Stat Methods Appl (2018) 27:333–362
https://doi.org/10.1007/s10260-017-0400-4

ORIGINAL PAPER

Bayes and maximum likelihood for L^1-Wasserstein deconvolution of Laplace mixtures

Catia Scricciolo[1]

Accepted: 3 September 2017 / Published online: 15 September 2017

Abstract We consider the problem of recovering a distribution function on the real line from observations additively contaminated with errors following the standard Laplace distribution. Assuming that the latent distribution is completely unknown leads to a nonparametric deconvolution problem. We begin by studying the rates of convergence relative to the L^2-norm and the Hellinger metric for the direct problem of estimating the sampling density, which is a mixture of Laplace densities with a possibly unbounded set of locations: the rate of convergence for the Bayes' density estimator corresponding to a Dirichlet process prior over the space of all mixing distributions on the real line matches, up to a logarithmic factor, with the $n^{-3/8}\log^{1/8} n$ rate for the maximum likelihood estimator. Then, appealing to an inversion inequality translating the L^2-norm and the Hellinger distance between general kernel mixtures, with a kernel density having polynomially decaying Fourier transform, into any L^p-Wasserstein distance, $p \geq 1$, between the corresponding mixing distributions, provided their Laplace transforms are finite in some neighborhood of zero, we derive the rates of convergence in the L^1-Wasserstein metric for the Bayes' and maximum likelihood estimators of the mixing distribution. Merging in the L^1-Wasserstein distance between Bayes and maximum likelihood follows as a by-product, along with an assessment on the stochastic order of the discrepancy between the two estimation procedures.

Keywords Deconvolution · Dirichlet process · Entropy · Hellinger distance · Laplace mixture · Maximum likelihood · Posterior distribution · Rate of convergence · Sieve · Wasserstein distance

✉ Catia Scricciolo
catia.scricciolo@univr.it

[1] Università degli Studi di Verona, Verona, Italy

1 Introduction

The problem of recovering a distribution function from observations additively contaminated with measurement errors is the object of study in this note. Assuming data are sampled from a convolution kernel mixture, the interest is in “estimating” the mixing or latent distribution from contaminated observations. The statement of the problem is as follows. Let X be a random variable (r.v.) with probability measure P_0 on the Borel-measurable space $(\mathbb{R},\, \mathscr{B}(\mathbb{R}))$, with Lebesgue density $p_0 := \mathrm{d}P_0/\mathrm{d}\lambda$. Suppose that

$$X = Y + Z,$$

where Y and Z are independent, unobservable random variables, Z having Lebesgue density f. We examine the case where the error has the standard Laplace distribution with density

$$f(z) = \frac{1}{2}e^{-|z|}, \quad z \in \mathbb{R}.$$

The r.v. Y has unknown distribution G_0 on some measurable space $(\mathscr{Y},\, \mathscr{B}(\mathscr{Y}))$, with $\mathscr{Y} \subseteq \mathbb{R}$ and $\mathscr{B}(\mathscr{Y})$ the Borel σ-field on $\mathscr{Y}$. The density p_0 is then the convolution of G_0 and f,

$$p_0(x) = (G_0 * f)(x) = \int_{\mathscr{Y}} f(x-y)\,\mathrm{d}G_0(y), \quad x \in \mathbb{R}.$$

In what follows, we also write $p_0 \equiv p_{G_0}$ to stress the dependence of p_0 on G_0. Letting $\mathscr{G}$ be the set of all probability measures G on $(\mathscr{Y},\, \mathscr{B}(\mathscr{Y}))$, the parameter space

$$\mathscr{P} := \left\{ p_G(\cdot) := \int_{\mathscr{Y}} f(\cdot - y)\,\mathrm{d}G(y),\ G \in \mathscr{G} \right\}$$

is the collection of all convolution Laplace mixtures and the model is nonparametric.

Suppose we observe n independent copies $X_1,\, \ldots,\, X_n$ of X. The r.v.’s $X_1,\, \ldots,\, X_n$ are independent and identically distributed (i.i.d.) according to the density $p_0 \equiv p_{G_0}$ on the real line. The interest is in recovering the mixing distribution $G_0 \in \mathscr{G}$ from indirect observations. Deconvolution problems may arise in a wide variety of contexts, the error distribution being typically modelled as a Gaussian, even if also the Laplace has relevant applications. Full density deconvolution, together with the related many normal means problem, has drawn attention in the literature since the late 1950’s and different deconvolution methods have been proposed and developed since then taking the frequentist approach, the most popular being based on nonparametric maximum likelihood and kernel methods. Rates of convergence have been mostly investigated for *density* deconvolution: Fan (1991a, b) showed that deconvolution kernel density estimators achieve global optimal rates for weighted L^p-risks, $p \geq 1$, when the smoothness of the density to be recovered is measured in terms of the number of its derivatives. Hall and Lahiri (2008) considered estimation of the *distribution function* using the cumulative distribution function corresponding to the deconvolution kernel density estimator and showed that it attains minimax-optimal pointwise and global rates for the integrated mean-squared error over different functional classes for the error and latent distributions, smoothness being described through the tail behaviour of their Fourier transforms. For a comprehensive account on the topic, the reader may refer to the monograph of Meister (2009). In this note, we do not assume that the probability measure G_0 possesses Lebesgue density. Wasserstein metrics are then particularly well-suited as global loss functions: convergence in L^p-Wasserstein metrics for discrete mixing distributions has, in fact, a natural interpretation in terms of convergence of the single supporting atoms of the probability measures involved. Dedecker and Michel (2015) have obtained a lower bound on the rate of convergence for the L^p-Wasserstein risk, $p \geq 1$, when no smoothness assumption, except for a moment condition, is imposed on the latent distribution and the error distribution is ordinary smooth, the Laplace being a special case.

