ESAIM: PS 27 (2023) 402–460
https://doi.org/10.1051/ps/2023004

ESAIM: Probability and Statistics
www.esaim-ps.org

ROBUST ESTIMATION IN FINITE MIXTURE MODELS*

Alexandre Lecestre**

Abstract. We observe a n-sample, the distribution of which is assumed to belong, or at least to be close enough, to a given mixture model. We propose an estimator of this distribution that belongs to our model and possesses some robustness properties with respect to a possible misspecification of it. We establish a non-asymptotic deviation bound for the Hellinger distance between the target distribution and its estimator when the model consists of a mixture of densities that belong to VC-subgraph classes. Under suitable assumptions and when the mixture model is well-specified, we derive risk bounds for the parameters of the mixture. Finally, we design a statistical procedure that allows us to select from the data the number of components as well as suitable models for each of the densities that are involved in the mixture. These models are chosen among a collection of candidate ones and we show that our selection rule combined with our estimation strategy result in an estimator which satisfies an oracle-type inequality.

Mathematics Subject Classification. 62G05, 62G35, 62F35, 62G07.

Received January 25, 2022. Accepted January 27, 2023.

1. Introduction

Mixture models are a flexible tool for modeling heterogeneous data, *e.g.* from a population consisting of multiple hidden homogeneous subpopulations. Finite mixture models are models containing distribution of the form

$$P_{w,F} = \sum_{k=1}^{K} w_k F_k, \tag{1.1}$$

where $K \geq 2$, each F_k belongs to a specific class of probability distributions (*e.g.* normal distributions in the case of Gaussian mixture models) and w belongs to the simplex $\mathcal{W}_K = \left\{w \in [0,1]^K; w_1 + \cdots + w_k = 1\right\}$. For a complete introduction to mixture models and an overview of the different applications we refer to the books of Mclachlan and Peel [22] and Frühwirth-Schnatter [12].

Assume we have a sample $\mathbf{X} := (X_1, \ldots, X_n)$ of i.i.d. data, each coordinate following the probability distribution P^*. The majority of the statistical methods based on finite mixture models aim to solve one of the

*This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement N° 811017

Keywords and phrases: Finite mixture model, robust estimation, supremum of an empirical process.

Department of Mathematics, University of Luxembourg, Maison du Nombre, 6 Avenue de la Fonte, 4364 Esch-sur-Alzette, Grand Duchy of Luxembourg.

** Corresponding author: alexandre.lecestre@uni.lu

following problems: density estimation (estimation of P^*), parameter estimation (estimation of w^* and/or F^* assuming $P^* = P_{w^*,F^*}$) and clustering. The monographs of Everitt and Hand [11] or Titterington *et al.* [27] provide a good overview of the different estimation methods that have been developed for mixture models such as maximum likelihood, minimum chi-square, moments method and Bayesian approaches. Although algorithms are numerous, theoretical guarantees are mostly asymptotic and restricted to very specific situations. To our knowledge, only a few non-asymptotic results have been established in the case of density estimation based on Gaussian Mixture Models (GMMs). The approximation and entropy properties of Gaussian mixture sieves have been investigated by Kruijer *et al.* [18], Ghosal and van der Vaart [15] and Genovese and Wasserman [14] where bounds on the convergence rate are given for the MLE and Bayesian estimators. Similarly, Maugis and Michel [21] use a penalized version of the MLE to build a Gaussian mixture estimator with non asymptotic adaptive properties proven in [20]. However, those results rely on relatively strong assumptions and estimators are not proved to be robust to small departures from those assumptions.

This paper aims to provide non-asymptotic results in a very general setting. In our framework, the data are assumed to be independent but not necessarily i.i.d. Our mixture model consists of probabilities of the form (1.1) where the F_k admit densities, called *emission densities*, that belong to classes of function that are VC-subgraph. We investigate the performances of ρ-estimators, as defined by Baraud and Birgé [3], on finite mixture models. This paper only focuses on the theoretical aspects and performances. We do not consider here the problem of computing estimators in practice. Our main result, Theorem 3.1, is an exponential deviation inequality for the risk of the estimator $\hat{P}$, which is measured with an Hellinger-type loss. We get an upper bound on the risk that is the sum of two terms. The first one is an approximation term which provides a measure of the distance between the true distribution of the data and our mixture model. The second term is a complexity term that depends on the classes containing the emission densities and which is proportional to the sum of their VC-indices. We deduce from this deviation bound that the estimator is not only robust with respect to model misspecification but also to contamination and the presence of outliers among the data set. Dealing with models that may be approximate allows to build estimators that possess properties over wider classes of distribution. Ghosal and Van der Vaart [15] used finite location-scale Gaussian mixtures to approximate general Gaussian mixtures with compactly supported mixing distribution. They consider mixtures with scale parameters lying between two constants that depend on the true distribution. By using a similar approximation (see Prop. 3.5), we show in Theorem 3.6 that our estimator achieves the same rate of convergence but without any restriction on the scale parameters so that the model we consider does not depend on the true mixing distribution. In particular, our result is insensitive to translation or rescaling.

Under suitable identifiability assumptions and when the distribution of the data belongs to our model, hence is of the form (1.1), we also analyze the performance of our estimators of the parameters $w_1, \ldots, w_K$ and $F_1, \ldots, F_K$. In order to establish convergence rates, we relate the Hellinger distance between the distribution of the data and its estimator to a suitable distance between the corresponding parameters. A general technique is using Fisher's information and results of Ibragimov and Has'minskiĭ [17] for regular parametric models. We can also use other results specific to parameter estimation in mixture models such as what Gadat *et al.* [13] proved in the context of two component mixtures with one known component. In both situations, we obtain, up to a logarithmic parameter, the usual $1/\sqrt{n}$-rate of convergence for regular parametric models. We also provide with Theorem 3.13 the example of a parametric model for which our techniques allow us to establish faster convergence rates while classical methods based on the likelihood or the least-squares fail to apply and hence give nothing.

In many applications, starting with a single mixture model may be restrictive and a more reasonable approach is to consider candidate ones for estimating the number of components of the mixture and proposing suitable models for the emission densities. To tackle this problem, we design a model selection procedure from which we establish, under suitable assumptions, an oracle-type inequality. We consider several illustrations of this strategy. For example, we use a penalized estimator to select the number of components of a Gaussian mixture estimator and obtain similar adaptivity results as Maugis and Michel [20]. We also consider a model with a fixed number of components but each emission density can either belong to the Gaussian or to the Cauchy location-scale family. We prove that if we know the number of components, we can estimate consistently the proportions of

Gaussian and Cauchy components as well as their location and scale parameters. To our knowledge, this result is the first of its kind.

The extension of the theory of ρ-estimation to mixture models is based on Proposition A.1 below. The proof of this result relies on an upper bound for the expectation of the supremum of an empirical process over a mixture of VC-subgraph classes. It generalizes the result that was previously established for a single VC-subgraph class. The key argument in the proof is the uniform entropy property of VC-subgraph classes that still holds for the overall density mixture model with lower bounded weights.

The paper is organized as follows. We describe our statistical framework in Section 2. In Section 3, we present the construction of the estimator on a single mixture model. We state the general result for density estimation on a single model and illustrate the performance of the estimator on the specific example of GMMs. The problem of estimating the parameters of the mixture is addressed in the Section 3.5. Finally, Section 4 is devoted to model selection criterion and the properties of the estimator on the selected model. The appendix contains all the proofs that are gathered in the same sections when they are related. Those sections include the main results, density estimation, the parametric estimation in regular parametric models, the case of two-component mixtures with one known component and the lemmas.

2. The statistical framework

We observe n independent random variables $X_1, X_2, \dots, X_n$ with respective marginal distributions $P_1^*, P_2^*, \dots, P_n^*$ on the measurable space $(\mathscr{X}, \mathcal{X})$. We model the joint distribution $\mathbf{P}^* = P_1^* \otimes P_2^* \otimes \cdots \otimes P_n^*$ of $\mathbf{X} = (X_1, X_2, \dots, X_n)$ by a probability of the form $\overline{P}^{\otimes n}$ doing as if the observations were i.i.d. with common distribution $\overline{P}$. We assume that $\overline{P}$ is a mixture of the form (1.1) where K is a positive integer, the w_k some positive weights that satisfy $\sum_{k=1}^K w_k = 1$, and F_k probability distributions. In order to model each of these probabilities we introduce a collection $\left\{\overline{\mathscr{F}}_{k,\lambda}; k \geq 1, \lambda \in \Lambda_k\right\}$ of possible models and assume that for each $k \in \{1, \dots, K\}$, F_k belongs to $\cup_{\lambda \in \Lambda_k} \overline{\mathscr{F}}_{k,\lambda}$. We denote by $\mathscr{Q}_K$ the family of distributions of the previous form. For each $k \geq 1$, we call F_k an emission probability, $\overline{\mathscr{F}}_{k,\lambda}$ an emission model, and $\mathscr{E}_k = \left\{\overline{\mathscr{F}}_{k,\lambda}; \lambda \in \Lambda_k\right\}$ an emission family. Based on the observation of $\mathbf{X}$, our aim is to design an estimator $\hat{P}$ of $\overline{P}$ of the form

$$\hat{P} = \sum_{k=1}^{\hat{K}} \hat{w}_k \hat{F}_k \in \bigcup_{K \geq 1} \mathscr{Q}_K \tag{2.1}$$

where $\hat{K}$, $(\hat{w}_k)_{1 \leq k \leq \hat{K}}$ and $(\hat{F}_k)_k$ are estimators of K, $(w_k)_k$ and $(F_k)_k$ respectively. There is a lot of possibilities for the collections Λ_k, depending on the estimation strategy (nonparametric, polynomial basis, wavelets, ...). We illustrate it in details with the following example of usual parametric models on $\mathbb{R}$.

Example 2.1. Let us take $\Lambda_k = \{1, 2, 3\}$ with

- the Gaussian location-scale family,

$$\overline{\mathscr{F}}_{k,1} = \mathscr{G} = \{\mathcal{N}(\mu, \sigma); \mu \in \mathbb{R}, \sigma > 0\}; \tag{2.2}$$

- the Cauchy location-scale family,

$$\overline{\mathscr{F}}_{k,2} = \mathscr{C} = \{\text{Cauchy}(\mu, \sigma); \mu \in \mathbb{R}, \sigma > 0\};$$

- and the Laplace location-scale family,

$$\overline{\mathscr{F}}_{k,3} = \mathscr{L} = \{\text{Laplace}(\mu, \sigma); \mu \in \mathbb{R}, \sigma > 0\}.$$

The classical situation that has been considered in the literature corresponds to the case where the collection $\{\overline{\mathscr{F}}_{k,\lambda}; k \geq 1, \lambda \in \Lambda_k\}$ reduces to a single emission model $\mathscr{F}$, for example the family of Gaussian distributions, and the problem is to estimate K and the emission probabilities F_k under the assumption that they all belong to $\mathscr{F}$. This assumption is quite restrictive and we rather consider a collection $\mathscr{E}_k$ of candidate models for F_k that may even depend on k. We say that $\mathscr{E}_k$ is simple when it reduces to a single emission model $\overline{\mathscr{F}}_k$ and composite otherwise.

In order to evaluate the performance of the estimator $\hat{P}$, we introduce on the set $\mathscr{P}$ of all product probabilities on $(\mathscr{X}^n, \mathcal{X}^{\otimes n})$ the Hellinger-type distance $\mathbf{h}$ defined by

$$\mathbf{h}(\mathbf{Q}, \mathbf{Q}') = \sqrt{\sum_{i=1}^{n} h^2(Q_i, Q_i')}, \quad \text{for } \mathbf{Q} = \bigotimes_{i=1}^{n} Q_i, \mathbf{Q}' = \bigotimes_{i=1}^{n} Q_i' \in \mathscr{P}, \tag{2.3}$$

where h is the Hellinger distance on the set $\mathscr{P}$ of probability distributions on $(\mathscr{X}, \mathcal{X})$. We recall that for Q, Q' in $\mathscr{P}$

$$h^2(Q, Q') = \frac{1}{2} \int \left(\sqrt{\frac{dQ}{d\mu}} - \sqrt{\frac{dQ'}{d\mu}} \right)^2 d\mu,$$

where μ is a measure that dominates both Q and Q', the result being independent of μ.

Assumption 1. For all $k \geq 1$, the set Λ_k is at most countable (which means finite or countable) and such that for all λ in Λ_k, $\overline{\mathscr{F}}_{k,\lambda}$ contains an at most countable subset $\mathscr{F}_{k,\lambda}$ which is dense in $\overline{\mathscr{F}}_{k,\lambda}$ with respect to the Hellinger distance h.

This assumption is only made for technical reasons, *i.e.* it ensures the measurability of the different objects considered in the proofs. But it is not really restrictive as, from a very practical point of view, one would only deal with rational numbers which already restrict to countable models. Moreover, one can check that $\mathscr{F}_{k,1} = \{\mathcal{N}(\mu, \sigma); \mu \in \mathbb{Q}, \sigma \in \mathbb{Q} \cap (0, \infty)\}$ satisfy our assumption in the context of Example 2.1. It holds as well for $\mathscr{F}_{k,2}$ and $\mathscr{F}_{k,3}$ with the same construction. Given Assumption 1 we can fix some notation. The countability condition implies that there exists a σ-finite measure μ that dominates all the $\overline{\mathscr{F}}_{k,\lambda}$ for $k \geq 1$ and $\lambda \in \Lambda_k$. Throughout this paper, we fix such a measure μ and associate to each emission model $\overline{\mathscr{F}}_{k,\lambda}$ a family of density distributions $\overline{\mathcal{F}}_{k,\lambda}$ such that $\overline{\mathscr{F}}_{k,\lambda} = \{f \cdot \mu; f \in \overline{\mathcal{F}}_{k,\lambda}\}$. In all the different examples considered in the rest of the paper μ is the Lebesgue measure. As explained, Assumption 1 is necessary for very technical reasons. Next assumption allows to bound the "dimension" of the model (see the introduction or Prop. A.1).

Assumption 2. For all $k \geq 1$ and $\lambda \in \Lambda_k$, the family of density distributions $\overline{\mathcal{F}}_{k,\lambda}$ is VC-subgraph with VC-index smaller than or equal to $V_{k,\lambda} \geq 1$.

In order to avoid too much technicality in the core of this paper, we dedicated Section F to VC-subgraph classes of functions with the definition and proofs of the different results. The next lemma shows that the VC-index corresponds to what we expect as the "dimension" of the model in the case of multivariate for normal distributions.

Lemma 2.2. *Let $d \geq 1$. Let $Cov_{+*}(d)$ be the set of $d \times d$ symmetric and positive-definite matrices. For $\mu \in \mathbb{R}^d$ and $\Sigma \in Cov_{+*}(d)$, we denote by $g_{\mu,\Sigma}$ the density function of $\mathcal{N}(\mu, \Sigma)$ with respect to the Lebesgue measure given by*

$$g_{\mu,\Sigma}(x) := \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right).$$

Let $\mathcal{G}_d$ be the location-scale family of densities given by $\mathcal{G}_d := \left\{g_{\mu,\Sigma}; \mu \in \mathbb{R}^d, \Sigma \in Cov_{+*}\right\}$. For a fixed Σ, we denote by $\mathcal{G}_{loc}(\Sigma)$ the associated location family given by $\mathcal{G}_{loc}(\Sigma) := \left\{g_{\mu;\Sigma}; \mu \in \mathbb{R}^d\right\}$. The sets $\mathcal{G}_d$ and $\mathcal{G}_{loc}(\Sigma)$ are VC-subgraph with VC-index bounded by $3 + \frac{d(d+3)}{2}$ and $3 + d$ respectively.

The dependence in d is linear and quadratic for the location family and location-scale family respectively, as for the number of parameters needed to describe each class. Throughout this paper we shall use the following notation. For $\mathbf{P} = P_1 \otimes \cdots \otimes P_n \in \boldsymbol{\mathscr{P}}$ and $\mathscr{A} \subset \mathscr{P}$, we write

$$\mathbf{h}^2(\mathbf{P}, \mathscr{A}) = \inf_{Q \in \mathscr{A}} \mathbf{h}^2\left(\mathbf{P}, Q^{\otimes n}\right) = \inf_{Q \in \mathscr{A}} \sum_{i=1}^{n} h^2(P_i, Q).$$

For $x \in \mathbb{R}$, $\lfloor x \rfloor$ is the only integer satisfying $\lfloor x \rfloor \leq x < \lfloor x \rfloor + 1$ and similarly $\lceil x \rceil$ denotes the integer satisfying $\lceil x \rceil - 1 < x \leq \lceil x \rceil$. Moreover, if $x > 0$ we write $\log_+(x) = \log(x) \vee 0$. If A is a finite set, we denote its cardinal by $|A|$ and if A is infinite, we write $|A| = \infty$. For k in $\mathbb{N}^*$, we denote by $[k]$ the set $\{1, 2, \ldots, k\}$. The notation $C(\theta)$ will mean that the constant $C = C(\theta)$ depends on the parameter or set of parameters θ.

3. Estimation on a mixture model based on simple emission families

In this section, we assume that the $\mathscr{E}_k = \left\{\overline{\mathscr{F}}_k\right\}$ are simple for all $k \geq 1$ and that $\overline{P}$ belongs to $\mathscr{Q}_K$ for some known value of $K \geq 1$. This means that we know that $\overline{P}$ is a mixture of at most K emission probabilities $F_1, \ldots, F_K$ and that F_k belongs to $\overline{\mathscr{F}}_k$ for all $k \in [K]$. Under Assumption 2, we denote by V_k the VC-index of $\overline{\mathscr{F}}_k$.

3.1. Construction of the estimator on $\mathscr{Q}_K$

For δ in $(0, 1/K]$, we define the subset $\mathscr{Q}_{K,\delta}$ of $\mathscr{Q}_K$ by

$$\mathscr{Q}_{K,\delta} := \left\{ \sum_{k=1}^{K} w_k F_k \in \mathscr{Q}_K; w \in \mathcal{W}_K \cap \left([\delta, 1] \cap \mathbb{Q}\right)^K, F_k \in \mathscr{F}_k \right\} \tag{3.1}$$

where the $\mathscr{F}_k$ are the countable and dense subsets of $\overline{\mathscr{F}}_k$ provided by Assumption 1. We associate to $\mathscr{Q}_{K,\delta}$ the family $\mathcal{Q}_{K,\delta}$ of densities with respect to μ and the ρ-estimator $\hat{P}_\delta$ of $\overline{P}$ based on the family $\mathcal{Q}_{K,\delta}$. We recall that $\hat{P}_\delta$ is defined as follows. Given

$$\psi : \left| \begin{array}{ccc} [0, +\infty] & \to & [-1, 1] \\ x & \mapsto & \frac{x-1}{x+1} \end{array} \right. , \tag{3.2}$$

we set for $\mathbf{x} = (x_1, ..., x_n) \in \mathscr{X}^n$ and $q, q' \in \mathcal{Q}_{K,\delta}$

$$\mathbf{T}(\mathbf{x}, q, q') := \sum_{i=1}^{n} \psi\left(\sqrt{\frac{q'(x_i)}{q(x_i)}}\right), \tag{3.3}$$

with the convention $0/0 = 1$ and $a/0 = +\infty$ for all $a > 0$, and

$$\boldsymbol{\Upsilon}(\mathbf{X}, q) := \sup_{q' \in \mathcal{Q}_{K,\delta}} \mathbf{T}(\mathbf{X}, q, q'). \tag{3.4}$$

The ρ-estimator $\hat{P}_\delta$ is any measurable element of the closure (with respect to the Hellinger distance) of the set

$$\mathscr{E}(\psi, \mathbf{X}) := \left\{ Q = q \cdot \mu; q \in \mathcal{Q}_{K,\delta}, \Upsilon(\mathbf{X}, q) < \inf_{q' \in \mathcal{Q}_{K,\delta}} \Upsilon(\mathbf{X}, q') + 11.36 \right\}. \tag{3.5}$$

This construction follows [3] and the constant 11.36 is given by (7) in [3]. This constant does not play an essential role and can be replaced by any smaller positive number. Ideally, one would take an estimator that achieves the infimum but it might happen that no minimizer exists. Using (3.5) allows to avoid this problem without significantly deteriorating the deviation bounds we obtain for our estimator.

As explained earlier, we only focus on the theoretical aspects in this paper. Although ρ-estimators have been developed to obtain theoretical rather than computational properties, it is possible to actually compute the estimators in practice for some models and to run simulations, as in Baraud and Chen [5] (Sect. 5).

3.2. The performance of the estimator

The following result holds.

Theorem 3.1. *Let $\delta \in (0, 1/K]$ and $\xi > 0$. Assume that Assumptions 1 and 2 hold and set $\overline{V} = V_1 + \cdots + V_K$. Any ρ-estimator $\hat{P}_\delta$ on $\mathcal{Q}_{K,\delta}$ satisfies with probability at least $1 - e^{-\xi}$,*

$$\begin{aligned} \mathbf{h}^2\left(\mathbf{P}^*, \left(\hat{P}_\delta\right)^{\otimes n}\right) \leq\ & c_0\left[\mathbf{h}^2\left(\mathbf{P}^*, \mathcal{Q}_K\right) + n(K-1)\delta\right] \\ & + c_1 116.1 \overline{V}\left[5.82 + \log\left(\frac{(K+1)^2}{\delta}\right) + \log_+\left(\frac{n}{\overline{V}}\right)\right] \\ & + c_1(1.49 + \xi), \end{aligned} \tag{3.6}$$

where $c_0 = 300$, $c_1 = 5014$. In particular, for the choice $\delta = \frac{\overline{V}}{n(K-1)} \bigwedge \frac{1}{K}$, the resulting estimator $\hat{P} = \hat{P}_\delta$ satisfies

$$C\mathbf{h}^2\left(\mathbf{P}^*, \hat{P}^{\otimes n}\right) \leq \mathbf{h}^2(\mathbf{P}^*, \mathcal{Q}_K) + \overline{V}\left[1 + \log\left(\frac{Kn}{\overline{V} \wedge n}\right)\right] + \xi, \tag{3.7}$$

with probability at least $1 - e^{-\xi}$, where C is a positive universal constant.

The proof of the theorem is postponed to Section B.2. One can notice that the bound we obtain does not depend on the space $\mathscr{X}$, *e.g.* on the dimension d in the case $\mathscr{X} = \mathbb{R}^d$, but only on the VC-indices $V_1, \ldots, V_K$ and on δ. Inequality (3.6) shows the influence of the choice of the parameter δ on the performance of the estimator $\hat{P}_\delta$. Hereafter, we shall choose δ as in the second part of Theorem 3.1 and therefore only comment on inequality (3.7). Given $\overline{P}$ in $\mathcal{Q}_K$, it follows from the triangle inequality and the fact that $(a+b)^2 \leq 2a^2 + 2b^2$ for all non-negative numbers a and b, that

$$nh^2\left(\overline{P}, \hat{P}\right) = \mathbf{h}^2\left(\overline{P}^{\otimes n}, \hat{P}^{\otimes n}\right) \leq 2\mathbf{h}^2\left(\mathbf{P}^*, \hat{P}^{\otimes n}\right) + 2\mathbf{h}^2\left(\mathbf{P}^*, \overline{P}^{\otimes n}\right).$$

We immediately derive from (3.7) that on a set of probability at least $1 - e^{-\xi}$

$$Ch^2\left(\overline{P}, \hat{P}\right) \leq \frac{1}{n}\sum_{i=1}^{n} h^2(P_i^*, \overline{P}) + \frac{\overline{V}\log\left(Kn/\overline{V}\right) + \xi}{n}. \tag{3.8}$$

In the ideal situation where the observations are i.i.d. with common distribution $\overline{P} \in \mathscr{Q}_K$, we obtain that

$$Ch^2\left(\overline{P}, \hat{P}\right) \leq \frac{\overline{V}\log\left(Kn/\overline{V}\right) + \xi}{n}.$$

Integrating this result with respect to ξ and the fact that $\overline{P}$ is arbitrary in $\mathscr{Q}_K$ leads to the uniform risk bound

$$\sup_{\overline{P}\in\mathscr{Q}_K} \mathbb{E}\left[h^2\left(\overline{P}, \hat{P}\right)\right] \leq C' \frac{\overline{V}\log\left(Kn/\overline{V}\right)}{n}, \tag{3.9}$$

where C' is a positive universal constant. This means that up to a logarithmic factor, the estimator $\hat{P}$ uniformly converges over $\mathscr{Q}_K$ at the rate $1/\sqrt{n}$ with respect to the Hellinger distance. One knows that when working with the Hellinger distance, no estimator can do better that this $1/\sqrt{n}$ rate (see (1.1) in [6]).

We can see that we only need to bound the quantity $\overline{V}$ to deduce deviation inequalities in specific cases. Therefore, we can already get a bound on the convergence rate for Gaussian mixtures with Lemma 2.2.

Corollary 3.2.

- *Let $\mathscr{Q}_K$ be the Gaussian location-scale mixture model, i.e. $\overline{\mathscr{F}}_1 = \cdots = \overline{\mathscr{F}}_K = \left\{\mathcal{N}(\mu,\Sigma); \mu \in \mathbb{R}^d, \Sigma \in Cov_{+*}\right\}$. There is a positive universal constant $C > 0$ such that, for any ρ-estimator $\hat{P} = \hat{P}_\delta$ on $\mathscr{Q}_{K,\delta}$, for all $\overline{P} \in \mathscr{Q}_K$ and for all $\xi > 0$, we have*

$$Ch^2\left(\overline{P}, \hat{P}\right) \leq \frac{Kd^2\left[1 + \log\left(\frac{n}{d^2} \vee K\right)\right] + \xi}{n},$$

  with probability at least $1 - e^{-\xi}$.
- *Let $\mathscr{Q}_K$ be the Gaussian location mixture model associated to a fixed covariance matrix $\Sigma \in Cov_{+*}(d)$, i.e. $\overline{\mathscr{F}}_1 = \cdots = \overline{\mathscr{F}}_K = \left\{\mathcal{N}(\mu,\Sigma); \mu \in \mathbb{R}^d\right\}$. There is a positive universal constant $C > 0$ such that, for any ρ-estimator $\hat{P} = \hat{P}_\delta$ on $\mathscr{Q}_{K,\delta}$, for all $\overline{P} \in \mathscr{Q}_K$ and for all $\xi > 0$, we have*

$$Ch^2\left(\overline{P}, \hat{P}\right) \leq \frac{Kd\left[1 + \log\left(\frac{n}{d} \vee K\right)\right] + \xi}{n},$$

  with probability at least $1 - e^{-\xi}$.

Those rates would be optimal if the logarithmic factor was necessary. Doss *et al.* [10] proved it is not the case for Gaussian location mixtures with known isotropic covariance matrix. They provide an estimator that achieves the minimax rate $\sqrt{d/n}$ with respect to the Hellinger distance. However, the dependency in K of their bound in (1.12) is worse than exponential when it is just linear for our estimator.

Our assumption that the families of density functions $\overline{\mathcal{F}}_k$ are VC-subgraph is actually weak since it includes situations where these models consist of unbounded densities or densities which are not in L_2 which to our knowledge have never been considered in the literature. A concrete example of such situations is the following one. Let g be some non-increasing function on $(0, +\infty)$ which is unbounded, nonnegative and satisfies $\int_0^{+\infty} g(x)\mathrm{d}x = \frac{1}{2}$ and $\overline{\mathscr{F}}_k$ is the translation model associated to the family of densities $\left\{x \mapsto g(|x-\theta|)\mathbb{1}_{|x-\theta|>0}; \theta \in \mathbb{R}\right\}$ for all $k \in \{1, \ldots, K\}$. It follows from Proposition 42-(vi) of Baraud *et al.* [4] that the VC-index of $\overline{\mathcal{F}}_k$ is not larger than 10.

When the data are independent but not i.i.d., we derive from inequality (3.8) that the estimator $\hat{P}$ performs almost as well as in the i.i.d. case as long as the marginals $P_1^*, \ldots, P_n^*$ are close enough to $\overline{P}$. This means that the estimator is robust with respect to a possible misspecification of the model and the departure from the assumption that the data are i.i.d. In particular, this includes the situations where the dataset contains some outliers or has been contaminated. Consider Hüber's contamination model where a proportion ϵ of the data is

contaminated, *i.e.* we have $P^* = (1-\epsilon)\overline{P} + \epsilon Q$, where $\overline{P}$ is the probability distribution we want to estimate and Q is the distribution of the contaminated data. In this situation, for any probability distribution Q, using (3.8) and the convexity property of the Hellinger distance we get

$$Ch^2\left(\overline{P}, \hat{P}\right) \leq \epsilon + \frac{\overline{V}\log(n) + \xi}{n}. \tag{3.10}$$

We can see that there is no perturbation of the convergence rate as long as the contamination rate ϵ remains small as compared to $\overline{V}\log(n)/n$. Contrary to other loss functions, the Hellinger distance does not allow to obtain a better rate than $\sqrt{\epsilon}$ in the general case (see Birgé [7]). Inequality (3.14), stated later, also allows to consider misspecification for the emission models for example.

