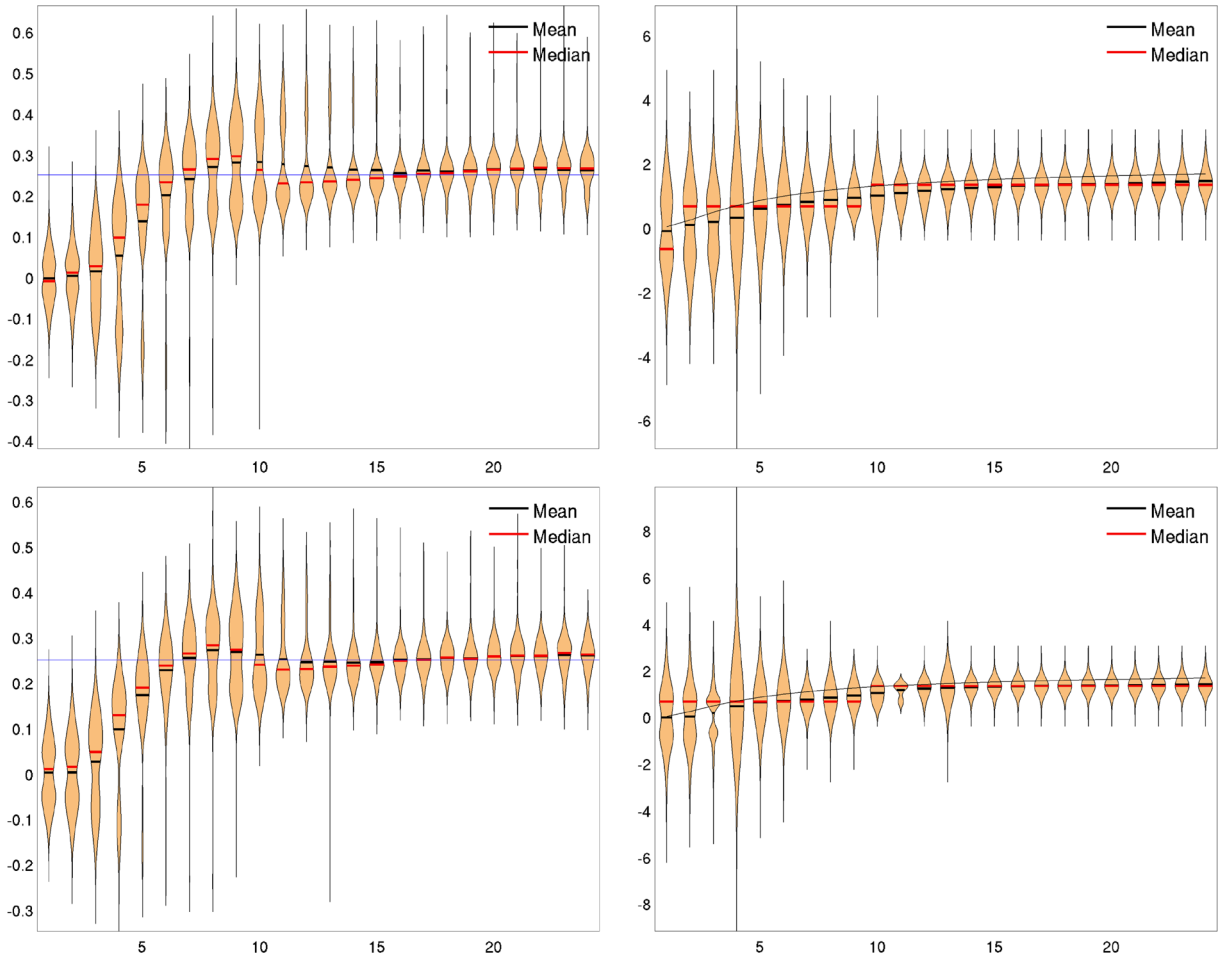

Fig. 2. Evaluation of $\lambda^\star$ (on the left) and $\mu^\star$ (on the right) for our estimators when *Gaussian* mixtures (top) and *Laplace* mixtures (bottom) are considered, for the 24 values of v in descending order.

7. Proofs of the upper bounds

7.1. *Preliminary oracle inequality*

We first establish a technical proposition that will be used to derive the proof of Theorem 2.1. For a given grid $\mathcal{M}_{\Lambda,\mathfrak{M}}$, we first introduce the theoretical minimizer of the $\mathbb{L}^2$-norm on this grid:

$$(\lambda_0, \mu_0) = \arg \min_{(\lambda,\mu)\in\mathcal{M}_{\Lambda,\mathfrak{M}}} \|f_{\lambda,\mu} - f^\star\|_2^2. \tag{7.1}$$

We then define $\mathcal{E}_n(\lambda, \mu)$ the empirical process indexed by $(\lambda, \mu) \in \mathcal{M}_{\Lambda,\mathfrak{M}}$ as:

$$\mathcal{E}_n(\lambda, \mu) = \frac{2}{n}\sum_{i=1}^{n}\big\{f_{\lambda,\mu}(X_i) - f_{\lambda_0,\mu_0}(X_i) - \big[\langle f_{\lambda,\mu} - f_{\lambda_0,\mu_0}, f^\star\rangle\big]\big\}.$$

For all $(\lambda, \mu) \in \mathcal{M}_{\Lambda,\mathfrak{M}}$, the term $\mathcal{E}_n(\lambda, \mu)$ can be rewritten as:

$$\mathcal{E}_n(\lambda, \mu) = \frac{1}{n}\sum_{i=1}^{n}\big(Y_i - \mathbb{E}[Y_i]\big) \quad \text{where } Y_i := 2\big[f_{\lambda,\mu}(X_i) - f_{\lambda_0,\mu_0}(X_i)\big]. \tag{7.2}$$

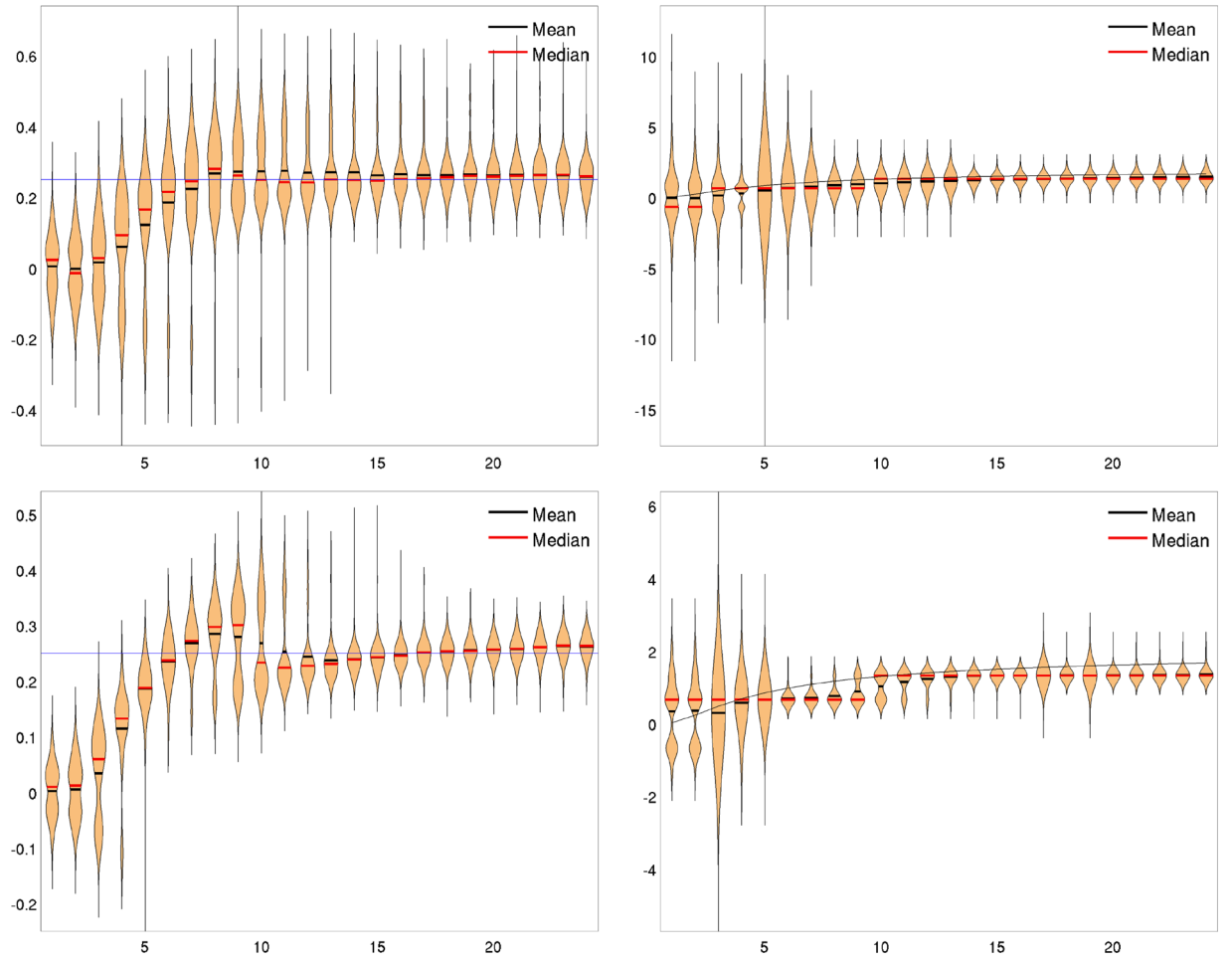

Fig. 3. Evaluation of $\lambda^\star$ (on the left) and $\mu^\star$ (on the right) for our estimators when *Cauchy* mixtures (top) and *skew Gaussian* mixtures (bottom) are considered, for the 24 values of v in descending order.

In particular, $\mathbb{E}[\mathcal{E}_n(\lambda,\mu)] = 0$ and:

$$\begin{aligned}\mathrm{Var}(Y_i) &\le \mathbb{E}\big[Y_i^2\big] = 4\mathbb{E}\big[\big(f_{\lambda,\mu}(X_i) - f_{\lambda_0,\mu_0}(X_i)\big)^2\big] \\ &= 4\int_{\mathbb{R}} \big[f_{\lambda,\mu}(x) - f_{\lambda_0,\mu_0}(x)\big]^2 f^\star(x)\,dx \\ &\le 4\|\phi\|_\infty \|f_{\lambda,\mu} - f_{\lambda_0,\mu_0}\|_2^2,\end{aligned}$$

since $\|f^\star\|_\infty \le \|\phi\|_\infty$. We will use a normalized version of this process below, which naturally leads to the introduction of $\mathcal{G}_n(\lambda,\mu)$:

$$\forall (\lambda,\mu) \in \mathcal{M}_{\Lambda,\mathfrak{M}} \setminus \big\{(\lambda_0,\mu_0)\big\} \quad \mathcal{G}_n(\lambda,\mu) = \frac{\mathcal{E}_n(\lambda,\mu)}{\|f_{\lambda,\mu} - f_{\lambda_0,\mu_0}\|_2}.$$

Our estimator $(\hat\lambda_n, \hat\mu_n)$ defined in (2.2) satisfies the following useful property.

Lemma 7.1.

(i) *For any* (λ,μ) *such that* $\|f_{\lambda,\mu} - f_{\lambda_0,\mu_0}\|_2 \ge n^{-1/2}$:

$$\forall s > 0 \quad \mathbb{P}\big(\big|\mathcal{G}_n(\lambda,\mu)\big| > s\big) \le \exp\left(-\frac{ns^2}{8\|\phi\|_\infty[1 + \frac{s\sqrt{n}}{3}]}\right). \tag{7.3}$$

(ii) *We can find $C > 0$ such that*:

$$\mathbb{E}\big[\mathcal{G}_n^2(\hat{\lambda}_n, \hat{\mu}_n)\mathbb{1}_{\mathcal{B}^c}\big] \leq \frac{C\log^2(|\mathcal{M}_{\Lambda,\mathfrak{M}}|)}{n}, \tag{7.4}$$

where $\mathcal{B}$ is the event defined as $\mathcal{B} = \{\|\hat{f}_n - f_{\lambda_0,\mu_0}\|_2 \leq \frac{1}{\sqrt{n}}\}$.

Proof. In this proof, C refers to a constant that is independent of n, whose value may change from line to line.