Deconvolution problems have only recently begun to be studied from a Bayesian perspective: the typical scheme considers the mixing distribution as a draw from a Dirichlet process prior. Posterior contraction rates for recovering the mixing distribution in L^p-Wasserstein metrics have been investigated in Nguyen (2013) and Gao and van der Vaart (2016), even though the upper bounds in these articles do not match with the lower bound in Dedecker and Michel (2015). Minimax-optimal adaptive recovery rates for mixing densities belonging to Sobolev spaces have been instead obtained by Donnet et al. (2018) in a fully Bayes as well as in an empirical Bayes approach to inference, the latter accounting for a data-driven choice of the prior hyperparameters of the Dirichlet process baseline measure.

In this note, we study nonparametric Bayes and maximum likelihood estimation of the mixing distribution G_0, when no smoothness assumption is imposed on it. The analysis begins with the estimation of the sampling density p_0: estimating the *mixed* density p_0 can, in effect, be the first step for recovering the *mixing* distribution G_0. Taking a Bayesian approach, if the random density p_G is modelled as a Dirichlet–Laplace mixture, then p_0 can be consistently estimated at a rate $n^{-3/8}$, up to a $(\log n)$-factor, if G_0 has tails matching with those of the baseline measure of the Dirichlet process, which essentially requires G_0 to be in the weak support of the process, see Proposition 1 and Proposition 2. This requirement allows to extend to a possibly unbounded set of locations the results of Gao and van der Vaart (2016), which take into account only the case of compactly supported mixing distributions. Taking a frequentist approach, p_0 can be estimated by the maximum likelihood still at a rate $n^{-3/8}$, up to a logarithmic factor. As far as we are aware, the result on the rate of convergence in the Hellinger metric for the maximum likelihood estimator (MLE) of a Laplace convolution mixture is new and is obtained taking the approach proposed by Van de Geer (1996), according to which it is the “dimension” of the class of kernels and the behaviour of p_0 near zero that determine the rate of convergence for the MLE. As previously mentioned, results on the estimation of p_0 are interesting in view of the fact that, appealing to an inversion inequality translating the Hellinger or the L^2-distance between kernel mixtures, with Fourier transform of the kernel density having polynomially decaying tails, into any L^p-Wasserstein distance, $p \geq 1$, between the corresponding mixing distributions, rates of convergence in the L^1-Wasserstein metric for the MLE and the Bayes' estimator of the mixing distribution can be assessed. Merging in the L^1-Wasserstein metric between Bayes and maximum likelihood for deconvolving Laplace mixtures follows as a by-product.

Organization. The note is organized as follows. Convergence rates in the Hellinger metric for Bayes and maximum likelihood density estimation of Laplace convolution mixtures are preliminarily studied in Sect. 2 and in Sect. 3, respectively, in view of their subsequent instrumental use for assessing the L^1-Wasserstein accuracy of the two estimation procedures in recovering the mixing distribution of the sampling density. Merging between Bayes and maximum likelihood follows, as shown in Sect. 4. Remarks and suggestions for possible refinements and extensions of the exposed results are presented in Sect. 5. Auxiliary lemmas, along with the proofs of the main results, are deferred to Appendices A–D.

Notation. We fix the notation and recall some definitions used throughout.

Calculus

- 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.
- For sequences of real numbers $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$, the notation $a_n \sim b_n$ means that $(a_n/b_n) \to 1$ as $n \to +\infty$. Analogously, for real-valued functions f and g, the notation $f \sim g$ means that $f/g \to 1$ in an asymptotic regime that is clear from the context.

Covering and entropy numbers

- Let (T, d) be a (subset of a) semi-metric space. For every $\varepsilon > 0$, the ε-*covering number* of (T, d), denoted by $N(\varepsilon, T, d)$, is defined as the minimum number of d-balls of radius ε needed to cover T. Take $N(\varepsilon, T, d) = +\infty$ if no finite covering by d-balls of radius ε exists. The logarithm of the ε-covering number, $\log N(\varepsilon, T, d)$, is called the ε-*entropy*.
- Let (T, d) be a (subset of a) semi-metric space. For every $\varepsilon > 0$, the ε-*packing number* of (T, d), denoted by $D(\varepsilon, T, d)$, is defined as the maximum number of points in T such that the distance between each pair is at least ε. Take $D(\varepsilon, T, d) = +\infty$ if no such finite ε-packing exists. The logarithm of the ε-packing number, $\log D(\varepsilon, T, d)$, is called the ε-*entropy*.