3.3. The case of totally bounded emission models

We might also consider emission models for which we do not have any bound on the VC-index. For a subset $\mathscr{N}$ of $\mathscr{P}$ and $\eta \in [0,1]$, the η-covering number $N(\eta, \mathscr{N}, h)$ of $\mathscr{N}$, with respect to the Hellinger distance, is the minimum number of balls $\mathcal{B}_h(P_i, \eta)$, $i = 1, \ldots, N$, necessary to cover $\mathscr{N}$. In that case, the set $\mathscr{N}[\eta] = \{P_i; i = 1, \ldots, N\}$ constitutes a finite approximation of $\mathscr{N}$, *i.e.* for all Q in $\mathscr{N}$ there exists $i \in \{1, \ldots, N\}$ such that $h(Q, P_i) \leq \eta$. We say that $\mathcal{N}$ is totally bounded (for the Hellinger distance) if its η-covering number is finite for all $\eta \in (0,1]$. A direct consequence of the definition of VC-subgraph classes is that any finite set $\mathcal{F}$ of real-valued functions is VC-subgraph with VC-index at most $V(\mathcal{F}) \leq \log_2(|\mathcal{F}|)$. Consequently, we can still use ρ-estimation for models that are not proven to satisfy Assumption 2 but still are such that emission models are totally bounded.

Theorem 3.3. *Let $\overline{\mathscr{F}}_k$ be a totally bounded class of distributions for all $k \in \{1, \ldots, K\}$ with $K \geq 2$. Let $\mathscr{Q}_K$ be the mixture model defined by*

$$\mathscr{Q}_K = \left\{ \sum_{k=1}^{K} w_k F_k; w \in \mathcal{W}_K, F_k \in \overline{\mathscr{F}}_k, \forall k \in \{1, \ldots, K\} \right\}.$$

Assume there are constants $A_k \geq 1$ and α_k such that $\log_2 N(\eta, \mathscr{F}_k, h) \leq \left(\frac{A_k}{\eta}\right)^{\alpha_k}$ for all k in $[K]$ and for all $\eta \in (0,1)$. Let ϵ be in $(0,1)$. For k in $[K]$, let $\mathscr{F}_k[\epsilon]$ be a minimal ϵ-net of $\overline{\mathscr{F}}_k$ such that $|\mathscr{F}_k[\epsilon]| = N(\epsilon, \mathscr{F}_k, h)$. Let $\mathscr{Q}_{K,\delta}[\epsilon]$ be the countable model defined by

$$\mathscr{Q}_{K,\delta}[\epsilon] = \{P_{w,F}; w \in \mathcal{W}_K, w_k \geq \delta, w_k \in \mathbb{Q}, F_k \in \mathscr{F}_k[\epsilon], \forall k \in \{1, \ldots, K\}\}.$$

Take $\delta = \frac{\overline{V}}{n(K-1)} \wedge \frac{1}{K}$ with

$$\overline{V} = \sum_{k=1}^{K} \log_2(|\mathscr{F}_k[\epsilon]|) \leq \sum_{k=1}^{K} \left(\frac{A_k}{\epsilon}\right)^{\alpha_k},$$

where $\epsilon = n^{-\frac{1}{\alpha_{\max}+2}}$ and $\alpha_{\max} = \max_{1 \leq k \leq K} \alpha_k$. There exists a positive constant C such that for any ρ-estimator $\hat{P} = \hat{P}_\delta$ on $\mathscr{Q}_{K,\delta}[\epsilon]$, for all $\xi > 0$, we have

$$C\mathbf{h}^2\left(\mathbf{P}^*, \left(\hat{P}_\delta\right)^{\otimes n}\right) \leq \mathbf{h}^2\left(\mathbf{P}^*, \mathscr{Q}_K\right) + n^{\frac{\alpha_{\max}}{\alpha_{\max}+2}} \sum_{k=1}^{K} A_k^{\alpha_k} \left[1 + \log(Kn)\right] + \xi,$$

with probability at least $1-e^{-\xi}$. *In particular, if the observations are i.i.d. with common distribution* $P^* \in \mathscr{P}$ *we have*

$$Ch^2\left(P^*, \hat{P}_\delta\right) \leq h^2\left(P^*, \mathscr{Q}_K\right) + n^{-\frac{2}{\alpha_{\max}+2}} \sum_{k=1}^{K} A_k^{\alpha_k} \left[1 + \log\left(Kn\right)\right] + \frac{\xi}{n},$$

This theorem is proved in Section B.3 (p. 436) and we illustrate it with the following example. Doss and Wellner [9] provide a bound on the entropy for classes of log-concave and s-concave densities. Let $\mathcal{C} = \{\varphi : \mathbb{R} \to [-\infty, \infty); \varphi \text{ is a closed, proper concave function}\}$ where *proper* and *closed* are defined in [24] (Sects. 4 and 7). For $0 < M < \infty$ and $s > -1$, let $\mathcal{P}_{M,s}$ be the class of densities defined by

$$\mathcal{P}_{M,s} = \left\{p \in \mathcal{P}_s; \sup_{x\in\mathbb{R}} p(x) \leq M, 1/M \leq p(x) \text{ for all } |x| \leq 1\right\},$$

where $\mathcal{P}_s = \left\{p : \int p d\lambda = 1\right\} \bigcap h_s \circ \mathcal{C}$, λ is the Lebesgue measure on $\mathbb{R}$ and $h_s : \mathbb{R} \to \mathbb{R}$ is defined by

$$h_s(y) = \begin{cases} e^y, & s = 0 \\ (-y)_+^{1/s}, & s \in (-1, 0), \\ y_+^{1/s}, & s > 0. \end{cases}$$

We fix such values of M and s. Let $\mathcal{Q}_K$ be the density model of mixtures of s-concave densities (or log-concave for $s = 0$) defined by

$$\mathcal{Q}_K = \left\{\sum_{k=1}^{K} w_k f_k; w \in \mathcal{W}_K, f_k \in \mathcal{P}_{M,s}\right\},$$

with $K \geq 2$. Let $\mathscr{Q}_K$ be the class of distributions associated to $\mathcal{Q}_K$. The class $\mathcal{P}_{M,s}$ is not proven to be VC-subgraph but it is totally bounded. As a direct consequence of Theorem 3.1 of Doss and Wellner [9], there exists a positive constant A, depending only on M and s, such that for all ϵ in $(0, 1]$, we have

$$\log_2 N(\epsilon, \mathcal{P}_{M,s}, h) \leq A\epsilon^{-1/2}.$$

In particular, it means there exists a ϵ-net $\mathcal{P}_{M,s}[\epsilon]$ such that $\log_2\left(|\mathcal{P}_{M,s}[\epsilon]|\right) \leq \left(A^2/\epsilon\right)^{1/2}$. Let $\mathcal{Q}_{K,\delta}[\epsilon]$ be the countable density model given by

$$\mathcal{Q}_{K,\delta}[\epsilon] = \left\{\sum_{k=1}^{K} w_k f_k; w \in \mathcal{W}_K, w_k \geq \delta, w_k \in \mathbb{Q}, f_k \in \mathcal{P}_{M,s}[\epsilon]\right\}.$$

One can check that $\mathscr{Q}_{K,\delta}[\epsilon]$ is also a ϵ-net of $\mathscr{Q}_{K,\delta}$ with respect to the Hellinger distance using inequality (3.14) hereafter page 412. The application of Theorem 3.3 on this example gives the following result.

Corollary 3.4. *Assume there exists* P^* *in* $\mathscr{P}$ *such that* $\mathbf{P}^* = (P^*)^{\otimes n}$. *Take* $\epsilon = n^{-2/5}$ *and* $\delta = n^{-4/5} \wedge K^{-1}$. *Let* $\hat{P} = \hat{P}_\delta$ *be a* ρ*-estimator on* $\mathscr{Q}_{K,\delta}[\epsilon]$. *There exists a constant* $C(M, s)$ *such that for all* $\xi > 0$, *we have*

$$C(M,s)h^2\left(P^*, \hat{P}\right) \leq h^2\left(P^*, \mathscr{Q}_K\right) + \frac{K}{n^{4/5}}\left[1 + \log\left(Kn\right)\right] + \frac{\xi}{n},$$

with probability at least $1-e^{-\xi}$.

This corollary provides a risk bound over the class of distributions associated to mixtures of s-concave densities. Up to a logarithmic factor, the estimator $\hat{P}$ uniformly converges over $\mathscr{Q}_K$ at the rate $n^{-2/5}$ with respect to the Hellinger distance, which is the same rate given in Theorem 3.2 of Doss and Wellner [9] for the MLE over the model $\mathcal{P}_{M,s}$, *i.e.* for $K = 1$.

3.4. Application to the estimation of a general Gaussian mixture

We denote by ϕ_σ the density function of the normal distribution (with respect to the Lebesgue measure on $\mathbb{R}$) with mean 0 and variance $\sigma^2 > 0$, *i.e.*

$$\phi_\sigma : x \mapsto \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}. \tag{3.11}$$

We assume P^* is of the following form or is close enough to a distribution of the form

$$p_H(x) = \int \phi_\sigma(x - z) \mathrm{d}H(z, \sigma), \forall x \in \mathbb{R}.$$

We say that p_H is the Gaussian mixture density with mixing distribution H. We want to approximate any distribution of this form with finite Gaussian mixtures, *i.e.* distribution with densities of the same form with mixing distribution supported on a finite set. For a mixing measure H on $\mathbb{R} \times \mathbb{R}^{+*}$, we denote by $\mathrm{supp}(H)$ its support. To obtain an approximation result, we need to consider mixing measures H that are supported on a compact set, *i.e.* there exist $A \geq 0$ and $R \geq 1$ such that $\mathrm{supp}(H) \subset [-A, A] \times [1, R]$. The Hellinger distance being invariant to translation and rescaling, we consider the following class of densities. For $A > 0$ and $R \geq 1$ we define

$$\mathcal{C}(A, R) = \left\{ p_H; \exists l \in \mathbb{R}, \exists s > 0, \mathrm{supp}(H) \subset [l - sA, l + sA] \times [s, sR] \right\}$$

and we denote by $\mathscr{C}(A, R)$ the associated class of distributions. We denote by $\mathscr{G}_{mix,K}$ the Gaussian mixture model with K components associated to the class of densities $\mathcal{G}_{mix,K}$ defined by

$$\mathcal{G}_{mix,K} := \left\{ \sum_{k=1}^{K} w_k \phi_{\sigma_k}(\cdot - z_k); w \in \mathcal{W}_K, \sigma_k \in (0, +\infty), z_k \in \mathbb{R}, \forall k \in \{1, \ldots, K\} \right\}. \tag{3.12}$$

This situation corresponds to $\overline{\mathcal{F}}_k = \mathcal{G}_1$ for all $k \in \{1, \ldots, K\}$. We can approximate the class $\mathscr{C}(A, R)$ with the model $\mathscr{G}_{mix,K}$ as indicated by the following result.

Proposition 3.5. *For $K \geq 2\left(24A^2 + 1\right)^2$, we have*

$$\sup_{P_H \in \mathscr{C}(A,R)} h^2\left(P_H, \mathscr{G}_{mix,K}\right) \leq \frac{1}{2} \exp\left(-\frac{K^{1/2}}{12\sqrt{6}R^2}\right) \left[K^{1/4} \frac{3\sqrt{2}}{\sqrt{e\pi}7^{1/4}} + R\right].$$

This proposition allows to obtain a deviation bound on the estimation over $\mathscr{C}(A, R)$, with Theorem 3.1. Its proof is postponed to Section C.2.

Theorem 3.6. *For $R \geq 1$ and $n \geq e$, we take $K = K(R, n) := \lceil 864R^4 \log^2(n) \rceil$. Let $\hat{P}$ be a ρ-estimator on $\mathscr{G}_{K,\delta}$ with δ as in (3.7) and assume the true distribution is i.i.d., i.e. $\mathbf{P}^* = (P^*)^{\otimes n}$. There exists a numeric constant*

$C > 0$, hence not depending on R, such that for all $\xi > 0$, with probability at least $1 - e^{-\xi}$, we have

$$Ch^2\left(P^*, \hat{P}\right) \leq h^2\left(P^*, \mathscr{C}(A, R)\right) + \frac{R^4 \log^3(n) + \xi}{n}, \tag{3.13}$$

for $A = A(R, n) := \sqrt{\frac{12\sqrt{3}-1}{24}} R \log^{1/2}(n)$.

This result is proven in Section C.1. Therefore, for a fixed R, we obtain a rate of $\log^{3/2}(n)/\sqrt{n}$ over $\mathscr{C}(\infty, R) := \bigcup_{A>0} \mathscr{C}(A, R)$ with respect to the Hellinger distance. We can also consider larger classes of distributions, with R increasing as n increases but it would deteriorate this rate. Our result is still an improvement of Theorem 4.2 from [15] as it requires weaker assumptions. Their result is sensitive to translation or scaling and they have to specify bounds $0 < \underline{\sigma} < \overline{\sigma}$ in the model such that H^* is supported on a compact set $[-a, a] \times [\underline{\sigma}, \overline{\sigma}]$. Moreover, our estimator is robust, to contamination for instance. Assume we have an ϵ contamination rate of our data, *i.e.* P^* is of the form $P^* = (1-\epsilon)\overline{P} + \epsilon Q$ with $\epsilon \in (0, 1)$, $\overline{P} \in \mathscr{C}(A(R, n), R)$ and Q is any probability distribution. Then, our estimator satisfies $Ch^2(P^*, \hat{P}) \leq \epsilon + \frac{R^4 \log^3(n) + \xi}{n}$ on an event of probability $1 - e^{-\xi}$. As long as ϵ remains small as compared to $R^4 \log^3(n)/n$, the rate is not deteriorated by the contamination.

3.5. Parameter estimation

We say that $\hat{w}$ and $\hat{F}$ are ρ-estimators if the resulting mixture distribution $\hat{P}$ given by

$$\hat{P} = \sum_{k=1}^{K} \hat{w}_k \hat{F}_k$$

is a ρ-estimator. We have a general result for the performance of $\hat{P}$ but not for $\hat{w}$ and $\hat{F}$. Hopefully we those parameter estimators would inherit the properties of $\hat{P}$ under additional assumptions. Some results about the robust estimation of parameters exist in the machine learning community, see Diakonikolas *et al.* [8] for instance. As before, the available results are all restricted to specific cases such as Gaussian mixture models. Convexity properties ensure that we always have the upper bound

$$h\left(P_{w,F}, P_{v,G}\right) \leq \inf_{\tau \in \mathscr{S}_K} \left\{ h(w, v \circ \tau) + \max_{k \in [K]} h\left(F_k, G_{\tau(k)}\right) \right\}, \tag{3.14}$$

for all mixing weights and emission distributions (see Lem. B.3), where $\mathscr{S}_K$ denotes the set of all permutations of $[K]$ and $\mathcal{W}_K$ is seen as the set of probability distributions on $[K]$ and justify the notation $h(w, v \circ \tau)$. Therefore, a good estimation of the mixing weights $w = (w_1, \ldots, w_K)$ and of the emission distributions $F = (F_1, \ldots, F_K)$ ensures a good estimation of the mixture distributions $P_{w,F}$. However the converse is not true as the parameters are not even identifiable in general.

Example 3.7. Let $\overline{\mathscr{F}}$ be the set of uniform distributions $\mathcal{U}(a, b)$ the uniform distribution on the interval (a, b) of positive lengths. Then the parameters w and F in the mixture model

$$\mathscr{Q}_2 = \left\{ w_1 F_1 + (1 - w_1) F_2; w_1 \in (0, 1), F_1, F_2 \in \overline{F} \right\}$$

are not identifiable since

$$\frac{3}{4}\mathcal{U}(0, 1) + \frac{1}{4}\mathcal{U}(1/3, 2/3) = \frac{1}{2}\mathcal{U}(0, 2/3) + \frac{1}{2}\mathcal{U}(1/3, 1).$$

We shall say that $P = P_{w,F}$ is identifiable (with respect to the model) if for all v in $\mathcal{W}_K$ and all G in $\mathscr{F}_1 \times \cdots \times \mathscr{F}_K$, we have

$$P_{w,F} = P_{v,G} \Rightarrow \exists \tau \in \mathscr{S}_K, \forall k \in [K], w_k = v_{\tau(k)} \text{ and } F_k = G_{\tau(k)},$$

There is a wide literature about identifiability that includes the works of Teicher [26] and Sapatinas [25] for example. Allman *et al.* [1] provides identifiability conditions in a nonparametric framework but this is quite unusual. In this section, we will consider a unique parametric model for the emission models, *i.e.* we have $\overline{\mathscr{F}}_1, \ldots, \overline{\mathscr{F}}_K \subset \{F_\theta; \theta \in \Theta\}$ with $\Theta \subset \mathbb{R}^d$ and assume $P_{w^*,F_{\theta^*}}$ is the true distribution or the best approximation within the model. Identifiability is a minimum requirement for the parameter estimators to be meaningful but we can hardly get more than consistency with it.

There is one approach that allows not to consider the identifiability issue is to consider the estimation of the mixing distribution instead of the parameters themselves, *i.e.* $w_1^*\delta_{\theta_1^*} + \cdots + w_K^*\delta_{\theta_K^*}$ where δ_x is the Dirac measure in x. Most results are given for the L_1-Wassertein metric W_1 which can be defined as follow for $\Theta \subset \mathbb{R}$. For probability distributions G_1, G_2 on Θ, we have

$$W_1(G_1, G_2) := \sup_{f \in Lip(1)} \int_\Theta f(\mathrm{d}G_1 - \mathrm{d}G_2), \tag{3.15}$$

where $Lip(1)$ is the class of Lipschitz functions with Lipschitz constant at most 1. Heinrich and Kahn [16] establish minimax rates of estimation for mixing distribution under some regularity and strong identifiability conditions. Wu and Yang [29] prove that their denoised method of moments for univariate Gaussian mixtures provides an estimator of the mixing distribution that reaches the optimal rate with respect to W_1. They also prove an oracle bound for density estimation in the case of misspecification similar to (3.10), for the total variation distance instead of the Hellinger distance. However, they only consider misspecified distributions that are sub-Gaussian and in dimension one.

Our approach is to derive bounds on the convergence rates for the parameter estimators from (3.7). Typically, we are looking for an inequality of the form

$$h(P_{w^*,F_{\theta^*}}, P_{w,F_\theta}) \geq C(w^*, \theta^*) \left[\sum_{k=1}^K d_\Theta(\theta_k^*, \theta_k) + d_{\mathcal{W}}(w^*, w) \right], \forall w \in \mathcal{W}_K, \forall \theta \in \Theta, \tag{3.16}$$

where $C(w^*, \theta^*)$ is positive, d_Θ is a distance on Θ and $d_{\mathcal{W}}$ is a distance on $\mathcal{W}_K$. Intuitively, if we can estimate each parameter individually we should be able to estimate the mixing distribution as well. Formally, for $\Theta \subset \mathbb{R}$, we have

$$W_1\left(\sum_{k=1}^K w_k^*\delta_{\theta_k^*}, \sum_{k=1}^K w_k\delta_{\theta_k} \right) \leq \sum_{k=1}^K |\theta_k^* - \theta_k| + \max_i |\theta_i^*| \cdot \sum_{k=1}^K |w_k^* - w_k|, \forall w \in \mathcal{W}_K, \forall \theta \in \Theta^K,$$

which is a direct consequence of (3.15). One can see that when d_Θ and $d_{\mathcal{W}_K}$ in (3.16) are the L_1 distance we can deduce a bound for the estimation of the mixing distribution. The main difficulty remains to obtain a lower bound on the Hellinger distance between mixtures. There are still some situations where we do have such a lower bound.

Regular parametric model

Let K be an integer larger than 1. We consider parametric emission models associated to density models $\overline{\mathcal{F}}_k = \{f_k(\cdot; \alpha), \alpha \in A_k\}$, where A_k is a subset of $\mathbb{R}^{d_k}$ for all $k \in \{1, \ldots, K\}$. It is always possible to find a countable dense subset of A_k with respect to the Euclidean distance on $\mathbb{R}^{d_k}$. We assume there is a reasonably

good connection between the Hellinger distance on the emission models and the Euclidean distances on the parameter spaces such that a dense subset of A_k would translate into a dense subset of the emission model with respect to the Hellinger distance. This assumption is very weak and does not seem to be restrictive in any way. In the different examples we consider we can always consider $A_k \cap \mathbb{Q}^{d_k}$ as a dense subset of A_k. Therefore Assumption 1 is satisfied with $\mathcal{F}_k = \{f_k(\cdot;\alpha), \alpha \in B_k\}$. We denote by $\mathscr{Q}_K$ the distribution model associated to the mixture density model

$$\mathcal{Q}_K = \left\{ p(\cdot;\theta) = \sum_{k=1}^{K-1} w_k f_k(\cdot;z_k) + (1 - w_1 - \cdots - w_{K-1}) f_K(\cdot;\alpha_K); \theta = (w,\alpha) \in \Theta \right\},$$

where Θ is an open convex subset of $\left\{ w \in (0,1)^{K-1}; \sum_{k=1}^{K-1} w_k < 1 \right\} \times A_1 \times \cdots \times A_K$. To be in the context of regular parametric models consider by Ibragimov and Has'minskiĭ [17] we need to make some assumptions.

Assumption 3. The classes of functions $\overline{\mathcal{F}}_1, \ldots, \overline{\mathcal{F}}_K$ satisfy the following regularity conditions.

a) The function $z \mapsto f_k(x;z)$ is continuous on A_k (with respect to the Euclidean distance) for μ-almost all $x \in \mathscr{X}$, for all $k \in \{1, \ldots, K\}$.
b) For all $k \in \{1, \ldots, K\}$, for μ-almost all $x \in \mathscr{X}$ the function $u \mapsto f_k(x;u)$ is differentiable at the point $u = \alpha$ and for all $j \in \{1, \ldots, d_k\}$, we have

$$\int_{\mathscr{X}} \left| \frac{\partial f_k(x;\alpha)}{\partial \alpha_j} \right|^2 \frac{\mu(\mathrm{d}x)}{f_k(x;\alpha)} < \infty.$$

c) The function $\theta \mapsto \psi(\cdot;\theta) = \frac{\partial}{\partial \theta} p^{1/2}(\cdot;\theta)$ is continuous in the space $L_2(\mu)$.
d) The class of densities $\overline{\mathcal{F}}_k$ is VC-subgraph with VC-index not larger than V_k for all $k \in \{1, \ldots, K\}$. We write $\overline{V} = V_1 = \cdots + V_k$.

The work of Ibragimov and Has'minskiĭ [17] allows to derive a deviation inequality on the Euclidean distance between parameters using Fisher's information.

Theorem 3.8. *Let $\overline{\theta}$ be in Θ. Assume the Fisher's information matrix*

$$I\left(\overline{\theta}\right) = \int_{\mathscr{X}} \frac{\partial p\left(x;\overline{\theta}\right)}{\partial \theta} \left(\frac{\partial p\left(x;\overline{\theta}\right)}{\partial \theta} \right)^T \frac{\mu(\mathrm{d}x)}{p\left(x;\overline{\theta}\right)}$$

is definite positive and $\inf_{\substack{||\overline{\theta}-\theta|| \geq a \\ \theta \in \Theta}} h^2\left(P_{\overline{\theta}}, P_\theta\right) > 0$ for all $a > 0$. Let $\hat{P} = P_{\hat{w},\hat{F}}$ be a ρ-estimator on $\mathscr{Q}_{K,\delta}$, with δ as in (3.7). There exists a positive constant $C\left(\overline{\theta}\right)$ such that for all $\xi > 0$, with probability at least $1 - e^{-\xi}$, we have

$$C\left(\overline{\theta}\right)\left(||\overline{w} - \hat{w}||^2 + \sum_{k=1}^{K} 1 \wedge ||\overline{\alpha}_k - \hat{\alpha}_k||^2 \right) \leq \frac{1}{n}\left[\mathbf{h}^2\left(\mathbf{P}^*, P_{\overline{\theta}}^{\otimes n}\right) + \overline{V}\log(n) + \xi \right]. \tag{3.17}$$

And assuming $P^* = P_{\overline{\theta}}$, we obtain the usual parametric convergence rate up to a logarithmic factor for the parameter estimators.

This result is proven in Section D.1. Following the proof and Theorem 7.6 [17], the best constant $C\left(\overline{\theta}\right)$ depends on the smallest eigenvalue of the Fisher's information matrix $I\left(\overline{\theta}\right)$ and the geometry induced by the Hellinger distance around $\overline{\theta}$ in Θ. Inequality (3.17) proves that even if "true parameters" might not exist the

parameter estimators can be meaningful as long as $\mathbf{P}^*$ is relatively close to the model. The Gaussian mixture model is the most common mixture model and it is a regular parametric model. Let $K \geq 2$ and for all k in $[K]$ take $\mathcal{F}_k = \mathcal{G}_1$, where $\mathcal{G}_1$ is given in Lemma 2.2. We define a binary relation on $\mathbb{R} \times (0, \infty)$ by

$$(z_1, \sigma_1) > (z_2, \sigma_2) \Leftrightarrow \begin{cases} \sigma_1 > \sigma_2; \\ \text{or } \sigma_1 = \sigma_2 \text{ and } z_1 > z_2. \end{cases} \tag{3.18}$$

We consider the parameters $\theta = (w_1, \ldots, w_{K-1}, z_1, \sigma_1^2, \ldots, z_K, \sigma_K^2)$ belonging to the set

$$\Theta = \left\{ \theta \in (0,1)^{K-1} \times (\mathbb{R} \times \mathbb{R}^*)^K ; \sum_{k=1}^{K-1} w_k < 1, (z_1, \sigma_1) > \cdots > (z_K, \sigma_K) \right\}.$$

Theorem 3.9. *Assume* $P^* = P_{\overline{\theta}} = \sum_{k=1}^{K} \overline{w}_k \mathcal{N}(\overline{z}_k, \overline{\sigma}_k^2)$ *such that* $(\overline{z}_1, \overline{\sigma}_1) > \cdots > (z_K, \sigma_K)$ *are all distinct and* $\inf_{1 \leq k \leq K} \overline{w}_k > 0$. *Let* $\hat{P}$ *be a* ρ*-estimator on* $\mathscr{G}_{K,\delta}$, *with* δ *as in (3.7). There exists a positive constant* $C\left(\overline{\theta}\right)$ *such that, for all* $\xi > 0$, *we have*

$$C\left(\overline{\theta}\right) \left(\sum_{k=1}^{K-1} ||\overline{w}_k - \hat{w}_k||^2 + \sum_{k=1}^{K} \left|\left| \left(\overline{z}_k, \overline{\sigma}_k^2\right) - \left(\hat{z}_k, \hat{\sigma}_k^2\right) \right|\right|^2 \wedge 1 \right) \leq \frac{5K \log(n) + \xi}{n}, \tag{3.19}$$

with probability at least $1 - e^{-\xi}$.

This result is proven in Section D.3. Our estimator reaches the optimal rate of convergence up to a logarithmic factor. One can notice that the assumption of ordered couples of parameters (z_j, σ_j^2) can be replace by considering distinct couples only and taking the infimum over permutation of the hidden states in (3.19).

