Bayesian Kantorovich Deconvolution in Finite Mixture Models

Catia Scricciolo

Abstract This chapter addresses the problem of recovering the mixing distribution in finite kernel mixture models, when the number of components is unknown, yet bounded above by a fixed number. Taking a step back to the historical development of the analysis of this problem within the Bayesian paradigm and making use of the current methodology for the study of the posterior concentration phenomenon, we show that, for general prior laws supported over the space of mixing distributions with at most a fixed number of components, under replicated observations from the mixed density, the mixing distribution is estimable in the Kantorovich or L^1-Wasserstein metric at the optimal pointwise rate $n^{-1/4}$ (up to a logarithmic factor), n being the sample size.

Keywords Dirichlet distribution · Kantorovich metric · Kolmogorov metric · Mixing distribution · Mixture model · Posterior distribution · Rate of convergence · Sieve prior · Wasserstein metric

1 Introduction

The Bayesian analysis of the problem of recovering the unknown mixing distribution in mixture models has recently attracted much attention and stimulated an active discussion encouraging new ideas. Several papers–including [Efron [4], Gao and van der Vaart [5], Heinrich and Kahn [9], Ishwaran et al. [11], Nguyen [14], Scricciolo [18]]–have been devoted to the investigation of this topic, with extensive comparisons with the frequentist solutions. In order to introduce the problem, suppose that $x \mapsto k(x \mid y)$ is a probability density for every $y \in \mathscr{Y} \subseteq \mathbb{R}$, where $(\mathscr{Y}, \mathscr{B})$ is a measurable space. If the mapping $(x, y) \mapsto k(x \mid y)$ is jointly measurable, then

C. Scricciolo (✉)
Dipartimento di Scienze Economiche, Università degli Studi di Verona, Via Cantarane 24, 37129 Verona (VR), Italy
e-mail: catia.scricciolo@univr.it

A. Petrucci et al. (eds.), *New Statistical Developments in Data Science*,
Springer Proceedings in Mathematics & Statistics 288,
https://doi.org/10.1007/978-3-030-21158-5_10

$$p_G(x) := \int_{\mathscr{Y}} k(x \mid y)\, \mathrm{d}G(y), \quad x \in \mathbb{R}, \tag{1}$$

defines a probability density on $\mathbb{R}$ for every probability measure G on $(\mathscr{Y}, \mathscr{B})$, whose collection is indicated by $\mathscr{G}$. The cumulative distribution function of the mixed density in (1) is denoted by

$$F_G(x) = \int_{-\infty}^{x} p_G(u)\, \mathrm{d}u, \quad x \in \mathbb{R}.$$

Suppose we observe n independent random variables $X_1, \ldots, X_n$ identically distributed according to the mixed density

$$p_0(x) \equiv p_{G_0}(x) = \int_{\mathscr{Y}} k(x \mid y)\, \mathrm{d}G_0(y), \quad x \in \mathbb{R}.$$

We denote by F_0 the cumulative distribution function of the density p_0, namely,

$$F_0(x) \equiv F_{G_0}(x) = \int_{-\infty}^{x} p_0(u)\, \mathrm{d}u, \quad x \in \mathbb{R}.$$

The interest is in recovering the unknown mixing distribution $G_0 \in \mathscr{G}$ from observations of the random sample $X^{(n)} := (X_1, \ldots, X_n)$. The formulation of the problem applies to both finite and infinite mixtures, but the focus of this chapter is primarily on the case when the sampling density is a mixture with an unknown, but bounded above number of components.

The problem has been initially studied from the frequenstist perspective by Chen [1], who established that, when p_0 has an unknown number of components d_0 such that $1 \leq d_0 \leq N$, for some fixed integer N, then the optimal rate for estimating the mixing distribution G_0 is only $n^{-1/4}$ and this rate is achievable, under identifiability conditions, by some minimum distance estimator. Even if Theorem 2 in Chen [1], p. 226, is not correct because of Lemma 2 it relies on, an emended version of Lemma 2 has been recently given by Heinrich and Kahn [9] in assertion (21) of their Theorem 6.3, p. 2857, by comparing a fixed mixture with all the mixtures having mixing distributions in an L^1-Wasserstein ball, instead of comparing all possible pairs of mixtures in a ball. As a consequence, Theorem 2 of Chen [1] remains valid by dropping uniformity over an L^1-Wasserstein ball and the statement is weakened to an assertion on the optimal pointwise rate of estimation: for any fixed mixing distribution, say G_0, the minimum distance estimator converges at $n^{-1/4}$-rate, but with a multiplicative constant that may depend on G_0. The first Bayesian analysis of the problem we are aware of traces back to Ishwaran et al. [11], who define a prior law over the space of all mixing distributions with at most N components, the mixture weights being assigned an N-dimensional Dirichlet distribution with a non-informative choice for the shape parameters that are all set equal to α/N for a positive constant α. Under conditions similar to those postulated by Chen [1], which, in particular, employ the

notion of strong identifiability in mixture models, they prove that Bayesian estimation of the mixing distribution in the Kantorovich metric is possible at the optimal rate $n^{-1/4}$, up to a $\log n$-factor. More recently, posterior convergence rates for estimating the mixing distribution in the L^2-Wasserstein metric for finite mixtures of multivariate distributions have been discussed by Nguyen [14] following a different line of reasoning. In this chapter, we show that, by combining the approach of Ishwaran et al. [11], which instrumentally uses posterior contraction rates in the sup-norm for the distribution function and strong identifiability to shift to the Kantorovich distance between mixing distributions, with the current methodology for the study of posterior contraction rates, which can by now count upon many refined results for small ball prior probability estimates, the mixing distribution is estimable in the Kantorovich or L^1-Wasserstein metric at the optimal rate $n^{-1/4}$ (up to a logarithmic factor) for a large class of prior laws over the space of mixing distributions with at most N components, under less stringent conditions than those used in Ishwaran et al. [11] or in Nguyen [14]. Many aspects of this fundamental statistical problem still remain unclear and we hope to contribute to a better understanding of it in a follow-up study.