Covering and packing numbers are related by the inequalities

$$N(\varepsilon, T, d) \le D(\varepsilon, T, d) \le N(\varepsilon/2, T, d).$$

Function spaces and probability

- For real number $1 \le p < +\infty$, let $L^p(\mathbb{R}) := \{f \,|\, f : \mathbb{R} \to \mathbb{C},\ f \text{ is Borel measurable}, \int |f|^p\,\mathrm{d}\lambda < +\infty\}$. For $f \in L^p(\mathbb{R})$, the L^p-norm of f is defined as $||f||_p := (\int |f|^p\,\mathrm{d}\lambda)^{1/p}$. The supremum norm of a function f is defined as $||f||_\infty := \sup_{x\in\mathbb{R}} |f(x)|$.
- For $f \in L^1(\mathbb{R})$, the complex-valued function $\hat{f}(t) := \int_{-\infty}^{+\infty} e^{itx} f(x)\,\mathrm{d}x$, $t \in \mathbb{R}$, is called the *Fourier transform of* f.

- 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, G (say), is used to denote a probability measure on a Borel-measurable space $(\mathscr{Y}, \mathscr{B}(\mathscr{Y}))$ and the corresponding cumulative distribution function (c.d.f.).
- The degenerate probability distribution putting mass one at a point $\theta \in \mathbb{R}$ is denoted by δ_θ.
- The notation Pf abbreviates 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.
- Given a r.v. Y with distribution G, the *moment generating function* of Y or the *Laplace transform of the probability measure* G is defined as

$$M_G(s) := E[e^{sY}] = \int_{\mathscr{Y}} e^{sy} \, \mathrm{d}G(y) \quad \text{for all } s \text{ for which the integral is finite.}$$

Metrics and divergences

- The *Hellinger distance* between any pair of probability density functions q_1 and q_2 on $\mathbb{R}$ is defined as $h(q_1, q_2) := \{\int (q_1^{1/2} - q_2^{1/2})^2 \, \mathrm{d}\lambda\}^{1/2}$, the L^2-distance between the square-root densities. The following inequalities, due to LeCam (1973), p. 40, relating the L^1-norm and the Hellinger distance hold:

$$h^2(q_1, q_2) \leq ||q_1 - q_2||_1 \tag{1}$$

and

$$||q_1 - q_2||_1 \leq 2h(q_1, q_2). \tag{2}$$

- For ease of notation, the same symbol d is used throughout to denote the L^1-norm, the L^2-norm or the Hellinger metric, the intended meaning being declared at each occurrence.
- For any probability measure Q on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ with density q, let

$$\mathrm{KL}(P_0 \| Q) := \begin{cases} \displaystyle\int \log \frac{\mathrm{d}P_0}{\mathrm{d}Q} \, \mathrm{d}P_0 = \int_{p_0 q > 0} p_0 \log \frac{p_0}{q} \, \mathrm{d}\lambda, & \text{if } P_0 \ll Q, \\ +\infty, & \text{otherwise,} \end{cases}$$

be the *Kullback–Leibler divergence* of Q from P_0 and, for $k \geq 2$, let

$$\mathrm{V}_k(P_0 \| Q) := \begin{cases} \displaystyle\int \left| \log \frac{\mathrm{d}P_0}{\mathrm{d}Q} \right|^k \mathrm{d}P_0 = \int_{p_0 q > 0} p_0 \left| \log \frac{p_0}{q} \right|^k \mathrm{d}\lambda, & \text{if } P_0 \ll Q, \\ +\infty, & \text{otherwise,} \end{cases}$$

be the kth absolute moment of $\log(\mathrm{d}P_0/\mathrm{d}Q)$. For any $\varepsilon > 0$ and a given $k \geq 2$, define a Kullback–Leibler type neighborhood of P_0 as

$$B_{\mathrm{KL}}(P_0;\ \varepsilon^k) := \left\{Q :\ \mathrm{KL}(P_0\|Q) \leq \varepsilon^2,\ \mathrm{V}_k(P_0\|Q) \leq \varepsilon^k\right\}.$$

- For any real number $p \geq 1$ 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 marginals G_1 and G_2 on the first and second arguments, respectively.

Stochastic order symbols

Let $(Z_n)_{n\in\mathbb{N}}$ be a sequence of real-valued random variables, possibly defined on entirely different probability spaces $(\Omega_n,\ \mathscr{F}_n,\ \mathbf{P}_n)_{n\in\mathbb{N}}$. Suppressing n in $\mathbf{P}$ causes no confusion if it is understood that $\mathbf{P}$ refers to whatever probability space Z_n is defined on. Let $(k_n)_{n\in\mathbb{N}}$ be a sequence of positive real numbers. We write

- $Z_n = O_{\mathbf{P}}(k_n)$ if $\lim_{T\to+\infty} \limsup_{n\to+\infty} \mathbf{P}(|Z_n| > Tk_n) = 0$. Then, $Z_n/k_n = O_{\mathbf{P}}(1)$,
- $Z_n = o_{\mathbf{P}}(k_n)$ if, for every $\varepsilon > 0$, $\lim_{n\to+\infty} \mathbf{P}(|Z_n| > \varepsilon k_n) = 0$. Then, $Z_n/k_n = o_{\mathbf{P}}(1)$.

Unless otherwise specified, in all stochastic order symbols used throughout, 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 $P_0^{\mathbb{N}}$.