Connection with the L_2*-distance*

We can use results from the literature that do not apply to the Hellinger distance but to other ones such as the L_2-distance between densities. There is a general inequality between the L_2 and Hellinger distances when the density functions are bounded, *i.e.*

$$||p - q||_2^2 \leq 4 \left(||p||_\infty + ||q||_\infty \right) h^2(P, Q). \tag{3.20}$$

Assume one can prove an inequality of the following type. For any w, v in $\mathcal{W}_K$ and any f_k, g_k in $\overline{\mathcal{F}}_k$ for all $k \in \{1, \ldots, K\}$ such that the resulting mixtures belong to our model, we have

$$\underline{c} \left(d_{\mathcal{W}}^2(w, v) + \max_{k \in [K]} d_F^2(f_k, g_k) \right) \leq \left|\left| \sum_{k=1}^{K} w_k f_k - \sum_{k=1}^{K} v_k g_k \right|\right|_2^2, \tag{3.21}$$

where $d_{\mathcal{W}}$ is a distance on $\mathcal{W}_K$ and d_F is a distance on $\bigcup_{1 \leq k \leq K} \overline{\mathcal{F}}_k$. Moreover, assuming the density models $\overline{\mathcal{F}}_k$ are uniformly bounded, *i.e.*

$$\sup_{k \in [K]} \sup_{f \in \overline{\mathcal{F}}_k} ||f||_\infty =: U < \infty, \tag{3.22}$$

we get

$$d^2_{\mathcal{W}}(w,v) + \max_{k\in[K]} d^2_F(f_k, g_{\tau(k)}) \leq \frac{8U}{\underline{c}} h^2 \left(\sum_{k=1}^K w_k F_k, \sum_{k=1}^K v_k G_k \right).$$

Here again, a density estimation result implies a result for the parameter estimation. We can apply this method to the special case of two-component mixture model with one known component. Let ϕ be a density function on $\mathbb{R}^d$ with respect to the Lebesgue measure. We consider the 2-component mixture model $\mathscr{Q}$ associated to the class of densities

$$\mathcal{Q} = \left\{ x \mapsto p_{w,z}(x) = (1-w)\phi(x) + w\phi(x-z); w \in [0,1], z \in \mathbb{R}^d \right\}, \tag{3.23}$$

with $\overline{\mathcal{F}}_1 = \{\phi\}$ and $\overline{\mathcal{F}}_2 = \left\{ x \mapsto \phi(x-z); z \in \mathbb{R}^d \right\}$. Gadat *et al.* [13] proved an inequality such as (3.21) in this situation. They still require the following assumptions on ϕ.

Assumption 4. The function ϕ belongs to $\mathcal{C}^3\left(\mathbb{R}^d\right) \cap \mathbb{L}^2\left(\mathbb{R}^d\right)$. For any $M > 0$, there exists a function g in $\mathbb{L}^2\left(\mathbb{R}^d\right)$ such that

$$\forall x \in \mathbb{R}^d, \forall z \in [-M, M]^d, |\phi(x) - \phi(x-z)| \leq ||z|| g(x)$$

and

$$\int g^2(x)\phi^{-1}(x)\mathrm{d}x < +\infty.$$

In this context, we have the desired inequality with respect to the L^2-distance.

Proposition 3.10. *(inequality (7.11) [13])*
Under Assumption 4, for all $M > 0$, there exists a positive constant $c(\phi, M)$ such that for all $z_1, z_2 \in [-M, M]^d$ and $w_1, w_2 \in (0,1)$,

$$c(\phi, M)||z_1||^2 \left(||z_2||^2 (w_1 - w_2)^2 + w_1^2 ||z_1 - z_2||^2 \right) \leq ||p_{w_1,z_1} - p_{w_2,z_2}||_2^2.$$

One can notice that Assumption 4 implies that ϕ is bounded (see Assm. ($\mathbf{H}_{\mathcal{S}}$) in [13]). Hence, we can deduce a deviation inequality for ρ-estimators of parameters.

Theorem 3.11. *We assume $\overline{\mathcal{F}}_2$ has a finite VC-index V, $w^* \in (0,1]$ and $z^* \neq 0$. For δ as in (3.7), there exists a positive constant $C(\phi, w^*, z^*)$ and an integer $n_0 = n_0(\phi, w^*, z^*)$ such that for any ρ-estimator $\hat{P} = P_{\hat{w},\hat{z}}$ on $\mathscr{Q}_\delta$, $n \geq n_0$ and for all $\xi \in (0, \xi_n)$, we have*

$$C(\phi, z^*, w^*) \left((w^* - \hat{w})^2 + \left(||z^* - \hat{z}||^2 \wedge 1 \right) \right) \leq \frac{\xi + (V+1)\log(n)}{n},$$

with probability at least $1 - e^{-\xi}$, where $\xi_n = (1+V)[1 + \log(2n/(1+V))]$).

This result is proven in Section E.1. It implies the consistency of $\hat{z}$ and consequently the consistency of $\hat{w}$ if $z^* \neq 0$, the parameter w^* being ill defined if $z^* = 0$. We can deduce a bound on the convergence rate for $\hat{z}$ and also for $\hat{\lambda}$ but only for n large enough. It is similar to Theorem 3.1 of Gadat *et al.* [13] with a smaller power for the logarithmic term. This slight improvement is allowed by the VC assumption. Furthermore, we do not need to know a value of M such that $z^* \in [-M, M]$ or to specify it in the model. The examples of translation families taken by Gadat *et al.* [13] (Sect. 6) all satisfy the VC assumption.

Lemma 3.12. *We have the following VC-subgraph classes of density functions.*

- *The Cauchy location-scale family $\mathcal{C}$ of density functions, given hereafter by (4.3), is VC-subgraph with VC-index $V(\mathcal{C}) \leq 5$.*
- *As a consequence of Lemma 2.2, the univariate normal location-scale family $\mathcal{G}_1$ is VC-subgraph with VC-index at most 5.*
- *The Laplace location family $\mathcal{L}$ of density functions defined by*

$$\mathcal{L} = \left\{ x \mapsto \frac{1}{2} e^{-|x-z|}; z \in \mathbb{R} \right\}$$

is VC-subgraph with VC-index $V(\mathcal{L}) \leq 29$.
- *The location family of densities $\mathcal{SG}_\alpha$ associated to the skew Gaussian density defined by*

$$\mathcal{SG}_\alpha = \left\{ x \mapsto 2\phi_1(x-z) \int_{-\infty}^{x-z} \phi_1(\alpha t) dt; z \in \mathbb{R} \right\}$$

is VC-subgraph with VC-index $V(\mathcal{SG}_\alpha) \leq 10$ for all $\alpha \in \mathbb{R}$, where ϕ_1 is given by (3.11).

This lemma is proven in Section F. By inclusion, if the bound holds for the location-scale family it also holds for the location family with fixed scale parameter.

Proving a lower bound for a specific example

In some specific situations, it is relatively easy to prove a lower bound on the Hellinger distance. This is what we do in the following example and it allows us to obtain faster rates than the usual parametric one. Let α be in $(0,1)$. We denote by s_α the probability density function with respect to the Lebesgue measure on $\mathbb{R}$ defined by

$$s_\alpha : x \in \mathbb{R} \mapsto \frac{1-\alpha}{2|x|^\alpha} \mathbb{1}_{|x| \in (0,1]}.$$

We consider $\mathcal{Q}$ as in (3.23) with $\phi = s_\alpha$ and for $w \in [0,1]$ and $z \in \mathbb{R}$, we write

$$p_{w,z} = (1-w)s_\alpha + w s_\alpha(\cdot - z).$$

We can prove that the Hellinger distance $h(P_{w,z}, P_{w',z'})$ is lower bounded by some distance between the parameters which leads to the following theorem.

Theorem 3.13. *For $w^* > 0$ and $z^* \neq 0$, there is a positive constant $C(\alpha, z^*, w^*)$ such that, for any ρ-estimator $\hat{P} = P_{\hat{w},\hat{z}}$ on $\mathcal{Q}_\delta$ with $\delta = 10/n$ and $n \geq 20$, for all $\xi > 0$, with probability at least $1 - e^{-\xi}$ we have*

$$C(\alpha, z^*, w^*) \left[1 \wedge |\hat{z} - z^*|^{1-\alpha} + (w^* - \hat{w})^2 \right] \leq \frac{\log(n) + \xi}{n}.$$

This result is proven in Section E.2. It implies rather directly that our estimators $\hat{w}$ and $\hat{z}$ estimate w^* and z^* at a rate which is at worst $\sqrt{(\log n)/n}$ and $(n^{-1} \log n)^{1/(1-\alpha)}$ respectively. This latter rate is faster than the usual $1/\sqrt{n}$-rate for all $\alpha \in (0,1)$. Up to the logarithmic factors, these rates are optimal. For $\hat{z}$, it a consequence of Theorem 1.1 in [17] (Chapter VI), noticing that s_α has a singularity of order $-\alpha$ in 0, and with the fact that we cannot do better than $1/\sqrt{n}$ for the Hellinger distance. One can notice that both maximum likelihood and least squares approaches do not apply here since we consider density functions that are unbounded, and not even square integrable for $\alpha \in [1/2, 1)$.

4. Model selection

In Section 3 we consider estimation on a model with a fixed order K and simple emission families. We use model selection to overcome this restriction in this section and consider composite emission families and/or models with different orders.

4.1. Construction of the estimator

Let Θ be a subset of

$$\bigcup_{K\geq 1}\{K\}\times\prod_{k=1}^{K}\Lambda_k.$$

Let $\delta:\Theta\to(0,1]$ be such that for $\theta=(K,\lambda_1,\dots,\lambda_K)\in\Theta$, $\delta(\theta)\in(0,1/K]$. We write

$$\mathscr{Q}_\delta(\theta)=\left\{\sum_{k=1}^{K}w_kF_k;\, w\in\mathcal{W}_K, w_k\geq\delta, w_k\in\mathbb{Q}, F_k\in\mathscr{F}_k,\forall k\in[K]\right\}.$$

We define $\mathscr{Q}_\delta$ by

$$\mathscr{Q}_\delta=\bigcup_{\theta\in\Theta}\mathscr{Q}_\delta(\theta).$$

We associate to $\mathscr{Q}_\delta$ the family $\mathcal{Q}_\delta$ of densities with respect to μ and the ρ-estimator $\hat{P}_\delta$ of $\overline{P}$ based on the family $\mathcal{Q}_\delta$. Assuming we have a penalty function $\mathbf{pen}:\mathcal{Q}_\delta\to\mathbb{R}$, we set

$$\boldsymbol{\Upsilon}(\mathbf{X},q)=\sup_{q'\in\mathcal{Q}_\delta}\left[\mathbf{T}(\mathbf{X},q,q')-\mathbf{pen}(q')\right]+\mathbf{pen}(q), \tag{4.1}$$

for all $q\in\mathcal{Q}_\delta$. The ρ-estimator $\hat{P}_\delta$ is any measurable element of the closure (with respect to the Hellinger distance) of the set $\mathscr{E}(\psi,\mathbf{X})$, as defined by (3.5). One can notice that a constant penalty function does not have any impact on the definition of $\boldsymbol{\Upsilon}$ and brings us back to the previous situation.

4.2. Estimation on a mixture model based on composite emission families

Let K be larger than or equal to 2. Let L be a subset of $\prod_{k=1}^{K}\Lambda_k$ and define Θ by $\Theta=\{K\}\times L$, *i.e.* K is fixed. For $\lambda=(\lambda_1,\dots,\lambda_K)\in L$, the model $\mathscr{Q}(\lambda)$ is a subset of

$$\left\{\sum_{k=1}^{K}w_kF_k;\, w\in\mathcal{W}_K, F_k\in\overline{\mathscr{F}}_{\lambda_k},\forall k\in[K]\right\}$$

and we define its countable subset $\mathscr{Q}_\delta(\lambda)$ by

$$\mathscr{Q}_\delta=\left\{\sum_{k=1}^{K}w_kF_k\in\mathscr{Q}(\lambda);\, w\in\mathcal{W}_K, w_k\geq\delta(\lambda), w_k\in\mathbb{Q}, F_k\in\mathscr{F}_{\lambda_k},\forall k\in[K]\right\},$$

where δ is any function $L\to(0,1/K]$, and $\mathscr{Q}_\delta=\bigcup_{\lambda\in L}\mathscr{Q}_\delta(\lambda)$. Under Assumption 2, we write $\overline{V}(\lambda)=V(\lambda_1)+\cdots+V(\lambda_K)$.

Theorem 4.1. *Let Δ be a mapping $L \to \mathbb{R}^+$ such that $\sum_{\lambda\in L} e^{-\Delta(\lambda)} \leq 1$. Let* **pen** *be the penalty function defined by*

$$\mathbf{pen}(q) = \kappa \inf_{\lambda\in L | Q\in\mathscr{Q}(\lambda)} \left[116.1\overline{V}(\lambda)\left[5.82 + \log\left(\frac{(K+1)^2}{\delta(\lambda)}\right) + \log_+\left(\frac{n}{\overline{V}(\lambda)}\right)\right] + \Delta(\lambda)\right], \tag{4.2}$$

where κ is given by (19) in [4]. Assume there is P^* in $\mathscr{P}$ such that $\mathbf{P}^* = (P^*)^{\otimes n}$. For the choice $\delta(\lambda) = \frac{\overline{V}(\lambda)}{n(K-1)} \bigwedge \frac{1}{K}$, there is a positive constant C such that the resulting estimator $\hat{P} = \hat{P}_\delta$ satisfies the following. For all $\xi > 0$, with probability at least $1 - e^{-\xi}$ we have

$$Ch^2(P^*, \hat{P}) \leq \inf_{\lambda\in L}\left\{h^2(P^*, \mathscr{Q}(\lambda)) + \frac{1}{n}\left(\overline{V}(\lambda)\left[1 + \log\left(\frac{Kn}{\overline{V}(\lambda)\wedge n}\right)\right] + \Delta(\lambda) + \xi\right)\right\}.$$

The constant C is universal, in particular it does not depend on K or on the choice of the model.

This proof of this theorem is postponed to Section B.4. It is a general result for the situation where you know the number K of subpopulations, or at least want to fix it for the estimation, but are hesitating on the models for the emission distributions. For instance, let us consider Gaussian and Cauchy location-scale families for the composite emission families, an example simpler than Example 2.1. For all $k \in \{1, \dots, K\}$, we take $\Lambda_k = \{1, 2\}$ with $\overline{\mathscr{F}}_1 = \mathscr{G}$ and $\overline{\mathscr{F}}_2 = \mathscr{C}$, where $\mathscr{C}$ is the Cauchy location-scale family of distributions associated to the density class

$$\mathcal{C} = \left\{x \mapsto \frac{1}{\pi\sigma}\frac{1}{1 + \left(\frac{x-z}{\sigma}\right)^2}; z \in \mathbb{R}, \sigma > 0\right\}. \tag{4.3}$$

We consider the model $\mathscr{Q} = \cup_{0\leq j\leq K}\mathscr{Q}_j$ with

$$\mathscr{Q}_j = \left\{\sum_{k=1}^{j} w_k\mathcal{N}(z_k, \sigma_k^2) + \sum_{k=j+1}^{K} w_k\text{Cauchy}(z_k, \sigma_k);\ \begin{array}{l}(z_1, \sigma_1) > \cdots > (z_j, \sigma_j),\\ (z_{j+1}, \sigma_{j+1}) > \cdots > (z_K, \sigma_K)\end{array}\right\},$$

where the order $>$ on the parameters (z_k, σ_k) is defined by (3.18) and allows to have identifiability properties again here. Lemma 3.12 gives the same bound on the VC-indices of $\mathcal{G}_1$ and $\mathcal{C}$ therefore (4.2) provides a constant penalty function, hence we will consider a null penalty function.

Theorem 4.2. *Assume $P^* = \sum_{k=1}^{j^*} \overline{w}_k\mathcal{N}(\overline{z}_k, \overline{\sigma}_k^2) + \sum_{k=j^*+1}^{K} \overline{w}_k\,Cauchy(\overline{z}_k, \overline{\sigma}_k) \in \mathscr{Q}_{j^*}$ with $(\overline{z}_1, \overline{\sigma}_1) > \cdots > (\overline{z}_{j^*}, \overline{\sigma}_{j^*})$ and $(\overline{z}_{j^*+1}, \overline{\sigma}_{j^*+1}) > \cdots > (\overline{z}_K, \overline{\sigma}_K)$. Let $\hat{P}$ be a ρ-estimator on $\mathscr{Q}_\delta$ with $\delta = \frac{5}{n} \bigwedge \frac{1}{K}$ and a null penalty. There exists an integer $n_0(P^*)$ and a positive constant $C(P^*)$ such that for $n \geq n_0(P^*)$ there exists an event of probability $1 - (n(K+1))^{-K}$ on which such that $\hat{P} \in \mathscr{Q}_{j^*}$ and*

$$C(P^*)\left(||\overline{w} - \hat{w}||^2 + \sum_{k=1}^{j^*}\left|\left|(\overline{z}_k, \overline{\sigma}_k^2) - (\hat{z}_k, \hat{\sigma}_k^2)\right|\right|^2 \wedge 1 + \sum_{k=j^*+1}^{K}||(\overline{z}_k, \overline{\sigma}_k) - (\hat{z}_k, \hat{\sigma}_k)||^2 \wedge 1\right) \leq \frac{K\log(n(K+1))}{n}.$$

This result is proven in Section D.2. Following the proof, the constant $C(P^*)$ depends both on the distance between P^* and the "wrong models" $\mathscr{Q}_j, j \neq j^*$ and on the smallest eigen value of the Fisher's information

matrix (within the regular parametric model $\mathscr{Q}_{j^*}$). Theorem 4.2 shows that it is possible to identify the true emission models for n large enough and if this identification is established we can also estimate the different parameters. This seems to be somehow original as we did not find any result of this kind in the literature.

4.3. Selection of the order K

We consider Θ of the form $\Theta = \bigcup_{K\in\mathscr{K}} \{K\} \times \{\lambda\}^K$, where $\mathscr{K}$ is a subset of $\{1,\dots,n\}$. For $K \in \mathscr{K}$, we write $\overline{\mathscr{F}} = \overline{\mathscr{F}}_\lambda$ and $\mathscr{F} = \mathscr{F}_\lambda$ its countable and dense subset given by Assumption 1. For $K \in \mathscr{K}$, the model $\mathscr{Q}(K)$ is a subset of

$$\left\{ \sum_{k=1}^{K} w_k F_k ; w \in \mathcal{W}_K, F_k \in \overline{\mathscr{F}}, \forall k \in [K] \right\}.$$

We define $\mathscr{Q}_\delta(K) := \left\{ \sum_{k=1}^{K} w_k F_k \in \mathscr{Q}(K); w \in \mathcal{W}_K, w_k \geq \delta, w_k \in \mathbb{Q}, F_k \in \mathscr{F}, \forall k \in [K] \right\}$ and $\mathscr{Q}_\delta = \bigcup_{K\in\mathscr{K}} \mathscr{Q}_\delta(K)$, where $\delta : \mathscr{K} \to (0,1]$ satisfies $\delta(K) \leq 1/K$. Under Assumption 2, we denote by V the VC-index of $\overline{\mathcal{F}}$, therefore $\overline{V}(K) = K \times V$. If $\hat{P} = \hat{P}_\delta$ is a ρ-estimator on $\mathscr{Q}_\delta$, we denote by $\hat{K}$ the smallest integer K in $\mathscr{K}$ such that $\hat{P} \in \mathscr{Q}_\delta(K)$.

Theorem 4.3. *Let Δ be a function $\mathscr{K} \to \mathbb{R}^+$ satisfying $\sum_{K\in\mathscr{K}} e^{-\Delta(K)} \leq 1$. We consider the penalty function defined by*

$$\mathbf{pen}(q) = \kappa \inf_{K\in\mathscr{K} | Q\in\mathscr{Q}(K)} \left[116.1KV \left[5.82 + \log\left(\frac{(K+1)^2}{\delta(K)} \right) + \log_+\left(\frac{n}{KV} \right) \right] + \Delta(K) \right], \tag{4.4}$$

where κ is given by (19) in [4]. Assume there exists P^* in $\mathscr{P}$ such that $\mathbf{P}^* = (P^*)^{\otimes n}$. For the choice $\delta(1) = 1$ and $\delta(K) = \frac{V}{n} \bigwedge \frac{1}{K}$ for $K \geq 2$, there is a positive constant C such that any ρ-estimator $\hat{P} = \hat{P}_\delta$ on $\mathscr{Q}_\delta$ satisfies the following. For all $\xi > 0$, with probability at least $1 - e^{-\xi}$ we have

$$Ch^2(P^*, \hat{P}) \leq \inf_{K\in\mathscr{K}} \left\{ h^2(P^*, \mathscr{Q}(K)) + \frac{KV\log(n) + \xi + \Delta(K)}{n} \right\}. \tag{4.5}$$

The constant C is universal, in particular it does not depend on $\mathscr{F}$ and therefore neither on V.

This result is proven in Section B.5. It gives an oracle inequality and it provides a way to determine the number of clusters if one wants to use mixture models in order to do clustering. It is also interesting in the context of density estimation. Once again, we take advantage of the approximation properties of GMMs to use our estimator for density estimation on a wider class. We use the approximation result proven by Maugis and Michel [20]. Let $\beta > 0$, $r = \lfloor \beta \rfloor$ and $k \in \mathbb{N}$ such that $\beta \in (2k, 2k+2]$. Let also $\mathcal{P}$ be the 8-tuple of parameters $(\gamma, l^+, L, \epsilon, C, \alpha, \xi, M)$ where L is a polynomial function on $\mathbb{R}$ and the other parameters are positive constants. We define the density class $\mathcal{H}(\beta, \mathcal{P})$ of all densities p satisfying the following conditions.

- For all x and y such that $|y - x| \leq \gamma$,

$$(\log p)^{(r)}(x) - (\log p)^{(r)}(y) \leq r! L(x) |y - x|^{\beta - r}.$$

  Furthermore for all $j \in \{0,\dots,r\}$,

$$|(\log p)^{(j)}(0)| \leq l^+.$$

- We have

$$\max_{1\le j\le r}\int_{\mathbb{R}}\left|(\log p)^{(j)}(x)\right|^{\frac{2\beta+\epsilon}{j}}p(x)\mathrm{d}x\vee\int_{\mathbb{R}}|L(x)|^{2+\frac{\epsilon}{\beta}}p(x)\mathrm{d}x\le C.$$

- For all $x\in\mathbb{R}, p(x)\le M\psi(x)$.
- The function f is strictly positive, non-decreasing on $(-\infty,-\alpha)$ and non-increasing on (α,∞). For all $x\in[-\alpha,\alpha]$ we have $p(x)\ge\xi$.

This class of functions can be approximated by Gaussian mixture models, the quality of the approximation depending on the regularity parameter β.

Lemma 4.4. *(Lemma 6.1 in [20])*
For $0<\underline{\beta}<\overline{\beta}$, there exists a set of parameters $\mathcal{P}(\underline{\beta},\overline{\beta})$ and a positive constant $c_{\underline{\beta},\overline{\beta}}$ such that for all $\beta\in[\underline{\beta},\overline{\beta}]$, all $p\in\mathcal{H}\left(\beta,\mathcal{P}(\underline{\beta},\overline{\beta})\right)$ and all $K\ge 2$, we have

$$h^2\left(P,\mathscr{G}_{mix,K}\right)\le c_{\underline{\beta},\overline{\beta}}\frac{(\log K)^{3\beta}}{K^{2\beta}},$$

where $\mathscr{G}_{mix,K}$ is given by (3.12).

We consider $\mathscr{K}=\{2,\dots,n\}$, $\Delta(K)=K$ and the penalty function **pen** as in (4.4).

Theorem 4.5. *Let $\hat{P}=\hat{P}_\delta$ be a ρ-estimator on $\mathscr{Q}_\delta$ with δ as in (4.5). For $0<\underline{\beta}<\overline{\beta}$, there exist a positive constant $C_{\underline{\beta},\overline{\beta}}$ such that for any p in $\mathcal{H}\left(\beta,\mathcal{P}(\underline{\beta},\overline{\beta})\right)$ with $\beta\in[\underline{\beta},\overline{\beta}]$, for all $\xi>0$, we have*

$$h^2(P^*,\hat{P})\le C_{\underline{\beta},\overline{\beta}}\left(\frac{(\log n)^{\frac{5\beta}{2\beta+1}}}{n^{\frac{2\beta}{2\beta+1}}}+\frac{\xi}{n}\right),$$

with probability at least $1-e^{-\xi}$.

This theorem is proven in Section C.3. It provides an upper bound on the convergence rate of our estimator of order $(\log n)^{5\beta/(4\beta+2)}n^{-\beta/(2\beta+1)}$. It is the same rate obtained in Theorem 2.9 of Maugis and Michel [20] and therefore our estimator as well is minimax adaptive to the regularity β, up to a power of $\log(n)$. Moreover, in our setting there is no need to specify $\underline{\beta}$ nor $\overline{\beta}$ in our model *i.e.* there is no condition on the location and scale parameters of each component. Intuitively, this would allow to obtain a better approximation bound but we did not have time to look into that direction.

Appendix A. Main result

In this section we prove the main result of this paper, Proposition A.1, which gives an upper bound on the ρ-dimension for finite mixture models. The ρ-dimension function is properly defined introduced in [3]. Bounding the ρ-dimension is the key element as it allows to obtain the general result Theorem B.1 as a direct application of Theorem 2 [3]. We recall definitions from [3] that we adapt to our context, in particular the function ψ defined by (3.2) satisfies Assumption 2 of Baraud and Birgé [3] with $a_0=4, a_1=3/8$ and $a_2^2=3\sqrt{2}$ (see Prop. 3 [3]) which gives the different constants. Let $\mathscr{M}$ be a countable subset of $\mathscr{P}$. For $y>0$ and $\overline{P}\in\mathscr{P}$ we write

$$\mathscr{B}^{\mathscr{M}}\left(\overline{P},y\right)=\left\{Q\in\mathscr{M};\mathbf{h}^2\left(\mathbf{P}^*,\overline{P}^{\otimes n}\right)+\mathbf{h}^2\left(\mathbf{P}^*,Q^{\otimes n}\right)<y^2\right\}.$$

If $\mathcal{Q}$ is a set of probability density functions with respect to a σ-finite measure ν such that $\mathscr{M} \cup \{\overline{P}\} = \{q \cdot \nu; q \in \mathcal{M}\}$, we write

$$w\left(\nu, \mathcal{M}, \mathscr{M}, \overline{P}, y\right) = \left[\sup_{Q \in \mathscr{B}^{\mathscr{M}}\left(\overline{P}, y\right)} \left|\mathbf{T}\left(\mathbf{X}, \overline{p}, q\right) - \mathbb{E}_{\mathbf{P}^*}\left[\mathbf{T}\left(\mathbf{X}, \overline{p}, q\right)\right]\right|\right].$$

Similarly, we define $\mathbf{w}^{\mathscr{M}}\left(\overline{P}, y\right) = \inf_{(\nu,\mathcal{M})} w\left(\nu, \mathcal{M}, \mathscr{M}, \overline{P}, y\right)$, where the infimum is taken over all couples $(\nu, \mathcal{M})$ such that $\mathcal{M}$ is the class of density functions associated to $\mathscr{M}$ with respect to ν, σ-finite measure. We can now define the ρ-dimension function of $\mathscr{M}$ by

$$D^{\mathscr{M}}\left(\mathbf{P}^*, \overline{P}^{\otimes n}\right) = \left[\beta^2 \sup\left\{y^2; \mathbf{w}^{\mathscr{Q}}\left(\overline{P}, y\right) > \frac{3y^2}{64}\right\}\right] \bigvee 1,$$

with $\beta = \frac{\sqrt{3}}{2^{5+1/4}}$. Following the notation established in Section 4, we need to bound the ρ-dimension function over each $\mathscr{Q}_\delta(\theta)$ in order to apply Theorem 2 [3].

Proposition A.1. *Under Assumption 2, for $\theta = (K, \lambda_1, \ldots, \lambda_K) \in \Theta$, we write*

$$\overline{V}(\theta) = V_{1,\lambda_1} + \cdots + V_{K,\lambda_K},$$

where V_{k,λ_k} is an upper bound on the VC-index of $\overline{\mathcal{F}}_{k,\lambda_k}$. For all $\mathbf{P} \in \mathscr{P}$ and $\overline{P} \in \mathscr{Q}_\delta$, we have the following bound

$$D^{\mathscr{Q}_\delta(\theta)}\left(\mathbf{P}, \overline{P}^{\otimes n}\right) \leq D_n\left(\delta, \theta\right) := 545.3\overline{V}(\theta)\left[5.82 + \log\left(\frac{(K+1)^2}{\delta(\theta)}\right) + \log_+\left(\frac{n}{\overline{V}(\theta)}\right)\right]. \tag{A.1}$$

A.1 Proof of Proposition A.1

The strategy of the proof is based on the following remark. One can notice that if for some pair $(\nu, \mathcal{Q})$ there is y_0 such that $w\left(\nu, \mathcal{Q}, \mathscr{Q}, \overline{P}, y\right) \leq \frac{3y^2}{64}$ for all $y \geq y_0$, then we have

$$D^{\mathscr{Q}}\left(\mathbf{P}^*, \overline{P}^{\otimes n}\right) \leq (\beta y_0)^2 \bigvee 1. \tag{A.2}$$

Let θ' be an element of Θ such that $\overline{P}$ belongs to $\mathscr{Q}_\delta(\theta')$. Following notation of Section 4, we prove such an inequality for the pair $(\mu, \mathcal{Q}_\delta(\theta) \cup \{\overline{p}\})$ where $\overline{p}$ is the density function in $\mathcal{Q}_\delta(\theta')$ associated to $\overline{P}$. To bound $w\left(\nu, \mathcal{Q}, \mathscr{Q}, \overline{P}, y\right)$, we are going to bound the entropy of $\mathscr{B}^{\mathscr{Q}_\delta(\theta)}\left(\overline{P}, y\right)$ which is possible since each emission models is associated to VC-subgraph classes of density functions (see Assm. 2). For a metric space $(\mathscr{A}, d)$ and $\epsilon > 0$, we denote by $N(\epsilon, \mathscr{A}, d)$ the minimal number of balls of radius ϵ needed to cover $\mathscr{A}$. The next lemma provides a bound on the covering number for our model, up to some modification.