Before introducing the notation, a remark on the use of the term "Bayesian deconvolution" is in order. This phrase has been recently introduced by Efron [4] to describe a maximum likelihood procedure for estimating the mixing distribution in general mixture models of the form in (1). Even if the mixtures herein considered are not necessarily convolution kernel mixtures, we liked the evocative power of the expression to recall the general inverse problem of recovering the unknown mixing distribution.

Notation. In this paragraph, we set out the notation and recall some definitions used throughout the chapter.

- The symbols "$\lesssim$" and "$\gtrsim$" indicate inequalities valid up to a constant multiple that is universal or fixed within the context, but anyway inessential for our purposes.
- All probability density functions are meant to be with respect to Lebesgue measure λ on $\mathbb{R}$ or on some subset thereof.
- The same symbol, say G, is used to denote a probability measure on $(\mathscr{Y}, \mathscr{B})$ as well as the corresponding cumulative distribution function.
- The degenerate probability distribution putting mass one at a point $y \in \mathbb{R}$ is denoted by δ_y.
- The notation Pf stands for the expected value $\int f \, \mathrm{d}P$, where the integral is understood to extend over the entire natural domain when, here and elsewhere, the domain of integration is omitted. With this convention, for the empirical measure $\mathbb{P}_n := n^{-1} \sum_{i=1}^n \delta_{X_i}$ associated with the random sample $X_1, \ldots, X_n$, namely, the discrete uniform distribution on the sample values that puts mass $1/n$ on each one of the observations, the notation $\mathbb{P}_n f$ abbreviates the formula $n^{-1} \sum_{i=1}^n f(X_i)$.
- For every pair $\mathbf{x}_N, \mathbf{y}_N \in \mathbb{R}^N$, $\|\mathbf{x}_N - \mathbf{y}_N\|_{\ell^1}$ stands for the ℓ^1-distance $\sum_{j=1}^N |x_j - y_j|$.
- For a probability measure Q on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$, let q denote its density. For any $\epsilon > 0$,

$$B_{\mathrm{KL}}(P_0;\,\epsilon^2) := \left\{ Q :\, P_0\left(\log\frac{p_0}{q}\right) \le \epsilon^2,\quad P_0\left(\log\frac{p_0}{q}\right)^2 \le \epsilon^2 \right\}$$

denotes a *Kullback-Leibler* type neighborhood of P_0 of radius ϵ^2. Defined, for every $\alpha \in (0,\,1]$, the divergence $\rho_\alpha(P_0\|Q) := (1/\alpha)[P_0(p_0/q)^\alpha - 1]$, see Wong and Shen [21], pp. 351–352,

$$B_{\rho_\alpha}(P_0;\,\epsilon^2) := \left\{ Q :\, \rho_\alpha(P_0\|Q) \le \epsilon^2 \right\}$$

is the ρ_α-neighborhood of P_0 of radius ϵ^2. The definition of ρ_α extends to negatives values of $\alpha \in (-1,\,0)$. In particular, for $\alpha = -1/2$, the divergence $\rho_{-1/2}(P_0\|Q) = -2\int p_0[(q/p_0)^{1/2} - 1]\,\mathrm{d}\lambda = \int (p_0^{1/2} - q^{1/2})^2\,\mathrm{d}\lambda$ is the squared Hellinger distance. We can thus define the following *Hellinger* type neighborhood of P_0 of radius ϵ^2:

$$B_{\rho_{-1/2}\|\cdot\|_\infty}(P_0;\,\epsilon^2) := \left\{ Q :\, \rho_{-1/2}(P_0\|Q) \left\| \frac{p_0}{q} \right\|_\infty \le \epsilon^2 \right\}.$$

– For any real number $p \ge 1$ and any pair of probability measures $G_1,\, G_2 \in \mathscr{G}$ with finite pth absolute moments, the *L^p-Wasserstein distance* between G_1 and G_2 is defined as

$$W_p(G_1,\,G_2) := \left(\inf_{\gamma\in\Gamma(G_1,\,G_2)} \int_{\mathscr{Y}\times\mathscr{Y}} |y_1 - y_2|^p\,\gamma(\mathrm{d}y_1,\,\mathrm{d}y_2) \right)^{1/p},$$

where $\Gamma(G_1,\,G_2)$ is the set of all joint probability measures on $(\mathscr{Y}\times\mathscr{Y}) \subseteq \mathbb{R}^2$, with marginal distributions G_1 and G_2 on the first and second arguments, respectively.

2 Main Results

This section is devoted to expose the main results of the chapter and is split into two parts. In the first one, preliminary results on Bayesian estimation of distribution functions in the Kolmogorov metric, which are valid for a large class of prior laws, are presented and some issues highlighted. In the second part, arguably the most relevant, attention is restricted to finite mixtures with an unknown, but bounded above number of components and Bayesian estimation of the mixing distribution in the Kantorovich metric at the optimal rate $n^{-1/4}$ (up to a logarithmic factor) is discussed.

Posterior Concentration of Kernel Mixtures in the Kolmogorov Metric

The following assumption will be hereafter in force.

Assumption **A**. Let

$$\epsilon_n := \left(\frac{\log n}{n}\right)^{1/2} L_n, \quad n \in \mathbb{N}, \tag{2}$$

where, depending on the prior concentration rate on small balls around P_0, the sequence of positive real numbers (L_n) can be either slowly varying at $+\infty$ or degenerate at an appropriate constant L_0.

Comments on the two possible specifications of (L_n) in connection with the prior concentration rate are postponed to Lemma 1, which provides sufficient conditions on the distribution function F_0 and the prior concentration rate ϵ_n for the posterior to contract at a nearly $\sqrt{n}$-rate on Kolmogorov neighborhoods of F_0. We warn the reader that, unless otherwise specified, in all stochastic order symbols used hereafter, the probability measure $\mathbf{P}$ is understood to be P_0^n, the joint law of the first n coordinate projections of the infinite product probability measure P_0^∞. Also, Π_n stands for a prior law, possibly depending on the sample size, over the space of probability measures $\{P_G,\ G \in \mathscr{G}\}$, with density p_G as defined in (1).