2 Rates of convergence for L^1-Wasserstein deconvolution of Dirichlet–Laplace mixtures

In this section, we present some results on the Bayesian recovery of a distribution function from data contaminated with an additive random error following the standard Laplace distribution: we derive rates of convergence for the L^1-Wasserstein deconvolution of Dirichlet–Laplace mixture densities. The density is modeled as a Dirichlet–Laplace mixture

$$p_G(\cdot) \equiv (G * f)(\cdot) = \int_{\mathscr{Y}} f(\cdot - y)\,\mathrm{d}G(y),$$

with the kernel density f being the standard Laplace and the mixing distribution G being any probability measure on $(\mathscr{Y},\ \mathscr{B}(\mathscr{Y}))$, with $\mathscr{Y} \subseteq \mathbb{R}$. As a prior for G, we consider a Dirichlet process with base measure α on $(\mathscr{Y},\ \mathscr{B}(\mathscr{Y}))$, denoted by $\mathscr{D}_\alpha$. We recall that a Dirichlet process on a measurable space $(\mathscr{Y},\ \mathscr{B}(\mathscr{Y}))$, with finite and positive base measure α on $(\mathscr{Y}, \mathscr{B}(\mathscr{Y}))$, is a random probability measure $\tilde{G}$ on $(\mathscr{Y}, \mathscr{B}(\mathscr{Y}))$ such that, for every finite partition $(B_1, \ldots, B_k)$ of $\mathscr{Y}$, $k \geq 1$, the vector of random probabilities $(\tilde{G}(B_1), \ldots, \tilde{G}(B_k))$ has Dirichlet distribution with parameters $(\alpha(B_1), \ldots, \alpha(B_k))$. A Dirichlet process mixture of Laplace densities can be structurally described as follows:

- $\tilde{G} \sim \mathscr{D}_\alpha$,
- given $\tilde{G} = G$, the r.v.'s $Y_1, \ldots, Y_n$ are i.i.d. according to G,
- given $(G, Y_1, \ldots, Y_n)$, the r.v.'s $Z_1, \ldots, Z_n$ are i.i.d. according to f,
- sampled values from p_G are defined as $X_i := Y_i + Z_i$ for $i = 1, \ldots, n$.

Let the sampling density p_0 be itself a Laplace mixture with mixing distribution G_0, that is, $p_0 \equiv p_{G_0} = G_0 * f$. In order to assess the rate of convergence in the L^1-Wasserstein metric for the Bayes' estimator of the true mixing distribution G_0, we appeal to an inversion inequality relating the L^2-norm or the Hellinger distance between Laplace mixed densities to any L^p-Wasserstein distance, $p \geq 1$, between the corresponding mixing distributions, see Lemma 4 in Appendix D. Therefore, we first derive rates of contraction in the L^2-norm and the Hellinger metric for the posterior distribution of a Dirichlet–Laplace mixture prior: convergence of the posterior distribution at a rate ε_n, in fact, implies the existence of Bayes' point estimators that converge at least as fast as ε_n in the frequentist sense. The same indirect approach has been taken by Gao and van der Vaart (2016), who deal with the case of compactly supported mixing distributions, while we extend the results to mixing distributions possibly supported on the whole real line or on some unbounded subset thereof. We present two results on posterior contraction rates for a Dirichlet–Laplace mixture prior. The first one, as stated in Proposition 1, is relative to the L^1-norm or the Hellinger metric; the second one, as stated in Proposition 2, is relative to the L^2-metric. Proofs are deferred to Appendix C.

Proposition 1 *Let $X_1, \ldots, X_n$ be i.i.d. observations from a density $p_0 \equiv p_{G_0} = G_0 * f$, with the kernel density f being the standard Laplace and the mixing distribution G_0 such that, for some decreasing function $A_0 : (0, +\infty) \to [0, 1]$ and a constant $0 < c_0 < +\infty$,*

$$G_0([-T, T]^c) \leq A_0(T) \lesssim \exp(-c_0 T) \quad \textit{for large } T > 0. \tag{3}$$

If the baseline measure α of the Dirichlet process is symmetric around zero and possesses density α' such that, for some constants $0 < b < +\infty$ and $0 < \tau \leq 1$,

$$\alpha'(y) \propto \exp(-b|y|^\tau), \quad y \in \mathbb{R}, \tag{4}$$

then there exists a sufficiently large constant $M > 0$ such that

$$\Pi(d(p_G, p_0) \geq M n^{-3/8} \log^{5/8} n \mid X^{(n)}) = o_{\mathbf{P}}(1),$$

where $\Pi(\cdot \mid X^{(n)})$ denotes the posterior distribution corresponding to a Dirichlet–Laplace process mixture prior after n observations and d can be either the Hellinger or the L^1-metric.

Remark 1 In virtue of the following inequality,

$$\forall\, G,\ G' \in \mathscr{G},\quad \|p_G - p_{G'}\|_2^2 \le 4\|f\|_\infty h^2(p_G,\ p_{G'}),$$

where $\|f\|_\infty = 1/2$ for the standard Laplace kernel density, see (28) in Lemma 3, the L^2-metric posterior contraction rate for a Dirichlet–Laplace mixture prior could, in principle, be derived from Proposition 1, which relies on Theorem 2.1 of Ghosal et al. (2000), p. 503, or Theorem 2.1 of Ghosal and van der Vaart (2001), p. 1239, but this would impose slightly stronger conditions on the density α' of the baseline measure than those required in Proposition 2 below, which is based on Theorem 3 of Giné and Nickl (2011), p. 2892, that is tailored for assessing posterior contraction rates in L^r-metrics, $1 < r < +\infty$, taking an approach that can only be used if one has sufficiently fine control of the approximation properties of the prior support in the L^r-metric considered.