Lemma A.2. *For $\theta = (K, \lambda_1, \ldots, \lambda_K)$, we write $\overline{V}(\theta) = V_{1,\lambda_1} + \cdots + V_{K,\lambda_K}$ and we define*

$$\mathcal{F}^{\mathcal{Q}_\delta(\theta)}\left(\overline{P}\right) := \left\{\psi\left(\sqrt{\frac{q}{\overline{p}}}\right); q \in \mathcal{Q}_\delta(\theta)\right\}.$$

For any probability distribution R, we have

$$\forall \epsilon \leq 2, \log N\left(\epsilon, \mathcal{F}^{\mathcal{Q}_\delta(\theta)}\left(\overline{P}\right), ||\cdot||_{L_2(R)}\right) \leq \overline{V}(\theta)\log\left(\frac{e^{1+1/e}8(K+1)^2}{\delta(\theta)}\right) + 2\overline{V}(\theta)\log(1/\epsilon). \tag{A.3}$$

The next lemma is an intermediate result in the proof of Theorem 2 [5]. It allows to bound the expectation of the supremum of an empirical process from a bound on the covering number on the considered space of functions.

Lemma A.3. *Let $\mathcal{F}$ be an at most countable set of measurable functions $\mathscr{X} \to \mathbb{R}$ such that for any probability distribution P on $(\mathscr{X}, \mathcal{X})$, we have*

$$\log(N(\epsilon, \mathcal{F}, ||\cdot||_{L_2(P)})) \leq a + b\log(1/\epsilon).$$

Let $X_1, \ldots, X_n$ be n independent random variables with values in $(\mathscr{X}, \mathcal{X})$. We define $Z(\mathcal{F})$ by

$$Z(\mathcal{F}) = \sup_{f\in\mathcal{F}} \left| \sum_{i=1}^{n} \left(f(X_i) - \mathbb{E}\left[f(X_i)\right]\right) \right|$$

and assume $\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[f^2(X_i)\right] \leq \sigma^2 \leq 1$. Let $q \in (0,1)$. We have

$$\mathbb{E}\left[Z(\mathcal{F})\right] \leq 32A^2 + A2\sqrt{2n\sigma^2},$$

with $A = \frac{1+q}{1-q}\left(1 + \frac{b}{\log 2 + 2a + b\log(1/q)}\right)\sqrt{\log 2 + 2a + b\log(1/q) + 2b\log(1/\sigma)}$.

Let y be a positive real number. We set

$$\mathcal{F}_{\delta,\theta,y}\left(\overline{P}\right) = \left\{ \psi\left(\sqrt{\frac{q}{\overline{p}}}\right); Q = q\cdot\mu \in \mathscr{Q}_\delta(\theta), \mathbf{h}^2\left(\mathbf{P}^*, \overline{\mathbf{P}}\right) + \mathbf{h}^2\left(\mathbf{P}^*, \hat{\mathbf{P}}\right) < y^2 \right\} \subset \mathcal{F}^{\mathcal{Q}_\delta(\theta)}\left(\overline{P}\right).$$

Since ψ satisfies Assumption 2 [3] and given Lemma A.2, we can apply Lemma A.3 with $\sigma^2 = (3\sqrt{2}y^2/n) \wedge 1$,

$$a = \overline{V}(\theta)\log\left(\frac{e^{1+1/e}8(K+1)^2}{\delta(\theta)}\right) \text{ and } b = 2\overline{V}(\theta).$$

We get

$$w^{\mathscr{Q}_\delta(\theta)}\left(\mu, \mathcal{Q}_\delta(\theta)\cup\{\overline{p}\}, \mathscr{Q}_\delta(\theta), \overline{P}, y\right) \leq \mathbb{E}\left[Z\left(\mathcal{F}_{\delta,\theta,y}\right)\right] \leq 32A^2 + A2\sqrt{2n\sigma^2},$$

with A given in Lemma A.3. Let us try to find a simple upper bound for it. In our situation, dropping the dependency on θ in the notation, we have

$$\begin{aligned}
\frac{b}{\log 2 + 2a + b\log(1/q)} &= \frac{2\overline{V}}{\log 2 + 2\overline{V}\log\left(\frac{e^{1+1/e}8(K+1)^2}{\delta}\right) + 2\overline{V}\log(1/q)} \\
&\leq \frac{1}{\log\left(\frac{e^{1+1/e}8(K+1)^2}{\delta q}\right)} \\
&\leq \frac{1}{\log\left(\frac{e^{1+1/e}8K(K+1)^2}{q}\right)} \leq \frac{1}{\log\left(\frac{e^{1+1/e}2^4\times 3^2}{q}\right)},
\end{aligned}$$

hence

$$A \leq \frac{1+q}{1-q}\left(1+\frac{1}{\log\left(\frac{e^{1+1/e}2^4\times 3^2}{q}\right)}\right)\sqrt{2\overline{V}\left[\log\left(\frac{e^{1+1/e}2^{13/4}}{q}\right)+\log\left(\frac{(K+1)^2}{\delta\sigma^2}\right)\right]}.$$

For $q = 1/9$, we have

$$\begin{aligned}
A &\leq \frac{5}{4}\left(1+\frac{1}{1+\frac{1}{e}+4\log(6)}\right)\sqrt{2\overline{V}\left[\frac{1}{e}+1+\log\left(2^{13/4}\times 9\right)+\log\left(\frac{(K+1)^2}{\delta\sigma^2}\right)\right]}\\
&\leq \frac{5}{4}\times 1.12\sqrt{2\overline{V}\left[5.82+\log\left(\frac{(K+1)^2}{\delta\sigma^2}\right)\right]}\\
&= 2.8\sqrt{2\overline{V}\left[5.82+\log\left(\frac{(K+1)^2}{\delta\sigma^2}\right)\right]}.
\end{aligned}$$

Finally,

$$w^{\mathscr{Q}_\delta(\theta)}\left(\mu, \mathcal{Q}_\delta(\theta)\cup\{\overline{p}\}, \overline{P}, y\right) \leq C_0\sqrt{n\overline{V}\sigma^2\mathscr{L}(\sigma,\delta,\theta)} + C_1\overline{V}\mathscr{L}(\sigma,\delta,\theta) \tag{A.4}$$

with $\mathscr{L}(\sigma,\delta,\theta) = 5.82+\log\left(\frac{(K+1)^2}{\delta\sigma^2}\right)$, $C_0 = 2.8\times 4 = 11.2$ and $C_1 = 2^6\times 2.8^2$. Then we follow the proof of Proposition 6 [5]. For $D \geq \frac{\beta^2}{3\sqrt{2}}\overline{V} = 2^{-11}\overline{V}$ and $y \geq \beta^{-1}\sqrt{D}$, we have

$$\begin{aligned}
\mathscr{L}(\sigma,\delta,\theta) &= 5.82+\log\left(\frac{(K+1)^2}{\delta}\right)+\log_+\left(\frac{n}{3\sqrt{2}y^2}\right)\\
&\leq 5.82+\log\left(\frac{(K+1)^2}{\delta}\right)+\log_+\left(\frac{\beta^2 n}{3\sqrt{2}D}\right)\\
&= 5.82+\log\left(\frac{(K+1)^2}{\delta}\right)+\log_+\left(\frac{n}{2^{11}D}\right)\\
&\leq 5.82+\log\left(\frac{(K+1)^2}{\delta}\right)+\log_+\left(\frac{n}{\overline{V}}\right) = L.
\end{aligned}$$

We combine it with (A.4) and since $y \geq \beta^{-1}\sqrt{D}$ we get

$$\begin{aligned}
w^{\mathscr{Q}_\delta(\theta)}(\mu, \mathcal{Q}_\delta(\theta)\cup\{\overline{p}\}, \overline{P}, y) &\leq 11.2\times\sqrt{3\sqrt{2}y}\sqrt{\overline{V}L}+2^6\times 2.8^2\overline{V}L\\
&= \frac{3y^2}{64}\left[\frac{64\times 11.2\times 2^{1/4}\sqrt{\overline{V}L}}{\sqrt{3}y}+\frac{2^{12}\times 2.8^2\overline{V}L}{3y^2}\right]\\
&\leq \frac{3y^2}{64}\left[\frac{64\times 11.2\times 2^{1/4}\sqrt{\overline{V}L}}{\sqrt{3}\beta^{-1}\sqrt{D}}+\frac{2^{12}\times 2.8^2\overline{V}L}{3\beta^{-2}D}\right]\\
&= \frac{3y^2}{64}\left[2\times 11.2\frac{\sqrt{\overline{V}L}}{\sqrt{D}}+2\sqrt{2}\times 2.8^2\frac{\overline{V}L}{D}\right].
\end{aligned}$$

For $D = 545.3\overline{V}L \geq \overline{V}L\left[\sqrt{11.2^2 + 2\sqrt{2}\times 2.8^2} + 11.2\right]^2$ we have $D \geq 2^{-11}\overline{V}$ since $L \geq 5.82$. Moreover, for all $y \geq y_0 = \beta^{-1}\sqrt{D}$, we have $w^{\mathcal{Q}_\delta(\theta)}(\mu, \mathcal{Q}_\delta(\theta)\cup\{\overline{p}\}, \overline{P}, y) \leq \frac{3y^2}{64}$ which allows to conclude with (A.2). We now turn to the proofs of the two lemmas.

Proof of Lemma A.3

The lemma is an intermediate result in the proof of Theorem 2 of Baraud and Chen [5]. We write $\overline{Z}(f) = \sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \epsilon_i f(X_i)\right|$ where $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. Rademacher random variables. We follow the proof with $h(x) = a + b\log(1/x)$ in (A.7) and everything stays the same up to equation (A.10). We get

$$\mathbb{E}\left[\overline{Z}(\mathcal{F})\right] \leq \sqrt{2n}\frac{1+q}{1-q}\int_0^B \sqrt{\log 2 + 2a + b\log(1/q) + 2b\log(1/u)}du,$$

with $B = \sqrt{\sigma^2 + \frac{8\mathbb{E}\left[\overline{Z}(\mathcal{F})\right]}{n}} \wedge 1$. With Lemma 2 [5], we have

$$\mathbb{E}\left[\overline{Z}(\mathcal{F})\right] \leq 16A^2 + A\sqrt{2n\sigma^2},$$

with $A = \frac{1+q}{1-q}\left(1 + \frac{b}{\log 2 + 2a + b\log(1/q)}\right)\sqrt{\log 2 + 2a + b\log(1/q) + 2b\log(1/\sigma)}$. Classical symmetrization arguments imply

$$\mathbb{E}\left[Z(\mathcal{F})\right] \leq 2\mathbb{E}\left[\overline{Z}(\mathcal{F})\right] \leq 32A^2 + A\sqrt{2n\sigma^2}.$$

□

Proof of Lemma A.2

We write $\phi = \psi\left(\sqrt{\cdot/\overline{p}}\right)$. We drop the dependency on θ in this proof.

Lemma A.4. *For any probability distribution R on $(\mathscr{X}, \mathcal{X})$, for $w, v \in \mathcal{W}_K$ such that $w_k, v_k \geq \delta$ for $k = 1, \ldots, K$ and for any probability densities $q_1, \ldots, q_K, r_1, \ldots, r_K$, we have*

$$\begin{aligned}&||\phi\circ(w_1q_1 + \cdots + w_kq_K) - \phi\circ(v_1r_1 + \cdots + v_Kr_K)||_{L_2(R)}\\ &\leq \frac{1}{\sqrt{\delta}}\sum_{k=1}^K ||\phi\circ q_k - \phi\circ r_k||_{L_2(R)} + \frac{2}{\delta}||w - v||_\infty,\end{aligned} \tag{A.5}$$

where $||w - v||_\infty = \max_{k\in[K]}|w_k - v_k|$.

This lemma implies that for any probability distribution R on $(\mathscr{X}, \mathcal{X})$, we have

$$\begin{aligned}\log N\left(\epsilon, \mathcal{F}^{\mathcal{Q}_\delta(\theta)}\left(\overline{P}\right), ||\cdot||_{L_2(R)}\right) &\leq \log N\left(\epsilon_{K+1}, \mathcal{W}_K, ||\cdot||_\infty\right)\\ &\quad + \sum_{k=1}^K \log N\left(\epsilon_k, \phi\circ\mathcal{F}_k, ||\cdot||_{L_2(R)}\right),\end{aligned} \tag{A.6}$$

where $\phi \circ \mathcal{F}_k := \left\{ \phi \circ f \middle| F \in \mathscr{F}_k \right\}$ for $k = 1, \dots, K$ and $\epsilon = \frac{\epsilon_1 + \dots + \epsilon_K}{\sqrt{\delta}} + \frac{2\epsilon_{K+1}}{\delta}$. Let us bound the covering numbers involved in the latter inequality. From Proposition 42 in [4] and Lemma 1 in [5], we have the following bound . For any probability measure R on $(\mathscr{X}, \mathcal{X})$ and for all $\epsilon_k \in (0, 2)$, we have

$$\log N\left(\epsilon_k, \phi \circ \mathcal{F}_k, ||\cdot||_{L_2(R)}\right) \leq \log\left(eV_k(8e)^{V_k - 1}\right) + 2(V_k - 1)\log(1/\epsilon_k). \tag{A.7}$$

We also need a bound on the covering number of $\mathcal{W}_K$. For $\epsilon_{K+1} > 0$, we have

$$\log N\left(\epsilon_{K+1}, \mathcal{W}_K, ||\cdot||_\infty\right) \leq K \log\left(\frac{3}{\epsilon_{K+1}}\right). \tag{A.8}$$

The proof comes at the end on page 428. We can now combine (A.6), (A.7) and (A.8). For $\epsilon \in (0, 2)$ and $\delta \in (0, 1/K]$, we take

$$\epsilon_{K+1} = \epsilon\frac{\delta}{2}\frac{K}{K + \sum_{k=1}^{K} 2(V_k - 1)} \text{ and } \epsilon_j = \epsilon\sqrt{\delta}\frac{2(V_j - 1)}{K + \sum_{k=1}^{K} 2(V_k - 1)}, j = 1, \dots, K.$$

We get

$$\begin{aligned}
\log N\left(\epsilon, \mathcal{F}^{\mathcal{Q}_\delta(\theta)}\left(\overline{P}\right), ||\cdot||_{L_2(R)}\right) &\leq K \log\left(\frac{3}{\epsilon_{K+1}}\right) + \log\left(e^K \left(\prod_k V_k\right)(8e)^{\sum_k (V_k - 1)}\right) \\
&\quad + \sum_{k=1}^{K} 2(V_k - 1)\log(1/\epsilon_k) \\
&= K \log\left(\frac{6}{\epsilon\delta}\frac{K + \sum_{k=1}^{K} 2(V_k - 1)}{K}\right) \\
&\quad + \log\left(e^K \left(\prod_k V_k\right)(8e)^{\overline{V} - K}\right) \\
&\quad + \sum_{k=1}^{K} 2(V_k - 1)\log\left(\frac{1}{\epsilon\sqrt{\delta}}\frac{K + \sum_{j=1}^{K} 2(V_j - 1)}{2(V_k - 1)}\right) \\
&= \log\left(\frac{\left[K + \sum_{j=1}^{K} 2(V_j - 1)\right]^{K + \sum_{j=1}^{K} 2(V_j - 1)}}{K^K \times \prod_{k=1}^{K}[2(V_k - 1)]^{2(V_k - 1)}}\right) \\
&\quad + \overline{V}\log\left(\left[\prod_k V_k\right]^{1/\overline{V}}\right)
\end{aligned}$$

$$+\log\left(\frac{6^K e^{\overline{V}} 8^{\overline{V}-K}}{\delta^{\overline{V}}}\right)+(2\overline{V}-K)\log(1/\epsilon).$$

The following inequalities allow to simplify this. For all $x_1,\dots,x_n\geq 0$ such that $x_1+\cdots+x_n>0$, we have

$$\log\left(\frac{(x_1+\cdots+x_n)^{x_1+\cdots+x_n}}{x_1^{x_1}\dots x_n^{x_n}}\right)\leq (x_1+\cdots+x_n)\log(n), \tag{A.9}$$

$$\left(\prod_{i=1}^n x_i\right)^{\frac{1}{x_1+\cdots+x_n}}\leq \left(e^{\frac{1}{e}}\right)^{\frac{1}{n^n}}\leq e^{1/e}. \tag{A.10}$$

Then, we get

$$\begin{aligned}\log N\left(\epsilon,\mathcal{F}^{\mathcal{Q}_\delta(\theta)}\left(\overline{P}\right),||\cdot||_{L_2(R)}\right)&\leq\left[K+\sum_{j=1}^K 2(V_j-1)\right]\log(K+1)+\overline{V}\log\left(e^{1/e}\right)\\&\quad+\log\left(\frac{e^{\overline{V}}8^{\overline{V}}}{\delta^{\overline{V}}}\right)+(2\overline{V}-K)\log(1/\epsilon)\\&\leq\overline{V}\log\left(\frac{e^{1+1/e}8(K+1)^2}{\delta}\right)+2\overline{V}\log(1/\epsilon).\end{aligned}$$

To conclude we need to prove (A.9), (A.10) and (A.8).

Proof of (A.9) and (A.10)

- In a first time, we assume $x_1+\cdots+x_n=1$, *i.e.* $x\in\mathcal{W}_n$. Then

$$\log\left(\frac{(x_1+\cdots+x_n)^{x_1+\cdots+x_n}}{x_1^{x_1}\dots x_n^{x_n}}\right)=-\sum_{i=1}^n x_i\log(x_i)\text{ and }\left(\prod_{i=1}^n x_i\right)^{\frac{1}{x_1+\cdots+x_n}}=\prod_{i=1}^n x_i.$$

Both functions $x\mapsto-\sum_{i=1}^n x_i\log(x_i)$ and $x\mapsto\prod_{i=1}^n x_i$ are bounded and attains a maximum on $\mathcal{W}_n$ for $x_1=\cdots=x_n=1/n$, such that

$$-\sum_{i=1}^n x_i\log(x_i)\leq\log(n)\text{ and }\prod_{i=1}^n x_i\leq\left(\frac{1}{n}\right)^n.$$

- In the generic case, we define $s(x):=x_1+\cdots+x_n>0$ and y in $\mathcal{W}_n$ by $y_i=x_i/s(x)$ for $i=1,\dots,n$. We have

$$\log\left(\frac{(x_1+\cdots+x_n)^{x_1+\cdots+x_n}}{x_1^{x_1}\dots x_n^{x_n}}\right)=s(x)\times\left[-\sum_{i=1}^n y_i\log(y_i)\right]\qquad\leq(x_1+\cdots+x_n)\log(n)$$

and

$$\left(\prod_{i=1}^n x_i\right)^{\frac{1}{x_1+\cdots+x_n}}=s(x)^{1/s(x)}\times\left(\prod_{i=1}^n y_i\right)^{1/s(x)}$$

$$\leq s(x)^{1/s(x)} \times \left(\frac{1}{n^n}\right)^{1/s(x)}$$
$$\leq \left(e^{1/e}\right)^{\frac{1}{n^n}}.$$

The last inequality comes from $\forall x > 0, x^{1/x} \leq e^{1/e}$ and we get (A.10) with $e \geq 1$ and $n^n \geq 1$. □

Proof of (A.8)

Let $\epsilon \in (0,1)$. Let N be an integer greater than $\frac{1}{\epsilon}$. We define

$$\mathcal{W}_{K,N} := \left\{ w \in \mathcal{W}_K \middle| \forall k \in [K], \exists d_k \in \mathbb{N}, w_k = \frac{d_k}{N} \right\}.$$

- One can easily see that there is a bijection between $\mathcal{M}_{K,N}$ and the set

$$\mathcal{D}_{K,N} := \left\{ d_1, \dots, d_K \in \mathbb{N} \middle| \sum_{k=1}^{K} d_k = N \right\}.$$

  We have the following bound $|\mathcal{D}_{K,N}| = \binom{N+K-1}{N} \leq (N+1)^K$.
- Let w be in $\mathcal{W}_K$. For $k \in [K]$, we write $a_k = \lfloor N w_k \rfloor$. We define $s(a) \in \mathbb{N}$ and $d \in \mathcal{D}_{K,N}$ by $s(a) := a_1 + \dots + a_K \leq N$ and

$$\forall k \in [K], d_k := a_k + \mathbb{1}_{s(a)+k \leq N} \in [\lfloor N w_k \rfloor, \lfloor N w_k \rfloor + 1].$$

  Therefore, we have $v \in \mathcal{W}_{K,N}$ defined by $v_k = \frac{d_k}{N}$, such that

$$\forall k \in [K], |w_k - v_k| \leq 1/N,$$

  i.e. $||w - v||_\infty \leq 1/N \leq \epsilon$.

Therefore $\mathcal{W}_{K,N}$ is a ϵ-net of $\mathcal{W}_K$ with respect to $||\cdot||_\infty$ and for $N = \lceil 1/\epsilon \rceil \geq 1/\epsilon$ we have

$$\log(N(\epsilon, \mathcal{W}_K, d)) \leq \log\left(|\mathcal{W}_{K,N}|\right) = \log\left(|\mathcal{D}_{K,N}|\right)$$
$$\leq K \log(1+N) \leq K \log\left(\frac{3}{\epsilon}\right).$$

□

This concludes the proof of Lemma A.2.

Proof of Lemma A.4

The result is just the combination of the two following claims and the triangle inequality.

- First claim: For any probability distribution R, any nonnegative measurable functions q_1, q_2, g and any $w \in (0,1)$ we have

$$||\phi \circ (w q_1 + (1-w)g) - \phi \circ (w q_2 + (1-w)g)||_{L_2(R)} \leq \frac{1}{\sqrt{w}} ||\phi \circ q_1 - \phi \circ q_2||_{L_2(R)}. \tag{A.11}$$

- Second claim: Let $g_1, \ldots, g_K$ be K densities. For $w, v \in \mathcal{W}_{K,\delta}$, we have

$$||\phi \circ (w_1 g_1 + \cdots + w_K g_K) - \phi \circ (v_1 g_1 + \cdots + v_K g_K)||_{L_2(R)} \leq \frac{2}{\delta} ||w - v||_\infty. \tag{A.12}$$

Combining those inequalities, we have

$$\begin{aligned}
\left|\left|\phi \circ \left(\sum_{k=1}^K w_k f_k\right) - \phi \circ \left(\sum_{k=1}^K v_k g_k\right)\right|\right|_{L_2(R)} &\leq \left|\left|\phi \circ \left(\sum_{k=1}^K w_k f_k\right) - \phi \circ \left(\sum_{k=1}^K v_k f_k\right)\right|\right|_{L_2(R)} \\
&\quad + \sum_{k=1}^K ||\phi \circ (h_{k-1}) - \phi \circ (h_k)||_{L_2(R)} \\
&\leq \frac{2}{\delta} ||w - v||_\infty + \sum_{k=1}^K \frac{1}{\sqrt{\nu_k}} ||\phi \circ (g_k) - \phi \circ (f_k)||_{L_2(R)} \\
&\leq \frac{2}{\delta} ||w - v||_\infty + \frac{1}{\sqrt{\delta}} \sum_{k=1}^K ||\phi \circ (g_k) - \phi \circ (f_k)||_{L_2(R)},
\end{aligned}$$

with $h_k = \sum_{j=1}^k v_j g_j + \sum_{j=k+1}^K v_j f_j$.

- Proof of (A.11).
For two probability densities f_1 and f_2, for x such that $\overline{p}(x) > 0$ and $f_1(x) + f_2(x) > 0$, computation gives

$$\begin{aligned}
|\phi \circ f_1(x) - \phi \circ f_2(x)| &= \left|\psi\left(\sqrt{f_1/\overline{p}}(x)\right) - \psi\left(\sqrt{f_2/\overline{p}}(x)\right)\right| \\
&= \left|\frac{\sqrt{\frac{f_1}{\overline{p}}}(x) - 1}{\sqrt{\frac{f_1}{\overline{p}}}(x) + 1} - \frac{\sqrt{\frac{f_2}{\overline{p}}}(x) - 1}{\sqrt{\frac{f_2}{\overline{p}}}(x) + 1}\right| \\
&= \frac{2\left|\sqrt{\frac{f_1}{\overline{p}}}(x) - \sqrt{\frac{f_2}{\overline{p}}}(x)\right|}{\left(\sqrt{\frac{f_1}{\overline{p}}}(x) + 1\right)\left(\sqrt{\frac{f_2}{\overline{p}}}(x) + 1\right)} \\
&= \frac{2\left|\frac{f_1}{\overline{p}}(x) - \frac{f_2}{\overline{p}}(x)\right|}{\left(\sqrt{\frac{f_1}{\overline{p}}}(x) + 1\right)\left(\sqrt{\frac{f_2}{\overline{p}}}(x) + 1\right)\left(\sqrt{\frac{f_1}{\overline{p}}}(x) + \sqrt{\frac{f_2}{\overline{p}}}(x)\right)}.
\end{aligned} \tag{A.13}$$

For $f_1 = wq_1 + (1-w)g$ and $f_2 = wq_2 + (1-w)g$, dropping x in the notation, we get

$$\begin{aligned}
&|\phi \circ (wq_1 + (1-w)g) - \phi \circ (wq_2 + (1-w)g)| \\
&= \frac{2w\left|\frac{q_1 - q_2}{\overline{p}}\right|}{\left(\sqrt{\frac{wq_1 + (1-w)g}{\overline{p}}} + 1\right)\left(\sqrt{\frac{wq_2 + (1-w)g}{\overline{p}}} + 1\right)\left(\sqrt{\frac{wq_1 + (1-w)g}{\overline{p}}} + \sqrt{\frac{wq_2 + (1-w)g}{\overline{p}}}\right)} \\
&= \frac{2\left|\frac{q_1 - q_2}{\overline{p}}\right|}{\left(\sqrt{\frac{q_1}{\overline{p}}} + 1\right)\left(\sqrt{\frac{q_2}{\overline{p}}} + 1\right)\left(\sqrt{\frac{q_1}{\overline{p}}} + \sqrt{\frac{q_2}{\overline{p}}}\right)}
\end{aligned}$$

$$\times \frac{w\left(\sqrt{\frac{q_1}{\overline{p}}}+1\right)\left(\sqrt{\frac{q_2}{\overline{p}}}+1\right)\left(\sqrt{\frac{q_1}{\overline{p}}}+\sqrt{\frac{q_2}{\overline{p}}}\right)}{\left(\sqrt{\frac{wq_1+(1-w)g}{\overline{p}}}+1\right)\left(\sqrt{\frac{wq_2+(1-w)g}{\overline{p}}}+1\right)\left(\sqrt{\frac{wq_1+(1-w)g}{\overline{p}}}+\sqrt{\frac{wq_2+(1-w)g}{\overline{p}}}\right)}$$
$$= |\phi\circ q_1-\phi\circ q_2|\times\frac{\sqrt{w}\left(\sqrt{\frac{q_1}{\overline{p}}}+1\right)}{\left(\sqrt{\frac{wq_1+(1-w)g}{\overline{p}}}+1\right)}\times\frac{\sqrt{w}\left(\sqrt{\frac{q_2}{\overline{p}}}+1\right)}{\sqrt{\frac{wq_2+(1-w)g}{\overline{p}}}+1}$$
$$\times\frac{\sqrt{\frac{q_1}{\overline{p}}}+\sqrt{\frac{q_2}{\overline{p}}}}{\sqrt{\frac{wq_1+(1-w)g}{\overline{p}}}+\sqrt{\frac{wq_2+(1-w)g}{\overline{p}}}}.$$

For $w\in(0,1)$ and any $y_1,y_2,z\geq 0$ such that $y_1+y_2+z>0$, we have

$$\frac{\sqrt{y_1}+\sqrt{y_2}}{\sqrt{wy_1+(1-w)z}+\sqrt{wy_2+(1-w)z}}\times\frac{\sqrt{w}\left(\sqrt{y_1}+1\right)}{\sqrt{wy_1+(1-w)z}+1}$$
$$\times\frac{\sqrt{w}\left(\sqrt{y_2}+1\right)}{\sqrt{wy_2+(1-w)z}+1}$$
$$\leq\frac{\sqrt{y_1}+\sqrt{y_2}}{\sqrt{wy_1}+\sqrt{wy_2}}\times\frac{\sqrt{w}\left(\sqrt{y_1}+1\right)}{\sqrt{wy_1}+1}\times\frac{\sqrt{w}\left(\sqrt{y_2}+1\right)}{\sqrt{wy_2}+1}\leq\frac{1}{\sqrt{w}}.$$