Lemma 1 *Let F_0 be a continuous distribution function. If, for a constant $C > 0$ and a sequence ϵ_n as defined in (2), we have*

$$\Pi_n(B_{\mathrm{KL}}(P_0;\ \epsilon_n^2)) \gtrsim \exp(-Cn\epsilon_n^2), \tag{3}$$

then, for $M_n \gtrsim \sqrt{(C + 1/2)}L_n$,

$$\Pi_n\left(\sqrt{n}\sup_x |(F_G - F_0)(x)| > M_n(\log n)^{1/2} \mid X^{(n)}\right) = o_{\mathbf{P}}(1). \tag{4}$$

Proof The posterior probability of the event

$$A_n^c := \left\{G :\ \sqrt{n}\sup_x |(F_G - F_0)(x)| > M_n(\log n)^{1/2}\right\}$$

is given by

$$\Pi_n(A_n^c \mid X^{(n)}) = \frac{\int_{A_n^c} \prod_{i=1}^n p_G(X_i)\, \Pi_n(\mathrm{d}G)}{\int_{\mathscr{G}} \prod_{i=1}^n p_G(X_i)\, \Pi_n(\mathrm{d}G)}.$$

We construct (a sequence of) tests (ϕ_n) for testing the hypothesis

$$H_0 :\ P = P_0 \quad \textit{versus} \quad H_1 :\ P = P_G,\ G \in A_n^c,$$

where $\phi_n :\equiv \phi_n(X^{(n)};\ P_0) :\ \mathscr{X}^n \to \{0,\ 1\}$ is the indicator function of the rejection region of H_0, such that

$$P_0^n\phi_n \to 0 \quad \text{as } n \to +\infty$$

$$\text{and} \quad \sup_{G \in A_n^c} P_G^n(1-\phi_n) \le 2\exp\left(-2(M_n - K)^2 \log n\right) \text{ for sufficiently large } n,$$

with a finite constant $K > 0$ and a sequence $M_n > K$ for every n large enough. Define the test

$$\phi_n := 1_{R_n}, \quad \text{with} \quad R_n := \left\{x^{(n)} : \sqrt{n} \sup_x |(F_n - F_0)(x)| > K(\log n)^{1/2}\right\},$$

where F_n is the empirical distribution function, that is, the distribution function associated with the empirical probability measure $\mathbb{P}_n$ of the sample $X^{(n)}$. Since $x \mapsto F_0(x)$ is *continuous* by assumption, in virtue of the Dvoretzky–Kiefer–Wolfowitz [3] (DKW for short) inequality, with the tight universal constant in Massart [13], the type I error probability $P_0^n\phi_n$ can be bounded above as follows

$$P_0^n\phi_n = P_0^n(R_n) \le 2\exp\left(-2K^2 \log n\right).$$

Then,

$$E_0^n[\Pi_n(A_n^c \mid X^{(n)})\phi_n] \le P_0^n\phi_n \le 2\exp\left(-2K^2 \log n\right), \tag{5}$$

where E_0^n denotes expectation with respect to P_0^n, and

$$\begin{aligned} E_0^n[\Pi_n(A_n^c \mid X^{(n)})] &= E_0^n[\Pi_n(A_n^c \mid X^{(n)})\phi_n] + E_0^n[\Pi_n(A_n^c \mid X^{(n)})(1-\phi_n)] \\ &\le 2\exp\left(-2K^2 \log n\right) + E_0^n[\Pi_n(A_n^c \mid X^{(n)})(1-\phi_n)]. \end{aligned}$$

It remains to control the term $E_0^n[\Pi_n(A_n^c \mid X^{(n)})(1-\phi_n)]$. Defined the set

$$D_n := \left\{x^{(n)} : \int_{\mathscr{G}} \prod_{i=1}^n \frac{p_G}{p_0}(x_i)\, \Pi_n(\mathrm{d}G) \le \Pi_n(B_{\mathrm{KL}}(P_0;\, \epsilon_n^2))\exp\left(-(C+1)n\epsilon_n^2\right)\right\},$$

consider the following decomposition

$$E_0^n[\Pi_n(A_n^c \mid X^{(n)})(1-\phi_n)] = E_0^n[\Pi_n(A_n^c \mid X^{(n)})(1-\phi_n)(1_{D_n} + 1_{D_n^c})].$$

It is known from Lemma 8.1 of Ghosal et al. [7], p. 524, that $P_0^n(D_n) \le (C^2 n\epsilon_n^2)^{-1}$. It follows that

$$E_0^n[\Pi_n(A_n^c \mid X^{(n)})(1-\phi_n)1_{D_n}] \le P_0^n(D_n) \le (C^2 n\epsilon_n^2)^{-1}. \tag{6}$$

By the assumption in (3) and Fubini's theorem,

$$E_0^n[\Pi_n(A_n^c \mid X^{(n)})(1-\phi_n)1_{D_n^c}] \lesssim \exp\left((2C+1)n\epsilon_n^2\right) \int_{A_n^c} P_G^n(1-\phi_n)\,\Pi_n(\mathrm{d}G). \tag{7}$$

The following arguments are aimed at finding an exponential upper bound on $\sup_{G\in A_n^c} P_G^n(1-\phi_n)$. By the triangular inequality, over the set R_n^c, for every $G \in A_n^c$,

$$\begin{aligned} M_n(\log n)^{1/2} &< \sqrt{n}\sup_x |(F_G - F_0)(x)| \\ &\le \sqrt{n}\sup_x |(F_G - F_n)(x)| + \sqrt{n}\sup_x |(F_n - F_0)(x)| \\ &\le \sqrt{n}\sup_x |(F_G - F_n)(x)| + K(\log n)^{1/2}, \end{aligned}$$

which implies that

$$\sqrt{n}\sup_x |(F_G - F_n)(x)| > (M_n - K)(\log n)^{1/2}.$$

Since $x \mapsto F_G(x) := \int_{-\infty}^x p_G(u)\,\mathrm{d}u$ is continuous, by applying again the DKW's inequality, we obtain that

$$\begin{aligned} \sup_{G\in A_n^c} P_G^n(1-\phi_n) &\le \sup_{G\in A_n^c} P_G^n\left(\sqrt{n}\sup_x |(F_G - F_n)(x)| > (M_n - K)(\log n)^{1/2}\right) \\ &\le 2\exp\left(-2(M_n - K)^2 \log n\right). \end{aligned}$$

Combining the above assertion with (7), we see that

$$E_0^n[\Pi_n(A_n^c \mid X^{(n)})(1-\phi_n)1_{D_n^c}] \lesssim 2\exp\left(-[2(M_n - K)^2 - (2C+1)L_n^2]\log n\right), \tag{8}$$

where the right-hand side of the above inequality tends to zero provided that $(M_n - K) > \sqrt{(C+1/2)}L_n$ for every sufficiently large n. The in-probability convergence in (4) follows from (5), (6) and (8). This concludes the proof. □

Some remarks and comments on Lemma 1 are in order.