Proposition 2 *Let $X_1,\ \ldots,\ X_n$ be i.i.d. observations from a density $p_0 \equiv p_{G_0} = G_0 * f$, with the kernel density f being the standard Laplace and the mixing distribution G_0 such that condition (3) holds as in Proposition 1. If the baseline measure α of the Dirichlet process possesses continuous and positive density α' such that, for some constants $0 < b < +\infty$ and $0 < \tau \le 1$,*

$$\alpha'(y) \gtrsim \exp\left(-b|y|^\tau\right) \quad \textit{for large } |y|, \tag{5}$$

then there exists a sufficiently large constant $M > 0$ such that

$$\Pi(\|p_G - p_0\|_2 \ge M n^{-3/8} \log^{5/8} n \mid X^{(n)}) = o_{\mathbf{P}}(1), \tag{6}$$

where $\Pi(\cdot \mid X^{(n)})$ denotes the posterior distribution corresponding to a Dirichlet–Laplace process mixture prior after n observations.

As previously mentioned, convergence of the posterior distribution at a rate ε_n implies the existence of point estimators that converge at least as fast as ε_n in the frequentist sense, see, for instance, Theorem 2.5 in Ghosal et al. (2000), p. 506, for the construction of a point estimator that applies to general statistical models and posterior distributions. The posterior expectation of the density p_G, which we refer to as the Bayes' density estimator,

$$\hat{p}_n^{\mathrm{B}}(\cdot) := \int_{\mathscr{G}} p_G(\cdot)\Pi(\mathrm{d}G \mid X^{(n)}),$$

has a similar property when jointly considered with bounded semi-metrics that are convex or whose square is convex in one argument. When the random mixing distribution $\tilde{G}$ is distributed according to a Dirichlet process, the expression of the Bayes' density estimator $\hat{p}_n^{\mathrm{B}}$ is given by formula (2.6) of Lo (1984), p. 353, replacing $K(\cdot,\ u)$ with $\frac{1}{2}\exp\{-|\cdot - u|\}$ at each occurrence.

Corollary 1 *Suppose that condition* (3) *holds for some decreasing function* $A_0 : (0,\ +\infty) \to [0,\ 1]$ *and a finite constant* $c_0 > (1/e)$ *such that*

$$G_0([-T,\ T]^c) \le A_0(T) \lesssim \exp\left(-e^{c_0 T}\right) \quad \textit{for large } T > 0 \tag{7}$$

and condition (4) *holds as in Proposition* 1. *Then,*

$$d(\hat{p}_n^{\mathrm{B}},\ p_0) = O_{\mathbf{P}}(n^{-3/8} \log^{1/2} n),$$

for d *being either the Hellinger or the* L^1*-metric.*

Proof In virtue of the inequality in (2), it suffices to prove the assertion for the Hellinger metric. The proof follows standard arguments as, for instance, in Ghosal et al. (2000), pp. 506–507. By convexity of h^2 in each argument and Jensen's inequality, for $\varepsilon_n := \max\{\bar{\varepsilon}_n,\ \tilde{\varepsilon}_n\} = n^{-3/8}(\log n)^{(3\vee 4)/8} = n^{-3/8}\log^{1/2} n$ and a sufficiently large constant $M > 0$,

$$\begin{aligned} h^2(\hat{p}_n^{\mathrm{B}},\ p_0) &\le \int_{\mathscr{G}} h^2(p_G,\ p_0)\Pi(\mathrm{d}G \mid X^{(n)}) \\ &= \left(\int_{h(p_G,\ p_0) < M\varepsilon_n} + \int_{h(p_G,\ p_0) \ge M\varepsilon_n}\right) h^2(p_G,\ p_0)\Pi(\mathrm{d}G \mid X^{(n)}) \\ &\lesssim M^2\varepsilon_n^2 + 2\Pi(h(p_G,\ p_0) \ge M\varepsilon_n \mid X^{(n)}). \end{aligned}$$

It follows that

$$P_0^n h^2(\hat{p}_n^{\mathrm{B}},\ p_0) \lesssim M^2\varepsilon_n^2 + 2P_0^n\Pi(h(p_G,\ p_0) \ge M\varepsilon_n \mid X^{(n)}) \lesssim \varepsilon_n^2 + o(\varepsilon_n^2)$$

because we can apply the almost sure version of Theorem 7 in Scricciolo (2007), p. 636 (see also Theorem A.1 in Scricciolo (2006), p. 2918), which, under the prior mass condition

$$\Pi(h^2(p_G,\ p_0)\|p_0/p_G\|_\infty \le \tilde{\varepsilon}_n^2) \gtrsim \exp\left(-Bn\tilde{\varepsilon}_n^2\right), \tag{8}$$