Finally, for x such that $\overline{p}(x)>0$ and $q_1(x)+q_2(x)+g(x)>0$, we have

$$|\phi\circ(wq_1+(1-w)g)(x)-\phi\circ(wq_2+(1-w)g)(x)|\leq\frac{1}{\sqrt{w}}|\phi\circ q_1(x)-\phi\circ q_2(x)|. \tag{A.14}$$

We now considered the atypical cases given the convention established in section 3.1. If $q_1(x)=q_2(x)=r(x)=0$, we have

$$|\phi\circ(wq_1+(1-w)g)(x)-\phi\circ(wq_2+(1-w)g)(x)|=0$$

whether $\overline{p}(x)$ is positive or not. This equality is also true when $\overline{p}(x)=0$, $q_1(x)+g(x)>0$ and $q_2(x)+g(x)>0$. The last case is for $\overline{p}(x)=q_1(x)=g(x)=0$ and $q_2(x)>0$ (q_1 and q_2 being interchangeable). We have

$$|\phi\circ(wq_1+(1-w)g)(x)-\phi\circ(wq_2+(1-w)g)(x)|$$
$$=1=|\phi\circ q_1(x)-\phi\circ q_2(x)|\leq\frac{1}{\sqrt{w}}|\phi\circ q_1(x)-\phi\circ q_2(x)|.$$

Therefore, inequality (A.14) is always valid and taking the $L_2(R)$ norm provides the desired result. □

- Proof of (A.12).
–We first prove an inequality for mixtures with fixed emission densities. Let r and q be any probability densities on $(\mathscr{X},\mathcal{X})$. Let w and v be in $(0,1)$. Using (A.13) and dropping x in the notation, for $r\neq q$ we have

$$|\phi\circ(wr+(1-w)q)-\phi\circ(vr+(1-v)q)|$$

$$= \frac{2\,|w-v|\left|\frac{r-q}{\overline{p}}\right|}{\left(\sqrt{\frac{wr+(1-w)q}{\overline{p}}}+1\right)\left(\sqrt{\frac{vr+(1-v)q}{\overline{p}}}+1\right)\left(\sqrt{\frac{wr+(1-w)q}{\overline{p}}}+\sqrt{\frac{vr+(1-v)q}{\overline{p}}}\right)}$$

$$\leq \begin{cases} \frac{2|w-v|\left|\frac{r-q}{\overline{p}}\right|}{\left(\sqrt{\frac{w|r-q|+(1-w)q}{\overline{p}}}+1\right)\left(\sqrt{\frac{v|r-q|+(1-v)q}{\overline{p}}}+1\right)\left(\sqrt{\frac{w|r-q|+(1-w)q}{\overline{p}}}+\sqrt{\frac{v|r-q|+(1-v)q}{\overline{p}}}\right)}, & \text{if } r>q \\ \frac{2|w-v|\left|\frac{r-q}{\overline{p}}\right|}{\left(\sqrt{\frac{(1-w)|q-r|+wr}{\overline{p}}}+1\right)\left(\sqrt{\frac{(1-v)|q-r|+vr}{\overline{p}}}+1\right)\left(\sqrt{\frac{(1-w)|q-r|+wr}{\overline{p}}}+\sqrt{\frac{(1-v)|q-r|+vr}{\overline{p}}}\right)}, & \text{if } r<q. \end{cases}$$

$$\leq \begin{cases} \frac{2\left|\sqrt{w}-\sqrt{v}\right|\sqrt{\left|\frac{r-q}{\overline{p}}\right|}}{\left(\sqrt{\frac{w|r-q|}{\overline{p}}}+1\right)\left(\sqrt{\frac{v|r-q|}{\overline{p}}}+1\right)}, & \text{if } r>q \\ \frac{2\left|\sqrt{1-w}-\sqrt{1-v}\right|\sqrt{\left|\frac{r-q}{\overline{p}}\right|}}{\left(\sqrt{\frac{(1-w)|q-r|}{\overline{p}}}+1\right)\left(\sqrt{\frac{(1-v)|q-r|}{\overline{p}}}+1\right)}, & \text{if } r<q. \end{cases}$$

One can easily check that the function $x \mapsto \frac{\sqrt{x}}{(\sqrt{\alpha x}+1)(\sqrt{\beta x}+1)}$ is bounded above by $\left(\alpha^{1/4}+\beta^{1/4}\right)^{-2}$. Therefore, we get

$$\begin{aligned} &|\phi\circ(wr+(1-w)q)-\phi\circ(vr+(1-v)q)| \\ &\leq 2\left(\frac{|\sqrt{w}-\sqrt{v}|}{\left(w^{1/4}+v^{1/4}\right)^2}\bigvee\frac{\left|\sqrt{1-w}-\sqrt{1-v}\right|}{\left((1-w)^{1/4}+(1-v)^{1/4}\right)^2}\right) \\ &= 2\left(\frac{\left|w^{1/4}-v^{1/4}\right|}{w^{1/4}+v^{1/4}}\bigvee\frac{\left|(1-w)^{1/4}-(1-v)^{1/4}\right|}{(1-w)^{1/4}+(1-v)^{1/4}}\right). \end{aligned}$$

The inequality obviously stands for x such that $r(x)=q(x)$. Therefore we can take the $L_2(R)$-norm and get

$$\begin{aligned} &||\phi\circ(wr+(1-w)q)-\phi\circ(vr+(1-v)q)|| \\ &\leq 2\left(\frac{\left|w^{1/4}-v^{1/4}\right|}{w^{1/4}+v^{1/4}}\bigvee\frac{\left|(1-w)^{1/4}-(1-v)^{1/4}\right|}{(1-w)^{1/4}+(1-v)^{1/4}}\right). \end{aligned} \tag{A.15}$$

–We can now prove (A.12). Let $g_1,\ldots,g_K$ be K probability densities. Let $w,v\in\mathcal{W}_{K,\delta}$. If $w=v$ the proof is obvious. Therefore we consider $w\neq v$. The idea is to rewrite $w_1g_1+\cdots+w_Kg_K$ and $v_1g_1+\cdots+v_Kg_K$ as 2 component mixtures with the same emission densities, allowing us to use (A.15). We define

$$t_1 := \max_{1\leq k\leq K}\frac{w_k-v_k}{\mathbb{1}_{w_k>v_k}-v_k}\in[0,1] \text{ and } t_2 := \max_{1\leq k\leq K}\frac{v_k-w_k}{\mathbb{1}_{v_k>w_k}-w_k}\in[0,1].$$

Since $w\neq v$, we have $t_1,t_2>0$. We define two probability densities f_1 and f_2 by

$$f_1 := \sum_{k=1}^{K}\left[v_k+\frac{w_k-v_k}{t_1}\right]g_k \text{ and } f_2 := \sum_{k=1}^{K}\left[w_k+\frac{v_k-w_k}{t_2}\right]g_k.$$

One can check that we have

$$\sum_{k=1}^{K} w_k g_k = \frac{t_1}{t_1 + t_2(1-t_1)} f_1 + \frac{t_2(1-t_1)}{t_1 + t_2(1-t_1)} f_2,$$
$$\sum_{k=1}^{K} \nu_k g_k = \frac{t_1(1-t_2)}{t_2 + t_1(1-t_2)} f_1 + \frac{t_2}{t_2 + t_1(1-t_2)} f_2.$$

We get straight from (A.15) that

$$\begin{aligned}
&||\phi \circ (w_1 g_1 + \cdots + w_k g_K) - \phi \circ (v_1 g_1 + \cdots + v_K g_K)||_{L_2(Q)} \\
&= \left|\left| \phi \circ \left(\frac{t_1}{t_1 + t_2(1-t_1)} f_1 + \frac{t_2(1-t_1)}{t_1 + t_2(1-t_1)} f_2 \right) \right.\right. \\
&\left.\left. - \ \phi \circ \left(\frac{t_1(1-t_2)}{t_2 + t_1(1-t_2)} f_1 + \frac{t_2}{t_2 + t_1(1-t_2)} f_2 \right) \right|\right|_{L_2(Q)} \\
&\leq 2 \left(\frac{\left| \left(\frac{t_2(1-t_1)}{t_1+t_2(1-t_1)} \right)^{1/4} - \left(\frac{t_2}{t_2+t_1(1-t_2)} \right)^{1/4} \right|}{\left(\frac{t_2(1-t_1)}{t_1+t_2(1-t_1)} \right)^{1/4} + \left(\frac{t_2}{t_2+t_1(1-t_2)} \right)^{1/4}} \bigvee \frac{\left| \left(\frac{t_1}{t_1+t_2(1-t_1)} \right)^{1/4} - \left(\frac{t_1(1-t_2)}{t_2+t_1(1-t_2)} \right)^{1/4} \right|}{\left(\frac{t_1}{t_1+t_2(1-t_1)} \right)^{1/4} + \left(\frac{t_1(1-t_2)}{t_2+t_1(1-t_2)} \right)^{1/4}} \right) \\
&= 2 \left(\frac{\left| (t_2(1-t_1))^{1/4} - (t_2)^{1/4} \right|}{(t_2(1-t_1))^{1/4} + (t_2)^{1/4}} \bigvee \frac{\left| (t_1)^{1/4} - (t_1(1-t_2))^{1/4} \right|}{(t_1)^{1/4} + (t_1(1-t_2))^{1/4}} \right) \\
&= 2 \left(\frac{\left| (1-t_1)^{1/4} - 1 \right|}{(1-t_1)^{1/4} + 1} \bigvee \frac{\left| 1 - (1-t_2)^{1/4} \right|}{1 + (1-t_2)^{1/4}} \right) \\
&= 2 \left(\frac{t_1}{\left((1-t_1)^{1/4} + 1\right)^2 \left((1-t_1)^{1/2} + 1\right)} \bigvee \frac{t_2}{\left((1-t_2)^{1/4} + 1\right)^2 \left((1-t_2)^{1/2} + 1\right)} \right) \\
&\leq 2(t_1 \vee t_2).
\end{aligned}$$

We end the proof of (A.12) with the following upper bound on $t_1 \vee t_2$. We have

$$\begin{aligned}
t_1 \vee t_2 &= \max_{1 \leq k \leq K} \left(\frac{w_k - v_k}{\mathbb{1}_{w_k > v_k} - v_k} \vee \frac{v_k - w_k}{\mathbb{1}_{v_k > w_k} - w_k} \right) \\
&= \max_{1 \leq k \leq K} \left\{ |w_k - v_k| \times \max\left((1-v_k)^{-1}, (1-w_k)^{-1}, v_k, w_k \right) \right\} \\
&\leq \delta^{-1} \, ||w - v||_\infty . \quad \square
\end{aligned}$$

The proof of Lemma A.4 is now complete.

Appendix B. Theorems

In this section we provide a very general result from which we will derive Theorems 3.1, 3.3, 4.1 and 4.3.

Theorem B.1. *Any ρ-estimator $\hat{P}$ on $\mathscr{Q}_\delta$ satisfies, with probability at least $1 - e^{-\xi}$,*

$$\mathbf{h}^2 \left(\mathbf{P}^*, \hat{P}^{\otimes n} \right) \leq \inf_{\theta \in \Theta} \left\{ c_0 \left(\mathbf{h}^2(\mathbf{P}^*, \mathscr{Q}(\theta)) + n(K-1)\delta(\theta) \right) \right. \tag{B.1}$$

$$+ c_1 \left(116.1 \overline{V}(\theta) \left[5.82 + \log\left(\frac{(K+1)^2}{\delta(\theta)} \right) + \log_+ \left(\frac{n}{\overline{V}(\theta)} \right) \right] + \Delta(\theta) \right) \Bigg\}$$
$$+ c_1 (1.49 + \xi).$$

with $c_0 = 300$ *and* $c_1 = 5014$. *Moreover, for* $K \geq 2$ *and* $\delta(\theta) = \frac{\overline{V}(\theta)}{n(K-1)} \wedge \frac{1}{K}$, *we have*

$$\log\left(\frac{(K+1)^2}{\delta(\theta)} \right) + \log_+ \left(\frac{n}{\overline{V}(\theta)} \right) \leq (2 + \log_2(9)) \log\left(\frac{Kn}{\overline{V}(\theta) \wedge n} \right) \tag{B.2}$$

and $n(K-1)\delta(\theta) \leq n \wedge \overline{V}(\theta)$.

B.1 Proof of Theorem B.1

We recall that the function ψ defined by (3.2) satisfies Assumption 2 of Baraud and Birgé [3] with $a_0 = 4, a_1 = 3/8$ and $a_2^2 = 3\sqrt{2}$ (see Prop. 3 [3]). Using Proposition A.1, we can apply Theorem 2 [3] with

$$D_n(\delta, \theta) = 545.3 \overline{V}(\theta) \left[5.82 + \log\left(\frac{(K+1)^2}{\delta} \right) + \log_+ \left(\frac{n}{\overline{V}(\theta)} \right) \right].$$

There exist constants γ and κ (given by (19) in [3]) such that, with probability $\geq 1 - e^{-\xi}$, we have

$$\mathbf{h}^2\left(\mathbf{P}^*, \hat{\mathbf{P}} \right) \leq \inf_{\theta \in \Theta} \left[\gamma \mathbf{h}^2(\mathbf{P}^*, \mathscr{Q}_\delta(\theta)) + \frac{4\kappa}{a_1} \left(\frac{D_n(\delta, \theta)}{4.7} + \Delta(\theta) \right) \right]$$
$$+ \frac{4\kappa}{a_1}(1.49 + \xi).$$

Lemma B.2. *For all* $K \geq 2$ *and* $\theta \in \Theta$, *we have*

$$\forall P \in \mathscr{P}, h\left(P, \mathscr{Q}_\delta(\theta)\right) \leq \sqrt{(K-1)\delta(\theta)} + h\left(P, \mathscr{Q}(\theta)\right). \tag{B.3}$$

Using this inequality, we get

$$\mathbf{h}^2\left(\mathbf{P}^*, \hat{\mathbf{P}} \right) \leq \inf_{\theta \in \Theta} \Big[2\gamma \left(\mathbf{h}^2(\mathbf{P}^*, \mathscr{Q}(\theta)) + n(K(\theta) - 1)\delta(\theta) \right)$$
$$+ \frac{4\kappa}{a_1} \left(116.1 \overline{V}(\theta) \left[5.82 + \log\left(\frac{(K+1)^2}{\delta} \right) + \log_+ \left(\frac{n}{\overline{V}(\theta)} \right) \right] + \Delta(\theta) \right) \Big]$$
$$+ \frac{4\kappa}{a_1}(1.49 + \xi).$$

From Baraud and Chen [5] (see proof of Thm. 1), we get that $\gamma < 150$ and $4\kappa/a_1 < 5014$. Let us now prove (B.2). We consider θ such that $K \geq 2$ and we take $\delta(\theta) = \frac{\overline{V}(\theta)}{n(K-1)} \wedge \frac{1}{K}$.

- If $\overline{V}(\theta) \leq n(K-1)/K$, then

$$\log\left(\frac{(K+1)^2}{\delta(\theta)} \right) + \log_+ \left(\frac{n}{\overline{V}(\theta)} \right) = \log\left(\frac{(K^2-1)(K+1)n^2}{\overline{V}(\theta)^2} \right)$$
$$= 3\log\left(\frac{Kn}{\overline{V}(\theta)} \right) + \log\left(\frac{(K^2-1)(K+1)\overline{V}(\theta)}{K^3 n} \right)$$

$$\leq 3\log\left(\frac{Kn}{\overline{V}(\theta)}\right) + \log\left(\frac{(K^2-1)^2}{K^4}\right)$$
$$\leq 3\log\left(\frac{Kn}{\overline{V}(\theta)\wedge n}\right).$$

- Otherwise $\overline{V}(\theta) > n(K-1)/K$ and

$$\log\left(\frac{(K+1)^2}{\delta(\theta)}\right) + \log_+\left(\frac{n}{\overline{V}(\theta)}\right) \leq \log\left(\frac{(K+1)^2K^2}{K-1}\right)$$
$$= 3\log(K) + \log\left(\frac{K^2+2K+1}{K(K-1)}\right)$$
$$\leq 3\log(K) + \log(9/2)$$
$$\leq \left[3 + \frac{\log(9)-\log(2)}{\log(2)}\right]\log(K)$$
$$\leq (2+\log_2(9))\log\left(\frac{Kn}{\overline{V}(\theta)\wedge n}\right).$$

Finally, one can check that $n(K-1)\delta(\theta) \leq n \wedge \overline{V}(\theta)$.

Proof of Lemma B.2

For $K \geq 2$ and $\delta \in (0, 1/K]$, we define $\mathcal{W}_{K,\delta}$ by

$$\mathcal{W}_{K,\delta} = \mathcal{W}_K \cap [\delta, 1]^K. \tag{B.4}$$

We prove by induction that

$$\forall \delta \in (0,1/K],\ \sup_{w\in\mathcal{W}_K} h^2\left(w, \mathcal{W}_{K,\delta}\right) \leq 1 - \sqrt{1-(K-1)\delta}. \tag{B.5}$$

• Assume (B.5) holds true for $K \geq 2$. Let δ be in $(0, 1/(K+1))$ and w be in $\mathcal{W}_{K+1}$. Without loss of generality we consider $w_1 \leq w_2 \leq \cdots \leq w_K \leq w_{K+1}$. We define the function r by

$$r : \left| \begin{array}{lll} \mathcal{W}_{K+1} & \to \mathcal{W}_K & \\ w & \mapsto \begin{cases} \left(\frac{w_2}{1-w_1}, \frac{w_3}{1-w_1}, \ldots, \frac{w_K}{1-w_1}\right) & \text{for } w_1 \neq 0, \\ \left(\frac{1}{K}, \frac{1}{K}, \ldots, \frac{1}{K}\right) & \text{for } w_1 = 1, \end{cases} \end{array} \right.$$

and informally r^{-1} by

$$r^{-1} : \left| \begin{array}{ll} \mathcal{W}_K \times [0,1) & \to \mathcal{W}_{K+1} \\ (w', a) & \mapsto (a, (1-a)w'_1, \ldots, (1-a)w'_K). \end{array} \right.$$

- If $w_1 \geq \delta$ then $w \in \mathcal{M}_{K+1,\delta}$ and $h\left(w, \mathcal{W}_{K+1,\delta}\right) = 0$.
- Otherwise $w_1 < \delta$ and we build a distribution $v \in \mathcal{W}_{K+1,\delta}$ to approximate w. Take $\eta = \delta/(1-\delta) \in (0, 1/K]$. From (B.5), there exists $v' \in \mathcal{M}_{K,\eta}$ such that $h^2(r(w), v') \leq 1 - \sqrt{1-(K-1)\eta}$. Now take $v = r^{-1}(\delta, v')$. We have $v_1 = \delta$ and for $j \geq 2$, $v_j = (1-\delta)v'_{j-1} \geq (1-\delta)\eta = \delta$. Therefore v belongs to $\mathcal{W}_{K+1,\delta}$. We also

have

$$\begin{aligned} h^2(w,v) &= \frac{1}{2}\left[\left(\sqrt{w_1}-\sqrt{\delta}\right)^2+\left(\sqrt{1-w_1}-\sqrt{1-\delta}\right)^2\right] \\ &\quad + \sqrt{1-w_1}\sqrt{1-\delta}h^2(r(w),v') \\ &\le \left[1-\sqrt{1-\delta}\right]+\sqrt{1-\delta}\left[1-\sqrt{1-(K-1)\eta}\right] \\ &= 1-\sqrt{1-\delta}\sqrt{1-(K-1)\delta/(1-\delta)} \\ &= 1-\sqrt{1-K\delta}. \end{aligned}$$

• We now prove (B.5) for $K=2$. Let w be in $\mathcal{W}_2$ and without loss of generality assume that $w_1 \le 1/2 \le w_2$. Once again we only need to consider $w_1 < \delta$. Then we take $v=(\delta, 1-\delta)$ and we get

$$\begin{aligned} h^2(w,\mathcal{W}_{2,\delta}) &\le h^2(w,v) \\ &= \frac{1}{2}\left[\left(\sqrt{w_1}-\sqrt{\delta}\right)^2+\left(\sqrt{1-w_1}-\sqrt{1-\delta}\right)^2\right] \\ &\le 1-\sqrt{1-\delta}. \end{aligned}$$

This ends the proof of (B.5). We can now prove Lemma B.2. Let $P \in \mathscr{P}$ and $P_{w,F} \in \mathscr{Q}(\theta)$. There is $v \in \mathcal{W}_{K,\delta}$ such that $P_{v,F} \in \mathscr{Q}_\delta(\theta)$ and

$$h^2\left(w,v\right) \le 1-\sqrt{1-(K-1)\delta} \le (K-1)\delta.$$

By a density argument we can assume that $v \in \mathbb{Q}^K$. Therefore,

$$\begin{aligned} h\left(P,\mathscr{Q}_\delta(\theta)\right) &\le h\left(P,P_{v,F}\right) \\ &\le h\left(P_{v,F},P_{w,F}\right)+h\left(P,P_{w,F}\right) \\ &\le \sqrt{(K-1)\delta}+h\left(P,P_{w,F}\right) \end{aligned}$$

where the last inequality comes from Lemma B.3. Then, taking the infimum over $\mathscr{Q}(\theta)$ ends the proof. □

B.2 Proof of Theorem 3.1

It is a direct application of Theorem B.1 in the specific situation where

$$\Theta = \{\theta = (K, \lambda_1, \lambda_2, \ldots, \lambda_K)\}.$$

Then, taking $\Delta(\theta)=0$, inequality (B.1) becomes

$$\begin{aligned} \mathbf{h}^2\left(\mathbf{P}^*,\left(\hat{P}_\delta\right)^{\otimes n}\right) &\le c_0\left(\mathbf{h}^2(\mathbf{P}^*,\mathscr{Q}_K)+n(K-1)\delta\right) \\ &\quad + c_1 116.1\overline{V}\left[5.82+\log\left(\frac{(K+1)^2}{\delta}\right)+\log_+\left(\frac{n}{\overline{V}}\right)\right] \\ &\quad + c_1(1.49+\xi). \end{aligned}$$

With (B.2), we have

$$\begin{aligned}\mathbf{h}^2\left(\mathbf{P}^*, \hat{P}^{\otimes n}\right) \leq\ & c_0\left(\mathbf{h}^2(\mathbf{P}^*, \mathscr{Q}) + n \wedge \overline{V}\right) \\ & + c_1 116.1\,(2 + \log_2(9))\,\overline{V}\left[5.82 + \log\left(\frac{Kn}{\overline{V} \wedge n}\right)\right] \\ & + c_1(1.49 + \xi),\end{aligned}$$

for $K \geq 2$. One can easily check that it still holds for $K = 1$ (see [3]). Therefore (3.7) is proven.

B.3 Proof of Theorem 3.3

Let $\mathscr{Q}_K[\epsilon]$ be the model defined by

$$\mathscr{Q}_K[\epsilon] = \left\{\sum_{k=1}^{K} w_k F_k; w \in \mathcal{W}_K, F_k \in \mathscr{F}_k[\epsilon], \forall k \in [K]\right\}.$$

Since the class $\overline{\mathscr{F}}_k$ is totally bounded, the set $\mathscr{F}_k[\epsilon]$ is finite for all $k \in [K]$. We satisfy Assumptions 1 and 2 and therefore can apply Theorem 3.1 with

$$\overline{V} = \sum_{k=1}^{K} \log_2(|\mathscr{F}_k[\epsilon]|) \leq \sum_{k=1}^{K}\left(\frac{A_k}{\epsilon}\right)^{\alpha_k}.$$

Let $\hat{P} = \hat{P}_\delta$ be a ρ-estimator on $\mathscr{Q}_{K,\delta}[\epsilon]$. For all $\xi > 0$, we have

$$\begin{aligned}\mathbf{h}^2\left(\mathbf{P}^*, \left(\hat{P}_\delta\right)^{\otimes n}\right) \leq\ & c_0\left[\mathbf{h}^2\left(\mathbf{P}^*, \mathscr{Q}_K[\epsilon]\right) + n(K-1)\delta\right] \\ & + c_1 116.1 \overline{V}\left[5.82 + \log\left(\frac{(K+1)^2}{\delta}\right) + \log_+\left(\frac{n}{\overline{V}}\right)\right] \\ & + c_1(1.49 + \xi),\end{aligned}$$

with probability at least $1 - e^{-\xi}$.

Lemma B.3. *Let w and v be in $\mathcal{W}_K$. Let F_k and G_k be in $\mathscr{P}$ for all $k \in \{1, \ldots, K\}$. We have*

$$h\left(\sum_{k=1}^{K} w_k F_k, \sum_{k=1}^{K} v_k G_k\right) \leq h(w, v) + \max_{k \in [K]} h\left(F_k, G_k\right).$$

This lemma implies that $\mathscr{Q}_K[\epsilon]$ is a ϵ-net of $\mathscr{Q}_K$ with respect to the Hellinger distance, and in particular

$$\mathbf{h}^2\left(\mathbf{P}^*, \mathscr{Q}_K[\epsilon]\right) \leq 2\mathbf{h}^2\left(\mathbf{P}^*, \mathscr{Q}_K\right) + 2n\epsilon^2.$$

Therefore, if we use (3.7) with $\overline{V} = \sum_{k=1}^{K}\left(\frac{A_k}{\epsilon}\right)^{\alpha_k}$ we get

$$C\mathbf{h}^2\left(\mathbf{P}^*, \left(\hat{P}_\delta\right)^{\otimes n}\right) \leq 2\mathbf{h}^2\left(\mathbf{P}^*, \mathscr{Q}_K\right) + 2n\epsilon^2 + \epsilon^{-\alpha_{\max}} \sum_{k=1}^{K} A_k^{\alpha_k}\left[1 + \log(Kn)\right] + \xi.$$

Finally, for $\epsilon = n^{-\frac{1}{\alpha_{\max}+2}}$, there exists a positive constant C such that for all $\xi > 0$, we have

$$C\mathbf{h}^2\left(\mathbf{P}^*, \left(\hat{P}_\delta\right)^{\otimes n}\right) \leq \mathbf{h}^2\left(\mathbf{P}^*, \mathscr{Q}_K\right) + n^{\frac{\alpha_{\max}}{\alpha_{\max}+2}} \sum_{k=1}^{K} A_k^{\alpha_k} \left[1 + \log\left(Kn\right)\right] + \xi,$$

with probability at least $1 - e^{-\xi}$.

Proof of Lemma B.3

With Young's inequality, we can easily prove the following inequality

$$\forall x, y, z \in \mathbb{R}_+^K, \left(\sqrt{\sum_{k\in[K]} x_k z_k} - \sqrt{\sum_{k\in[K]} x_k y_k}\right)^2 \leq \sum_{k\in[K]} x_k (\sqrt{z_k} - \sqrt{y_k})^2.$$

Therefore, we get an upper bound on the Hellinger distance between mixture distributions. For $w, v \in \mathcal{W}_K$ and $F_k, G_k \in \mathscr{P}$ for all $k \in [K]$, we have

$$\begin{aligned} h\left(\sum_{k\in[K]} w_k F_k, \sum_{k\in[K]} v_k G_k\right) &\leq h\left(\sum_{k\in[K]} w_k F_k, \sum_{k\in[K]} w_k G_k\right) + h\left(\sum_{k\in[K]} w_k G_k, \sum_{k\in[K]} v_k G_k\right) \\ &\leq \sqrt{\sum_{k\in[K]} w_k h^2\left(F_k, G_k\right)} + h\left(w, v\right) \\ &\leq \max_{k\in[K]} h(F_k, G_k) + h(w, v). \end{aligned}$$

□

B.4 Proof of Theorem 4.1

Applying Theorem B.1 in the described setting, we get

$$\begin{aligned} h^2\left(P^*, \hat{P}\right) \leq \inf_{\lambda\in L} \Bigg[& c_0\left(h^2(P^*, \mathscr{Q}(\lambda)) + (K-1)\delta(\lambda)\right) \\ & + c_2\left\{\frac{116.1\overline{V}(\lambda)}{n}\left[5.82 + \log\left(\frac{(K+1)^2}{\delta(\lambda)}\right) + \log_+\left(\frac{n}{\overline{V}(\lambda)}\right)\right] + \Delta(\lambda)\right\}\Bigg] \\ & + c_2\frac{1.49 + \xi}{n}, \end{aligned}$$

with probability at least $1 - e^{-\xi}$. As $K \geq 2$ and $\delta(\lambda) = \frac{\overline{V}(\lambda)}{n(K-1)} \wedge \frac{1}{K}$ we have the following with (B.2). and finally we have

$$\begin{aligned} \mathbf{h}^2\left(\mathbf{P}^*, \hat{P}^{\otimes n}\right) \leq \inf_{\lambda\in L} \Big\{ & c_0\left(\mathbf{h}^2(\mathbf{P}^*, \mathscr{Q}(\lambda)) + n \wedge \overline{V}(\lambda)\right) \\ & + c_1\left(116.1\overline{V}(\lambda)\left[5.82 + (2 + \log_2(9))\log\left(\frac{Kn}{\overline{V}(\lambda) \wedge n}\right)\right] + \Delta(\lambda)\right)\Big\} \\ & + c_1(1.49 + \xi) \end{aligned}$$

$$\leq C \inf_{\lambda \in L} \left\{ \mathbf{h}^2 \left(\mathbf{P}^*, \mathscr{Q}(\lambda) \right) + \overline{V}(\lambda) \left[1 + \log \left(\frac{Kn}{\overline{V}(\lambda) \wedge n} \right) \right] + \Delta(\lambda) \right\} + \xi,$$

where C is a positive numeric constant that does not depend on L.