- The first one aims at spelling out the assumptions used in the proof, some of which could otherwise erroneously seem to be confined to the context of finite mixture models, as well as at clarifying their role. Given the prior concentration rate ϵ_n as defined in (3), which depends on the prior distribution Π_n and the "point" P_0, the only further assumption used is the *continuity* of the distribution functions F_0 and F_G, which is satisfied for Lebesgue dominated probability measures P_0 and P_G. This condition is used to control the type I and type II error probabilities of the (sequence of) tests (ϕ_n) by the DKW's inequality. It is, instead, in no way used the assumption that the density p_G is modeled as a mixture, so that, even if the result has its origin in the context of finite mixtures, it applies to general dominated models and a nearly parametric (up to a logarithmic factor) prior concentration

rate is the only driver and determinant of posterior contraction.

- Lemma 1 has its roots in Theorem 2 of Ishwaran et al. [11], p. 1324 (see pp. 1330–1331 for the proof), which deals with *finite* mixtures having an *unknown* number of components d_0, yet *bounded* above by an integer N, namely, $1 \leq d_0 \leq N < +\infty$, while the prior is supported over the space of all mixing distributions with at most N components, the mixture weights being assigned an N-dimensional Dirichlet distribution with a non-informative choice for the shape parameters that are all set equal to α/N for a positive constant α. Nonetheless, as previously remarked, Lemma 1 has a broader scope of validity and applies also to *infinite* kernel mixtures with other prior laws for the mixing distribution than the Dirichlet process, which "locally" attain an almost parametric prior concentration rate. This is the case for Dirichlet location or location-scale mixtures of normal densities and, more in general, for location-scale mixtures of exponential power densities with an even integer shape parameter, when the sampling density is of the same form as the assumed model, with mixing distribution being either compactly supported or having sub-exponential tails, see Ghosal and van der Vaart [8], Scricciolo [16], Theorems 4.1, 4.2 and Corollary 4.1, pp. 285–290. In all these cases, the prior concentration rate is (at worst) $\epsilon_n = n^{-1/2} \log n$, where $L_n = (\log n)^{1/2}$. An extension of the previous results to convolution mixtures of super-smooth kernels, with Pitman-Yor or normalized inverse-Gaussian processes as priors for the mixing distribution, for which Lemma 1 also holds, is considered in Scricciolo [17], see Theorem 1, pp. 486–487. Another class of priors on kernel mixtures to which Lemma 1 applies is that of *sieve* priors. For a given kernel, a sieve prior is defined by combining single priors on classes of kernel mixtures, each one indexed by the number of mixture components, with a prior on such random number. A probability measure with kernel mixture density is then generated in two steps: first the model index, i.e., the number of mixture components, is selected; then a probability measure is generated from the chosen model according to a prior on it. When the true density p_0 is itself a kernel mixture, the prior concentration rate can be assessed by bounding below the probability of Kullback-Leibler type neighborhoods of P_0 by the probability of ℓ^1-balls of appropriate dimension. In fact, approximation properties of mixtures under consideration can be exploited to find a good fitting distribution of the sampling density in a proper subclass. More precisely, any finite kernel mixture can be approximated arbitrarily well (in the distance induced by the L^1-norm) by mixtures having the same number of components, the mixture components and weights taking values in ℓ^1-neighborhoods of the corresponding true elements. The number of mixture components is then constant, this leading to the prior concentration rate $\epsilon_n \propto (n/\log n)^{-1/2}$, where $L_n \equiv L_0$. Examples of sieve priors in which, for every choice of the model index, the mixture weights are jointly distributed according to a Dirichlet distribution, are provided by the Bernstein polynomials, see Theorem 2.2 of Ghosal [6], pp. 1268–1269, by histograms and polygons, see Theorem 1 of Scricciolo [15], pp. 629–630 (see pp. 636–637 for the proof). If, as a special case, a single prior distribution on kernel mixtures with a sample size-dependent number $N \equiv L_n$ of mixture components is considered,

then the prior concentration rate is $\epsilon_n = (n/\log n)^{-1/2} L_n$ for any arbitrary slowly varying sequence $L_n \to +\infty$.

We now state sufficient conditions on the kernel density and the prior distributions for the mixture atoms and weights so that the overall prior on kernel mixtures with (at most) N components verifies condition (3) for $\epsilon_n \propto (n/\log n)^{-1/2}$, when the sampling density is itself a kernel mixture with $1 \leq d_0 \leq N$ components. The aim of this analysis is twofold: first, to provide less stringent requirements on the kernel density than those postulated in condition (b) employed in Theorem 2 of Ishwaran et al. [11], p. 1324, which relies on Lemma 4 of Ishwaran [10], pp. 2170–2171; second, to generalize the aforementioned result to a class of prior distributions on the mixture weights that comprises the Dirichlet distribution as a special case. The density p_G is modeled as

$$p_G(\cdot) = \sum_{j=1}^{N} w_j k(\cdot \mid y_j),$$

with a discrete mixing distribution $G = \sum_{j=1}^{N} w_j \delta_{y_j}$. The vector $\mathbf{w}_N := (w_1, \ldots, w_N)$ of mixing weights has a prior distribution $\tilde{\pi}_N$ on the $(N-1)$-dimensional simplex $\Delta_N := \{\mathbf{w}_N \in \mathbb{R}^N : 0 \leq w_j \leq 1,\ j = 1, \ldots, N,\ \sum_{j=1}^{N} w_j = 1\}$ and the atoms $y_1, \ldots, y_N$ are independently and identically distributed according to a prior distribution π. We shall also use the notation $\mathbf{y}_N$ for $(y_1, \ldots, y_N)$. The model can be thus described:

- the random vectors $\mathbf{y}_N$ and $\mathbf{w}_N$ are independent;
- given $(\mathbf{y}_N, \mathbf{w}_N)$, the random variables $X_1, \ldots, X_n$ are conditionally independent and identically distributed according to p_G.