with $\tilde{\varepsilon}_n := n^{-3/8}\log^{1/2} n$ and a constant $0 < B < +\infty$, yields exponentially fast convergence of the posterior distribution since $P_0^n\Pi(h(p_G,\ p_0) \ge M\varepsilon_n \mid X^{(n)}) \lesssim \exp\left(-B_1 n\tilde{\varepsilon}_n^2\right)$ for a suitable constant $0 < B_1 < +\infty$. To verify that condition (8) is satisfied, we can proceed as in the proof of Proposition 2: for any G satisfying (27), not only is $h(p_G,\ p_0) \lesssim \varepsilon$, but, under assumption (7) which guarantees that $M_{G_0}(-1) < +\infty$ and $M_{G_0}(1) < +\infty$, it also is

$$\begin{aligned} &||p_0/p_G||_\infty \le e^{a_\varepsilon}[M_{G_0}(-1) + M_{G_0}(1)] \lesssim \log(1/\varepsilon), \\ &\quad \text{for } a_\varepsilon := A_0^{-1}(\varepsilon^2) \lesssim \log\log(1/\varepsilon). \end{aligned}$$

Then,

$$\log\Pi(h^2(p_G,\ p_0)\|p_0/p_G\|_\infty \le \varepsilon^2\log(1/\varepsilon)) \gtrsim -\varepsilon^{-2/3}\log(1/\varepsilon).$$

Condition (8) is thus verified for $\tilde{\varepsilon}_n := \varepsilon \log^{1/2}(1/\varepsilon) = n^{-3/8}\log^{1/2} n$. Conclude that $h(\hat{p}_n^{\mathrm{B}},\ p_0) = O_{\mathbf{P}}(\varepsilon_n)$. □

Remark 2 Admittedly, condition (7) imposes a stringent constraint on the tail decay rate of G_0. An alternative sufficient condition for concluding that

$$P_0^n \Pi(d(p_G,\ p_0) \geq M\varepsilon_n \mid X^{(n)}) = o(\varepsilon_n^2), \quad \text{for } d = h \text{ or } d = \|\cdot\|_1, \tag{9}$$

is a prior mass condition involving the kth absolute moment of $\log(p_0/p_G)$ for a suitable value of k, in place of the sup-norm $\|p_0/p_G\|_\infty$, which can possibly induce a lighter condition on G_0. For $\tilde{\varepsilon}_n := n^{-3/8}\log^{\omega} n$, with $\omega > 0$, let $\varepsilon_n := \max\{\bar{\varepsilon}_n,\ \tilde{\varepsilon}_n\} = n^{-3/8}(\log n)^{(3/8)\vee\omega}$. It is known from Lemma 10 of Ghosal and van der Vaart (2007b), p. 220, that if

$$\Pi(B_{\mathrm{KL}}(P_0;\ \tilde{\varepsilon}_n^k)) \gtrsim \exp(-Bn\tilde{\varepsilon}_n^2), \quad k \geq 2, \tag{10}$$

then

$$P_0^n \Pi(d(p_G,\ p_0) \geq M\varepsilon_n \mid X^{(n)}) \lesssim (n\tilde{\varepsilon}_n^2)^{-k/2}. \tag{11}$$

Thus, if condition (10) holds for some $k \geq 6$ so that $(n\tilde{\varepsilon}_n^2)^{-k/2} = o(\varepsilon_n^2)$, the value $k = 6$ would suffice for the purpose, then condition (9) is satisfied.

We now state a result on the rate of convergence for the Bayes' estimator, denoted by $\hat{G}_n^{\mathrm{B}}$, of the mixing distribution G_0 for the L^1-Wasserstein deconvolution of Dirichlet–Laplace mixtures. The Bayes' estimator is the posterior expectation of the random probability measure $\tilde{G}$, that is, $\hat{G}_n^{\mathrm{B}}(\cdot) := E[\tilde{G}(\cdot) \mid X^{(n)}]$ and its expression can be derived from the expression of the posterior distribution, cf. Ghosh and Ramamoorthi (2003), pp. 144–146. In order to state the result, let $M_{\hat{G}_n^{\mathrm{B}}}(s) := \int_{-\infty}^{+\infty} e^{sy}\,\mathrm{d}\hat{G}_n^{\mathrm{B}}(y)$, $s \in \mathbb{R}$, whose expression can be obtained from formula (2.6) of Lo (1984), p. 353, replacing $K(x,\ u)$ with e^{su} at all occurrences (s playing the role of x).

Proposition 3 *Suppose that the assumptions of Corollary 1 hold. If, in addition,* $\bar{\alpha} := \alpha/\alpha(\mathbb{R})$ *has finite moment generating function on some interval* $(-s_0,\ s_0)$*, with* $0 < s_0 < 1$*, and*

$$\begin{aligned} &\forall\, 0 < s < s_0, \quad \limsup_{n\to+\infty} P_0^n M_{\hat{G}_n^{\mathrm{B}}}(-s) \leq M_{G_0}(-s) \\ &\textit{and} \quad \limsup_{n\to+\infty} P_0^n M_{\hat{G}_n^{\mathrm{B}}}(s) \leq M_{G_0}(s), \end{aligned} \tag{12}$$

then

$$W_1(\hat{G}_n^{\mathrm{B}},\ G_0) = O_{\mathbf{P}}(n^{-1/8}(\log n)^{2/3}). \tag{13}$$