B.5 Proof of Theorem 4.3

Applying Theorem B.1, we get

$$h^2 \left(P^*, \hat{P} \right) \leq \inf_{K \in \mathscr{K}} \Bigg[c_0 \left(h^2(P^*, \mathscr{Q}(K)) + (K-1)\delta(K) \right) + c_2 \left\{ \frac{116.1KV}{n} \left[5.82 + \log \left(\frac{(K+1)^2}{\delta(K)} \right) + \log_+ \left(\frac{n}{KV} \right) \right] + \Delta(K) \right\} \Bigg] + c_2 \frac{1.49 + \xi}{n},$$

with probability at least $1 - e^{-\xi}$. For $K = 1$ and $\delta(K) = 1$ we have $(K-1)\delta(K) = 0 \leq KV \wedge n$ and

$$\log \left(\frac{(K+1)^2}{\delta(K)} \right) + \log_+ \left(\frac{n}{KV} \right) = 2\log(2) + \log \left(\frac{n}{KV \wedge n} \right).$$

Combining this inequality with (B.2), we have

$$5.82 + \log \left(\frac{(K+1)^2}{\delta(K)} \right) + \log_+ \left(\frac{n}{KV} \right) \leq (5.82 + 2\log(2)) + (2 + \log_2(9)) \log \left(\frac{Kn}{KV \wedge n} \right)$$

for all $K \geq 1$. Finally, there is a numeric constant $C > 0$ that is universal, such that for all $\xi > 0$ we have

$$Ch^2 \left(P^*, \hat{P} \right) \leq \inf_{K \in \mathscr{K}} \left[h^2(P^*, \mathscr{Q}(K)) + \frac{1}{n} \left\{ KV \left[1 + \log \left(\frac{Kn}{KV \wedge n} \right) \right] + \Delta(K) \right\} \right] + \frac{\xi}{n},$$

with probability at least $1 - e^{-\xi}$.

Appendix C. Density estimation

This section gathers the proofs of density estimation results, namely Theorems 3.6 and 4.5.

C.1 Proof of Theorem 3.6

The Gaussian location-scale family of density functions is VC-subgraph with VC-index $V(\mathcal{C}) \leq 5$ (see Lem. 3.12). Proposition 3.5 provides an approximation bound for $\mathscr{C}(A, R)$. The proof can be found on page 439. We can now apply Theorem 3.1 with those two propositions. With (3.7), there exists a universal constant C such that for $\mathbf{P}^* = (P^*)^{\otimes n}$, $\xi > 0$, with probability at least $1 - e^{-\xi}$, we have

$$Ch^2 \left(P^*, \hat{P} \right) \leq h^2 \left(P^*, \mathscr{C}(A, R) \right) + \exp \left(-\frac{K^{1/2}}{12\sqrt{6}R^2} \right) \left[K^{1/4} \frac{3\sqrt{2}}{\sqrt{e\pi}7^{1/4}} + R \right] + \frac{K \log(n) + \xi}{n}$$

$$\leq h^2\left(P^*, \mathscr{C}(A,R)\right) + \frac{1}{n}\left[\frac{3\sqrt{2}}{(e\pi)^{1/2}7^{1/4}}\left(864R^4\log^2(n)+1\right)^{1/4} + R\right]$$
$$+ \frac{\left(864R^4\log^2(n)+1\right)\log(n)+\xi}{n}$$

One can check that the assumptions ensure that $\log(n) \geq 1$ and therefore

$$Ch^2\left(P^*, \hat{P}\right) \leq h^2\left(P^*, \mathscr{C}(A,R)\right) + \frac{R\log^{1/2}(n)}{n}\left[\frac{3\sqrt{2}865^{1/4}}{(e\pi)^{1/2}7^{1/4}} + 1\right] + \frac{865R^4\log^3(n)+\xi}{n}.$$

Finally, there exists a numeric constant $C > 0$ such that, for $K = \left\lceil 864R^4\log^2(n)\right\rceil \geq 2\left(24A^2+1\right)^2$, for all $\xi > 0$, with probability at least $1 - e^{-\xi}$, we have

$$Ch^2\left(P^*, \hat{P}\right) \leq h^2(P^*, \mathscr{C}(A,R)) + \frac{R^4\log^3(n)+\xi}{n}.$$

The different conditions are satisfied for $n \geq \exp\left(\frac{A^2}{R^2}\frac{25}{12\sqrt{3}}\right)$.

C.2 Proof of Proposition 3.5

We first need the following result.

Lemma C.1. *Let k be a positive integer. For any probability distribution H on $[-a,a]\times[\underline{\sigma},\overline{\sigma}]$, there is a discrete probability distribution H' supported by $k(2k-1)+1$ points in $[-a,a]\times[\underline{\sigma},\overline{\sigma}]$ such that*

$$d_{TV}\left(P_H, P_{H'}\right) \leq \inf_{m>1}\left\{\frac{\sqrt{2/\pi}}{\underline{\sigma}}am\left(\frac{ea^2(1+m)^2}{2k\underline{\sigma}^2}\right)^k + \frac{\overline{\sigma}}{2\underline{\sigma}}\exp\left(-\frac{(m-1)^2a^2}{2\overline{\sigma}^2}\right)\right\}.$$

The proof is postponed at the end of this one. Let A and R be two real numbers respectively greater than 0 and 1. As a direct consequence of this lemma, for any $l \in \mathbb{R}$, any probability distribution H on $[l \pm \underline{\sigma}A]\times[\underline{\sigma}, R\underline{\sigma}]$ and for $K \geq k(2k-1)+1$, we have

$$h^2(P_H, \mathscr{G}_K) \leq \inf_{m>0}\left\{\sqrt{2/\pi}A(1+m)\left(\frac{eA^2(2+m)^2}{2k}\right)^k + \frac{R}{2}\exp\left(-\frac{m^2A^2}{2R^2}\right)\right\}.$$

The goal is to have an upper bound without an infimum. For that we are going to take a value of m given by the parameters A and R. Now

$$h^2(P_H, \mathscr{G}_K) \leq \inf_{m\geq 2}\left\{\sqrt{2/\pi}A\frac{3}{2}m\left(\frac{eA^2 4m^2}{2k}\right)^k + \frac{R}{2}\exp\left(-\frac{m^2A^2}{2R^2}\right)\right\}$$
$$= \inf_{m\geq 2}\left\{\frac{3}{\sqrt{2\pi}}Am\left(\frac{2eA^2m^2}{k}\right)^k + \frac{R}{2}\exp\left(-\frac{m^2A^2}{2R^2}\right)\right\}.$$

Let W denote the Lambert W function restricted to $(0;\infty)$ such that $W(x)$ is the only positive number such that $W(x)e^{W(x)} = x$. For $m = \frac{\sqrt{2W(1/4eR^2)}R}{A}k^{1/2}$ and $k \geq \frac{2A^2}{W(1/4eR^2)R^2}$, to ensure that $m \geq 2$, we get

$$\begin{aligned} h^2(P_H, \mathscr{G}_K) &\leq \frac{3}{\sqrt{2\pi}}\sqrt{2W(1/4eR^2)}Rk^{1/2}\left(4eR^2W(1/4eR^2)\right)^k + \frac{R}{2}\exp\left(-kW(1/4eR^2)\right) \\ &= R\exp\left(-kW(1/4eR^2)\right)\left[k^{1/2}3\sqrt{W(1/4eR^2)/\pi} + 1/2\right]. \end{aligned}$$

Let us simplify this bound using simple properties of the function W.

- For all $x > 0$, $0 < W(x) < x$.
- For all $x \in (0,1), x(1-x) < W(x)$. Therefore,

$$\begin{aligned} W(1/4eR^2) &\geq \frac{1}{4eR^2}\left(1 - \frac{1}{4eR^2}\right) \\ &\geq \frac{(1-1/4e)}{4eR^2} = \frac{4e-1}{16e^2R^2} \geq \frac{1}{12R^2}. \end{aligned}$$

Therefore, we have

$$h^2(P_H, \mathscr{G}_K) \leq R\exp\left(-\frac{k}{12R^2}\right)\left[k^{1/2}\frac{3}{2R\sqrt{e\pi}} + 1/2\right].$$

Since $K \geq 2\left(24A^2+1\right)^2$, one can check that the set

$$B = \left\{k \in \mathbb{N} : K \geq k(2k-1)+1 \text{ and } k \geq \frac{2A^2}{R^2W(1/4eR^2)}\right\}$$

is not empty, *e.g.* $\lceil 24A^2 \rceil \in B$. We set $k = \max B \geq 1$, *i.e.* $k = \left\lfloor \frac{1}{4} + \sqrt{(K-7/8)/2} \right\rfloor \leq \sqrt{K}\frac{2}{\sqrt{7}}$, we have

$$K \in \{n(2n-1)+1, \ldots, (2n+1)(n+1)\} \Rightarrow k = n \geq \sqrt{K}\frac{n}{\sqrt{(2n+1)(n+1)}}.$$

Since $x \mapsto \frac{x}{\sqrt{(2x+1)(x+1)}}$ is non-decreasing on $[1,+\infty)$, we have $k \geq \sqrt{K}/\sqrt{6}$ for all $K \geq 2$. Finally, we have

$$\begin{aligned} h^2(P_H, \mathscr{G}_K) &\leq R\exp\left(-\frac{k}{12R^2}\right)\left[k^{1/2}\frac{3}{2R\sqrt{e\pi}} + 1/2\right] \\ &\leq R\exp\left(-\frac{K^{1/2}}{12\sqrt{6}R^2}\right)\left[K^{1/4}\frac{3\sqrt{2}}{2R\sqrt{e\pi}7^{1/4}} + 1/2\right] \\ &= \frac{1}{2}\exp\left(-\frac{K^{1/2}}{12\sqrt{6}R^2}\right)\left[K^{1/4}\frac{3\sqrt{2}}{\sqrt{e\pi}7^{1/4}} + R\right]. \end{aligned}$$

One can see that $\underline{\sigma}$ does not play a role here and is equivalent to s in the definition of $\mathcal{C}(A,R)$.

Proof of Lemma C.1

The bound is obtained following the proofs of lemmas in Ghosal and van der Vaart [15]

- 1st step:
  For $|x| > a$ we have,

$$p_H(x) = \int \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-z)^2}{2\sigma^2}\right) \mathrm{d}H(z,\sigma)$$
$$\leq \frac{1}{\sqrt{2\pi\underline{\sigma}^2}} \exp\left(-\frac{(|x|-a)^2}{2\overline{\sigma}^2}\right). \tag{C.1}$$

- 2nd step:
  See Lemma A.1 in Ghosal and van der Vaart [15]. Take $N = k(2k-1)+1$. There is a discrete distribution H' with at most K support points in $[-a,a] \times [\underline{\sigma}, \overline{\sigma}]$ such that

$$\int z^l \sigma^{-(2j+1)} \mathrm{d}H(z,\sigma) = \int z^l \sigma^{(2j+1)} \mathrm{d}H'(z,\sigma) \tag{C.2}$$

  for $l = 0, \ldots, 2k-2$ and $j = 0, \ldots, k-1$. Because of (C.2) we get

$$\int \sum_{j=0}^{k-1} \frac{(-1)^j \sigma^{-(2j+1)} (x-z)^{2j}}{j!} \mathrm{d}H(z,\sigma) = \int \sum_{j=0}^{k-1} \frac{(-1)^j \sigma^{-(2j+1)} (x-z)^{2j}}{j!} \mathrm{d}H'(z,\sigma),$$

  for $x \in \mathbb{R}$. Taylor's expansion of the exponential function ([15]),

$$\left| \exp\left(-\frac{(x-z)^2}{2\sigma^2}\right) - \sum_{j=0}^{k-1} \frac{\left(-\frac{(x-z)^2}{2\sigma^2}\right)^j}{j!} \right| \leq \left(\frac{e(x-z)^2}{k2\sigma^2}\right)^k.$$

  Therefore,

$$\sqrt{2\pi} \sup_{|x|\leq M} |p_H(x) - p_{H'}(x)|$$
$$= \sup_{|x|\leq M} \left| \int \frac{1}{\sigma} \exp\left(-\frac{(x-z)^2}{2\sigma^2}\right) \mathrm{d}H(z,\sigma) \right.$$
$$\left. - \int \frac{1}{\sigma} \exp\left(-\frac{(x-z)^2}{2\sigma^2}\right) \mathrm{d}H'(z,\sigma) \right|$$
$$= \sup_{|x|\leq M} \left| \int \frac{1}{\sigma} \left[\exp\left(-\frac{(x-z)^2}{2\sigma^2}\right) - \sum_{j=0}^{k-1} \frac{\left(-\frac{(x-z)^2}{2\sigma^2}\right)^j}{j!} \right] \mathrm{d}H(z,\sigma) \right.$$
$$\left. - \int \frac{1}{\sigma} \left[\exp\left(-\frac{(x-z)^2}{2\sigma^2}\right) - \sum_{j=0}^{k-1} \frac{\left(-\frac{(x-z)^2}{2\sigma^2}\right)^j}{j!} \right] \mathrm{d}H'(z,\sigma) \right|$$
$$\leq 2 \sup_{\substack{|x|\leq M \\ |z|\leq a \\ \underline{\sigma}\leq\sigma\leq\overline{\sigma}}} \frac{1}{\sigma} \left| \exp\left(-\frac{(x-z)^2}{2\sigma^2}\right) - \sum_{j=0}^{k-1} \frac{\left(-\frac{(x-z)^2}{2\sigma^2}\right)^j}{j!} \right|$$

$$\leq 2 \sup_{\substack{|x|\leq M \\ |z|\leq a \\ \underline{\sigma}\leq\sigma\leq\overline{\sigma}}} \frac{1}{\sigma}\left(\frac{e(x-z)^2}{k2\sigma^2}\right)^k$$
$$\leq \frac{2}{\underline{\sigma}}\left(\frac{e(M+a)^2}{k2\underline{\sigma}^2}\right)^k.$$

Obviously, the inequality (C.1) holds also for $p_{H'}$. We combine it with the last one we obtained in order to bound the total variation distance. Therefore, for $M = ma$, $m > 1$, we have

$$\begin{aligned} d_{TV}\left(P_H, P_{H'}\right) &= \frac{1}{2}\int |p_H(x) - p_{H'}(x)|\mathrm{d}x \\ &\leq M \sup_{|x|\leq M} |p_H(x) - p_{H'}(x)| + \frac{1}{2}\int_{|x|>M} p_H(x) \vee p_{H'}(x)\mathrm{d}x \\ &\leq \frac{\sqrt{2/\pi}}{\underline{\sigma}} M\left(\frac{e(M+a)^2}{2k\underline{\sigma}^2}\right)^k + \frac{1}{2}\int_{|x|>M} \frac{1}{\sqrt{2\pi\underline{\sigma}^2}}\exp\left(-\frac{(|x|-a)^2}{2\overline{\sigma}^2}\right)\mathrm{d}x \\ &\leq \frac{\sqrt{2/\pi}}{\underline{\sigma}} M\left(\frac{e(M+a)^2}{2k\underline{\sigma}^2}\right)^k + \frac{\overline{\sigma}}{\underline{\sigma}}\int_{x>M} \frac{1}{\sqrt{2\pi\overline{\sigma}^2}}\exp\left(-\frac{(x-a)^2}{2\overline{\sigma}^2}\right)\mathrm{d}x \\ &\leq \frac{\sqrt{2/\pi}}{\underline{\sigma}} am\left(\frac{ea^2(1+m)^2}{2k\underline{\sigma}^2}\right)^k + \frac{\overline{\sigma}}{2\underline{\sigma}}\exp\left(-\frac{(m-1)^2a^2}{2\overline{\sigma}^2}\right). \end{aligned}$$

Finally, writing $A = a/\underline{\sigma}$ and $R = \overline{\sigma}/\underline{\sigma}$, we have

$$d_{TV}\left(P_H, P_{H'}\right) \leq \inf_{m>1}\left\{\frac{\sqrt{2/\pi}}{\underline{\sigma}} am\left(\frac{ea^2(1+m)^2}{2k\underline{\sigma}^2}\right)^k + \frac{\overline{\sigma}}{2\underline{\sigma}}\exp\left(-\frac{(m-1)^2a^2}{2\overline{\sigma}^2}\right)\right\}.$$

□

This concludes the proof of Proposition 3.5.

C.3 Proof of Theorem 4.5

We firs provide the proof of Lemma 4.4 which provides the necessary bound for the approximation.

Proof of Lemma 4.4

We will use notation from [20]. With Lemma 7.23 [19] and an inclusion argument, we have

$$h^2\left(P, \mathscr{G}_K\right) \leq h^2\left(P, \mathcal{S}_K\right) \leq \frac{1}{2} D_{KL}\left(P||\mathcal{S}_K\right).$$

Combined with Lemma 6.1 [20], we get

$$\begin{aligned} h^2\left(P, \mathscr{G}_K\right) &\leq \frac{c_{\underline{\beta},\overline{\beta}}}{2}\lambda(K)^{2\beta} \\ &= \frac{c_{\underline{\beta},\overline{\beta}}}{2}\left(a_{\overline{\beta}} K^{-1}\left(\ln K\right)^{3/2}\right)^{2\beta} \\ &\leq C_{\underline{\beta},\overline{\beta}}\frac{\left(\ln K\right)^{3\beta}}{K^{2\beta}}, \end{aligned}$$

with $C_{\underline{\beta},\overline{\beta}} = c_{\underline{\beta},\overline{\beta}} a_{\overline{\beta}}^{2\overline{\beta}}/2$. □

The Gaussian location-scale family of density functions is VC-subgraph (see Lem. 3.12). For $0 < \underline{\beta} < \overline{\beta}$ and $\beta \in [\underline{\beta}, \overline{\beta}]$, let $\mathcal{H}\left(\beta, \mathcal{P}(\underline{\beta}, \overline{\beta})\right)$ be the class of density functions defined in Maugis and Michel [20]. One can check that

$$\sum_{k \in \mathscr{K}} e^{-\Delta(K)} \leq 1,$$

for $\Delta(K) = K$. Applying Theorem 4.3, for $\xi > 0$, with probability at least $1 - e^{-\xi}$, we have

$$\begin{aligned} Ch^2(P^*, \hat{P}) &\leq \inf_{K \in \mathscr{K}} \left\{ h^2(P^*, \mathscr{G}_K) + \frac{K(5\log(n)+1) + \xi}{n} \right\} \\ &\leq 2h^2(P^*, \mathcal{H}\left(\beta, \mathcal{P}(\underline{\beta}, \overline{\beta})\right)) + \inf_{K \in \mathscr{K}} \left\{ 2c_{\underline{\beta},\overline{\beta}} \frac{(\log K)^{3\beta}}{K^{2\beta}} + \frac{K(5\log(n)+1)}{n} \right\} + \frac{\xi}{n}. \end{aligned}$$

Therefore, following the proof of Theorem 2.9 of Maugis and Michel [20], we have

$$\begin{aligned} \inf_{K \in \mathscr{K}} \left\{ 2c_{\underline{\beta},\overline{\beta}} \frac{(\log K)^{3\beta}}{K^{2\beta}} + \frac{K(5\log(n)+1)}{n} \right\} &\lesssim c_{\underline{\beta},\overline{\beta}} \inf_{K \in \mathscr{K}} \left\{ \frac{(\log K)^{3\beta}}{K^{2\beta}} + \frac{K\log(n)}{n} \right\} \\ &\lesssim c_{\underline{\beta},\overline{\beta}} \frac{(\log n)^{\frac{5\beta}{2\beta+1}}}{n^{\frac{2\beta}{2\beta+1}}}. \end{aligned}$$

Finally, there exists $C_{\underline{\beta},\overline{\beta}}$ such that for all $\xi > 0$, with probability at least $1 - e^{-\xi}$, we have

$$h^2(P^*, \hat{P}) \leq C_{\underline{\beta},\overline{\beta}} \left(\frac{(\log n)^{\frac{5\beta}{2\beta+1}}}{n^{\frac{2\beta}{2\beta+1}}} + \frac{\xi}{n} \right).$$

Appendix D. Regular parametric models

This section gathers the proof of Theorems 3.8, 4.2 and 3.9.

D.1 Proof of Theorem 3.8

We apply the results of Ibragimov and Has'minskiĭ [17] (Chap. 1, Sect. 7.1 and 7.3) to parametric mixture models. We recall the notation

$$p(\cdot;\theta) = \sum_{k=1}^{K-1} w_k f_k(\cdot;\alpha_k) + (1 - w_1 - \cdots - w_{K-1}) f_K(\cdot;\alpha_K)$$

and $\Theta = \left\{ w \in (0,1)^{K-1}, \sum_{k=1}^{K-1} w_k < 1 \right\} \times A_1 \times \cdots \times A_K$. Obviously, Θ is an open convex subset of $\mathbb{R}^{K-1} \times \mathbb{R}_1^d \times \cdots \times \mathbb{R}^{d_K}$. We first check that Assumption 3 implies that the model is regular.

- a) $\Rightarrow \theta \mapsto p(x;\theta)$ is continuous on Θ for μ-almost all $x \in \mathscr{X}$.

- b) $\Rightarrow$ For μ-almost all $x \in \mathscr{X}$ the function $u \mapsto p(x;u)$ is differentiable at the point $u = \theta$. For all $k \in \{1, \dots, K\}$ and $j \in \{1, \dots, d_k\}$, we have

$$\begin{aligned}\int_{\mathscr{X}} \left|\frac{\partial p(x;\theta)}{\partial \alpha_{k,j}}\right|^2 \frac{\mu(\mathrm{d}x)}{p(x;\theta)} &= \int_{\mathscr{X}} \left|\frac{\partial f_k(x;\alpha_k)}{\partial \alpha_{k,j}}\right|^2 \frac{w_k^2}{p(x;\theta)}\mu(\mathrm{d}x) \\ &\leq \int_{\mathscr{X}} \left|\frac{\partial f_k(x;\alpha_k)}{\partial \alpha_{k,j}}\right|^2 \frac{\mu(\mathrm{d}x)}{f_k(x;\alpha_k)} < \infty.\end{aligned}$$

  It also works with $k = K$ since w is fixed here. For $k \in \{1, \dots, K-1\}$ we get

$$\begin{aligned}\int_{\mathscr{X}} \left|\frac{\partial p(x;\theta)}{\partial w_k}\right|^2 \frac{\mu(\mathrm{d}x)}{p(x;\theta)} &= \int_{\mathscr{X}} \left(f_k(x;\alpha_k) - f_K(x,\alpha_K)\right)^2 \frac{\mu(\mathrm{d}x)}{p(x;\theta)} \\ &\leq \frac{2}{w_k}\int_{\mathscr{X}} f_k^2(x;\alpha_k)\frac{\mu(\mathrm{d}x)}{f_k(x;\alpha_k)} \\ &+ \frac{2}{1-w_1-\cdots-w_k}\int_{\mathscr{X}} f_K^2(x;\alpha_k)\frac{\mu(\mathrm{d}x)}{f_K(x;\alpha_K)} \\ &= \frac{2}{w_k} + \frac{2}{1-w_1-\cdots-w_k} < \infty.\end{aligned}$$

Therefore, we have a regular statistical experiment (see [17]). Since the Fisher's information matrix

$$I\left(\overline{\theta}\right) = \int_{\mathscr{X}} \frac{\partial p\left(x;\overline{\theta}\right)}{\partial \theta}\left(\frac{\partial p\left(x;\overline{\theta}\right)}{\partial \theta}\right)^T \frac{\mu(\mathrm{d}x)}{p\left(x;\overline{\theta}\right)}$$

is definite positive. We can apply Theorem 7.6 of Ibragimov and Has'minskiĭ [17] which says that we have

$$\liminf_{t\to 0} ||t||^{-2} h^2(P_{\overline{\theta}}, P_{\overline{\theta}+t}) \geq \lambda(\overline{\theta})/4.$$

where $\lambda(\overline{\theta})$ is the smallest eigen value of the Fisher's information matrix $I(\overline{\theta})$. Therefore there exists $a > 0$ such that

$$\inf_{\theta\in\Theta:||\overline{\theta}-\theta||<a} \left|\left|\overline{\theta}-\theta\right|\right|^{-2} h^2\left(P_{\overline{\theta}}, P_\theta\right) \geq \lambda\left(\overline{\theta}\right)/8.$$

Finally, there exists a positive constant $C\left(\overline{\theta}\right) = \frac{\lambda(\overline{\theta})}{8} \wedge \inf_{\substack{||\theta-\overline{\theta}||\geq a \\ \theta\in\Theta}} h^2\left(P_{\overline{\theta}}, P_\theta\right) > 0$ such that

$$\forall \theta \in \Theta, \left(1 + \left|\left|\overline{\theta}-\theta\right|\right|^{-2}\right) h^2\left(P_{\overline{\theta}}, P_\theta\right) \geq C\left(\overline{\theta}\right).$$

We apply Theorem 3.1 so that with probability at least $1 - e^{-\xi}$ we have

$$\frac{1}{n}\left[\mathbf{h}^2\left(\mathbf{P}^*, P_{\overline{\theta}}^{\otimes n}\right) + \overline{V}\log(n) + \xi\right] \geq Ch^2\left(P_{\overline{\theta}}, P_{\hat{\theta}}\right) \geq \frac{\left|\left|\overline{\theta}-\hat{\theta}\right|\right|^2}{1+\left|\left|\overline{\theta}-\hat{\theta}\right|\right|^2} C \times C\left(\overline{\theta}\right)$$

$$\geq \frac{\left|\left|\overline{\theta}-\hat{\theta}\right|\right|^2 \wedge b}{1+b} C \times C(\overline{\theta}),$$

for any $b \geq 0$. Since $||\overline{w}-\hat{w}||^2 \leq K \sum_{k=1}^{K-1} (\overline{w}_k - \hat{w}_k)^2$ and

$$\begin{aligned}\sum_{k=1}^{K-1} (\overline{w}_k - \hat{w}_k)^2 + \sum_{k=1}^{K} \left[||\overline{\alpha}_k - \hat{\alpha}_K||^2 \wedge 1\right] &\leq \sum_{k=1}^{K-1} (\overline{w}_k - \hat{w}_k)^2 + \left[\sum_{k=1}^{K} ||\overline{\alpha}_k - \hat{\alpha}_K||^2\right] \wedge K \\ &\leq \left[\sum_{k=1}^{K-1} (\overline{w}_k - \hat{w}_k)^2 + \sum_{k=1}^{K} ||\overline{\alpha}_k - \hat{\alpha}_K||^2\right] \wedge (K+1) \\ &= \left|\left|\overline{\theta}-\hat{\theta}\right|\right|^2 \wedge (K+1),\end{aligned}$$

we get, with $b = K+1$,

$$\frac{1}{n}\left[\mathbf{h}^2\left(\mathbf{P}^*, P_{\overline{\theta}}^{\otimes n}\right) + \overline{V}\log(n) + \xi\right] \geq \left[\frac{1}{K}||\overline{w}-\hat{w}||^2 + \sum_{k=1}^{K} ||\overline{\alpha}_k - \hat{\alpha}_k||^2 \wedge 1\right] \frac{C \times C(\overline{\theta})}{K+2},$$

with probability at least $1-e^{-\xi}$.