Proof of (i): thanks to the Bennett inequality, we obtain for all $s > 0$:

$$\begin{aligned}
&\mathbb{P}\big(\big|\mathcal{G}_n(\lambda,\mu)\big| > s\big) \\
&\quad\leq \exp\bigg(-\frac{n^2 s^2 \|f_{\lambda,\mu} - f_{\lambda_0,\mu_0}\|_2^2}{8n\|\phi\|_\infty \|f_{\lambda,\mu} - f_{\lambda_0,\mu_0}\|_2^2 + 8n\|\phi\|_\infty s\|f_{\lambda,\mu} - f_{\lambda_0,\mu_0}\|_2/3}\bigg), \\
&\quad= \exp\bigg(-\frac{ns^2}{8\|\phi\|_\infty[1 + s\|f_{\lambda,\mu} - f_{\lambda_0,\mu_0}\|_2^{-1}/3]}\bigg).
\end{aligned}$$

Using the fact that $\|f_{\lambda,\mu} - f_{\lambda_0,\mu_0}\|_2 \geq n^{-1/2}$, we obtain:

$$\mathbb{P}\big(\big|\mathcal{G}_n(\lambda,\mu)\big| > s\big) \leq \exp\bigg(-\frac{ns^2}{8\|\phi\|_\infty[1 + \frac{s\sqrt{n}}{3}]}\bigg),$$

which is the desired Inequality (7.3).

Proof of (ii): observe that for all $t > 0$,

$$\begin{aligned}
\mathbb{E}\big[\mathcal{G}_n^2(\hat{\lambda}_n, \hat{\mu}_n)\mathbb{1}_{\mathcal{B}^c}\big] &\leq t^2 + \mathbb{E}\big[\mathcal{G}_n^2(\hat{\lambda}_n, \hat{\mu}_n)\mathbb{1}_{\{|\mathcal{G}_n(\hat{\lambda}_n,\hat{\mu}_n)|>t\}}\mathbb{1}_{\mathcal{B}^c}\big], \\
&\leq t^2 + \mathbb{E}\Big[\sup_{(\lambda,\mu):\|f_{\lambda,\mu} - f_{\lambda_0,\mu_0}\| \geq n^{-1/2}} \big\{\mathcal{G}_n^2(\lambda,\mu)\mathbb{1}_{\{|\mathcal{G}_n(\lambda,\mu)|>t\}}\big\}\Big], \\
&\leq t^2 + \sum_{(\lambda,\mu):\|f_{\lambda,\mu} - f_{\lambda_0,\mu_0}\| \geq n^{-1/2}} \mathbb{E}\big[\mathcal{G}_n^2(\lambda,\mu)\mathbb{1}_{\{|\mathcal{G}_n(\lambda,\mu)|>t\}}\big].
\end{aligned} \tag{7.5}$$

Integrating by parts, we can remark that:

$$\mathbb{E}\big[\mathcal{G}_n^2(\lambda,\mu)\mathbb{1}_{\{|\mathcal{G}_n(\lambda,\mu)|>t\}}\big] = t^2\mathbb{P}\big(\big|\mathcal{G}_n(\lambda,\mu)\big| > t\big) + \int_{t^2}^{+\infty} \mathbb{P}\big(\big|\mathcal{G}_n(\lambda,\mu)\big| > \sqrt{x}\big)\,dx.$$

Thus, if we choose $t = (\frac{16\|\phi\|_\infty \log(|\mathcal{M}_{\Lambda,\mathfrak{M}}|)}{3} \vee 3)n^{-1/2}$, then $t\sqrt{n}/3 \geq 1$, so that for any $s \geq t$ and for a fixed (λ,μ), (7.3) yields:

$$\begin{aligned}
\mathbb{E}\big[\mathcal{G}_n^2(\lambda,\mu)\mathbb{1}_{\{|\mathcal{G}_n(\lambda,\mu)|>t\}}\big] &\leq t^2\exp\big(-\log\big(|\mathcal{M}_{\Lambda,\mathfrak{M}}|\big)\big) + \int_{t^2}^{+\infty} \exp\bigg(-\frac{3\sqrt{nx}}{16\|\phi\|_\infty}\bigg)\,dx \\
&\leq C\frac{\log^2(|\mathcal{M}_{\Lambda,\mathfrak{M}}|)}{n} \times \frac{1}{|\mathcal{M}_{\Lambda,\mathfrak{M}}|} + 2\int_t^{+\infty} u\exp\bigg(-\frac{3\sqrt{n}u}{16\|\phi\|_\infty}\bigg)\,du,
\end{aligned}$$

for large enough C, where the last line comes from the size of t for the left-hand side, and from the change of variable $u = \sqrt{x}$ in the integral. The remaining integral may be integrated by parts, which in turn leads to:

$$\mathbb{E}\big[\mathcal{G}_n^2(\lambda,\mu)\mathbb{1}_{\{|\mathcal{G}_n(\lambda,\mu)|>t\}}\big] \leq C\frac{\log^2(|\mathcal{M}_{\Lambda,\mathfrak{M}}|)}{n} \times \frac{1}{|\mathcal{M}_{\Lambda,\mathfrak{M}}|}.$$

If we plug the above upper bound into (7.5), we then obtain that a sufficiently large constant C exists such that:

$$\mathbb{E}\big[\mathcal{G}_n^2(\hat{\lambda}_n, \hat{\mu}_n)\mathbb{1}_{\mathcal{B}^c}\big] \leq C\frac{\log^2(|\mathcal{M}_{\Lambda,\mathfrak{M}}|)}{n} \times \frac{|\mathcal{M}_{\Lambda,\mathfrak{M}}|}{|\mathcal{M}_{\Lambda,\mathfrak{M}}|} = C\frac{\log^2(|\mathcal{M}_{\Lambda,\mathfrak{M}}|)}{n}.$$

□

We are now interested in the proof of the oracle inequality.

Proof of Theorem 2.1. The best approximation term (λ_0, μ_0) over the grid $\mathcal{M}_{\Lambda,\mathfrak{M}}$ is defined in (7.1) and the event $\mathcal{B} = \{\|\hat{f}_n - f_{\lambda_0,\mu_0}\|_2 \leq \sqrt{\frac{1}{n}}\}$ is introduced in Proposition 7.1. On the event $\mathcal{B}$, the situation is easy using the Young inequality $2ab \leq \alpha a^2 + \alpha^{-1} b^2$ so that for all $\alpha > 0$,

$$\begin{aligned}\mathbb{E}\big[\big\|\hat{f}_n - f^\star\big\|_2^2 \mathbb{1}_{\mathcal{B}}\big] &\leq (1+\alpha)\big\|f_{\lambda_0,\mu_0} - f^\star\big\|_2^2 + \big(1+\alpha^{-1}\big)\mathbb{E}\big[\|\hat{f}_n - f_{\lambda_0,\mu_0}\|_2^2 \mathbb{1}_{\mathcal{B}}\big],\\ &\leq (1+\alpha)\big\|f_{\lambda_0,\mu_0} - f^\star\big\|_2^2 + \frac{1+\alpha^{-1}}{n}. \end{aligned} \tag{7.6}$$

We provide below a similar control on the event $\mathcal{B}^c$. First, observe that according to the definition of $(\hat{\lambda}_n, \hat{\mu}_n)$, for all $(\lambda, \mu) \in \mathcal{M}_{\Lambda,\mathfrak{M}}$, we have:

$$\begin{aligned}&\gamma_n(\hat{\lambda}_n, \hat{\mu}_n) + \big\|f^\star\big\|_2^2 \leq \gamma_n(\lambda,\mu) + \big\|f^\star\big\|_2^2,\\ &\Leftrightarrow \quad \big\|\hat{f}_n - f^\star\big\|_2^2 \leq \big\|f_{\lambda,\mu} - f^\star\big\|_2^2 + 2\left[\frac{1}{n}\sum_{i=1}^n \hat{f}_n(X_i) - \big\langle \hat{f}_n, f^\star\big\rangle\right] - 2\left[\frac{1}{n}\sum_{i=1}^n f_{\lambda,\mu}(X_i) - \big\langle f_{\lambda,\mu}, f^\star\big\rangle\right].\end{aligned}$$

This inequality being true for $(\lambda, \mu) = (\lambda_0, \mu_0)$, we obtain:

$$\big\|\hat{f}_n - f^\star\big\|_2^2 \mathbb{1}_{\mathcal{B}^c} \leq \big\|f_{\lambda_0,\mu_0} - f^\star\big\|_2^2 + \mathcal{E}_n(\hat{\lambda}_n, \hat{\mu}_n)\mathbb{1}_{\mathcal{B}^c}.$$

This implies that for all $0 < \alpha < 1$:

$$\begin{aligned}&\big\|\hat{f}_n - f^\star\big\|_2^2 \mathbb{1}_{\mathcal{B}^c} \leq \big\|f_{\lambda_0,\mu_0} - f^\star\big\|_2^2 + \|\hat{f}_n - f_{\lambda_0,\mu_0}\|_2 \frac{\mathcal{E}_n(\hat{\lambda}_n, \hat{\mu}_n)}{\|\hat{f}_n - f_{\lambda_0,\mu_0}\|_2}\mathbb{1}_{\mathcal{B}^c},\\ &\Rightarrow \quad \big\|\hat{f}_n - f^\star\big\|_2^2 \mathbb{1}_{\mathcal{B}^c} \leq \big\|f_{\lambda_0,\mu_0} - f^\star\big\|_2^2 + \frac{\alpha}{2}\|\hat{f}_n - f_{\lambda_0,\mu_0}\|_2^2 \mathbb{1}_{\mathcal{B}^c} + \frac{1}{2\alpha}\mathcal{G}_n^2(\hat{\lambda}_n, \hat{\mu}_n)\mathbb{1}_{\mathcal{B}^c}.\end{aligned}$$

Using $\|u+v\|^2 \leq 2\|u\|^2 + 2\|v\|^2$, we then deduce that:

$$\big\|\hat{f}_n - f^\star\big\|_2^2 \mathbb{1}_{\mathcal{B}^c} \leq \frac{(1+\alpha)}{(1-\alpha)}\big\|f_{\lambda_0,\mu_0} - f^\star\big\|_2^2 + \frac{1}{2\alpha}\mathcal{G}_n^2(\hat{\lambda}_n, \hat{\mu}_n)\mathbb{1}_{\mathcal{B}^c}. \tag{7.7}$$

We can conclude the proof taking (7.4) in (7.7), and (7.6) together. □

7.2. *Proof of Theorem 3.1*

We aim to apply the oracle inequality established in Theorem 2.1. First, we need an upper bound on the approximation term given by $\|f_{\lambda_0,\mu_0} - f^\star\|_2^2$ when (λ_0, μ_0) belongs to our grid $\mathcal{M}_n$. We can observe that for all $(\lambda, \mu) \in (0,1) \times \mathbb{R}^d$,

$$\begin{aligned}\big\|f_{\lambda,\mu} - f^\star\big\|_2^2 &= \big\|(1-\lambda)\phi + \lambda\phi_\mu - \big(1-\lambda^\star\big)\phi - \lambda^\star\phi_{\mu^\star}\big\|_2^2\\ &= \big\|\big(\lambda^\star - \lambda\big)\{\phi - \phi_\mu\} + \lambda^\star\{\phi_\mu - \phi_{\mu^\star}\}\big\|_2^2\\ &\leq 2\big(\lambda^\star - \lambda\big)^2\|\phi - \phi_\mu\|_2^2 + 2\big\{\lambda^\star\big\}^2\|\phi_\mu - \phi_{\mu^\star}\|_2^2.\end{aligned} \tag{7.8}$$

Using Proposition A.1, we can find two positive constants $\overline{\kappa}$ and $\underline{\kappa}$ such that:

$$\forall(\mu, \tilde{\mu}) \in \mathbb{R}^d \times \mathbb{R}^d \quad \underline{\kappa}\|\mu - \tilde{\mu}\|^2 \leq \|\phi_\mu - \phi_{\tilde{\mu}}\|_2^2 \leq \overline{\kappa}\|\mu - \tilde{\mu}\|^2, \tag{7.9}$$

which in turn implies that:

$$\big\|f_{\lambda,\mu} - f^\star\big\|_2^2 \leq 8\|\phi\|_2^2\big(\lambda^\star - \lambda\big)^2 + 2\overline{\kappa}\big\{\lambda^\star\big\}^2\big\|\mu - \mu^\star\big\|^2.$$

In particular, the definition of $\mathcal{M}_n$ given in (3.2) makes it possible to find a constant $C > 0$ such that:

$$\left\| f_{\lambda_0,\mu_0} - f^\star \right\|_2^2 = \inf_{(\lambda,\mu)\in\mathcal{M}_n} \left\| f_{\lambda,\mu} - f^\star \right\|_2^2 \leq \frac{C}{n}. \tag{7.10}$$

At the same time, observe that (7.8) leads to:

$$\left\| \hat{f}_n - f^\star \right\|_2^2 = \left(\lambda^\star - \hat{\lambda}_n\right)^2 \|\phi - \phi_{\hat{\mu}_n}\|_2^2 + \left\{\lambda^\star\right\}^2 \|\phi_{\hat{\mu}_n} - \phi_{\mu^\star}\|_2^2 + 2\left(\lambda^\star - \hat{\lambda}_n\right)\lambda^\star \langle \phi - \phi_{\hat{\mu}_n}, \phi_{\hat{\mu}_n} - \phi_{\mu^\star}\rangle.$$

Then the following lemma (proved in [12]), which can be viewed as a refinement of the Cauchy–Schwarz inequality, is required.