Proof Let $\rho_n := n^{-1/8}(\log n)^{2/3}$ and, for a suitable finite constant $c_1 > 0$, $M_n = c_1(\log n)$. Fix numbers s and u such that $0 < u < s < s_0 < 1$. For sufficiently large constants $0 < T,\ T',\ T'' < +\infty$, reasoning as in Lemma 4,

$$\begin{aligned} P_0^n(W_1(\hat{G}_n^{\mathrm{B}},\ G_0) > T\rho_n) &\le P_0^n(h(\hat{p}_n^{\mathrm{B}},\ p_0) > T'\rho_n^3(\log n)^{-3/2}) \\ &\quad + P_0^n(M_{\hat{G}_n^{\mathrm{B}}}(-s) + M_{\hat{G}_n^{\mathrm{B}}}(s) > T''e^{uM_n}\rho_n) =: P_1 + P_2. \end{aligned}$$

By Corollary 1, $h(\hat{p}_n^{\mathrm{B}},\ p_0) = O_{\mathbf{P}}(n^{-3/8}\log^{1/2} n)$. Hence, $P_1 \to 0$ as $n \to +\infty$. By Markov's inequality, for some real $\nu > 0$,

$$\begin{aligned} P_2 &\lesssim e^{-uM_n}\rho_n^{-1}[P_0^n M_{\hat{G}_n^{\mathrm{B}}}(-s) + P_0^n M_{\hat{G}_n^{\mathrm{B}}}(s)] \\ &\lesssim \frac{1}{n^{\nu}}[P_0^n M_{\hat{G}_n^{\mathrm{B}}}(-s) + P_0^n M_{\hat{G}_n^{\mathrm{B}}}(s)] \to 0 \quad \text{as } n \to +\infty \end{aligned}$$

by assumption (12). Thus, $P_2 \to 0$ as $n \to +\infty$. The assertion follows. □

Some remarks are in order. There are two main reasons why we focus on deconvolution in the L^1-Wasserstein metric. The first one is related to the inversion inequality in (30), where the upper bound on the L^p-Wasserstein metric, as a function of the order $p \ge 1$, increases as p gets larger, thus making it advisable to begin the analysis from the smallest value of p. The second reason is related to the interpretation of the assertion in (13): the L^1-Wasserstein distance between any two probability measures G_1 and G_2 on some Borel-measurable space $(\mathscr{Y},\ \mathscr{B}(\mathscr{Y}))$, $\mathscr{Y} \subseteq \mathbb{R}$, with finite first absolute moments, is by itself an interesting distance because it metrizes weak convergence plus convergence of the first absolute moments, but it is even more interesting in view of the fact that, letting $G_1^{-1}(\cdot)$ and $G_2^{-1}(\cdot)$ denote the left-continuous inverse or quantile functions, $G_i^{-1}(u) := \inf\{y \in \mathscr{Y} :\ G_i(y) \ge u\}$, $u \in (0,\ 1)$, $i = 1,\ 2$, it can be written as the L^1-distance between the quantile functions or, equivalently, as the L^1-distance between the cumulative distribution functions,

$$\begin{aligned} W_1(G_1,\ G_2) &= \int_0^1 |G_1^{-1}(u) - G_2^{-1}(u)|\,\mathrm{d}u \\ &= \int_{\mathscr{Y}} |G_1(y) - G_2(y)|\,\mathrm{d}y = ||G_1 - G_2||_1, \end{aligned} \tag{14}$$

see, *e.g.*, Shorack and Wellner (1986), pp. 64–66. The representation in (14) was obtained by Dall'Aglio (1956). Thus, by rewriting $W_1(\hat{G}_n^{\mathrm{B}},\ G_0)$ as the L^1-distance between the c.d.f.'s $\hat{G}_n^{\mathrm{B}}$ and G_0, the assertion of Proposition 3,

$$W_1(\hat{G}_n^{\mathrm{B}},\, G_0) = ||\hat{G}_n^{\mathrm{B}} - G_0||_1 = O_{\mathbf{P}}(n^{-1/8}(\log n)^{2/3}),$$

becomes more transparent and meaningful.

3 Rates of convergence for ML estimation and L^1-Wasserstein deconvolution of Laplace mixtures

In this section, we first study the rate of convergence in the Hellinger metric for the MLE $\hat{p}_n$ of a Laplace mixture density $p_0 \equiv p_{G_0} = G_0 * f$, with unknown mixing distribution $G_0 \in \mathscr{G}$. We then derive the rate of convergence in the L^1-Wasserstein metric for the MLE $\hat{G}_n$ of the mixing distribution G_0, which corresponds to the MLE $\hat{p}_n$ of the mixed density p_0, by appealing to an inversion inequality relating the Hellinger distance between Laplace mixture densities to any L^p-Wasserstein distance, $p \geq 1$, between the corresponding mixing distributions (see Lemma 4 in Appendix D).

A MLE $\hat{p}_n$ of p_0 is a measurable function of the observations taking values in $\mathscr{P} := \{p_G : G \in \mathscr{G}\}$ such that

$$\hat{p}_n \in \underset{p_G \in \mathscr{P}}{\arg\max} \frac{1}{n}\sum_{i=1}^{n} \log p_G(X_i) = \underset{p_G \in \mathscr{P}}{\arg\max} \int (\log p_G)\,\mathrm{d}\mathbb{P}_n,$$

where $\mathbb{P}_n := n^{-1}\sum_{i=1}^{n} \delta_{X_i}$ is the empirical measure 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. We assume that the MLE exists, but do not require it to be unique, see Lindsay (1995), Theorem 18, p. 112, for sufficient conditions ensuring uniqueness.