The overall prior is then $\Pi = \tilde{\pi}_N \times \pi^{\otimes N}$. Let the sampling density p_0 be itself a finite kernel mixture, with $1 \leq d_0 \leq N$ components,

$$p_0(\cdot) \equiv p_{G_0}(\cdot) = \sum_{j=1}^{d_0} w_j^0 k(\cdot \mid y_j^0),$$

where the mixing distribution is $G_0 = \sum_{j=1}^{d_0} w_j^0 \delta_{y_j^0}$ for weights $\mathbf{w}_{d_0}^0 := (w_1^0, \ldots, w_{d_0}^0) \in \Delta_{d_0}$ and support points $\mathbf{y}_{d_0}^0 := (y_1^0, \ldots, y_{d_0}^0) \in \mathbb{R}^{d_0}$. A caveat applies: if d_0 is strictly smaller than N, that is, $1 \leq d_0 < N$, then the vectors $\mathbf{w}_{d_0}^0$ and $\mathbf{y}_{d_0}^0$ are viewed as degenerate elements of Δ_N and $\mathbb{R}^N$, respectively, with coordinates $w_{d_0+1} = \cdots = w_N = 0$ and $y_{d_0+1} = \cdots = y_N = 0$.

We assume that

(i) there exists a constant $c_k > 0$ such that

$$\|k(\cdot \mid y_1) - k(\cdot \mid y_2)\|_1 \leq c_k |y_1 - y_2| \quad \text{for all } y_1,\ y_2 \in \mathscr{Y};$$

(ii) for every $\epsilon > 0$ small enough and a constant $c_0 > 0$,

$$\tilde{\pi}_N(\{\mathbf{w}_N \in \Delta_N : \|\mathbf{w}_N - \mathbf{w}_N^0\|_{\ell^1} \leq \epsilon\}) \gtrsim \epsilon^{c_0 N};$$

(iii) the prior distribution π for the atoms has a continuous and positive Lebesgue density (also denoted by π) on an interval containing the support of G_0.

Some remarks and comments on the previously listed assumptions are in order. Condition (i) requires the kernel density $k(\cdot \mid y)$ to be globally Lipschitz continuous on $\mathscr{Y}$. Condition (ii) is satisfied for a Dirichlet prior distribution $\tilde{\pi}_N = \mathrm{Dir}(\alpha_1, \ldots, \alpha_N)$, with parameters $\alpha_1, \ldots, \alpha_N$ such that, for constants a, $A > 0$, $D \geq 1$ and, for $0 < \epsilon \leq 1/(DN)$,

$$A\epsilon^a \leq \alpha_j \leq D, \quad j = 1, \ldots, N.$$

Using Lemma A.1 of Ghosal [6], pp. 1278–1279, we find that $\tilde{\pi}_N(N(\mathbf{w}_N^0; \epsilon)) \gtrsim \exp(-c_0 N \log(1/\varepsilon))$ for a constant $c_0 > 0$ depending only on a, A, D and $\sum_{j=1}^N \alpha_j$.

Proposition 1 *Under assumptions* (i)–(iii), *condition* (3) *is verified for*

$$\epsilon_n \propto (n/\log n)^{-1/2}.$$

Proof For every density p_G, with mixing distribution $G = \sum_{j=1}^N w_j \delta_{y_j}$ having support points $\mathbf{y}_N \in \mathbb{R}^N$ and mixture weights $\mathbf{w}_N \in \Delta_N$, by assumption (i) we have

$$\begin{aligned}\|p_G - p_0\|_1 &\lesssim \sum_{j=1}^N w_j^0 \|k(\cdot \mid y_j) - k(\cdot \mid y_j^0)\|_1 + \sum_{j=1}^N |w_j - w_j^0| \|k(\cdot \mid y_j)\|_1 \\ &\lesssim \|\mathbf{y}_N - \mathbf{y}_N^0\|_{\ell^1} + \|\mathbf{w}_N - \mathbf{w}_N^0\|_{\ell^1}.\end{aligned}$$

Let $0 < \epsilon \leq [(1/2) \wedge (1 - e^{-1})/\sqrt{2}]$ be fixed. For $\mathbf{y}_N \in \mathbb{R}^N$ and $\mathbf{w}_N \in \Delta_N$ such that $\|\mathbf{y}_N - \mathbf{y}_N^0\|_{\ell^1} \leq \epsilon$ and $\|\mathbf{w}_N - \mathbf{w}_N^0\|_{\ell^1} \leq \epsilon$, by LeCam [12] inequalities, p. 40, relating the L^1-norm and the Hellinger metric, the squared Hellinger distance between p_0 and p_G can be bounded above by a multiple of ϵ:

$$\rho_{-1/2}(P_0 \| P_G) = \int (p_G^{1/2} - p_0^{1/2})^2 \, \mathrm{d}\lambda \leq \|p_G - p_0\|_1 \lesssim \epsilon.$$

Then, by Lemma A.10 in Scricciolo [16], p. 305, for a suitable constant $c_1 > 0$,

$$\left\{ p_G : G = \sum_{j=1}^N w_j \delta_{y_j}, \ \|\mathbf{w}_N - \mathbf{w}_N^0\|_{\ell^1} \leq \epsilon, \ \|\mathbf{y}_N - \mathbf{y}_N^0\|_{\ell^1} \leq \varepsilon \right\} \subseteq B_{\mathrm{KL}}\left(P_0; \, c_1 \epsilon \left(\log \tfrac{1}{\epsilon} \right)^2 \right).$$