Lemma 7.2. *If ϕ satisfies $(\mathbf{H}_{\mathcal{S}})$ and $(\mathbf{H}_{\mathbf{Lip}})$, then a constant $c > 0$ exists such that $\forall (a, b) \in \mathbb{R}^d \times \mathbb{R}^d$:*

$$\left|\langle \phi - \phi_a, \phi_{a+b} - \phi_a\rangle\right| \leq \|\phi - \phi_a\|_2 \|\phi_{a+b} - \phi_a\|_2 \left(1 - c\|\phi - \phi_{a+b}\|_2^2\right). \tag{7.11}$$

Using Lemma 7.2 with $a = \hat{\mu}_n$ and $b = \mu^\star - \hat{\mu}_n$ and (7.9), a positive constant c exists such that:

$$\begin{aligned}
&\left\| \hat{f}_n - f^\star \right\|_2^2 \\
&\quad \geq \left(\lambda^\star - \hat{\lambda}_n\right)^2 \|\phi - \phi_{\hat{\mu}_n}\|_2^2 + \left\{\lambda^\star\right\}^2 \|\phi_{\hat{\mu}_n} - \phi_{\mu^\star}\|_2^2 \\
&\qquad - 2\left|\lambda^\star - \hat{\lambda}_n\right|\lambda^\star \|\phi - \phi_{\hat{\mu}_n}\|_2 \|\phi_{\hat{\mu}_n} - \phi_{\mu^\star}\|_2 \left(1 - c\|\phi - \phi_{\mu^\star}\|_2^2\right) \\
&\quad \geq \left(\lambda^\star - \hat{\lambda}_n\right)^2 \|\phi - \phi_{\hat{\mu}_n}\|_2^2 + \left\{\lambda^\star\right\}^2 \|\phi_{\hat{\mu}_n} - \phi_{\mu^\star}\|_2^2 \\
&\qquad - \left[\left(\lambda^\star - \hat{\lambda}_n\right)^2 \|\phi - \phi_{\hat{\mu}_n}\|_2^2 + \left\{\lambda^\star\right\}^2 \|\phi_{\hat{\mu}_n} - \phi_{\mu^\star}\|_2^2\right]\left(1 - c\|\phi - \phi_{\mu^\star}\|_2^2\right) \\
&\quad \geq c\left(\lambda^\star - \hat{\lambda}_n\right)^2 \|\phi - \phi_{\hat{\mu}_n}\|_2^2 \|\phi - \phi_{\mu^\star}\|_2^2 + c\left\{\lambda^\star\right\}^2 \|\phi_{\hat{\mu}_n} - \phi_{\mu^\star}\|_2^2 \|\phi - \phi_{\mu^\star}\|_2^2.
\end{aligned}$$

We then obtained the crucial inequality:

$$\left\| \hat{f}_n - f^\star \right\|_2^2 \geq c\underline{\kappa}^2 \left(\lambda^\star - \hat{\lambda}_n\right)^2 \|\hat{\mu}_n\|^2 \left\|\mu^\star\right\|^2 + c\underline{\kappa}^2 \left\{\lambda^\star\right\}^2 \left\|\mu^\star\right\|^2 \left\|\hat{\mu}_n - \mu^\star\right\|^2. \tag{7.12}$$

We see here the central role of the refinement of the Cauchy–Schwarz inequality to obtain a tractable bound that involves the parameters of the mixture themselves, from the bound on the $\mathbb{L}^2$-norm of $\hat{f}_n - f^\star$. We now use the oracle inequality on $\|\hat{f}_n - f^\star\|_2^2$ to deduce that a constant $C > 0$ exists such that:

$$\mathbb{E}\left[\left(\lambda^\star - \hat{\lambda}_n\right)^2 \|\hat{\mu}_n\|^2 \left\|\mu^\star\right\|^2 + \left\{\lambda^\star\right\}^2 \left\|\mu^\star\right\|^2 \left\|\hat{\mu}_n - \mu^\star\right\|^2\right] \leq \frac{C \log^2 n}{n}. \tag{7.13}$$

In particular, we immediately deduce from (7.13) that:

$$\mathbb{E}\left[\left\{\lambda^\star\right\}^2 \left\|\mu^\star\right\|^2 \left\|\hat{\mu}_n - \mu^\star\right\|^2\right] \leq \frac{C \log^2 n}{n}.$$

This result is uniform in $(\lambda^\star, \mu^\star)$, we obtain the proof of Theorem 3.1.

Unfortunately, we cannot directly use a similar approach for the estimation of $\lambda^\star$. Indeed, we have to first ensure that $\hat{\mu}_n$ is close to $\mu^\star$ with a large enough probability.

7.3. *Proof of Theorem 3.2*

Let $\mathcal{B}$ and $\mathcal{D}$ be the events respectively defined as:

$$\mathcal{B} = \left\{ \|\hat{f}_n - f_{\lambda_0,\mu_0}\|_2 \leq \sqrt{\frac{1}{n}} \right\} \tag{7.14}$$

and

$$\mathcal{D} = \left\{ \left|\mathcal{G}_n(\hat{\lambda}_n, \hat{\mu}_n)\right| \leq \frac{16\|\phi\|_\infty \log(n|\mathcal{M}_n|)}{3\sqrt{n}} \right\}. \tag{7.15}$$

Below, the control of the quadratic risk of $\hat{\mu}_n$ will be investigated according to the partition $\mathcal{B}, \mathcal{B}^c \cap \mathcal{D}$ and $\mathcal{B}^c \cap \mathcal{D}^c$.

Control of the risk on $\mathcal{B}$: Equation (7.6) together with (7.10) indicates that:

$$\|\hat{f}_n - f^\star\|_2^2 \mathbb{1}_{\mathcal{B}} \leq \frac{C}{n}.$$

Then, Equation (7.12) implies that:

$$\|\hat{\mu}_n - \mu^\star\|^2 \mathbb{1}_{\mathcal{B}} \leq \frac{C}{n\{\lambda^\star\}^2 \|\mu^\star\|^2} \leq \frac{C\|\mu^\star\|^2}{\ell_n^2}. \tag{7.16}$$

Control of the risk on $\mathcal{B}^c \cap \mathcal{D}$: On the set $\mathcal{B}^c \cap \mathcal{D}$, we apply Inequality (7.7), which yields:

$$\begin{aligned}\|\hat{f}_n - f^\star\|_2^2 \mathbb{1}_{\mathcal{B}^c \cap \mathcal{D}} &\leq \frac{(1+\alpha)}{(1-\alpha)} \|f_{\lambda_0,\mu_0} - f^\star\|_2^2 + \frac{1}{2\alpha} |\mathcal{G}_n(\hat{\lambda}_n, \hat{\mu}_n)|^2 \mathbb{1}_{\mathcal{B}^c \cap \mathcal{D}} \\ &\leq C \frac{\log^2(n|\mathcal{M}_n|)}{n}\end{aligned}$$

for some positive constant C. Since the size of $|\mathcal{M}_{\Lambda_n, \mathfrak{M}_n}|$ is a polynomial of n, we can find a constant C such that Equation (7.12) leads to:

$$\|\hat{\mu}_n - \mu^\star\|^2 \mathbb{1}_{\mathcal{B}^c \cap \mathcal{D}} \leq C \frac{\log^2 n}{n\{\lambda^\star\}^2 \|\mu^\star\|^2} \leq C \frac{\log^2 n}{\ell_n^2} \|\mu^\star\|^2. \tag{7.17}$$

Since we assume that $(\lambda^\star, \mu^\star) \in \Theta_n(M, (\ell_n)_n, \overline{\lambda})$ with $\ell_n / \log n \longrightarrow +\infty$ when $n \longrightarrow +\infty$, Equations (7.16) and (7.17) imply that for large enough n,

$$\|\hat{\mu}_n - \mu^\star\|^2 [\mathbb{1}_{\mathcal{B}} + \mathbb{1}_{\mathcal{B}^c \cap \mathcal{D}}] \leq \frac{\|\mu^\star\|^2}{4}.$$

Remark that for any x and y: $\|x - y\| \leq \frac{\|y\|}{2}$ implies that $\|y\| \geq 2\|y\| - 2\|x\|$ (using the triangle inequality), which in turns yields $\|y\| \leq 2\|x\|$. Applying this simple remark to the former inequality leads to:

$$\|\mu^\star\|^2 [\mathbb{1}_{\mathcal{B}} + \mathbb{1}_{\mathcal{B}^c \cap \mathcal{D}}] \leq 4\|\hat{\mu}_n\|^2 [\mathbb{1}_{\mathcal{B}} + \mathbb{1}_{\mathcal{B}^c \cap \mathcal{D}}]. \tag{7.18}$$

Control of the risk on $\mathcal{B}^c \cap \mathcal{D}^c$: Applying (7.3) we can check that:

$$\mathbb{P}\big(\mathcal{B}^c \cap \mathcal{D}^c\big) \leq \mathbb{P}\big(\mathcal{D}^c\big) \leq \frac{C}{n}$$

for some positive constant C.

Synthesis: Using (7.18), a large enough N exists such that for $n \geq N$:

$$\begin{aligned}\mathbb{E}\big[\big(\hat{\lambda}_n - \lambda^\star\big)^2 \|\mu^\star\|^4\big] &= \mathbb{E}\big[\big(\hat{\lambda}_n - \lambda^\star\big)^2 \|\mu^\star\|^4 (\mathbb{1}_{\mathcal{B}} + \mathbb{1}_{\mathcal{B}^c \cap \mathcal{D}})\big] + \mathbb{E}\big[\big(\hat{\lambda}_n - \lambda^\star\big)^2 \|\mu^\star\|^4 \mathbb{1}_{\mathcal{B}^c \cap \mathcal{D}^c}\big], \\ &\leq 4\mathbb{E}\big[\big(\hat{\lambda}_n - \lambda^\star\big)^2 \|\mu^\star\|^2 \|\hat{\mu}_n\|^2\big] + d^2 M^4 \mathbb{P}\big(\mathcal{D}^c\big), \\ &\leq \frac{C \log^2(n)}{n},\end{aligned}$$

for some constant $C > 0$, according to (7.13). This result being uniform in $(\lambda^\star, \mu^\star)$, we obtain the proof of Theorem 3.2.

8. Link between the $\|\cdot\|_2$ norm and the Wasserstein distance(s)

Proof of Theorem 5.1. Below, we will establish that the following inequality (stated in Theorem 5.1) holds:

$$W_2^4(G_{\lambda,\mu}, G_{\lambda',\mu'}) \lesssim \|f_{\lambda,\mu} - f_{\lambda',\mu'}\|_2^2. \tag{8.1}$$

Expression of W_2: Since the role played by (λ, μ) and (λ', μ') is symmetric, in the following, we assume without loss of generality that $\lambda \leq \lambda'$. After some calculations (see [12] for more details), it yields

$$W_2^2(G_{\lambda,\mu}, G_{\lambda',\mu'}) = \begin{cases} (\lambda' - \lambda)\|\mu'\|^2 + \lambda\|\mu - \mu'\|^2 & \text{if } \|\mu\|^2 + \|\mu'\|^2 \geq \|\mu - \mu'\|^2, \\ \lambda\|\mu\|^2 + \lambda'\|\mu'\|^2 & \text{if } \|\mu\|^2 + \|\mu'\|^2 < \|\mu - \mu'\|^2 \text{ and } \lambda + \lambda' \leq 1, \\ (1 - \lambda')\|\mu\|^2 + (1 - \lambda)\|\mu'\|^2 & \\ \quad + (\lambda + \lambda' - 1)\|\mu - \mu'\|^2 & \text{if } \|\mu\|^2 + \|\mu'\|^2 < \|\mu - \mu'\|^2 \text{ and } \lambda + \lambda' > 1. \end{cases} \tag{8.2}$$

Upper bound on W_2: The previous expression for $W_2(G_{\lambda,\mu}, G_{\lambda',\mu'})$ allows to prove that

$$W_2^2(G_{\lambda,\mu}, G_{\lambda',\mu'}) \leq (\lambda' - \lambda)\|\mu'\|^2 + \lambda\|\mu - \mu'\|^2. \tag{8.3}$$