D.2 Proof of Theorem 4.2

Assumption 2 is satisfied with Lemma 3.12. For all j in $\{0,\dots,K\}$, we have $\overline{V}_j = 5K$. We apply Theorem 4.1 with $\Delta_j = \log(K+1)$ for all $j \in \{0,\dots,K\}$. This induces a constant penalty function and one can check that this does not modify the definition of ρ-estimators compared to a null penalty function. Therefore, the estimator can be computed with a null penalty. There exists a positive constant that does not depend on P^* such that for $n \geq 5K$, any ρ-estimator $\hat{P}_\delta$ on $\mathscr{Q}_\delta$ satisfies, with probability at least $1-e^{-\xi}$,

$$Ch^2(P^*, \hat{P}) \leq \frac{K\log\left(n(K+1)\right)+\xi}{n}.$$

The following lemma allows to prove that for n large enough, the estimator $\hat{P}$ belongs to the true model $\mathscr{Q}_{j^*}$ with high probability.

Lemma D.1. *Let $j \in \{0,\dots,K\}$ and assume there is a sequence*

$$(P_n)_n = \left(\sum_{k=1}^{j} w_{k,n}\mathcal{N}(z_{k,n},\sigma_{k,n}^2) + \sum_{k=j+1}^{K} w_{k,n}\, Cauchy(z_{k,n},\sigma_{k,n})\right)_n \in \mathscr{Q}_j^{\mathbb{N}}$$

such that $\lim_{n\to\infty} h(P_n, P^*) = 0$. Then, $j = j^*$ and there is a subsequence $\left(P_{\psi(n)}\right)_n$ such that $\lim_{n\to\infty} (z_{k,\psi(n)}, \sigma_{k,\psi(n)})_{1\leq k\leq K} = (\overline{z}_k, \overline{\sigma}_k)_{1\leq k\leq K}$.

This implies that $\alpha = \inf_{j\neq j^*} h\left(P^*, \mathscr{Q}_j\right) > 0$. For $n \geq n_0 = \inf\{n \geq 1 : C^{-1}\alpha^{-1}K < n/\log(n(K+1))\}$ and $0 < \xi < \frac{Cn\alpha}{K\log(n(K+1))}$, there is an event $\Omega_{\xi,n}$ of probability $1-e^{-\xi}$ such that

$$Ch^2(P^*, \hat{P}) \leq \frac{K\log\left(n(K+1)\right)+\xi}{n} \text{ and } \hat{P} \in \mathscr{Q}_{j^*}.$$

From now, we follow the proof of Theorem 3.6 to prove a lower bound on the Hellinger distance $h(P^*, P)$ for $P \in \mathscr{Q}_{j^*}$.

Lemma D.2. *There exists a positive constant $\overline{a}$ such that for all $P_\theta = \sum_{k=1}^{j^*} w_k \mathcal{N}(z_k, \sigma_k^2) + \sum_{k=j^*+1}^{K} w_k \, Cauchy(z_k, \sigma_k) \in \mathscr{Q}_{j^*}$,*

$$h^2(P^*, P_\theta) \geq \overline{a}\Bigg(||w - \overline{w}||^2 + \sum_{k=1}^{j^*} \left|\left|\left(z_k, \sigma_k^2\right) - \left(\overline{z}_k, \overline{\sigma}_k^2\right)\right|\right|_2^2 \wedge 1 + \sum_{k=j^*+1}^{K} ||(z_k, \sigma_k) - (\overline{z}_k, \overline{\sigma}_k)||_2^2 \wedge 1 \Bigg).$$

Finally, there is a constant $\overline{C}$ such that for ξ and n, on the event $\Omega_{\xi,n}$, we have

$$\overline{C}\left(||\hat{w} - \overline{w}||^2 + \sum_{k=1}^{j^*} \left|\left|\left(\hat{z}_k, \hat{\sigma}_k^2\right) - \left(\overline{z}_k, \overline{\sigma}_k^2\right)\right|\right|_2^2 \wedge 1 + \sum_{k=j^*+1}^{K} ||(\hat{z}_k, \hat{\sigma}_k) - (\overline{z}_k, \overline{\sigma}_k)||_2^2 \wedge 1 \right) \leq \frac{K \log(n(K+1)) + \xi}{n}.$$

We still have to prove Lemmas D.1 and D.2.

Proof of Lemma D.1

Let $j \in \{0, \dots, K\}$ and assume there is a sequence

$$(P_n)_n = \left(\sum_{k=1}^{j} w_{k,n} \mathcal{N}(z_{k,n}, \sigma_{k,n}^2) + \sum_{k=j+1}^{K} w_{k,n} \text{Cauchy}(z_{k,n}, \sigma_{k,n}) \right) \in \mathscr{Q}_j^{\mathbb{N}}$$

such that $\lim_{n\to\infty} h(P_n, P^*) = 0$. The mixing weights are bounded so we can assume we are already considering a sequence such that $w_{k,n} \xrightarrow[n\to\infty]{} w_{k,\infty}$ for all $k \in \{1, \dots, K\}$. For the other parameters, it is always possible to extract a subsequence $P_{\psi(n)}$ such that for all k

$$z_{k,\psi(n)} \xrightarrow[n\to\infty]{} \begin{cases} z_{k,\infty} \in \mathbb{R}, \\ \text{or } \pm\infty, \end{cases} \quad \text{and } \sigma_{k,\psi(n)} \xrightarrow[n\to\infty]{} \begin{cases} \sigma_{k,\infty} \in \mathbb{R}^+, \\ \text{or } +\infty. \end{cases}$$

We now consider the different cases possible (dropping the dependency on ψ in the notation).

- If $z_{k,n} \xrightarrow[n\to\infty]{} \pm\infty$ (without loss of generality we consider $+\infty$ in the proof), for $b \in \mathbb{R}$, we have

$$\begin{aligned} P_n([b, +\infty[) &\geq w_{k,n} \left[\mathbb{1}_{k \leq j} \mathcal{N}(z_{k,n}, \sigma_{k,n}^2)([b, +\infty[) \right. \\ &\quad \left. + \mathbb{1}_{k>j} \text{Cauchy}(z_{k,n}, \sigma_{k,n})([b, +\infty[) \right] \\ &\geq \frac{w_{k,n}}{2} \text{ for } n \text{ large enough.} \end{aligned}$$

Assume $w_{k,\infty} > 0$. Since $P^*([b,+\infty[) \xrightarrow[b\to\infty]{} 0$, there exists b such that $P^*([b,+\infty[) \leq w_{j,\infty}/4$. On the other hand we have $P^*([b,+\infty[) = \lim_{n\to\infty} P_{\theta_n}([b,+\infty[) \geq w_{k,\infty}/2$. Therefore, it means that $w_{k,\infty} = 0$ and it also holds for $z_{k,n} \to -\infty$.

- If $z_{k,n} \xrightarrow[n\to\infty]{} z_{k,\infty} \in \mathbb{R}$ and $\sigma_{k,n} \xrightarrow[n\to\infty]{} 0$, for $b > 0$ we have

$$P_n([z_{k,\infty} - b, z_{k,\infty} + b]) \geq w_{k,n} \left(\mathbb{1}_{k\leq j} \mathcal{N}(z_{k,n}, \sigma_{k,n}^2)([b,+\infty[) \right. \\ \left. + \mathbb{1}_{k>j} \text{Cauchy}(z_{k,n}, \sigma_{k,n})([b,+\infty[) \right) \to w_{k,\infty}.$$

Assume $w_{k,\infty} > 0$. Since $P^*([z_{k,\infty} - b, z_{k,\infty} + b]) \xrightarrow[b\to 0]{} 0$, there exists $b > 0$ such that $P^*([z_{k,\infty} - b, z_{k,\infty} + b]) \leq w_{j,\infty}/2$. On the other hand we have $P^*([z_{k,\infty} - b, z_{k,\infty} + b]) = \lim_{n\to\infty} P_n([z_{k,\infty} - b, z_{k,\infty} + b]) \geq w_{k,\infty}$. Therefore, it means that $w_{k,\infty} = 0$.

- If $z_{k,n} \to z_{k,\infty} \in \mathbb{R}$ and $\sigma_{k,n} \to \infty$, for $a > 0$ we have

$$\begin{aligned} P_n([-a,a]) &\leq (1 - w_{k,n}) \\ &\quad + w_{k,n} \left(\mathbb{1}_{k\leq j} \mathcal{N}(z_{k,n}, \sigma_{k,n}^2)([-a,a]) + \mathbb{1}_{k>j} \text{Cauchy}(z_{k,n}, \sigma_{k,n})([-a,a]) \right) \\ &\xrightarrow[n\to\infty]{} (1 - w_{k,\infty}). \end{aligned}$$

Since $P^*([-a,+a]) \xrightarrow[a\to+\infty]{} 1$, we get $w_{k,\infty} = 0$

This proves that P_n converges to

$$P_\infty = \sum_{\substack{k\leq j(\lambda) \\ w_{k,\infty}>0}} w_{k,\infty} \mathcal{N}(z_{k,\infty}, \sigma_{k,\infty}^2) + \sum_{\substack{k>j(\lambda) \\ w_{k,\infty}>0}} w_{k,\infty} \text{Cauchy}(z_{k,\infty}, \sigma_{k,\infty}),$$

and necessarily $P^* = P_\infty$. Lemma D.1 with the assumptions on P^* implies $j = j^*$ and there exist two permutations τ_g, τ_c respectively on $\{1,\dots,j^*\}$ and $\{j^*+1,\dots,K\}$ such that $(\overline{\pi}_k, \overline{z}_k, \overline{\sigma}_k) = (w_{\tau_g(k)}, z_{\tau_g(k)}, \sigma_{\tau_g(k)})$ for k in $\{1,\dots,j^*\}$ and $(\overline{\pi}_k, \overline{z}_k, \overline{\sigma}_k) = (w_{\tau_c(k)}, z_{\tau_c(k)}, \sigma_{\tau_c(k)})$ for k in $\{j^*+1,\dots,K\}$. □

Proof of Lemma D.2

- The map $(z,\sigma^2) \mapsto g(x;z,\sigma^2) = \phi_\sigma(x-z)$ is continuous and differentiable on $\mathbb{R} \times \mathbb{R}^{+*}$ with

$$\partial_z \phi_\sigma(x-z) = \phi_\sigma(x-z) \frac{(x-z)}{\sigma^2}$$
$$\partial_{\sigma^2} \phi_\sigma(x-z) = \phi_\sigma(x-z) \left[\frac{(x-z)^2}{2\sigma^4} - \frac{1}{2\sigma^2} \right].$$

Similarly $(z,\sigma) \mapsto f(x;z,\sigma) = \frac{1}{\pi\sigma} \frac{1}{c(x;z,\sigma)}$ is continuous and differentiable on $\mathbb{R} \times \mathbb{R}^{+*}$ with

$$\partial_z f(x;z,\sigma) = \frac{1}{\pi\sigma^3} \frac{x-z}{c^2(x;z,\sigma)}$$
$$\partial_\sigma f(x;z,\sigma) = \frac{1}{\pi\sigma^2 c(x;z,\sigma)} \left[1 - \frac{2}{c(x;z,\sigma)} \right].$$

Moreover, on can check that we have

$$\int_{\mathbb{R}} \left|\partial_z g(x;z,\sigma^2)\right|^2 \frac{\mathrm{d}x}{g(x;z,\sigma^2)} = \int_{\mathbb{R}} \frac{(x-z)^2}{\sigma^4}\phi_\sigma(x-z)\mathrm{d}x < \infty$$
$$\int_{\mathbb{R}} \left|\partial_{\sigma^2} g(x;z,\sigma^2)\right|^2 \frac{\mathrm{d}x}{g(x;z,\sigma^2)} = \int_{\mathbb{R}} \left[\frac{(x-z)^2}{2\sigma^4} - \frac{1}{2\sigma^2}\right]^2 \phi_\sigma(x-z)\mathrm{d}x < \infty$$
$$\int_{\mathbb{R}} \left|\partial_z f(x;z,\sigma)\right|^2 \frac{\mathrm{d}x}{f(x;z,\sigma)} = \int_{\mathbb{R}} \frac{(x-z)^2}{\pi\sigma^5 c^3(x;z,\sigma)}\mathrm{d}x < \infty$$
$$\int_{\mathbb{R}} \left|\partial_{\sigma^2} f(x;z,\sigma^2)\right|^2 \frac{\mathrm{d}x}{f(x;z,\sigma)} = \int_{\mathbb{R}} \frac{1}{\pi\sigma^3 c(x;z,\sigma)}\left[1 - \frac{2}{c(x;z,\sigma)}\right]^2 \mathrm{d}x < \infty.$$

- The function $\theta \mapsto \psi(\cdot;\theta) = \frac{\partial}{\partial\theta} p^{1/2}(\cdot;\theta)$, where

$$p(x;\theta) = \sum_{k=1}^{j^*} w_k \phi_{\sigma_k}(x - z_k) + \sum_{k=j^*+1}^{K} \frac{1}{\pi\sigma c(x;z,\sigma)}$$

and

$$\theta = (w_1, \ldots, w_{K-1}, z_1, \ldots, z_K, \sigma_1^2, \ldots, \sigma_{j^*}^2, \sigma_{j^*+1}, \ldots, \sigma_K),$$

is continuous in the space $L_2(\mu)$.

- We apply Theorem 1 of Meijer and Ypma [23]. For $j^* < K$,

$$\begin{aligned}
&\det(I(\theta)) = 0 \\
&\Rightarrow \exists \lambda \neq 0, \sum_{k=1}^{j^*} \phi_{\sigma_k}(x - z_k)\left(\frac{w_k \lambda_{z_k}(x - z_k)}{\sigma_k^2} + w_k \lambda_{\sigma_k^2}\left[\frac{(x-z)^2}{2\sigma_k^4} - \frac{1}{2\sigma_k^2}\right] + \lambda_{w_k}\right) \\
&+ \sum_{k=j^*+1}^{K-1}\left(\frac{w_k \lambda_{z_k}(x - z_k)}{\pi\sigma^3 c^2(x;z_k,\sigma_k)} + \frac{w_k \lambda_{\sigma_k}}{\pi\sigma_k^2}\left[\frac{1}{c(x;z_k,\sigma_k)} - \frac{2}{c^2(x;z_k,\sigma_k)}\right] + \frac{\lambda_{w_k}}{\pi\sigma_k c(x;z_k,\sigma_k)}\right) \\
&+ (1 - w_1 - \cdots - w_{K-1})\left(\frac{\lambda_{z_K}(x - z_K)}{\pi\sigma_K^3 c^2(x;z_K,\sigma_K)} + \frac{\lambda_{\sigma_K}}{\pi\sigma_K^2}\left[\frac{1}{c(x;z_K,\sigma_K)} - \frac{2}{c^2(x;z_K,\sigma_K)}\right]\right) \\
&- \frac{1}{\pi\sigma_K c(x;z_K,\sigma_K)} \sum_{k=1}^{K-1} \lambda_{w_k} = 0 \text{ for } \mu\text{-almost all } x.
\end{aligned}$$

For $j^* = K$,

$$\begin{aligned}
&\det(I(\theta)) = 0 \\
&\Rightarrow \exists \lambda \neq 0, \sum_{k=1}^{K-1} \phi_{\sigma_k^2}(x - z_k)\left(w_k \lambda_{z_k}\frac{(x - z_k)}{\sigma_k^2} + w_k \lambda_{\sigma_k^2}\left[\frac{(x-z)^2}{2\sigma_k^4} - \frac{1}{2\sigma_k^2}\right] + \lambda_{w_k}\right) \\
&+ \phi_{\sigma_K}(x - z_K)\left\{(1 - w_1 - \cdots - w_{K-1})\left(\lambda_{z_K}\frac{(x - z_K)}{\sigma_K^2} + \lambda_{\sigma_K^2}\left[\frac{(x-z)^2}{2\sigma_K^4} - \frac{1}{2\sigma_K^2}\right]\right)\right. \\
&\left. - \sum_{k=1}^{K-1} \lambda_{w_k}\right\} = 0 \text{ for } \mu\text{-almost all } x.
\end{aligned}$$

Lemma D.3. *Let $(z_1,\sigma_1),\dots,(z_K,\sigma_K)$ be distinct elements of $\mathbb{R}\times\mathbb{R}^{+*}$. For any integer n, the families*

$$A=\left\{x\mapsto x^j\phi_{\sigma_i}(x-z_i);i\in\{1,\dots,K\},j\in\{0,\dots,n\}\right\}$$

and

$$B=\left\{x\mapsto\frac{x^j}{c^l(x;z_i,\sigma_i)};i\in\{1,\dots,K\},l\in\{1,2\},j\in\{0,1\}\right\}$$

are linearly independent. Moreover, the linear spaces $\boldsymbol{Span}_{\mathbb{R}}(A)$ and $\boldsymbol{Span}_{\mathbb{R}}(B)$ are orthogonal.

This proves that $I(\overline{\theta})$ is non singular.

- We now check $\inf_{\substack{||\overline{\theta}-\theta||\geq a\\P_\theta\in\mathscr{Q}_{j^*}}}h^2(P_{\overline{\theta}},P_\theta)>0,\forall a>0$. It is a direct consequence of Lemma D.1.
- $\mathscr{Q}(\lambda^*)$ is a regular parametric model. We consider the parameter to be σ for the Cauchy distribution and σ^2 for the Gaussian distribution. Obviously, $(z,\sigma)\mapsto g(x;z,\sigma)=\frac{1}{\pi\sigma}\frac{1}{c(x;z,\sigma)}$, with $c(x;z,\sigma)=1+\left(\frac{x-z}{\sigma}\right)^2$ is continuous and differentiable on $\mathbb{R}\times\mathbb{R}^{+*}$ with

$$\begin{aligned}\partial_z g(x;z,\sigma)&=\frac{2(x-z)}{\pi\sigma^3c^2(x;z,\sigma)}\\ \partial_\sigma g(x;z,\sigma)&=\frac{1}{\pi\sigma^2c(x;z,\sigma)}-\frac{2}{\pi\sigma^2c^2(x;z,\sigma)}.\end{aligned}$$

Moreover, on can check that we have

$$\int_{\mathbb{R}}|\partial_z g(x;z,\sigma)|^2\frac{\mathrm{d}x}{g(x;z,\sigma)}=\int_{\mathbb{R}}\frac{4(x-z)^2}{\pi\sigma^3c^3(x;z,\sigma)}\mathrm{d}x<\infty$$

and

$$\int_{\mathbb{R}}|\partial_\sigma g(x;z,\sigma)|^2\frac{\mathrm{d}x}{g(x;z,\sigma)}=\int_{\mathbb{R}}\frac{1}{\pi\sigma^3c(x;z,\sigma)}\left[1-\frac{2}{c(x;z,\sigma)}\right]^2\mathrm{d}x<\infty.$$

- With the results of [17], we get that there is a constant $a^*>0$ such that

$$\forall P_\theta\in\mathscr{Q}(\lambda^*),a^*\frac{||\theta-\overline{\theta}||^2}{1+||\theta-\overline{\theta}||^2}\leq h^2(P^*,P_\theta).$$

□

Proof of Lemma D.3

- Let f be any function in $\mathbf{Span}_{\mathbb{R}}(A)\cap\mathbf{Span}_{\mathbb{R}}(B)$. Therefore there are constants $(\lambda_{g,i,j})_{\substack{1\leq i\leq K,\\0\leq j\leq n}}$ and $(\lambda_{c,i,l,j})_{\substack{1\leq i\leq K,\\0\leq j\leq 1\leq l\leq 2}}$ such that

$$f(x)=\sum_{i=1}^K\sum_{j=0}^n\lambda_{g,i,j}x^j\phi_{\sigma_i}(x-z_i)=\sum_{i=1}^K\sum_{l=1}^2\sum_{j=0}^1\lambda_{c,i,l,j}\frac{x^j}{c^l(x;z_i,\sigma_i)}.$$

Since $f \in \mathbf{Span}_{\mathbb{R}}(A)$, we have $f(x) = o_{\pm\infty}(x^{-k}), \forall k \in \mathbb{N}$. Therefore $\lambda_{c,i,l,j} = 0$ for all i, j, l and $f = 0$. This proves $\mathbf{Span}_{\mathbb{R}}(A) \cap \mathbf{Span}_{\mathbb{R}}(B) = \{0\}$.

- One can check that $>$ is a strict total order such that

$$(z_1, \sigma_1) > (z_2, \sigma_2) \Rightarrow x^j \phi_{\sigma_2}(x - z_2)/\phi_{\sigma_1}(x - z_1) \xrightarrow[x \to +\infty]{} 0,$$

for any $j \in \mathbb{N}$. Let λ be such that $\sum_{i,j} \lambda_{i,j} x^j \phi_{\sigma_i}(x - z_i) = 0$ for all x. Without loss of generality, we assume $(z_1, \sigma_1) > \cdots > (z_K, \sigma_K)$. Therefore,

$$\begin{aligned} 0 &= \sum_{i,j} \lambda_{i,j} x^j \phi_{\sigma_i}(x - z_i) \\ &= \sum_{i,j} \lambda_{i,j} x^j \phi_{\sigma_i}(x - z_i)/\phi_{\sigma_1}(x - z_1) + \sum_j \lambda_{1,j} x^j \\ &= \sum_j \lambda_{1,j} x^j + o_{+\infty}(1). \end{aligned}$$

It implies that $\lambda_{1,j} = 0$ for all j. Then, we have $\sum_{i \geq 2, j} \lambda_{i,j} x^j \phi_{\sigma_i}(x - z_i) = 0$. By induction, we get that $\lambda = 0$ which proves that the family is indeed linearly independent.

- The partial fraction decomposition theorem implies that B is linearly independent. □

This concludes the proof of Theorem 4.2.

D.3 Proof of Theorem 3.9

We apply Theorem 3.1 and Lemma D.2 (see page 446) with $j^* = K$.

Appendix E. Two-component mixture models

This section gathers the proofs of the results for the two-component mixture model with one known component, namely Theorems 3.11 and 3.13.

E.1 Proof of Theorem 3.11

We take $M = ||z^*||_\infty + 1$ to have (E.1). With Proposition 3.10, there exists a positive constant C (depending on ϕ and M) such that for all $z \in [-M, M]^d$, and all $\lambda \in [0, 1]$, we have

$$C(\phi, M)||z^*||^2 \left(||z||^2 (\lambda^* - \lambda)^2 + (\lambda^*)^2 ||z^* - z||^2 \right) \leq ||p_{\lambda^*, z^*} - p_{\lambda, z}||^2.$$

One can prove (using Prop. 2.1 in [13] and $\lambda^* \neq 0$) that we have

$$\inf_{\substack{z \notin [-M,M]^d, \\ \lambda \in [0,1]}} ||p_{\lambda^*, z^*} - p_{\lambda, z}||^2 > 0. \tag{E.1}$$

Therefore, there is a constant $C(\phi, \lambda^*, z^*)$ such that for all $z \in \mathbb{R}^d$ and all $\lambda \in [0, 1]$,

$$C(\phi, \lambda^*, z^*) \left(\left(||z||^2 \wedge 1 \right) (\lambda^* - \lambda)^2 + (\lambda^*)^2 \left(||z^* - z||^2 \wedge 1 \right) \right) \leq ||p_{\lambda^*, z^*} - p_{\lambda, z}||_2^2.$$

Since ϕ is bounded, with inequality (3.20), there is another constant $C(\phi, \lambda^*, z^*)$ such that for all $z \in \mathbb{R}^d$ and $\lambda \in [0,1]$ we have

$$C(\phi,\lambda^*,z^*)\left(\left(||z||^2 \wedge 1\right)(\lambda^*-\lambda)^2 + (\lambda^*)^2\left(||z^*-z||^2 \wedge 1\right)\right) \leq h^2\left(P_{\lambda^*,z^*}, P_{\lambda,z}\right).$$

One can check the following

$$h^2(P_{\lambda^*,z^*}, P_{\hat{\lambda},\hat{z}}) \leq C(\phi,\lambda^*,z^*)(\lambda^*)^2\left(||z^*||^2 \wedge 1\right)/2 \Rightarrow ||z^*-\hat{z}||^2 \wedge 1 \leq (||z^*||^2 \wedge 1)/4$$
$$\Rightarrow ||\hat{z}|| \wedge 1 \geq \frac{||z^*||}{2} \wedge 1.$$

We use Theorem 3.1 for an upper bound on $h^2(P_{\lambda^*,z^*}, P_{\hat{\lambda},\hat{z}})$. For $n \geq n_0(\phi,\lambda^*,z^*)$, with

$$n_0(\phi,\lambda^*,z^*) := \inf\left\{n \geq 1+V \middle| \frac{4(1+V)[1+\log(2n/(1+V))]}{nC(\lambda^*)^2\left(||z^*||^2 \wedge 1\right)} \leq C(\phi,\lambda^*,z^*)\right\},$$

for $0 < \xi \leq \xi_n = (1+V)[1+\log(2n/(1+V))]$, with probability at least $1-e^{-\xi}$ we have

$$Ch^2\left(P_{\lambda^*,z^*}, P_{\hat{\lambda},\hat{z}}\right) \leq \frac{1}{n}\left\{(1+V)\left[1+\log\left(\frac{2n}{(V+1)}\right)\right]+\xi\right\}$$
$$\leq C \times C(\phi,\lambda^*,z^*)(\lambda^*)^2\left(||z^*||^2 \wedge 1\right)/2,$$

where C is the constant given in Theorem 3.1. Therefore, there is a new constant $C(\phi,\lambda^*,z^*)$ such that for $n \geq n_0$ and $\xi \in (0,\xi_n)$, with probability at least $1-e^{-\xi}$ we have

$$C(\phi,\lambda^*,z^*)\left((\lambda^*-\lambda)^2 + \left(||z^*-z||^2 \wedge 1\right)\right) \leq \frac{(1+V)\left[1+\log(2n/(1+V))\right]+\xi}{n}.$$

E.2 Proof of Theorem 3.13

We need some preliminary results before applying Theorem 3.1.

Proposition E.1. *For $\lambda^* \in (0,1]$ and $z^* \neq 0$, there is a positive constant $C(\alpha,\lambda^*,z^*)$ such that for all $z \in \mathbb{R}$ and all $\lambda \in [0,1]$, we have*

$$h^2\left(P_{\lambda^*,z^*}, P_{\lambda,z}\right) \geq C(\alpha,z^*,\lambda^*)\left[(\lambda^*)^{1/\alpha}\left(1 \wedge |z-z^*|^{1-\alpha}\right) + (\lambda^*-\lambda)^2\left(1 \wedge |z^*|\right)\right].$$

Since s_α is unimodal, the class of densities $\{x \mapsto s_\alpha(x-z), z \in \mathbb{R}\}$ is VC-subgraph with VC-dimension not larger than 10 (see Sect. 3.2). With Theorem 3.1 and Proposition E.1, there exists a positive constant $C(\alpha,\lambda^*,z^*)$ such that for all $\xi > 0$, we have

$$C(\alpha,z^*,\lambda^*)\left[1 \wedge |\hat{z}-z^*|^{1-\alpha} + \left(\lambda^*-\hat{\lambda}\right)^2\right] \leq \frac{\log(n)+\xi}{n},$$

with probability at least $1-e^{-\xi}$.

Proof of Proposition E.1

We write

$$f_z(x) = s_\alpha(x-z) = \frac{1-\alpha}{2|x-z|^\alpha}\mathbb{1}_{|x-z|\in(0,1]}.$$

We define g by

$$g(x) = \frac{2}{1-\alpha}\left(\sqrt{(1-\lambda^*)f_0(x)+\lambda^* f_{z^*}(x)} - \sqrt{(1-\lambda)f_0(x)+\lambda f_z(x)}\right)^2$$

such that

$$2h^2\left(P_{\lambda^*,z^*}, P_{\lambda,z}\right) = \frac{1-\alpha}{2}\int_{-\infty}^{+\infty} g(x)\mathrm{d}x.$$

Lemma E.2. *Assuming $z\cdot z^* > 0$ and $|z^*-z| \le \frac{1}{(1-\alpha)^{2/\alpha}}$. There exists $C(\alpha, z^*, \lambda^*) > 0$ such that*

$$\int g(x)\mathrm{d}x \ge C(\alpha, z^*, \lambda^*)\left[(\lambda^*)^{1/\alpha}\left(1\wedge|z-z^*|^{1-\alpha}\right) + (\lambda^*-\lambda)^2\left(1\wedge|z^*|\right)\right].$$

Lemma E.3. *For $z\cdot z^* \le 0$, we have*

$$\int g(x)\mathrm{d}x \ge \lambda^*\alpha^2\frac{1\bigwedge\left[(\lambda^*)^{(1-\alpha)/\alpha}(1-\alpha)^{2(1-\alpha)/\alpha}|z^*|^{1-\alpha}\right]}{1-\alpha}.$$

Lemma E.4. *For $|z-z^*| > \frac{1}{(1-\alpha)^{2/\alpha}}$ and $z^*\cdot z > 0$, we have*

$$\int g(x)\mathrm{d}x = \lambda^*(1\wedge|z^*|).$$

Combining those three lemmas, there exists a positive constant $C(\alpha, z^*, \lambda^*)$ such that

$$h^2\left(P_{\lambda^*,z^*}, P_{\lambda,z}\right) \ge C'(\alpha, z^*, \lambda^*)\left[(\lambda^*)^{1/\alpha}\left(1\wedge|z-z^*|^{1-\alpha}\right) + (\lambda^*-\lambda)^2\left(1\wedge|z^*|\right)\right],$$

for all λ in $[0,1]$ and z in $\mathbb{R}$. Without loss of generality, we assume $z^* > 0$ through the proof of the lemmas. □

Proof of Lemma E.2

Without loss of generality, we consider $z^* > 0$ for now.