Next, define the set $N(\mathbf{w}_N^0;\ \epsilon) := \{\mathbf{w}_N \in \Delta_N :\ \|\mathbf{w}_N - \mathbf{w}_N^0\|_{\ell^1} \le \epsilon\}$. For $\epsilon > 0$ small enough, by assumption (ii),

$$\tilde{\pi}_N(N(\mathbf{w}_N^0;\ \epsilon)) \gtrsim \exp\left(-c_0 N \log(1/\varepsilon)\right)$$

with an appropriate constant $c_0 > 0$. Denoted by $B(\mathbf{y}_N^0;\ \epsilon)$ the $\mathbf{y}_N^0$-centered ℓ^1-ball of radius $\epsilon > 0$,

$$B(\mathbf{y}_N^0;\ \epsilon) := \{\mathbf{y}_N \in \mathbb{R}^N :\ \|\mathbf{y}_N - \mathbf{y}_N^0\|_{\ell^1} \le \epsilon\},$$

by condition (iii) the prior probability of $B(\mathbf{y}_N^0;\ \epsilon)$ under the N-fold product measure $\pi^{\otimes N}$ can be bounded below as follows:

$$\begin{aligned}\pi^{\otimes N}(B(\mathbf{y}_N^0;\ \epsilon)) &\ge \prod_{j=1}^N \pi([y_j^0 - (\epsilon/N),\ y_j^0 + (\epsilon/N)]) \\ &= \prod_{j=1}^N \int_{y_j^0-(\epsilon/N)}^{y_j^0+(\epsilon/N)} \pi(y)\,\mathrm{d}y \gtrsim \exp\left(-d_1 N \log(1/\epsilon)\right)\end{aligned}$$

for a positive constant d_1. Therefore, for appropriate constants $c_1,\ d_2 > 0$,

$$\Pi(B_{\mathrm{KL}}(P_0;\ c_1\epsilon|\log\epsilon|^2)) \gtrsim \tilde{\pi}_N(N(\mathbf{w}_N^0;\ \epsilon))\,\pi^{\otimes N}(B(\mathbf{y}_N^0;\ \epsilon)) \gtrsim \exp\left(-d_2 N \log(1/\epsilon)\right).$$

Set $\xi := (c_1\epsilon)^{1/2}\log(1/\epsilon)$, since $\log(1/\epsilon) \lesssim \log(1/\xi)$, we have $\Pi(B_{\mathrm{KL}}(P_0;\ \xi^2)) \gtrsim \exp\left(-c_2 \log(1/\xi)\right)$ for a real constant $c_2 > 0$ (possibly depending on p_0). Replacing ξ with ϵ_n, we get $\Pi(B_{\mathrm{KL}}(P_0;\ \epsilon_n^2)) \gtrsim \exp\left(-c_2 n\epsilon_n^2\right)$ for sufficiently large n, and the proof is complete. □

Inspection of the proof of Lemma 1 reveals that, under the small ball prior probability estimate in (3), we have

$$E_0^n[\Pi_n(A_n^c \mid X^{(n)})] = O((n\epsilon_n^2)^{-1}).$$

The assertion of Lemma 1 can be enhanced to have

$$E_0^n[\Pi_n(A_n^c \mid X^{(n)})] = O(\exp\left(-B_1 n\epsilon_n^2\right))$$

by employing a small ball prior probability estimate involving stronger divergences. The convergence in (4) then becomes almost-sure. Besides, due to the fact that the posterior probability vanishes exponentially fast, namely, along almost all sample sequences, for a finite constant $B > 0$, we have

$$\Pi_n(A_n^c \mid X^{(n)}) \lesssim \exp\left(-Bn\epsilon_n^2\right) \text{ for all but finitely many } n,$$

the stochastic order of the maximum absolute difference between F_0 and the posterior expected distribution function can be assessed, see Corollary 1 below.

Lemma 2 *Under the conditions of Lemma 1, if the small ball prior probability estimate in (3) is replaced by*

$$\Pi_n(B_{\rho_\alpha}(P_0;\, \epsilon_n^2)) \gtrsim \exp(-Cn\epsilon_n^2), \quad \text{for } \alpha \in (0,\, 1], \tag{9}$$

then, for $M_n \gtrsim \sqrt{(C+1/2)}L_n$,

$$\Pi_n\left(\sqrt{n}\sup_x |(F_G - F_0)(x)| > M_n(\log n)^{1/2} \mid X^{(n)}\right) \to 0 \quad P_0^\infty\text{-almost surely.}$$

Proof The proof is an adaptation of that of Lemma 1. We therefore highlight only the main changes. Taking a sequence $K_n = \theta M_n$ for any $\theta \in (0,\, 1)$, we have

$$E_0^n[\Pi_n(A_n^c \mid X^{(n)})\phi_n] \le P_0^n\phi_n \le 2\exp(-2\theta^2 M_n^2 \log n)$$

and

$$E_0^n[\Pi_n(A_n^c \mid X^{(n)})(1-\phi_n)1_{D_n^c}] \lesssim 2\exp(-[2(1-\theta)^2 M_n^2 - (2C+1)L_n^2]\log n),$$

with

$$M_n > (1-\theta)^{-1}\sqrt{(C+1/2)}L_n \tag{10}$$

for every sufficiently large n. A straightforward extension of Lemma 2 in Shen and Wasserman [19], p. 691 (and pp. 709–710 for the proof), yields that, for every $\xi \in (0,\, 1)$,

$$P_0^n\big(D_n \le \xi\,\Pi_n(B_{\rho_\alpha}(P_0;\, \epsilon_n^2))\exp(-(C+1)n\epsilon_n^2)\big) \le (1-\xi)^{-1}\exp(-\alpha Cn\epsilon_n^2). \tag{11}$$