Indeed, according to (8.2), this bound turns to be an equality when $\|\mu\|^2 + \|\mu'\|^2 \geq \|\mu - \mu'\|^2$. When, $\|\mu\|^2 + \|\mu'\|^2 < \|\mu - \mu'\|^2$ and $\lambda + \lambda' \leq 1$, we have

$$\begin{aligned} W_2^2(G_{\lambda,\mu}, G_{\lambda',\mu'}) &= (\lambda' - \lambda)\|\mu'\|^2 + \lambda\|\mu - \mu'\|^2 + \lambda\big(\|\mu'\|^2 + \|\mu\|^2 - \|\mu - \mu'\|^2\big) \\ &\leq (\lambda' - \lambda)\|\mu'\|^2 + \lambda\|\mu - \mu'\|^2. \end{aligned}$$

In the last case displayed in (8.2), namely when $\|\mu\|^2 + \|\mu'\|^2 < \|\mu - \mu'\|^2$ and $\lambda + \lambda' > 1$, we obtain

$$\begin{aligned} W_2^2(G_{\lambda,\mu}, G_{\lambda',\mu'}) &= (1 - \lambda')\|\mu\|^2 + (1 - \lambda)\|\mu'\|^2 + (\lambda + \lambda' - 1)\|\mu - \mu'\|^2 \\ &= (\lambda' - \lambda)\|\mu'\|^2 + \lambda\|\mu - \mu'\|^2 + (1 - \lambda')\big[\|\mu'\|^2 + \|\mu\|^2 - \|\mu - \mu'\|^2\big]. \\ &\leq (\lambda' - \lambda)\|\mu'\|^2 + \lambda\|\mu - \mu'\|^2. \end{aligned}$$

This entails (8.3). We get from this inequality, still assuming $\lambda \leq \lambda'$

$$\begin{aligned} W_2^2(G_{\lambda,\mu}, G_{\lambda',\mu'}) &\leq (\lambda' - \lambda)\|\mu'\|^2 + \lambda\|\mu - \mu'\|^2 \\ &\leq (\lambda' - \lambda)\|\mu'\|^2 + \lambda\big(\|\mu\| + \|\mu'\|\big)\|\mu - \mu'\|, \\ &\leq (\lambda' - \lambda)\|\mu'\|^2 + \big(\lambda\|\mu\| + \lambda'\|\mu'\|\big)\|\mu - \mu'\|, \\ &\leq (\lambda' - \lambda)\|\mu'\|\|\mu\| + (\lambda' - \lambda)\|\mu'\|\|\mu - \mu'\| + \big(\lambda\|\mu\| + \lambda'\|\mu'\|\big)\|\mu - \mu'\|, \\ &\leq (\lambda' - \lambda)\|\mu'\|\|\mu\| + 2\big(\lambda\|\mu\| + \lambda'\|\mu'\|\big)\|\mu - \mu'\|. \end{aligned}$$

From this latter inequality, we obtain

$$W_2^4(G_{\lambda,\mu}, G_{\lambda',\mu'}) \leq 8\big[(\lambda' - \lambda)^2\|\mu'\|^2\|\mu\|^2 + \big(\lambda\|\mu\| + \lambda'\|\mu'\|\big)^2\|\mu - \mu'\|^2\big]. \tag{8.4}$$

In the other hand, Inequality (7.12) indicates that

$$\|f_{\lambda,\mu} - f_{\lambda',\mu'}\|_2^2 \geq c_{\underline{K}}^2(\lambda' - \lambda)^2\|\mu\|^2\|\mu'\|^2 + c_{\underline{K}}^2\{\lambda'\}^2\|\mu'\|^2\|\mu - \mu'\|^2.$$

Since the role played by (λ, μ) and (λ', μ') is symmetric, we obtain in fact

$$\|f_{\lambda,\mu} - f_{\lambda',\mu'}\|_2^2 \geq c_{\underline{K}}^2(\lambda' - \lambda)^2\|\mu\|^2\|\mu'\|^2 + \frac{c_{\underline{K}}^2}{2}\big(\{\lambda'\}^2\|\mu'\|^2 + \{\lambda\}^2\|\mu\|^2\big)\|\mu - \mu'\|^2,$$

which together with (8.4) implies (8.1). Using this inequality with $f_{\hat{\lambda}_n,\hat{\mu}_n}$ and $f_{\lambda^\star,\mu^\star}$, and according to Theorem 2.1, we conclude the proof of Theorem 5.1. □

Appendix A: Technical results

A.1. Identifiability result

Proof of Proposition 2.1. We assume that two parameters $\theta_1 = (\lambda_1, \mu_1)$ and $\theta_2 = (\lambda_2, \mu_2)$ exist such that $f_{\theta_1} = f_{\theta_2}$. In that case, consider the Fourier transform of X whose density is f_{θ_1}. This Fourier transform is given by

$$\varphi_X(\xi) = \mathbb{E}\big[e^{\mathrm{i}\xi\bullet X}\big] = \big[(1-\lambda_1) + \lambda_1 e^{\mathrm{i}\xi\bullet\mu_1}\big]\hat{\phi}(\xi),$$

where $\hat{\phi}$ is the Fourier transform of ϕ and i is the complex number such that $\mathrm{i}^2 = -1$. Since $f_{\theta_1} = f_{\theta_2}$, we then deduce that:

$$\forall \xi \in \mathbb{R}^d \quad \big[(1-\lambda_1) + \lambda_1 e^{\mathrm{i}\xi\bullet\mu_1}\big]\hat{\phi}(\xi) = \big[(1-\lambda_2) + \lambda_2 e^{\mathrm{i}\xi\bullet\mu_2}\big]\hat{\phi}(\xi).$$

Since $\phi \in \mathbb{L}^1(\mathbb{R}^d)$, $\hat{\phi}$ is continuous and cannot be zero everywhere. Thus, we can find an open set $I \subset \mathbb{R}^d$ such that $\hat{\phi}(\xi) \neq 0$ in I and the Lebesgue measure of I is strictly positive. Hence,

$$\forall \xi \in I \quad (1-\lambda_1) + \lambda_1 e^{\mathrm{i}\xi\bullet\mu_1} = (1-\lambda_2) + \lambda_2 e^{\mathrm{i}\xi\bullet\mu_2},$$

and from the analytical property of the exponential map, we deduce that:

$$\forall \xi \in I \quad (1-\lambda_1) + \lambda_1\big[\cos(\xi\bullet\mu_1) + \mathrm{i}\sin(\xi\bullet\mu_1)\big] = (1-\lambda_2) + \lambda_2\big[\cos(\xi\bullet\mu_2) + \mathrm{i}\sin(\xi\bullet\mu_2)\big]$$

Identifying now the imaginary parts yields:

$$\forall \xi \in I \quad \lambda_1 \sin(\xi\bullet\mu_1) = \lambda_2 \sin(\xi\bullet\mu_2).$$

If we write $\mu_1 = (\mu_1^{(1)}, \ldots, \mu_1^{(d)})$ and $\mu_2 = (\mu_2^{(1)}, \ldots, \mu_2^{(d)})$, we deduce that

$$\begin{aligned}\forall \xi = (\xi_1, \ldots, \xi_d): \quad \lambda_1 &\left[\sin\big(\xi_1\mu_1^{(1)}\big)\cos\left(\sum_{j=2}^d \xi_j\mu_1^{(j)}\right) + \cos\big(\xi_1\mu_1^{(1)}\big)\sin\left(\sum_{j=2}^d \xi_j\mu_1^{(j)}\right)\right] \\ &= \lambda_2\left[\sin\big(\xi_1\mu_2^{(1)}\big)\cos\left(\sum_{j=2}^d \xi_j\mu_2^{(j)}\right) + \cos\big(\xi_1\mu_2^{(1)}\big)\sin\left(\sum_{j=2}^d \xi_j\mu_2^{(j)}\right)\right].\end{aligned}$$

Considering now the function of the variable ξ_1, it is classical that the family of functions $(\xi_1 \mapsto \sin(\alpha_1\xi_1), \xi_1 \mapsto \sin(\alpha_2\xi_1))$ is linearly independent if and only if $|\alpha_1| \neq |\alpha_2|$. We can deduce that, necessarily, $\mu_1^{(1)} = \pm\mu_2^{(1)}$ and therefore $\cos(\xi_1\mu_1^{(1)}) = \cos(\xi_1\mu_2^{(1)})$, which shows that $\lambda_1 \sin(\sum_{j=2}^d \xi_j\mu_1^{(j)}) = \lambda_2 \sin(\sum_{j=2}^d \xi_j\mu_2^{(j)})$ for all $\xi \in I$. We then end the argument with an easy recursion: we obtain that $\lambda_1 \sin(\xi_d\mu_1^{(d)}) = \lambda_2 \sin(\xi_d\mu_2^{(d)})$ so that $\mu_1^{(d)} = \pm\mu_2^{(d)}$. Since λ_1 and λ_2 are positive, then $\mu_1^{(d)} = \mu_2^{(d)}$, which in turn implies that $\mu_1^{(j)} = \mu_2^{(j)}$ for all the coordinates $j \in \{1, \ldots, d\}$. □

A.2. Connection between $\|\phi - \phi_\mu\|_2$ and $|\mu|$

Proposition A.1. *Let any $M > 0$ be given and assume that ϕ satisfies $(\mathbf{H}_\mathcal{S})$ and $(\mathbf{H_{Lip}})$, then two constants $0 < \underline{\kappa} < \overline{\kappa} < +\infty$ exist such that*:

$$\forall (\mu, \tilde{\mu}) \in [-M, M]^d \times [-M, M]^d \quad \underline{\kappa}\|\mu - \tilde{\mu}\|^2 \leq \|\phi_\mu - \phi_{\tilde{\mu}}\|_2^2 \leq \overline{\kappa}\|\mu - \tilde{\mu}\|^2. \tag{A.1}$$

Proof. We prove the upper and lower bounds separately. According to the shift invariance of the $\mathbb{L}^2$ norm, we only establish these inequalities when $\tilde{\mu} = 0$. Using $(\mathbf{H_{Lip}})$, the upper bound simply derives from:

$$\|\phi - \phi_\mu\|_2^2 = \int_{\mathbb{R}^d}\big[\phi(x) - \phi(x-\mu)\big]^2\,dx \leq \int_{\mathbb{R}^d}\|\mu\|^2 g^2(x)\,dx = \|\mu\|^2\|g\|_2^2,$$

which is the desired inequality if we choose $\overline{\kappa} = \|g\|^2$. Concerning the lower bound, we have:

$$\frac{\|\phi(\cdot) - \phi(\cdot - \mu)\|_2^2}{\|\mu\|^2} = \int_{\mathbb{R}^d}\left[\frac{\phi(x) - \phi(x-\mu)}{\|\mu\|}\right]^2 dx.$$

We write $\mu = \|\mu\| e$ where e is a unit vector of the sphere. Inequality (3.1) brought by Assumption ($\mathbf{H_{Lip}}$) makes it possible to apply the Lebesgue convergence theorem, which implies:

$$\lim_{\|\mu\|\longrightarrow 0} \frac{\|\phi(\cdot)-\phi(\cdot-\mu)\|^2}{\|\mu\|^2} = \int_{\mathbb{R}^d} \lim_{\|\mu\|\to 0} \left[\frac{\phi(x)-\phi(x-\mu)}{\|\mu\|}\right]^2 dx,$$
$$= \|\nabla\phi \bullet e\|^2 = \left\|d_e[\phi]\right\|^2 > 0.$$

Indeed, ϕ being differentiable ($\phi \in \mathcal{C}^1(\mathbb{R}^d)$), $\frac{\phi(x)-\phi(x-\mu)}{\|\mu\|} \longrightarrow d_e[\phi](x)$ almost surely when $\|\mu\| \longrightarrow 0$. Now, ϕ is continuous and $\psi : \mu \longrightarrow \frac{\|\phi-\phi_\mu\|_2^2}{\|\mu\|^2} \in \mathcal{C}^0([-M, M]^d, \mathbb{R})$ from the Lebesgue convergence theorem. This continuous map ψ attains its lower bound on $[-M, M]^d$ and the identifiability result of Proposition 2.1 implies that this lower bound is positive. This leads to the existence of $\underline{\kappa} > 0$ such that:

$$\|\phi - \phi_\mu\|_2^2 \geq \underline{\kappa}\|\mu\|^2. \qquad \square$$

A.3. Log-concave distributions

In this section, we establish that most of the log-concave real distributions satisfy the assumptions ($\mathbf{H_S}$), ($\mathbf{H_{Lip}}$) and ($\mathbf{H_D}$). For this purpose, we introduce the associated class of probability measures:

$$\mathcal{LC} := \left\{\phi(\cdot) = e^{-u(\cdot)} : u \text{ is convex, } u \in \mathcal{C}^2\left(\mathbb{R}^d\right) \text{ and } \|\nabla u\| + \left\|D^2 u\right\| = o_\infty(u)\right\}.$$

The set of possible densities is rich and contains Gaussian or Gamma distributions. However, the set $\mathcal{LC}$ does not capture the situation where $u(x) = e^{|x|}$ or $u(x) = e^{x^2}$ since u exhibits variations that are too great for large values of x.