• For $x \in]-1, 0[$, we have

$$g(x) = \frac{1}{|x|^\alpha}\left(\sqrt{1-\lambda^*+\lambda^*\frac{|x|^\alpha}{|x-z^*|^\alpha}\mathbb{1}_{|x-z^*|\in(0,1]}} - \sqrt{1-\lambda+\lambda\frac{|x|^\alpha}{|x-z|^\alpha}\mathbb{1}_{|x-z|\in(0,1]}}\right)^2.$$

If $z^*\wedge z \ge 1$ then,

$$g(x) = \frac{1}{|x|^\alpha}\left(\sqrt{1-\lambda^*} - \sqrt{1-\lambda}\right)^2$$

and

$$\int_{-1}^{0} g(x)\mathrm{d}x \geq \left(\sqrt{1-\lambda^*} - \sqrt{1-\lambda}\right)^2 \frac{1}{1-\alpha}.$$

Otherwise $z^* \wedge z \in (0,1)$ then for $x \in]-1, z^* \wedge z - 1[$,

$$\int_{-1}^{z^* \wedge z - 1} g(x)\mathrm{d}x \geq \left(\sqrt{1-\lambda^*} - \sqrt{1-\lambda}\right)^2 \frac{1-(1-z\wedge z^*)^{1-\alpha}}{1-\alpha}.$$

Finally,

$$\int_{-1}^{0} g(x)\mathrm{d}x \geq \left(\sqrt{1-\lambda^*} - \sqrt{1-\lambda}\right)^2 \frac{1-(1-z\wedge z^*)_+^{1-\alpha}}{1-\alpha}.$$

• For $x \in]z^* \vee z, z^* \vee z + 1[$, we have

$$g(x) = \frac{1}{|x - z^* \vee z|^\alpha} \left(\sqrt{(1-\lambda^*)\frac{|x - z^* \vee z|^\alpha}{|x|^\alpha} \mathbb{1}_{|x|\in(0,1]} + \lambda^* \frac{|x - z^* \vee z|^\alpha}{|x - z^*|^\alpha} \mathbb{1}_{|x-z^*|\in(0,1]}} \right.$$
$$\left. - \sqrt{(1-\lambda)\frac{|x - z^* \vee z|^\alpha}{|x|^\alpha} \mathbb{1}_{|x|\in(0,1]} + \lambda \frac{|x - z^* \vee z|^\alpha}{|x - z|^\alpha} \mathbb{1}_{|x-z|\in(0,1]}} \right)^2.$$

- If $z < z^*$, with $V < \frac{1}{|z-z^*|}$, for $x \in]z^*, z^* + V|z - z^*|[$, we have

$$\bullet \frac{|x - z^*|}{|x|} \leq V \frac{|z^* - z|}{z^*} \leq V,$$
$$\bullet \frac{|x - z^*|}{|x - z|} \leq \frac{V|z^* - z|}{(1+V)|z^* - z|} \leq V.$$

We get

$$\int_{z^*}^{z^* + V|z-z^*|} g(x)\mathrm{d}x \geq \left(\sqrt{\lambda^*} - \sqrt{V^\alpha}\right)^2 \int_{z^*}^{z^* + V|z-z^*|} \frac{\mathrm{d}x}{|x - z^* \vee z|^\alpha}$$
$$= \left(\sqrt{\lambda^*} - \sqrt{V^\alpha}\right)^2 \frac{(V|z^* - z|)^{1-\alpha}}{1-\alpha}.$$

We take $V = (\lambda^*)^{1/\alpha}(1-\alpha)^{2/\alpha} \leq \frac{(\lambda^*)^{1/\alpha}}{|z^*-z|} \leq \frac{1}{|z^*-z|}$, and we have

$$\int_{z^*}^{z^* + V|z-z^*|} g(x)\mathrm{d}x \geq \lambda^* \alpha^2 \frac{(\lambda^*)^{(1-\alpha)/\alpha}(1-\alpha)^{2(1-\alpha)/\alpha}|z^* - z|^{1-\alpha}}{1-\alpha}$$
$$= \frac{(\lambda^*)^{1/\alpha}\alpha^2(1-\alpha)^{2(1-\alpha)/\alpha}|z^* - z|^{1-\alpha}}{1-\alpha}.$$

- If $z \geq z^*$, we obtain the same way

$$\int_z^{z+1} g(x)\mathrm{d}x \geq \frac{\lambda^{1/\alpha}\alpha^2(1-\alpha)^{2(1-\alpha)/\alpha}|z^*-z|^{1-\alpha}}{1-\alpha}.$$

Finally, for any z^* in $\mathbb{R}$, using the following inequalities

$$\forall x, y \in [0,1], 1-(1-|x|)_+^{1-\alpha} \geq (1-\alpha)(1\wedge|x|) \text{ and } \left(\sqrt{x}-\sqrt{y}\right)^2 \geq (x-y)^2/4, \tag{E.2}$$

we get

$$\begin{aligned}\int g(x)\mathrm{d}x \geq\ & \mathbb{1}_{|z|\geq|z^*|}\left[\frac{(\lambda)^{1/\alpha}\alpha^2(1-\alpha)^{2(1-\alpha)/\alpha}|z^*-z|^{1-\alpha}}{1-\alpha}+(\lambda^*-\lambda)^2(1\wedge|z^*|)\right]\\ & \mathbb{1}_{|z|<|z^*|}\left[\frac{(\lambda^*)^{1/\alpha}\alpha^2(1-\alpha)^{2(1-\alpha)/\alpha}|z^*-z|^{1-\alpha}}{1-\alpha}+(\lambda^*-\lambda)^2(1\wedge|z|)\right].\end{aligned}$$

- If $|z| \geq |z^*|$:
  - if $\lambda > c\lambda^*$, then

$$\begin{aligned}\int g(x)\mathrm{d}x &\geq \frac{(\lambda^*)^{1/\alpha}c^{1/\alpha}\alpha^2(1-\alpha)^{2(1-\alpha)/\alpha}|z^*-z|^{1-\alpha}}{1-\alpha}+(\lambda^*-\lambda)^2(1\wedge|z^*|)\\ &\geq C_1(\alpha,c)\left[(\lambda^*)^{1/\alpha}|z^*-z|^{1-\alpha}+(1\wedge|z^*|)(\lambda^*-\lambda)^2\right]\end{aligned}$$

    with $C_1(\alpha,c) = 1\bigwedge \frac{c^{1/\alpha}\alpha^2(1-\alpha)^{2(1-\alpha)/\alpha}}{1-\alpha}$;
  - otherwise $\int g(x)\mathrm{d}x \geq (\lambda^*)^2(1-c)^2(1\wedge|z^*|)$,

$$(\lambda^*)^{1/\alpha}|z^*-z|^{1-\alpha}+(1\wedge|z^*|)(\lambda^*-\lambda)^2 \leq (\lambda^*)^{1/\alpha}\frac{1}{(1-\alpha)^{2(1-\alpha)/\alpha}}+(1\wedge|z^*|)$$

    and finally

$$\begin{aligned}\int g(x)\mathrm{d}x \geq & \frac{(\lambda^*)^2(1-c)^2(1\wedge|z^*|)}{(\lambda^*)^{1/\alpha}\frac{1}{(1-\alpha)^{2(1-\alpha)/\alpha}}+(1\wedge|z^*|)}\\ & \times\left[(\lambda^*)^{1/\alpha}|z^*-z|^{1-\alpha}+(1\wedge|z^*|)(\lambda^*-\lambda)^2\right].\end{aligned}$$

- If $|z| < |z^*|$:
  - if $|z| \geq d|z^*|$, then

$$\begin{aligned}\int g(x)\mathrm{d}x &\geq \frac{(\lambda^*)^{1/\alpha}\alpha^2(1-\alpha)^{2(1-\alpha)/\alpha}|z^*-z|^{1-\alpha}}{1-\alpha}+(\lambda^*-\lambda)^2 d(1\wedge|z^*|)\\ &\geq C_2(\alpha,d)\left[(\lambda^*)^{1/\alpha}|z-z^*|^{1-\alpha}+(\lambda^*-\lambda)^2(1\wedge|z^*|)\right],\end{aligned}$$

    with $C_2(\alpha,d) = d\wedge\frac{\alpha^2(1-\alpha)^{2(1-\alpha)/\alpha}}{1-\alpha}$;

- otherwise $\int g(x)\mathrm{d}x \geq \frac{(\lambda^*)^{1/\alpha}\alpha^2(1-\alpha)^{2(1-\alpha)/\alpha}|z^*|^{1-\alpha}(1-d)^{1-\alpha}}{1-\alpha}$ and

$$(\lambda^*)^{1/\alpha}|z-z^*|^{1-\alpha} + (\lambda^*-\lambda)^2(1\wedge|z^*|) \leq (\lambda^*)^{1/\alpha}\frac{1}{(1-\alpha)^{2(1-\alpha)/\alpha}} + (1\wedge|z^*|)$$

and finally

$$\int g(x)\mathrm{d}x \geq \frac{(\lambda^*)^{1/\alpha}\alpha^2(1-\alpha)^{2(1-\alpha)/\alpha}|z^*|^{1-\alpha}(1-d)^{1-\alpha}/(1-\alpha)}{(\lambda^*)^{1/\alpha}\frac{1}{(1-\alpha)^{2(1-\alpha)/\alpha}} + (1\wedge|z^*|)} \times \left[(\lambda^*)^{1/\alpha}|z^*-z|^{1-\alpha} + (1\wedge|z^*|)(\lambda^*-\lambda)^2\right].$$

Finally,

$$\int g(x)\mathrm{d}x \geq C(\alpha,z^*,\lambda^*)\left[(\lambda^*)^{1/\alpha}\left(1\wedge|z-z^*|^{1-\alpha}\right) + (\lambda^*-\lambda)^2(1\wedge|z^*|)\right],$$

with

$$\begin{aligned} C(\alpha,z^*,\lambda^*) = \min\Bigg(&1, \frac{c^{1/\alpha}\alpha^2(1-\alpha)^{2(1-\alpha)/\alpha}}{1-\alpha}\\ &\frac{(\lambda^*)^2(1-c)^2(1\wedge|z^*|)}{(\lambda^*)^{1/\alpha}\frac{1}{(1-\alpha)^{2(1-\alpha)/\alpha}} + (1\wedge|z^*|)},\\ &d, \frac{\alpha^2(1-\alpha)^{2(1-\alpha)/\alpha}}{1-\alpha},\\ &\frac{(\lambda^*)^{1/\alpha}\alpha^2(1-\alpha)^{2(1-\alpha)/\alpha}|z^*|^{1-\alpha}(1-d)^{1-\alpha}/(1-\alpha)}{(\lambda^*)^{1/\alpha}\frac{1}{(1-\alpha)^{2(1-\alpha)/\alpha}} + (1\wedge|z^*|)}\Bigg)\\ = \min\Bigg(&1, \frac{c^{1/\alpha}\alpha^2(1-\alpha)^{2(1-\alpha)/\alpha}}{1-\alpha}, d,\\ &\frac{(\lambda^*)^2(1-c)^2(1\wedge|z^*|)}{(\lambda^*)^{1/\alpha}\frac{1}{(1-\alpha)^{2(1-\alpha)/\alpha}} + (1\wedge|z^*|)},\\ &\frac{(\lambda^*)^{1/\alpha}\alpha^2(1-\alpha)^{2(1-\alpha)/\alpha}|z^*|^{1-\alpha}(1-d)^{1-\alpha}/(1-\alpha)}{(\lambda^*)^{1/\alpha}\frac{1}{(1-\alpha)^{2(1-\alpha)/\alpha}} + (1\wedge|z^*|)}\Bigg). \end{aligned}$$

□

Proof of Lemma E.3

Without loss of generality, we take $z^* > 0$.

- For $x \in]z^*, z^*(1+a)[$, $a < (z^*)^{-1}$ we have

$$g(x) = \frac{1}{|x-z^*|^\alpha}\left(\sqrt{(1-\lambda^*)\frac{|x-z^*|^\alpha}{|x|^\alpha}\mathbb{1}_{|x|\in(0,1]} + \lambda^*} - \sqrt{(1-\lambda)\frac{|x-z^*|^\alpha}{|x|^\alpha}\mathbb{1}_{|x|\in(0,1]} + \lambda\frac{|x-z^*|^\alpha}{|x-z|^\alpha}\mathbb{1}_{|x-z|\in(0,1]}}\right)^2.$$

and

$$\frac{|x-z^*|}{|x-z|} \leq \frac{|x-z^*|}{|x|} \leq \frac{a}{1+a} \leq a.$$

We get

$$\begin{aligned}\int_{z^*}^{z^*+a} g(x)\mathrm{d}x &\geq \left(\sqrt{\lambda^*} - \sqrt{a^\alpha}\right)^2 \int_{z^*}^{z^*+a} \frac{\mathrm{d}x}{|x-z^*|^\alpha} \\ &= \left(\sqrt{\lambda^*} - \sqrt{a^\alpha}\right)^2 \frac{(az^*)^{1-\alpha}}{1-\alpha}.\end{aligned}$$

We take $a = (\lambda^*)^{1/\alpha}(1-\alpha)^{2/\alpha} \leq \frac{1}{z^*}$, and we have

$$\int_{z^*}^{z^*+a} g(x)\mathrm{d}x \geq \lambda^* \alpha^2 \frac{(\lambda^*)^{(1-\alpha)/\alpha}(1-\alpha)^{2(1-\alpha)/\alpha}(z^*)^{1-\alpha}}{1-\alpha}.$$

Otherwise $a = 1/z^* \leq (\lambda^*)^{1/\alpha}(1-\alpha)^{2/\alpha}$ and

$$\int_{z^*}^{z^*+a} g(x)\mathrm{d}x \geq \lambda^* \alpha^2 \frac{1}{1-\alpha}.$$

Finally,

$$\int_{z^*}^{z^*+1} g(x)\mathrm{d}x \geq \lambda^* \alpha^2 \frac{1 \bigwedge \left[(\lambda^*)^{(1-\alpha)/\alpha}(1-\alpha)^{2(1-\alpha)/\alpha}(z^*)^{1-\alpha}\right]}{1-\alpha}. \quad \square$$

Proof of Lemma E.4

Without loss of generality, we take $z^* \geq 0$.

- If $z \geq z^* + \frac{1}{(1-\alpha)^{2/\alpha}}$. For $x \in]z^* \vee 1, (z^*+1) \wedge (z-1)[$, we have

$$g(x) = \frac{\lambda^*}{|x-z^*|^\alpha}.$$

  One can prove that

$$|z-z^*| - 1 \geq \frac{1}{(1-\alpha)^{2/\alpha}} - 1 \geq 1.$$

  - If $z^* \geq 1$, then We get

$$\begin{aligned}\int_{z^*}^{z^*+1} g(x)\mathrm{d}x &\geq \frac{\lambda^*}{1-\alpha}\left[1 \bigwedge |z-z^*| - 1\right]^{1-\alpha} \\ &\geq \frac{\lambda^*}{1-\alpha}.\end{aligned}$$

◦ If $z^* \leq 1$, then

$$\int_1^{(z^*+1)\wedge(z-1)} g(x)\mathrm{d}x \geq \frac{\lambda^*}{1-\alpha}\left[1 \bigwedge (|z-z^*|-1)^{1-\alpha} - (1-z^*)^{1-\alpha}\right]$$
$$\geq \frac{\lambda^*}{1-\alpha}\left[1-(1-z^*)^{1-\alpha}\right].$$

- If $z^* \geq z + \frac{1}{(1-\alpha)^{2/\alpha}}$, we get

$$\int_{z^*}^{z^*+1} g(x)\mathrm{d}x = \frac{\lambda^*}{1-\alpha}.$$

Finally,

$$\int_{z^*}^{z^*+1} g(x)\mathrm{d}x = \frac{\lambda^*}{1-\alpha}\left[1-(1-z^*)_+^{1-\alpha}\right] \geq \lambda^*(1\wedge z^*).$$

□

Appendix F. VC-subgraph classes of functions

For more detailed introductions to VC-subgraph classes we refer the reader to Van der Vaart and Wellner [28] (Sect. 2.6.5) and Baraud *et al.* [4] (Sect. 8).

Definition F.1. Definition 41 [4]
Let $\mathscr{C}$ be a non-empty class of subsets of a set Ξ. If $A \subset \Xi$ with $|A| = n$, then

$$\Delta_n(\mathscr{C}, A) = |\{A \cap B, B \in \mathscr{C}\}| \text{ and } \Delta_n(\mathscr{C}) = \max_{A\subset\Xi, |A|=n} \Delta_n(\mathscr{C}, A).$$

If $V = \sup\{n \in N | \Delta_n(\mathscr{C}) = 2n\} < +\infty$, then $\mathscr{C}$ is a VC-class with VC-dimension V and VC-index $\overline{V} = \inf\{n \in N | \Delta_n(\mathscr{C}) < 2n\} = V + 1$. A class $\mathscr{F}$ of functions from a set $\mathscr{X}$ with values in $(-\infty, +\infty]$ is VC-subgraph with dimension V and index V if the class of subgraphs $\{(x,u) \in \mathscr{X} \times \mathbb{R}, f(x) > u\}$ as f varies among $\mathscr{F}$ is a VC-class of sets in $\mathscr{X} \times \mathbb{R}$ with dimension V and index $\overline{V}$.

It immediately follows from this definition the following:

- if $\mathscr{F}$ is VC-subgraph with dimension V, then any subset $\mathscr{G} \subset \mathscr{F}$ is VC-subgraph with dimension at most V,
- if $\mathscr{F}$ is a finite set, $\mathscr{F}$ is VC-subgraph and its dimension is not larger than $V = \log_2(|\mathscr{F}|) \vee 1$.

The main reason for using VC-subgraph theory is the uniform entropy property. Namely, if $\mathscr{F}$ is a VC-subgraph set of measurable functions on $(\mathscr{X}, \mathcal{X})$ with VC-dimension V and $||f||_\infty \leq 1$ for all $f \in \mathscr{F}$, it follows from Lemma 1 in Baraud and Chen [5] that, for any probability P on $(\mathscr{X}, \mathcal{X})$ we have

$$N(\epsilon, \mathscr{F}, L_r(P)) \leq e(V+1)(2e)^V \left(\frac{2}{\epsilon}\right)^{rV}.$$

F.1 Proof of Lemma 2.2

Let $\mathrm{Cov}_{+*}(d)$ be the set of $d \times d$ symmetric and positive-definite matrices. The normal distributions on $\mathbb{R}^d$ with mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathrm{Cov}_{+*}(d)$ admits $g_{\mu,\Sigma}$, defined by

$$g_{\mu,\Sigma} : x \mapsto \frac{\exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)}{\sqrt{(2\pi)^k |\Sigma|}},$$

as a density with respect to the Lebesgue measure on $\mathbb{R}^d$, where $|\Sigma|$ denotes the determinant of $|\Sigma|$. We have

$$\begin{aligned}\log(g_{\mu,\Sigma}(x)) &= -\frac{1}{2}\log\left((2\pi)^k |\Sigma|\right) - \frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu) \\ &= -\frac{1}{2}\log\left((2\pi)^k |\Sigma|\right) - \frac{1}{2}\mu^T\Sigma^{-1}\mu - \mu^T\Sigma^{-1}x - \frac{1}{2}x^T\Sigma x.\end{aligned}$$

For the location-scale family $\mathcal{G}_d := \left\{g_{\mu;\Sigma}; \mu \in \mathbb{R}^d, \Sigma \in \mathrm{Cov}_{+*}\right\}$, we have

$$\mathcal{G}_d \subset \exp \circ \left\{ x \mapsto a + \sum_{i\leq j} b_{i,j}x_i x_j + \sum_{i=1}^d c_i x_i ; a \in \mathbb{R}, (b_{ij})_{i\leq j} \in \mathbb{R}^{d(d+1)/2}, c \in \mathbb{R}^d \right\}.$$

Since $\left\{ x \mapsto a + \sum_{i\leq j} b_{i,j}x_i x_j + \sum_{i=1}^d c_i x_i ; a \in \mathbb{R}, (b_{ij}) \in \mathbb{R}^{d(d+1)/2}, c \in \mathbb{R}^d \right\}$ is a vector space of dimension $1 + d(d+3)/2$ and exp is monotone, we get that $V(\mathscr{G}_d) \leq 3 + \frac{d(d+3)}{2}$. For $\Sigma \in \mathrm{Cov}_{+*}(d)$ fixed, the location family $\mathcal{G}_{loc}(\Sigma) := \left\{g_{\mu;\Sigma}; \mu \in \mathbb{R}^d\right\}$, we have

$$\mathcal{G}_{loc}(\Sigma) \subset \exp \circ \left(x \mapsto -\frac{x^T\Sigma x}{2} + \left\{ x \mapsto a + \sum_{i=1}^d b_i x_i ; a \in \mathbb{R}, b \in \mathbb{R}^d \right\} \right).$$

With similar arguments and the fact that $x \mapsto -\frac{x^T\Sigma x}{2}$ is a fixed function, we have $V\left(\mathcal{G}_{loc}(\Sigma)\right) \leq 3 + d$.

F.2 Proof of Lemma 3.12

The different arguments used in this proof are from Proposition 42 of Baraud *et al.* [4] and Lemmas 2.6.15 and 2.6.16 from van der Vaart and Wellner [28]. We remind the reader that the VC-index is the VC-dimension plus 1.

- For the Cauchy location-scale family, we have

$$\mathcal{C} = \square^{-1} \circ \left\{ x \mapsto \pi\sigma\left[1 + \left(\frac{x-z}{\sigma}\right)^2\right] ; \sigma > 0, z \in \mathbb{R} \right\},$$

  where $\square^{-1}$ is the inverse function on $(0, +\infty)$. Since

$$\left\{ x \mapsto \pi\sigma\left[1 + \left(\frac{x-z}{\sigma}\right)^2\right] ; \sigma > 0, z \in \mathbb{R} \right\} \subset \mathbb{R}_2[x] = \left\{ x \mapsto ax^2 + bx + c ; (a,b,c) \in \mathbb{R}^3 \right\}$$

  and $\square^{-1}$ is monotone, we get that $V(\mathcal{C}) \leq 3 + 2$.

- For univariate normal distribution, it is a direct consequence of Lemma 2.2.
- We have

$$\begin{aligned}\mathcal{L} &= \left\{ x \mapsto \frac{1}{2b} \exp\left(-\frac{|x-z|}{b}\right); z \in \mathbb{R}, b > 0 \right\} \\ &= \exp \circ \left\{ x \mapsto -\log(2b) + b^{-1}[(x-z) \wedge (z-x)]; z \in \mathbb{R}, b > 0 \right\} \\ &\subset \exp \circ \left(\{ x \mapsto ax + b; a, b \in \mathbb{R} \} \wedge \{ x \mapsto ax + b; a, b \in \mathbb{R} \} \right).\end{aligned}$$

  Since exp is monotone and $\{x \mapsto ax + b; a, b \in \mathbb{R}\}$ is a vector space of dimension 2, we get that $\mathcal{L}$ is VC-subgraph with VC-index not larger than $V(\mathcal{L}) \leq 4.701 \times 2(2+1) + 1 = 29.206$.
- Azzalini and Capitanio [2] proved that the probability density function of the skew-normal distribution is unimodal, therefore the translation family $\mathcal{SG}_\alpha$ is VC-subgraph with VC-index at most 10 (see Sect. 3.2).

Acknowledgements. The author would like to thank Yannick Baraud for his guidance in the redaction of this article.

References

[1] E.S. Allman, C. Matias and J.A. Rhodes, Identifiability of parameters in latent structure models with many observed variables. *Ann. Stat.* **37** (2009) 3099–3132.
[2] A. Azzalini and A. Capitanio, The Skew-Normal and Related Families. Institute of Mathematical Statistics Monographs, Cambridge University Press (2013).
[3] Y. Baraud and L. Birgé, Rho-estimators revisited: General theory and applications. *Ann. Stat.* **46** (2018) 3767–3804.
[4] Y. Baraud, L. Birgé and M. Sart, A new method for estimation and model selection: rho-estimation. *Invent. Math.* **207** (2017) 425–517.
[5] Y. Baraud and J. Chen, Robust estimation of a regression function in exponential families (2020).
[6] L. Birgé, Approximation dans les espaces métriques et théorie de l'estimation. *Zeitsch. Wahrscheinlichkeitstheorie Verwand. Gebiete* **65** (1983).
[7] L. Birgé, On estimating a density using Hellinger distance and some other strange facts. *Prob. Theory Related Fields* **71** (1986).
[8] I. Diakonikolas, D.M. Kane and A. Stewart, List-decodable robust mean estimation and learning mixtures of spherical Gaussians, in *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018*. Association for Computing Machinery, New York, NY, USA (2018) pp. 1047–1060.
[9] C.R. Doss and J.A. Wellner, Global rates of convergence of the MLEs of log-concave and s-concave densities. *Ann. Stat.* **44** (2016) 954–981.
[10] N. Doss, Y. Wu, P. Yang and H.H. Zhou, Optimal estimation of high-dimensional location Gaussian mixtures (2020).
[11] B. Everitt and D.J. Hand, Finite mixture distributions. Chapman and Hall London; New York (1981).
[12] S. Frühwirth-Schnatter, Finite Mixture and Markov Switching Models. Springer Series in Statistics, Springer New York (2006).
[13] S. Gadat, C. Marteau and C. Maugis-Rabusseau, Parameter recovery in two-component contamination mixtures: The $\mathbb{L}^2$ strategy. *Ann. l'Institut Henri Poincaré, Prob. Stat.* **56** (2020) 1391–1418.
[14] C. Genovese and L. Wasserman, Convergence rates for the Gaussian mixture sieve, *Ann. Stat.* **28** (2000) 10.1214/aos/1015956709.
[15] 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** (2001) 1233–1263.
[16] P. Heinrich and J. Kahn, Strong identifiability and optimal minimax rates for finite mixture estimation. *Ann. Stat.* **46** (2018) 2844–2870.
[17] I.A. Ibragimov and H.R. Z., Statistical Estimation. Springer, New York (1981).
[18] W. Kruijer, J. Rousseau and A. van der Vaart, Adaptive Bayesian density estimation with location-scale mixtures. *Electr. J. Stat.* **4** (2010) 1225–1257.
[19] P. Massart, Concentration Inequalities and Model Selection. Vol. 1896 of *Lect. Notes Math.*. Springer, Berlin, Heidelberg (2007).
[20] C. Maugis and B. Michel, A non asymptotic penalized criterion for Gaussian mixture model selection. *ESAIM: PS* **15** (2011) 41–68.
[21] C. Maugis-Rabusseau and B. Michel, Adaptive density estimation for clustering with Gaussian mixtures. *ESAIM: PS* **17** (2013) 698–724.
[22] G. McLachlan and D. Peel, Finite mixture models. Vol. 44 of *Wiley Series in Probability and Statistics*. Wiley (2000).
[23] E. Meijer and J.Y. Ypma, A simple identification proof for a mixture of two univariate normal distributions. *J. Classif.* **25** (2008) 113–123.
[24] R.T. Rockafellar, Convex Analysis. Princeton University Press (2015).

[25] T. Sapatinas, Identifiability of mixtures of power-series distributions and related characterizations. *Ann. Inst. Stat. Math.* **47** (1995) 447–459.
[26] H. Teicher, Identifiability of mixtures. *Ann. Math. Stat.* **32** (1961) 244–248.
[27] D. Titterington, A. Smith and U. Makov, Statistical Analysis of Finite Mixture Distributions, Applied section. Wiley (1985).
[28] A.W. van der Vaart and J.A. Wellner, Weak Convergence and Empirical Processes. Springer, New York (1996).
[29] Y. Wu and P. Yang, Optimal estimation of Gaussian mixtures via denoised method of moments. *Ann. Stat.* **48** (2020) 1981–2007.