Considering $M_n = IL_n$ for a finite constant $I > (1-\theta)^{-1}\sqrt{(C+1/2)}$ so that condition (10) is satisfied, by combining partial bounds we obtain that

$$E_0^n[\Pi_n(A_n^c \mid X^{(n)})] \lesssim \exp(-B_1 n\epsilon_n^2)$$

for an appropriate finite constant $B_1 > 0$. For a constant $B > 0$,

$$P_0^n\big(\Pi_n(A_n^c \mid X^{(n)}) \ge \exp(-Bn\epsilon_n^2)\big) \lesssim \exp\big(-(B_1 - B)n\epsilon_n^2\big).$$

Choose $0 < B < B_1$. Since $\sum_{n=1}^\infty \exp(-(B_1 - B)n\epsilon_n^2) < +\infty$, almost sure convergence follows from the first Borel-Cantelli lemma. □

Remark 1 The assertion of Lemma 2 still holds if the small ball prior probability estimate in (9) is replaced by the requirement

$$\Pi_n(B_{\rho_{-1/2}\|\cdot\|_\infty}(P_0;\, \epsilon_n^2)) \gtrsim \exp(-Cn\epsilon_n^2), \tag{12}$$

which involves a Hellinger type neighborhood of P_0. Then, a bound similar to that in (11) is given in Lemma 8.4 of Ghosal et al. [7], pp. 526–527.

As previously mentioned, Lemma 2 allows to derive the stochastic order of the maximum absolute difference between F_0 and its Bayes' estimator

$$F_n^{\mathrm{B}}(\cdot) := \int_{\mathscr{G}} F_G(\cdot)\Pi(\mathrm{d}G \mid X^{(n)}),$$

namely, the posterior expected distribution function.

Corollary 1 *Under the conditions of Lemma 2, we have*

$$\sqrt{n}\sup_x |(F_n^{\mathrm{B}} - F_0)(x)| = O_{\mathbf{P}}(M_n(\log n)^{1/2}).$$

Proof By standard arguments,

$$\begin{aligned}
\sqrt{n}\sup_x |(F_n^{\mathrm{B}} - F_0)(x)| &= \sqrt{n}\sup_x \left| \int_{\mathscr{G}} F_G(x)\, \Pi_n(\mathrm{d}G \mid X^{(n)}) - F_0(x) \right| \\
&\le \int_{\mathscr{G}} \sqrt{n}\sup_x |(F_G - F_0)(x)|\, \Pi_n(\mathrm{d}G \mid X^{(n)}) \\
&= \left(\int_{A_n} + \int_{A_n^c} \right) \sqrt{n}\sup_x |(F_G - F_0)(x)|\, \Pi_n(\mathrm{d}G \mid X^{(n)}) \\
&\le M_n(\log n)^{1/2} + 2\sqrt{n}\Pi_n(A_n^c \mid X^{(n)}) \\
&\lesssim M_n(\log n)^{1/2} \quad \text{for sufficiently large } n
\end{aligned}$$

because condition (9) yields that, with probability one, for a finite constant $B > 0$, the posterior probability $\sqrt{n}\Pi_n(A_n^c \mid X^{(n)}) \lesssim \sqrt{n}\exp(-Bn\epsilon_n^2)$ for all but finitely many n. The assertion follows. □

Posterior Concentration of the Mixing Distribution in the Kantorovich Metric

In this section, we deal with the case where the prior distribution Π is supported over the collection of finite kernel mixtures with at most N components. Sufficient conditions are stated in Theorem 1 below so that the posterior rate of convergence, relative to the Kantorovich or L^1-Wasserstein metric, for the mixing distribution of overfitted mixtures is, up to a slowly varying sequence, (at worst) equal to $(n/\log n)^{-1/4}$, the optimal pointwise rate being $n^{-1/4}$, *cf.* Chen [1], Sect. 2, pp. 222–224.

In order to state the result, we need to introduce some more notation. For every $y \in \mathscr{Y}$, we denote by $K(x \mid y)$ the cumulative distribution function at x of the kernel density $k(\cdot \mid y)$,

$$K(x \mid y) := \int_{-\infty}^{x} k(u \mid y)\,\mathrm{d}u.$$

For clarity of exposition, we recall that F_0 is the distribution function of the mixture density $p_0 \equiv p_{G_0}$ corresponding to the mixing distribution G_0 having an *unknown* number of components d_0 *bounded* above by a fixed integer N.

Theorem 1 *Under the conditions of Lemma 1, if, in addition,*

(a) $\mathscr{Y}$ *is compact,*

(b) for all $x \in \mathbb{R}$, $K(x \mid y)$ *is* 2*-differentiable with respect to* y,

(c) $\{K(\cdot \mid y) : y \in \mathscr{Y}\}$ *is strongly identifiable in the sense of Definition 2 in Chen [1], p. 225, equivalently,* 2*-strongly identifiable in the sense of Definition 2.2 in Heinrich and Kahn [9], p. 2848,*

(d) there exists a uniform modulus of continuity $\omega(\cdot)$ *such that*

$$\sup_x |K^{(2)}(x \mid y) - K^{(2)}(x \mid y')| \leq \omega(|y - y'|) \quad \textit{with } \lim_{h \to 0} \omega(h) = 0,$$

then, for $M_n \gtrsim \sqrt{(C + 1/2)}L_n$,

$$\Pi\left(n^{1/4} W_1(G, G_0) > \sqrt{M_n}(\log n)^{1/4} \mid X^{(n)}\right) = o_{\mathbf{P}}(1).$$

Proof Since Lemma 1 holds, we have

$$\Pi\left(\sqrt{n} \sup_x |(F_G - F_0)(x)| > M_n(\log n)^{1/2} \mid X^{(n)}\right) = o_{\mathbf{P}}(1). \tag{13}$$

Consistently with the notation introduced in Lemma 1, we set

$$A_n := \left\{G : \sqrt{n} \sup_x |(F_G - F_0)(x)| \leq M_n(\log n)^{1/2}\right\}.$$

Under assumptions (a)–(d), assertion (21) of Theorem 6.3 of Heinrich and Kahn [9], p. 2857, holds true, this implying that, for every $G \in A_n$, the Kolmogorov distance between the distribution functions F_G and F_0 is bounded below (up to a constant) by the squared L^1-distance between the mixing distributions G and G_0, respectively: there exists a constant $C_0 > 0$ (possibly depending on G_0) such that, for every $G \in A_n$,

$$C_0 \|G - G_0\|_1^2 < \sup_x |(F_G - F_0)(x)| \leq M_n n^{-1/2}(\log n)^{1/2}. \tag{14}$$

Taking into account the following representation of the L^1-Wasserstein distance

$$W_1(G, G_0) = \|G - G_0\|_1,$$

see, e.g., Shorack and Wellner [20], pp. 64–66, which was obtained by Dall'Aglio [2], the assertion follows by combining (13) with (14). This concludes the proof. □

Some comments on the applicability and consequences of Theorem 1 are in order.