Proposition A.2. *Assume that μ varies in $[-M, M]^d$ and that $\phi \in \mathcal{LC}$. Let $\varepsilon \in (0, M)$. If we set*:

$$g(x) := g_1(x) \vee g_2(x) \vee g_3(x)$$

with

$$g_1(x) := \sqrt{\frac{\sup_{e\in\mathcal{S}^1} \int_{[x-Me,x]} \langle\nabla\phi(t), e\rangle^2\, dt}{\varepsilon}}, \qquad g_2(x) := \sqrt{\frac{\sup_{e\in\mathcal{S}^1} \int_{[x,x+Me]} \langle\nabla\phi(t), e\rangle^2\, dt}{\varepsilon}},$$

and

$$g_3(x) := \sup_{t\in B(x,\varepsilon)} \left\|\nabla\phi(t)\right\|.$$

Then, ($\mathbf{H_{Lip}}$) *and* ($\mathbf{H_D}$) *hold*:

(i) $\forall\mu \in [-M, M]^d\ \forall x \in \mathbb{R}^d\ |\phi(x) - \phi_\mu(x)| \leq \|\mu\| g(x)$
(ii) $g\phi^{-1/2} \in \mathbb{L}^2(\mathbb{R}^d)$
(iii) $D^2\phi\phi^{-1/2} \in \mathbb{L}^2(\mathbb{R}^d)$

Proof. We provide a proof in the case when $\phi \in \mathcal{C}^2$. This proof can be extended when $\phi \in \mathcal{C}_p^2$ according to some small modifications that are left to the reader, it then makes possible to extend our results to the Laplace distributions for example.

Proof of (*i*): Remark first that $\forall\mu \in [-M, M]^d$, a unit vector $e \in \mathcal{S}^1$ exists such that $\mu = \|\mu\|e$ and in that case

$$\forall x \in \mathbb{R}^d \quad \left|\phi(x) - \phi_\mu(x)\right| = \left|\int_{[x-\mu,x]} \left\langle\nabla\phi(t), e\right\rangle dt\right| \leq \sqrt{\|\mu\|}\sqrt{\int_{[x-\mu,x]} \langle\nabla\phi, e\rangle^2},$$

where $[x - \mu, x]$ refers to the segment that joins $x - \mu$ to x in $\mathbb{R}^d$ and the last upper bound comes from the Cauchy–Schwarz inequality. Let $\varepsilon \in (0, M)$. If $\|\mu\| \in [\varepsilon, M]$, we obtain that:

$$\left|\phi(x) - \phi_\mu(x)\right| \leq \|\mu\|\left(g_1(x) \vee g_2(x)\right),$$

where g_1 and g_2 are defined in the statement of the Proposition. Finally, we should remark that if $\|\mu\| \in [0, \varepsilon)$, then

$$\left|\phi(x) - \phi_\mu(x)\right| \leq \|\mu\| \sup_{t \in B(x,\varepsilon)} \left\|\nabla\phi(t)\right\| := \|\mu\| g_3(x).$$

It proves that $g = g_1 \vee g_2 \vee g_3$ satisfies the desired inequality.

Proof of (ii): In order to prove that $g\phi^{-1/2} \in \mathbb{L}^2(\mathbb{R}^d)$, we separately prove that $g_1^2\phi^{-1}$, $g_2^2\phi^{-1}$ and $g_3^2\phi^{-1}$ belong to $\mathbb{L}^1(\mathbb{R}^d)$. We should remark that since g_1, g_2 and g_3 are continuous functions, then we only have to check the integrability when $\|x\| \longrightarrow +\infty$. g_1 and g_2 are rather similar and we only handle the integrability of $g_1^2\phi^{-1}$.

We write

$$\begin{aligned} g_1^2(x)\phi^{-1}(x) &= \varepsilon^{-1} e^{u(x)} \sup_{e \in \mathcal{S}^1} \int_{[x-Me,x]} \left\langle \nabla\phi(t), e \right\rangle^2 dt \\ &= \varepsilon^{-1} \sup_{e \in \mathcal{S}^1} e^{u(x)} \int_{[x-Me,x]} \left\langle \nabla\phi(t), e \right\rangle^2 dt \\ &= \varepsilon^{-1} \sup_{e \in \mathcal{S}^1} \underbrace{e^{u(x)} \int_{[x-Me,x]} \left\langle \nabla u(t), e \right\rangle^2 e^{-2u(t)}\, dt}_{:= G_e(x)}. \end{aligned}$$

At this stage, we are driven to consider the 1-dimensional fonction $u_e(t) = u(x + (t - M)e)$, which is a convex function. We then have

$$G_e(x) = e^{u_e(M)} \int_0^M u_e'(s)^2 e^{-2u_e(s)}\, ds.$$

We shall now produce a 1-dimension argument with the convex function u_e. We assume that $u_e(M) \geq u_e(0)$, and know that u_e' is an increasing map and positive:

$$\begin{aligned} G_e(x) &\leq u_e'(M) e^{u_e(M)} \int_0^M u_e'(s) e^{-2u_e(s)}\, ds \\ &\leq \left\langle \nabla u(x), e \right\rangle e^{u(x)} \frac{e^{-2u(x-Me)} - e^{-2u(x)}}{2} \\ &\leq \frac{\langle \nabla u(x), e \rangle}{2} e^{-2u(x-Me)+u(x)}. \end{aligned}$$

The mean value theorem leads to:

$$\exists \xi \in [x - Me, x] \quad u(x - Me) = u(x) - M\left\langle \nabla u(\xi), e \right\rangle \geq u(x) - M\left\langle \nabla u(x), e \right\rangle.$$

Consequently, we obtain:

$$G_e(x) \leq \frac{\langle u(x), e \rangle}{2} e^{-u(x)+2M\|\nabla u(x)\|}.$$

The density $\phi \in \mathcal{LC}$ and we can find K large enough such that:

$$\forall \|x\| \geq K \ \forall e \in \mathcal{S}^1 \quad -u(x) + 2M\left\|\nabla u(x)\right\| \leq -(1 - \eta)u(x)$$

For such an x, we have $G_e(x) \leq \frac{\langle \nabla u(x), e \rangle}{2} e^{-(1-\eta)u(x)} \in \mathbb{L}^1(\mathbb{R}^d)$.

Concerning $g_2(x)\phi(x)^{-1}$, we can produce an almost identical argument left to the reader. We now consider $g_{3,\varepsilon}^2\phi^{-1}$:

$$g_{3,\varepsilon}^2(x)\phi^{-1}(x) = \sup_{t \in B(x,\varepsilon)} \left\|\nabla u(t)\right\|^2 e^{-2u(t)+u(x)}.$$

If $t \in [x - \varepsilon, x]$, the mean value theorem leads to:

$$\begin{aligned} u(t) &= u(x) - \left\langle (x - t), \nabla u(\xi) \right\rangle \quad \text{with } \xi \in]t, x[\\ &\geq u(x) - \varepsilon \sup_{B(x,\epsilon)} \|\nabla u\|. \end{aligned}$$

Using the fact that $\|D^2u\| + \|\nabla u\| = o_\infty(u)$, we can find a positive constant $C > 0$, a parameter $\eta \in (0, 1)$ and for K large enough such that $\forall \|x\| \geq K$:

$$\|u\|(t)^2 e^{-2u(t)+u(x)} \leq C\big\|u(x)\big\|e^{-(1-\eta)u(x)}. \tag{A.2}$$

Thus, (A.2) imply that $g_{3,\varepsilon}^2\phi^{-1} \in \mathbb{L}^1(\mathbb{R}^d)$. As a maximum of three functions in $\mathbb{L}^1(\mathbb{R}^d)$, we deduce that $g^2\phi^{-1} \in \mathbb{L}^1(\mathbb{R}^d)$.

Proof of (iii): A direct computation shows that, almost surely:

$$\{d_{jj}\phi\}^2\phi^{-1} = \big[d_{jj}u - \{d_ju\}^2\big]^2 e^{-u} \leq 2\{d_{jj}u\}^2 e^{-u} + 2\{d_ju\}^4 e^{-u}.$$

Again, using the fact that $\|D^2u\| + \|\nabla u\| = o_\infty(u)$, we can find a positive constant $C > 0$, a parameter $\eta \in (0, 1)$ and a large enough K such that $\forall \|x\| \geq K$:

$$\begin{aligned}
\{d_{jj}u\}^2(x)e^{-u(x)} &\leq Cd_{jj}u(x)e^{-(1-\eta)u(x)} \\
&\leq Cd_j\big(d_ju(x)e^{-(1-\eta)u(x)}\big) + C(1-\eta)\big\{d_ju(x)\big\}^2 e^{-(1-\eta)u(x)} \\
&\leq Cd_j\big(d_ju(x)e^{-(1-\eta)u(x)}\big) + C^2(1-\eta)d_ju(x)e^{-(1-\eta)^2u(x)},
\end{aligned}$$

which is integrable when $\|x\| \longrightarrow +\infty$. A similar argument leads to $d_ju^4e^{-u} \leq Cd_jue^{-(1-\eta)u}$. We can repeat the same argument when $\|x\| \longrightarrow -\infty$ with an adaptation of the sign of $d_ju(x)$. We can conclude that $\{d_{jj}\phi\}^2\phi^{-1} \in \mathbb{L}^1(\mathbb{R}^d)$. □

Appendix B: Proofs of the lower bounds

B.1. *Asymmetric risk*

We begin by a useful lemma, which is a generalization of the Le Cam method for proving lower bounds if the loss involved in the statistical model is not symmetric, meaning that $\rho(\theta_1, \theta_2)$ is generally not equal to $\rho(\theta_2, \theta_1)$, but still satisfies a weak triangle inequality. Hence, the Le Cam Lemma requires a small modification in the spirit of the remark of [30] (Example 2, Section 3).

In the sequel, $d_{\mathrm{TV}}(\mathbb{P}, \mathbb{Q})$ and $\mathrm{KL}(\mathbb{P}, \mathbb{Q})$ denote the total variation distance and the Kullback–Leibler divergence between two measures, $\mathbb{P}$ and $\mathbb{Q}$, respectively.

Lemma B.1. *Let $(\mathbb{P}_\theta)_{\theta\in\Theta}$ be a family of measures indexed by Θ and assume that $\rho : (\theta_1, \theta_2) \in \Theta^2 \mapsto \rho(\theta_1, \theta_2) \in \mathbb{R}^+$ satisfies the weak triangle inequality*:

$$\forall(\theta_1, \theta_2, \theta_3) \in \Theta^3, \quad \rho(\theta_1, \theta_3) + \rho(\theta_2, \theta_3) \geq \rho(\theta_1, \theta_2) \wedge \rho(\theta_2, \theta_1). \tag{B.1}$$

Let $\Phi : \mathbb{R}^+ \to \mathbb{R}^+$ be a non-decreasing function. Let $\delta > 0$ and $(\theta_1, \theta_2) \in \Theta^2$ such that $\rho(\theta_1, \theta_2) \wedge \rho(\theta_2, \theta_1) \geq 2\delta$. Then,

$$\begin{aligned}
\inf_{\hat\theta}\sup_{\theta\in\Theta} \mathbb{E}\big[\Phi\big(\rho(\theta, \hat\theta)\big)\big] &\geq \frac{\Phi(\delta)}{2}\big\{1 - d_{\mathrm{TV}}\big(\mathbb{P}_{\theta_1}^{\otimes^n}, \mathbb{P}_{\theta_2}^{\otimes^n}\big)\big\} \\
&\geq \frac{\Phi(\delta)}{2}\left\{1 - \sqrt{\frac{n}{2}\mathrm{KL}(\mathbb{P}_{\theta_1}, \mathbb{P}_{\theta_2})}\right\},
\end{aligned}$$

where the infimum is taken over all estimators $\hat\theta$.

Proof. First, we observe that:

$$\mathbb{E}\big[\Phi\big(\rho(\theta, \hat\theta)\big)\big] \geq \Phi(\delta)\mathbb{P}\big(\rho(\theta, \hat\theta) \geq \delta\big),$$

since Φ is a non-decreasing function. Let $\mathcal{V} = \{1, 2\}$ and $\Psi(\hat\theta) = \mathrm{argmin}_{v\in\mathcal{V}}\, \rho(\theta_v, \hat\theta)$.

We can show that $\rho(\theta_v, \hat\theta) < \delta$ implies that $\Psi(\hat\theta) = v$. According to Condition (B.1), we have:

$$\rho(\theta_v, \hat\theta) \geq \rho(\theta_v, \theta_{v'}) \wedge \rho(\theta_{v'}, \theta_v) - \rho(\theta_{v'}, \hat\theta) > 2\delta - \rho(\theta_{v'}, \hat\theta).$$

Now, if $\rho(\theta_v,\hat\theta)<\delta$, then $\delta>2\delta-\rho(\theta_{v'},\hat\theta)$, so that $\rho(\theta_{v'},\hat\theta)>\delta$, which is necessarily larger than $\rho(\theta_v,\hat\theta)$. Hence, we obtain $\Psi(\hat\theta)=v$.

Equivalently, for $v\in\{1,2\}$, we have $\Psi(\hat\theta)\neq v\Longrightarrow\rho(\theta_v,\hat\theta)>\rho(\theta_{v'},\hat\theta)$ since:

$$2\delta\le\rho(\theta_v,\theta_{v'})\wedge\rho(\theta_{v'},\theta_v)\le\rho(\theta_v,\hat\theta)+\rho(\theta_{v'},\hat\theta)\le 2\rho(\theta_v,\hat\theta).$$

The rest of the proof proceeds from the standard Le Cam argument: Φ is non decreasing so that:

$$\begin{aligned}\sup_{\theta\in\Theta}\mathbb{E}\big[\Phi\big(\rho(\theta,\hat\theta)\big)\big]&\ge\Phi(\delta)\sup_{\theta\in\Theta}\mathbb{P}\big(\rho(\theta,\hat\theta)\ge\delta\big)\\&\ge\frac{\Phi(\delta)}{2}\big\{\mathbb{P}\big(\rho(\theta_1,\hat\theta)\ge\delta\big)+\mathbb{P}\big(\rho(\theta_2,\hat\theta)\ge\delta\big)\big\}\\&\ge\frac{\Phi(\delta)}{2}\big\{\mathbb{P}_{\theta_1}^{\otimes^n}\big(\Psi(\hat\theta)\neq1\big)+\mathbb{P}_{\theta_2}^{\otimes^n}\big(\Psi(\hat\theta)\neq2\big)\big\}.\end{aligned}$$

Taking an infimum over all tests Ψ (see, *e.g.*, [22]) we obtain:

$$\begin{aligned}\inf_{\hat\theta}\sup_{\theta\in\Theta}\mathbb{E}\big[\Phi\big(\rho(\theta,\hat\theta)\big)\big]&\ge\frac{\Phi(\delta)}{2}\inf_{\Psi}\big\{\mathbb{P}_{\theta_1}^{\otimes^n}(\Psi\neq1)+\mathbb{P}_{\theta_2}^{\otimes^n}(\Psi\neq2)\big\}\\&\ge\frac{\Phi(\delta)}{2}\big\{1-d_{\mathrm{TV}}\big(\mathbb{P}_{\theta_1}^{\otimes^n},\mathbb{P}_{\theta_2}^{\otimes^n}\big)\big\}.\end{aligned}$$

Pinsker's inequality:

$$d_{\mathrm{TV}}\big(\mathbb{P}_{\theta_1}^{\otimes^n},\mathbb{P}_{\theta_2}^{\otimes^n}\big)\le\sqrt{\frac{1}{2}\mathrm{KL}\big(\mathbb{P}_{\theta_1}^{\otimes^n},\mathbb{P}_{\theta_2}^{\otimes^n}\big)}=\sqrt{\frac{n}{2}\mathrm{KL}(\mathbb{P}_{\theta_1},\mathbb{P}_{\theta_2})}$$

ends the proof. □

B.2. Lower bound for the strong contamination model

We now study the lower bounds in the first regime, namely when $\|\mu\|$ is lower bounded by a constant m that is independent of n.

Proof of Theorem 4.1.

Item (i): We apply Lemma B.1 with $\Phi(t)=t^2$ and the loss function ρ defined as:

$$\forall(\theta_1,\theta_2)\in\Theta_n(m,c)^2\quad\rho(\theta_1,\theta_2)=\lambda_1\|\mu_1-\mu_2\|.$$

Remark that ρ satisfies the weak triangle inequality (B.1). Indeed, for all $(\theta_1,\theta_2,\theta_3)\in\Theta_n(m,c)^3$, we have:

$$\begin{aligned}\rho(\theta_1,\theta_3)+\rho(\theta_2,\theta_3)&=\lambda_1\|\mu_1-\mu_3\|+\lambda_2\|\mu_2-\mu_3\|\\&\ge\min(\lambda_1,\lambda_2)\|\mu_1-\mu_2\|\\&\ge\rho(\theta_1,\theta_2)\wedge\rho(\theta_2,\theta_1).\end{aligned}$$

We introduce the subset

$$\Theta_n(m,M,c,\bar\lambda):=\left\{\theta=(\lambda,\mu):\frac{c}{\|\mu\|^2\sqrt{n}}\le\lambda\le\bar\lambda,m\le\|\mu\|\le M\right\}$$

where $0<m<M$ and $0<\frac{c}{m^2\sqrt{n}}<\bar\lambda<1$. Then, $\Theta_n(m,M,c,\bar\lambda)\subset\Theta_n(m,c)$. We consider $\theta_1=(\lambda,\mu_1)$ and $\theta_2=(\lambda,\mu_2)$; their values will be chosen later to ensure that $(\theta_1,\theta_2)\in\Theta_n(m,M,c,\bar\lambda)^2$. According to Lemma B.1 applied with $\delta=\frac{\lambda\|\mu_1-\mu_2\|}{2}$, we can write:

$$\begin{aligned}\inf_{\hat\theta}\sup_{\theta\in\Theta_n(m,c)}\mathbb{E}\big[\lambda^2\|\hat\mu-\mu\|^2\big]&\ge\inf_{\hat\theta}\sup_{\theta\in\Theta_n(m,M,c,\bar\lambda)}\mathbb{E}\big[\lambda^2\|\hat\mu-\mu\|^2\big]\\&\ge\frac{\delta^2}{2}\left\{1-\sqrt{\frac{n}{2}\mathrm{KL}(\mathbb{P}_{\theta_1},\mathbb{P}_{\theta_2})}\right\}.\end{aligned}\tag{B.2}$$

We can compute the Kullback–Leibler divergence between the two mixtures $\mathbb{P}_{\theta_1}$ and $\mathbb{P}_{\theta_2}$: if $f_1 = (1-\lambda)\phi + \lambda\phi_{\mu_1}$ (resp. $f_2 = (1-\lambda)\phi + \lambda\phi_{\mu_2}$) is the density of $\mathbb{P}_{\theta_1}$ (resp. $\mathbb{P}_{\theta_2}$) w.r.t. the Lebesgue measure, we have:

$$\begin{aligned}\mathrm{KL}(\mathbb{P}_{\theta_1}, \mathbb{P}_{\theta_2}) &= \int \log\left[1 + \frac{f_1(x) - f_2(x)}{f_2(x)}\right] f_1(x)\,dx \\ &\leq \int \frac{f_1(x) - f_2(x)}{f_2(x)} f_1(x)\,dx,\end{aligned}$$

where we used the inequality $\log(1+t) \leq t$. If we once again write $f_1 = f_2 + f_1 - f_2$, we obtain:

$$\begin{aligned}\mathrm{KL}(\mathbb{P}_{\theta_1}, \mathbb{P}_{\theta_2}) &\leq \int \frac{f_1(x) - f_2(x)}{f_2(x)}\big[f_2(x) + f_1(x) - f_2(x)\big]\,dx \\ &\leq \int \frac{[f_1(x) - f_2(x)]^2}{f_2(x)}\,dx \\ &\leq \lambda^2 \int \frac{[\phi_{\mu_1}(x) - \phi_{\mu_2}(x)]^2}{(1-\lambda)\phi(x) + \lambda\phi_{\mu_2}(x)}\,dx\end{aligned}$$

since $f_2(x) \geq (1-\lambda)\phi(x)$ and $f_1(x) - f_2(x) = \lambda[\phi_{\mu_1}(x) - \phi_{\mu_2}(x)]$. On the basis of Assumption ($\mathbf{H_{Lip}}$), we know that $|\phi_{\mu_1} - \phi_{\mu_2}| \leq \|\mu_1 - \mu_2\| g$ and we obtain:

$$\mathrm{KL}(\mathbb{P}_{\theta_1}, \mathbb{P}_{\theta_2}) \leq \frac{\lambda^2 \|\mu_1 - \mu_2\|^2 \mathcal{J}}{1 - \bar{\lambda}}, \tag{B.3}$$

where $\mathcal{J} := \|g\phi^{-1/2}\|_2^2$ is the constant involved in ($\mathbf{H_{Lip}}$).

We now choose λ, μ_1 and μ_2 so that we obtain the largest possible value in (B.2), while satisfying the constraints given in $\Theta_n(m, M, c, \bar{\lambda})$. Without loss of generality, we set $\mu_1^{(1)} < \mu_2^{(1)}$ and we need to find a choice of these parameters such that $m \leq \mu_1^{(1)} < \mu_2^{(1)} \leq M$ and $\frac{c}{(\mu_1^{(1)})^2\sqrt{n}} \leq \lambda \leq \bar{\lambda}$. We set $\mu_1 = (\mu_1^{(1)}, 0, \ldots, 0)$ and $\mu_2 = (\mu_2^{(1)}, 0, \ldots, 0)$ so that

$$\mu_1^{(1)} = m \quad \text{and} \quad \lambda = \frac{c}{m^2\sqrt{n}} < \bar{\lambda}.$$

For a given $\epsilon > 0$, we choose $\mu_2^{(1)}$ such that $\frac{n}{2}\mathrm{KL}(\mathbb{P}_{\theta_1}, \mathbb{P}_{\theta_2}) \leq 1 - \epsilon$. Using (B.3), we arrive at the calibration:

$$\mu_2^{(1)} - \mu_1^{(1)} = \sqrt{\frac{2(1-\bar{\lambda})(1-\epsilon)}{\lambda^2 \mathcal{J} n}}.$$

It remains to check that $\mu_2^{(1)} \leq M$. From our choice of λ and $\mu_1^{(1)}$, we see that:

$$\mu_2^{(1)} = m\left[1 + \sqrt{\frac{2(1-\bar{\lambda})m^2}{c^2\mathcal{J}}(1-\epsilon)}\right] \leq m\left[1 + \sqrt{\frac{2m^2(1-\epsilon)}{c^2\mathcal{J}}}\right],$$

which can be made smaller than M if $1 - \epsilon \leq \frac{c^2\mathcal{J}(M-m)^2}{2m^4}$. If we plug these choices of λ, μ_1 and μ_2 into (B.2), we obtain:

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta_n(m,M,c,\bar{\lambda})} \mathbb{E}\big[\lambda^2 \|\hat{\mu} - \mu\|^2\big] \geq \frac{(1-\bar{\lambda})(1-\epsilon)\epsilon}{8\mathcal{J}n},$$

which is the desired lower bound of the minimax risk (4.1).