- Theorem 1, like Lemma 1, has its roots in Theorem 2 of Ishwaran et al. [11], p. 1324, which is tailored for finite Dirichlet mixtures. However, thanks to Proposition 1, which implies the conclusion of Lemma 1, meanwhile ensuring applicability to a larger family of prior distributions, under conditions (a)–(d), the assertion that, for sufficiently large constant $M > 0$, the convergence

$$\Pi\left(n^{1/4} W_1(G,\, G_0) > M(\log n)^{1/4} \mid X^{(n)}\right) \to 0 \quad \text{in } P_0^n\text{-probability}$$

  takes place, still holds. The present result differs from that of Theorem 5 in Nguyen [14], pp. 383–384, under various respects: the latter gives an assessment of posterior contraction in the L^2-Wasserstein, as opposed to the L^1-Wasserstein metric, for finite mixtures of multivariate distributions, under more stringent conditions and following a completely different line of reasoning.

- As previously observed on the occasion of the transition from Lemma 1 to Lemma 2, if the small ball prior probability estimate in (3) is replaced with either that in (9) or in (12), then the almost-sure version of Theorem 1

$$\Pi\left(n^{1/4} W_1(G,\, G_0) > \sqrt{M_n}(\log n)^{1/4} \mid X^{(n)}\right) \to 0 \quad P_0^\infty\text{-almost surely}$$

  holds and the rate of convergence for the Bayes' estimator of the mixing distribution can be assessed as follows.

Corollary 2 *Under the conditions of Theorem 1, with the small ball prior probability estimate in (9), we have*

$$W_1(G_n^{\mathrm{B}},\, G_0) = O_{\mathbf{P}}(\sqrt{M_n}(n/\log n)^{-1/4}),$$

where $G_n^{\mathrm{B}}(\cdot) := \int_{\mathscr{G}} G(\cdot)\Pi(\mathrm{d}G \mid X^{(n)})$ *is the Bayes' estimator of the mixing distribution.*

Acknowledgements The author gratefully acknowledges financial support from MIUR, grant n° 2015SNS29B "Modern Bayesian nonparametric methods".

References

1. Chen, J.: Optimal rate of convergence for finite mixture models. Ann. Stat. **23**(1), 221–233 (1995)
2. Dall'Aglio, G.: Sugli estremi dei momenti delle funzioni di ripartizione doppia. (Italian) Ann. Scuola Norm. Sup. Pisa **3**(10), 35–74 (1956)
3. Dvoretzky, A., Kiefer, J., Wolfowitz, J.: Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. Ann. Math. Stat. **27**(3), 642–669 (1956)
4. Efron, B.: Empirical Bayes deconvolution estimates. Biometrika **103**(1), 1–20 (2016)
5. Gao, F., van der Vaart, A.: Posterior contraction rates for deconvolution of Dirichlet-Laplace mixtures. Electron. J. Stat. **10**(1), 608–627 (2016)
6. Ghosal, S.: Convergence rates for density estimation with Bernstein polynomials. Ann. Stat. **29**(5), 1264–1280 (2001)
7. Ghosal, S., Ghosh, J.K., van der Vaart, A.W.: Convergence rates of posterior distributions. Ann. Stat. **28**(2), 500–531 (2000)
8. Ghosal, S., van der Vaart, A.W.: Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. Ann. Stat. **29**(5), 1233–1263 (2001)
9. Heinrich, P., Kahn, J.: Strong identifiability and optimal minimax rates for finite mixture estimation. Ann. Stat. **46**(6A), 2844–2870 (2018)
10. Ishwaran, H.: Exponential posterior consistency via generalized Pólya urn schemes in finite semiparametric mixtures. Ann. Stat. **26**(6), 2157–2178 (1998)
11. Ishwaran, H., James, L.F., Sun, J.: Bayesian model selection in finite mixtures by marginal density decompositions. J. Am. Stat. Assoc. **96**(456), 1316–1332 (2001)
12. LeCam, L.: Convergence of estimates under dimensionality restrictions. Ann. Stat. **1**(1), 38–53 (1973)
13. Massart, P.: The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. Ann. Probab. **18**(3), 1269–1283 (1990)
14. Nguyen, X.: Convergence of latent mixing measures in finite and infinite mixture models. Ann. Stat. **41**(1), 370–400 (2013)
15. Scricciolo, C.: On rates of convergence for Bayesian density estimation. Scand. J. Stat. **34**(3), 626–642 (2007)
16. Scricciolo, C.: Posterior rates of convergence for Dirichlet mixtures of exponential power densities. Electron. J. Stat. **5**, 270–308 (2011)
17. Scricciolo, C.: Adaptive Bayesian density estimation in L^p-metrics with Pitman-Yor or Normalized Inverse-Gaussian process kernel mixtures. Bayesian Anal. **9**(2), 475–520 (2014)
18. Scricciolo, C.: Bayes and maximum likelihood for L^1-Wasserstein deconvolution of Laplace mixtures. Stat. Methods Appl. **27**(2), 333–362 (2018)
19. Shen, X., Wasserman, L.: Rates of convergence of posterior distributions. Ann. Stat. **29**(3), 687–714 (2001)
20. Shorack, G.R., Wellner, J.A.: Empirical Processes with Applications to Statistics. Wiley, New York (1986)
21. Wong, W.H., Shen, X.: Probability inequalities for likelihood ratios and convergence rates of sieve MLES. Ann. Stat. **23**(2), 339–362 (1995)