Item (*ii*): We keep the same Φ and define $\rho(\theta_1, \theta_2) = |\lambda_1 - \lambda_2| = \rho(\theta_2, \theta_1)$. We consider $\theta_1 = (\lambda_1, \mu)$ and $\theta_2 = (\lambda_2, \mu)$ such that $|\lambda_1 - \lambda_2| = \frac{\epsilon}{\sqrt{n}}$ and

$$\frac{c}{m^2\sqrt{n}} = \lambda_1 < \lambda_2 \leq \bar{\lambda},$$

μ and ϵ have to be chosen hereafter. Since $\lambda_2 = \lambda_1 + \frac{\epsilon}{\sqrt{n}} \leq \bar{\lambda}$, we must choose ϵ such that:

$$\epsilon \leq \bar{\lambda}\sqrt{n} - \frac{c}{m^2}, \tag{B.4}$$

which is possible since we assumed that $\frac{c}{m^2\sqrt{n}} < \bar{\lambda}$. From Lemma B.1,

$$\begin{aligned}\inf_{\hat{\theta}} \sup_{\theta\in\Theta_n(m,c)} \mathbb{E}\big[(\lambda - \hat{\lambda})^2\big] &\geq \inf_{\hat{\theta}} \sup_{\theta\in\Theta_n(m,M,c,\bar{\lambda})} \mathbb{E}\big[(\lambda - \hat{\lambda})^2\big] \\ &\geq \frac{\epsilon^2}{2n}\left\{1 - \sqrt{\frac{n}{2}\mathrm{KL}(\mathbb{P}_{\theta_1}, \mathbb{P}_{\theta_2})}\right\}.\end{aligned}$$

We can upper bound the Kullback–Leibler divergence as:

$$\begin{aligned}\mathrm{KL}(\mathbb{P}_{\theta_1}, \mathbb{P}_{\theta_2}) &\leq \int \big[f_1(x) - f_2(x)\big]^2 f_2(x)^{-1}\,dx \\ &\leq (\lambda_1 - \lambda_2)^2 \int \big[\phi_\mu(x) - \phi(x)\big]^2 f_2(x)^{-1}\,dx \\ &\leq \frac{(\lambda_1 - \lambda_2)^2\|\mu\|^2}{1 - \bar{\lambda}} \int g(x)^2\phi(x)^{-1}\,dx \\ &\leq \frac{\|\mu\|^2\epsilon^2\mathcal{J}}{(1-\bar{\lambda})n}.\end{aligned}$$

By choosing $\mu = (\mu^{(1)}, 0, \ldots, 0)$ with

$$\mu^{(1)} = \frac{m+M}{2} \quad \text{and} \quad \epsilon \leq \sqrt{\frac{2(1-\bar{\lambda})}{\mathcal{J}(m+M)^2}}, \tag{B.5}$$

we obtain $\frac{n}{2}\mathrm{KL}(\mathbb{P}_{\theta_1}, \mathbb{P}_{\theta_2}) \leq \frac{1}{4}$. Considering the minimal admissible value of ϵ in (B.4) and (B.5) now leads to a choice of the parameters θ_1 and θ_2 such that:

$$\inf_{\hat{\theta}} \sup_{\theta\in\Theta_n(m,c)} \mathbb{E}\big[(\lambda - \hat{\lambda})^2\big] \geq \frac{\epsilon^2}{4n}.$$

This last inequality is the second lower bound (4.2). □

B.3. Lower bound for the weak contamination model

Proof of Theorem 4.2.

Point (i): We consider $\Phi(t) = t^2$ and the loss function ρ defined as:

$$\rho(\theta_1, \theta_2) = \|\mu_1\|^2|\lambda_1 - \lambda_2|.$$

Note that ρ satisfies (B.1) since $\forall(\theta_1, \theta_2, \theta_3) \in \Theta_n(c)^3$,

$$\begin{aligned}\rho(\theta_1, \theta_3) + \rho(\theta_2, \theta_3) &= \|\mu_1\|^2|\lambda_1 - \lambda_3| + \|\mu_2\|^2|\lambda_2 - \lambda_3| \\ &\geq \min\big(\|\mu_1\|^2, \|\mu_2\|^2\big)|\lambda_1 - \lambda_2| \\ &\geq \rho(\theta_1, \theta_2) \wedge \rho(\theta_2, \theta_1).\end{aligned}$$

To obtain a convenient lower bound, we need to use Lemma B.1 and find a couple of parameters (θ_1, θ_2) that belongs to the admissible set and such that $\mathrm{KL}(\mathbb{P}_{\theta_1}, \mathbb{P}_{\theta_2})$ is small enough. In particular, the proximity between $\mathbb{P}_{\theta_1}$ and $\mathbb{P}_{\theta_2}$ will be obtained by a careful matching of the first moments of the two distributions, which is a good method for obtaining efficient lower bounds in mixture models (see, *e.g.*, [3] or [15]). We give an example of this method below. First, remark that:

$$\mathrm{KL}(\mathbb{P}_{\theta_1}, \mathbb{P}_{\theta_2}) = \int \log\left[\frac{f_1(x)}{f_2(x)}\right] f_1(x)\,dx.$$

Since ϕ satisfies $(\mathbf{H}_{\mathcal{S}})$, then ϕ is a $\mathcal{C}^3$ function on $\mathbb{R}^d$, considering a shift $\mu = (\mu^{(1)}, 0, \dots, 0) = o(1)$, we can write a third order Taylor expansion:

$$\forall x \in \mathbb{R}^d \quad \phi_\mu(x) = \phi(x) - \mu^{(1)} d_1\phi(x) + \frac{\{\mu^{(1)}\}^2 d_{11}\phi(x)}{2} - \frac{\{\mu^{(1)}\}^3}{6} d_{111}\phi(\xi_{x,\mu}),$$

where $\xi_{x,\mu}$ belongs to the interval defined by x and $x-\mu$ and $d_1\phi$ (resp. $d_{11}\phi$ and $d_{111}\phi$) denotes the first (resp. second and third) partial derivative of ϕ w.r.t. the first coordinate of x. In particular, assuming that $d_{111}\phi$ is bounded on $\mathbb{R}^d$ leads to:

$$\forall x \in \mathbb{R}^d \quad \phi_\mu(x) = \phi(x) - \mu^{(1)} d_1\phi(x) + \frac{\{\mu^{(1)}\}^2}{2} d_{11}\phi(x) + o\big(\|\mu\|^2\big).$$

This Taylor expansion permits us to write, for small values of $\mu_1^{(1)}$:

$$\begin{aligned}
\log\big[f_1(x)\big] &= \log\big[(1-\lambda_1)\phi(x) + \lambda_1\phi_{\mu_1}(x)\big] \\
&= \log\bigg[(1-\lambda_1)\phi(x) + \lambda_1\phi(x) - \lambda_1\mu_1^{(1)} d_1\phi(x) + \frac{1}{2}\lambda_1\big\{\mu_1^{(1)}\big\}^2 d_{11}\phi(x) + o\big(\|\mu_1\|^2\big)\bigg] \\
&= \log\big[\phi(x)\big] + \log\bigg[1 - \lambda_1\mu_1^{(1)}\frac{d_1\phi(x)}{\phi(x)} + \frac{1}{2}\lambda_1\big\{\mu_1^{(1)}\big\}^2\frac{d_{11}\phi(x)}{\phi(x)} + o\big(\|\mu_1\|^2\big)\bigg] \\
&= \log\big[\phi(x)\big] - \lambda_1\mu_1^{(1)}\frac{d_1\phi(x)}{\phi(x)} + \frac{1}{2}\lambda_1\big\{\mu_1^{(1)}\big\}^2\frac{d_{11}\phi(x)}{\phi(x)} \\
&\quad - \frac{1}{2}\lambda_1^2\big\{\mu_1^{(1)}\big\}^2\bigg(\frac{d_1\phi(x)}{\phi(x)}\bigg)^2 + o\big(\|\mu_1\|^2\big).
\end{aligned}$$

In the same way, for small values of μ_2:

$$\begin{aligned}
\log\big[f_2(x)\big] &= \log\big[(1-\lambda_2)\phi(x) + \lambda_2\phi_{\mu_2}(x)\big] \\
&= \log\big[\phi(x)\big] - \lambda_2\mu_2^{(1)}\frac{d_1\phi(x)}{\phi(x)} + \frac{1}{2}\lambda_2\big\{\mu_2^{(1)}\big\}^2\frac{d_{11}\phi(x)}{\phi(x)} \\
&\quad - \frac{1}{2}\lambda_2^2\big\{\mu_2^{(1)}\big\}^2\bigg(\frac{d_1\phi(x)}{\phi(x)}\bigg)^2 + o\big(\|\mu_2\|^2\big).
\end{aligned}$$

We thus obtain:

$$\begin{aligned}
\log\big[f_1(x)\big] - \log\big[f_2(x)\big] &= \big(\lambda_2\mu_2^{(1)} - \lambda_1\mu_1^{(1)}\big)\frac{d_1\phi(x)}{\phi(x)} + \frac{1}{2}\big(\lambda_1\big\{\mu_1^{(1)}\big\}^2 - \lambda_2\big\{\mu_2^{(1)}\big\}^2\big)\frac{d_{11}\phi(x)}{\phi(x)} \\
&\quad + \frac{1}{2}\big(\lambda_2^2\big\{\mu_2^{(1)}\big\}^2 - \lambda_1^2\big\{\mu_1^{(1)}\big\}^2\big)\bigg(\frac{d_1\phi(x)}{\phi(x)}\bigg)^2 + o\big(\|\mu_1\|^2\big) + o\big(\|\mu_2\|^2\big).
\end{aligned}$$

In particular, we observe that the term above can be considered as a "second order term" if θ_1 and θ_2 are chosen such that $\lambda_1\mu_1^{(1)} = \lambda_2\mu_2^{(1)}$, which corresponds to the first moment of $\mathbb{P}_{\theta_1}$ and $\mathbb{P}_{\theta_2}$. If $\lambda_1\mu_1^{(1)} = \lambda_2\mu_2^{(1)}$, we obtain:

$$\log\big[f_1(x)\big] - \log\big[f_2(x)\big] = \frac{1}{2}\big(\lambda_1\big\{\mu_1^{(1)}\big\}^2 - \lambda_2\big\{\mu_2^{(1)}\big\}^2\big)\frac{d_{11}\phi(x)}{\phi(x)} + o\big(\|\mu_1\|^2\big) + o\big(\|\mu_2\|^2\big).$$

We deduce that:

$$\begin{aligned}
\mathrm{KL}(\mathbb{P}_{\theta_1}, \mathbb{P}_{\theta_2}) &= \int\bigg[\frac{1}{2}\big(\lambda_1\big\{\mu_1^{(1)}\big\}^2 - \lambda_2\big\{\mu_2^{(1)}\big\}^2\big)\frac{d_{11}\phi(x)}{\phi(x)} + o\big(\|\mu_1\|^2\big) + o\big(\|\mu_2\|^2\big)\bigg] f_1(x)\,dx \\
&= \frac{1}{2}\big(\lambda_1\big\{\mu_1^{(1)}\big\}^2 - \lambda_2\big\{\mu_2^{(1)}\big\}^2\big)\bigg[(1-\lambda_1)\int d_{11}\phi(x)\,dx + \lambda_1\int\frac{d_{11}\phi(x)\phi(x-\mu_1)}{\phi(x)}\,dx\bigg] \\
&\quad + o\big(\|\mu_1\|^2\big) + o\big(\|\mu_2\|^2\big).
\end{aligned}$$

The smoothness of ϕ leads to $\int d_{11}\phi(x)\,dx = 0$. We deduce that:

$$\begin{aligned}\int \frac{d_{11}\phi(x)\phi(x-\mu_1)}{\phi(x)}\,dx &= \int \frac{d_{11}\phi(x)}{\phi(x)}\Big[\phi(x) - \mu_1^{(1)} d_1\phi(x) + \frac{\{\mu_1^{(1)}\}^2}{2} d_{11}\phi(x) + o\big(\|\mu_1\|^2\big)\Big]\,dx \\ &= \int d_{11}\phi(x)\,dx - \mu_1^{(1)} \int \frac{d_{11}\phi(x) d_1\phi(x)}{\phi(x)}\,dx \\ &\quad + \frac{1}{2}\big\{\mu_1^{(1)}\big\}^2 \int \frac{\{d_{11}\phi(x)\}^2}{\phi(x)}\,dx + o\big(\mu_1^2\big)\,dx.\end{aligned}$$

Now, we choose for the density ϕ an even function ($\phi(x) = \phi(-x)$ for all $x \in \mathbb{R}^d$) and we obtain that

$$\mathrm{KL}(\mathbb{P}_{\theta_1}, \mathbb{P}_{\theta_2}) = \frac{1}{2}\big\{\mu_1^{(1)}\big\}^2 \mathcal{I}_\phi + \underset{n\to+\infty}{o}\big(\|\mu_1\|^2\big),$$

where the last line comes from the fact that $x \mapsto d_{11}\phi(x) d_1\phi(x)/\phi(x)$ is an odd function and the definition of $\mathcal{I}_\phi$ (see (4.3)). Finally, since $\lambda_1\mu_1^{(1)} = \lambda_2\mu_2^{(1)}$, we deduce that:

$$\begin{aligned}\mathrm{KL}(\mathbb{P}_{\theta_1}, \mathbb{P}_{\theta_2}) &= \frac{1}{4}\big(\lambda_1\big\{\mu_1^{(1)}\big\}^2 - \lambda_2\big\{\mu_2^{(1)}\big\}^2\big)\lambda_1\|\mu_1\|^2\mathcal{I}_\phi + o\big(\|\mu_1\|^4\big) \\ &= \frac{1}{4}\bigg(1 - \frac{\lambda_1}{\lambda_2}\bigg)\lambda_1^2\|\mu_1\|^4\mathcal{I}_\phi + o\big(\|\mu_1\|^4\big). \qquad \text{(B.6)}\end{aligned}$$

Next, let $\bar\lambda \in (0,1)$. Choosing $\lambda_2 = \frac{\bar\lambda}{2} < \bar\lambda$ and $\lambda_1 = \frac{1}{\alpha}\lambda_2$ with $\alpha = \frac{1+\sqrt{5}}{2}$, we have:

$$\bigg(1 - \frac{\lambda_1}{\lambda_2}\bigg)\lambda_1^2 = (\lambda_1 - \lambda_2)^2.$$

Thus,

$$\mathrm{KL}(\mathbb{P}_{\theta_1}, \mathbb{P}_{\theta_2}) = \frac{1}{4}(\lambda_2 - \lambda_1)^2\|\mu_1\|^4\mathcal{I}_\phi + o\big(\|\mu_1\|^4\big).$$

In order to apply Lemma B.1, let $\delta > 0$ such that $2\delta = \rho(\theta_1, \theta_2) \wedge \rho(\theta_2, \theta_1)$. According to our constraint $\lambda_1\mu_1^{(1)} = \lambda_2\mu_2^{(1)}$ and $\lambda_2 = \alpha\lambda_1 > \lambda_1$, we observe that $\mu_2^{(1)} < \mu_1^{(1)}$ so that:

$$2\delta = \|\mu_2\|^2|\lambda_1 - \lambda_2|.$$

We deduce that:

$$|\lambda_1 - \lambda_2|\|\mu_1\|^2 = |\lambda_1 - \lambda_2|\bigg(\frac{\lambda_2}{\lambda_1}\bigg)^2\|\mu_2\|^2 = 2\delta\alpha^2$$

and

$$\|\mu_1\|^2 = \bigg(\frac{\lambda_2}{\lambda_1}\bigg)^2\|\mu_2\|^2 = \alpha^2\frac{4\alpha}{(\alpha-1)\bar\lambda}\delta.$$

Thus,

$$\mathrm{KL}(\mathbb{P}_{\theta_1}, \mathbb{P}_{\theta_2}) = \delta^2\alpha^4\mathcal{I}_\phi + o\big(\delta^2\big),$$

and according to Lemma B.1, we obtain:

$$\inf_{\hat\theta}\sup_{\theta\in\Theta_n(c)} \mathbb{E}\big[\|\mu\|^4(\lambda - \hat\lambda)^2\big] \geq \frac{\delta^2}{2}\bigg\{1 - \sqrt{\frac{n}{2}\delta^2\big[\alpha^4\mathcal{I}_\phi + o(1)\big]}\bigg\}.$$

The choice of δ is determined by the right brackets that should be non-negative. We can choose:

$$\delta = \left[2n\alpha^4\mathcal{I}_\phi\right]^{-\frac{1}{2}},$$

so that $\frac{n}{2}\delta^2[\alpha^4\mathcal{I}_\phi + o(1)] = \frac{1}{4}(1+o(1))$. Thus, an integer N exists such that:

$$\forall n \geq N \quad \inf_{\hat{\theta}} \sup_{\theta\in\Theta_n(c)} \mathbb{E}\left[\|\mu\|^4(\lambda-\hat{\lambda})^2\right] \geq \frac{\delta^2}{6} = \frac{1}{12\alpha^4\mathcal{I}_\phi n}.$$

This ends the proof of the first point.

Point (ii): We define the loss function $\rho(\theta_1,\theta_2) = \lambda_1\|\mu_1\|\|\mu_1-\mu_2\|$ and $\Phi(t) = t^2$. The function ρ satisfies the weak triangle inequality (B.1):

$$\begin{aligned}
&\forall(\theta_1,\theta_2,\theta_3) \in \Theta_n(c)^3 : \\
&\rho(\theta_1,\theta_3) + \rho(\theta_2,\theta_3) = \lambda_1\|\mu_1\|\|\mu_1-\mu_3\| + \lambda_2\|\mu_2\|\|\mu_2-\mu_3\| \\
&\qquad\qquad \geq \min\left(\lambda_1\|\mu_1\|, \lambda_2\|\mu_2\|\right)\|\mu_1-\mu_2\| \\
&\qquad\qquad \geq \rho(\theta_1,\theta_2) \wedge \rho(\theta_2,\theta_1).
\end{aligned}$$

The proof follows the same lines as the ones of (i) and our starting point is once again the Kullback–Leibler divergence asymptotics given in Equation (B.6). Our baseline relationship $\lambda_1\mu_1 = \lambda_2\mu_2$ is still necessary and we obtain while choosing $\mu_1 = (\mu_1^{(1)}, 0, \ldots, 0)$ and $\mu_2 = (\mu_2^{(1)}, 0, \ldots, 0)$:

$$\mathrm{KL}(\mathbb{P}_{\theta_1}, \mathbb{P}_{\theta_2}) = \frac{\mathcal{I}_\phi}{4}\left(1 - \frac{\lambda_2}{\lambda_1}\right)\lambda_1^2\mu_1^4 + o\left(\|\mu_1\|^4\right).$$

We choose $\mu_1 = 2\mu_2$ so that $\lambda_2 = 2\lambda_1$ and:

$$\rho(\theta_1,\theta_2) \wedge \rho(\theta_2,\theta_1) = \lambda_1\|\mu_1\|\|\mu_1-\mu_2\| = \frac{1}{2}\lambda_1\|\mu_1\|^2 := 2\delta.$$

The coefficients λ_1 and λ_2 can be made explicit, e.g., $\lambda_1 = \bar{\lambda}/2$ and $\lambda_2 = \bar{\lambda}$. This choice implies that $\mu_1^{(1)} = 2\sqrt{2\delta/\bar{\lambda}}$. These settings can be used in the result of Lemma B.1 and we obtain:

$$\inf_{\hat{\theta}} \sup_{\theta\in\Theta_n(c)} \mathbb{E}\left[\lambda^2\mu^2(\mu-\hat{\mu})^2\right] \geq \frac{\delta^2}{2}\left\{1 - \sqrt{\frac{n\delta^2}{2}\left[2\mathcal{I}_\phi + o(1)\right]}\right\}.$$

We can obtain an efficient lower bound by choosing:

$$\delta_n := \frac{1}{2\sqrt{n\mathcal{I}_\phi}},$$

which implies, of course, that $\mu_1 = o(1)$ and $\mu_2 = o(1)$. According to this choice, an integer N exists such that $\forall n \geq N$:

$$\inf_{\hat{\theta}} \sup_{\theta\in\Theta_n(c)} \mathbb{E}\left[\lambda^2\|\mu\|^2\|\mu-\hat{\mu}\|^2\right] \geq \frac{1}{8n\mathcal{I}_\phi} \times \left(1-\frac{1}{2}\right)/2 = \frac{1}{32n\mathcal{I}_\phi}.$$

This ends the proof of the second point. □

Acknowledgements

This work was partially supported by the French Agence Nationale de la Recherche (ANR-13-JS01-0001-01, project MixStatSeq).

References

[1] A. Azzalini. A class of distributions which includes the normal ones. *Scand. J. Stat.* (1985) 171–178. MR0808153

[2] S. Balakrishnan, M. Wainwright and B. Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *Ann. Statist.* **45** (2017) 77–120. MR3611487 https://doi.org/10.1214/16-AOS1435

[3] D. Bontemps and S. Gadat. Bayesian methods for the shape invariant model. *Electron. J. Stat.* **8** (1) (2014) 1522–1568. MR3263130 https://doi.org/10.1214/14-EJS933

[4] L. Bordes, S. Mottelet and P. Vandekerkhove. Semiparametric estimation of a two-component mixture model. *Ann. Statist.* **34** (3) (2006) 1204–1232. MR2278356 https://doi.org/10.1214/009053606000000353

[5] F. Bunea, A. B. Tsybakov, M. H. Wegkamp and A. Barbu. Spades and mixture models. *Ann. Statist.* **38** (4) (2010) 2525–2558. MR2676897 https://doi.org/10.1214/09-AOS790

[6] C. Butucea and P. Vandekerkhove. Semiparametric mixtures of symmetric distributions. *Scand. J. Stat.* **41** (1) (2014) 227–239. MR3181141 https://doi.org/10.1111/sjos.12015

[7] T. T. Cai, X. J. Jeng and J. Jin. Optimal detection of heterogeneous and heteroscedastic mixtures. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** (5) (2011) 629–662. https://doi.org/10.1111/j.1467-9868.2011.00778.x. MR2867452 https://doi.org/10.1111/j.1467-9868.2011.00778.x

[8] T. T. Cai, J. Jin and M. G. Low. Estimation and confidence sets for sparse normal mixtures. *Ann. Statist.* **35** (6) (2007) 2421–2449. MR2382653 https://doi.org/10.1214/009053607000000334

[9] J. H. Chen. Optimal rate of convergence for finite mixture models. *Ann. Statist.* **23** (1) (1995) 221–233. MR1331665 https://doi.org/10.1214/aos/1176324464

[10] A. P. Dempster, N. M. Laird and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** (1) (1977) 1–38. With discussion. MR0501537

[11] S. Frühwirth-Schnatter. *Finite Mixture and Markov Switching Models*. *Springer Series in Statistics*, xx+492. Springer, New York, 2006. MR2265601

[12] S. Gadat, J. Kahn, C. Marteau and C. Maugis-Rabusseau Parameter recovery in two-component contamination mixtures: The L^2 strategy. Preprint, hal-01713035, 2018. Available at https://hal.archives-ouvertes.fr/hal-01713035.

[13] C. R. Genovese and L. Wasserman. Rates of convergence for the Gaussian mixture sieve. *Ann. Statist.* **28** (4) (2000) 1105–1127. MR1810921

[14] S. Ghosal and A. W. van der Vaart. Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Statist.* **29** (5) (2001) 1233–1263. MR1873329

[15] P. Heinrich and J. Kahn. Strong identifiability and optimal minimax rates for finite mixture estimation. *Ann. Statist.* **46** (6A) (2018) 2844–2870. MR3851757

[16] N. Ho and X. Nguyen. Convergence rates of parameter estimation for some weakly identifiable finite mixtures. *Ann. Statist.* **44** (6) (2016) 2726–2755. MR3576559 https://doi.org/10.1214/16-AOS1444

[17] N. Ho and X. Nguyen. On strong identifiability and convergence rates of parameter estimation in finite mixtures. *Electron. J. Stat.* **10** (1) (2016) 271–307. https://doi.org/10.1214/16-EJS1105. MR3466183 https://doi.org/10.1214/16-EJS1105

[18] P. J. Huber et al. Robust estimation of a location parameter. *Ann. Math. Stat.* **35** (1) (1964) 73–101. MR161415

[19] D. R. Hunter, S. Wang and T. P. Hettmansperger. Inference for mixtures of symmetric distributions. *Ann. Statist.* **35** (1) (2007) 224–251. MR2332275 https://doi.org/10.1214/009053606000001118

[20] W. Kruijer, J. Rousseau and A. W. van der Vaart. Adaptive Bayesian density estimation with location-scale mixtures. *Electron. J. Stat.* **4** (2010) 1225–1257. MR2735885 https://doi.org/10.1214/10-EJS584

[21] B. Laurent, C. Marteau and C. Maugis-Rabusseau. Non asymptotic detection of two component mixtures with unknown means. *Bernoulli* **22** (2016) 242–274. MR3449782 https://doi.org/10.3150/14-BEJ657

[22] L. Le Cam and G. Yang. *Asymptotics in Statistics: Some Basic Concepts*. *Springer Series in Statistics*. Springer Verlag, New-York, 2000. MR1784901 https://doi.org/10.1007/978-1-4612-1166-2

[23] C. Maugis and B. Michel. A non asymptotic penalized criterion for Gaussian mixture model selection. *ESAIM Probab. Stat.* **15** (2011) 41–68. MR2870505 https://doi.org/10.1051/ps/2009004

[24] C. Maugis-Rabusseau and B. Michel. Adaptive density estimation for clustering with Gaussian mixtures. *ESAIM Probab. Stat.* **17** (2013) 698–724. https://doi.org/10.1051/ps/2012018. MR3126158 https://doi.org/10.1051/ps/2012018

[25] G. McLachlan and D. Peel. *Finite Mixture Models*. Wiley series in Probability and Statistics, 2000. MR1789474 https://doi.org/10.1002/0471721182

[26] X. Nguyen. Convergence of latent mixing measures in finite and infinite mixture models. *Ann. Statist.* **41** (1) (2013) 370–400. MR3059422 https://doi.org/10.1214/12-AOS1065

[27] R. K. Patra and B. Sen. Estimation of a two-component mixture model with applications to multiple testing. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** (4) (2016) 869–893. MR3534354 https://doi.org/10.1111/rssb.12148

[28] C. Stein. Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9** (1981) 1135–1151. MR0630098

[29] C. F. J. Wu. On the convergence properties of the EM algorithm. *Ann. Statist.* **11** (1983) 95–103. MR684867

[30] B. Yu Assouad, Fano, and Le Cam. *Festschrift for Lucien Le Cam*. Springer Verlag, 1997. MR1462931