Results on rates of convergence in the Hellinger metric for the MLE of a density can be found in Birgé and Massart (1993), Van de Geer (1993) and Wong and Shen (1995); it can, however, be difficult to calculate the L^2-metric entropy *with bracketing* of the square-root densities that is employed in these articles. Taking instead into account that a mixture model $\{\int_{\mathscr{Y}} K(\cdot,\, y)\,\mathrm{d}G(y) : G \in \mathscr{G}\}$ is the closure of the convex hull of the collection of kernels $\{K(\cdot,\, y) : y \in \mathscr{Y} \subseteq \mathbb{R}\}$, which is typically a much smaller class, a bound on a form of metric entropy *without bracketing* of the class of mixtures can be derived from a covering number of the class of kernels (a result on metric entropy *without bracketing* of convex hulls that is deducible from Ball and Pajor 1990), so that a relatively simple "recipe" can be given to obtain (an upper bound on) the rate of convergence in the Hellinger metric for the MLE of a density in terms of the "dimension" of the class of kernels and the behaviour of p_0 near zero, cf. Corollary 2.3 of Van de Geer (1996), p. 298.

Proposition 4 *Let the sampling density $p_0 \equiv p_{G_0} = G_0 * f$, with the kernel density f being the standard Laplace and the mixing distribution $G_0 \in \mathscr{G}$. Suppose that, for a sequence of non-negative real numbers $\sigma_n = O(n^{-3/8}\log^{1/8} n)$, we have*

(a) $\int_{p_0 \le \sigma_n} p_0 \, \mathrm{d}\lambda \lesssim \sigma_n^2$,
(b) $\int_{p_0 > \sigma_n} (1/p_0) \, \mathrm{d}\lambda \lesssim \log(1/\sigma_n)$.

Then,

$$h(\hat{p}_n,\ p_0) = O_{\mathbf{P}}(n^{-3/8} \log^{1/8} n).$$

Proof We begin by spelling out the remark mentioned in the introduction concerning the fact that a mixture model is the closure of the convex hull of the collection of kernels. Recall that the convex hull of a class $\mathscr{K}$ of functions, denoted by $\mathrm{conv}(\mathscr{K})$, is defined as the set of all finite convex combinations of functions in $\mathscr{K}$,

$$\mathrm{conv}(\mathscr{K}) := \left\{ \sum_{j=1}^{r} \theta_j K_j : \theta_j \ge 0,\ K_j \in \mathscr{K},\ j = 1,\ \ldots,\ r,\ \sum_{j=1}^{r} \theta_j = 1,\ r \in \mathbb{N} \right\}.$$

In our case,

$$\mathscr{K} := \{ f(\cdot - y) : \ y \in \mathscr{Y} \subseteq \mathbb{R} \}$$

is the collection of kernels with f the standard Laplace density. The class $\mathscr{P} := \{p_G : G \in \mathscr{G}\}$ of all Laplace convolution mixtures $p_G = G * f$ is the closure of the convex hull of $\mathscr{K}$,

$$\mathscr{P} = \overline{\mathrm{conv}}(\mathscr{K}).$$

Clearly, $\mathscr{P}$ is itself a convex class. This remark enables us to apply Theorem 2.2 and Corollary 2.3 of Van de Geer (1996), pp. 297–298 and 310, or, equivalently, Theorem 7.7 of Van de Geer (2000), pp. 104–105, whose conditions are hereafter shown to be satisfied. To the aim, we define the class

$$\mathscr{K}/p_0 := \left\{ \frac{f(\cdot - y)}{p_0(\cdot)} \mathbf{1}\{p_0 > \sigma_n\} : \ y \in \mathscr{Y} \right\}$$

and the envelope function

$$\bar{K}(\cdot) := \sup_{y \in \mathscr{Y}} \frac{f(\cdot - y)}{p_0(\cdot)} \mathbf{1}\{p_0 > \sigma_n\},$$

where we have suppressed the subscript n in $\mathscr{K}/p_0$ and $\bar{K}(\cdot)$ stressing possible dependence on σ_n when $\sigma_n > 0$. Since, by assumption (a),

$$\int_{p_0 \le \sigma_n} \mathrm{d}P_0 = \int_{p_0 \le \sigma_n} p_0 \, \mathrm{d}\lambda \lesssim \sigma_n^2$$

and, by assumption (b), together with the fact that $\|f\|_\infty = 1/2$,

$$\int \bar{K}^2 \, \mathrm{d}P_0 \lesssim \int_{p_0 > \sigma_n} \frac{1}{p_0} \, \mathrm{d}\lambda \lesssim \log(1/\sigma_n), \tag{15}$$

we can take the sequence $\delta_n^2 \propto \sigma_n^2$ in condition (7.21) of Theorem 7.7 of Van de Geer (2000), p. 104. Because the (standard) Laplace kernel density f is Lipschitz,

$$\forall\, y_1,\ y_2 \in \mathscr{Y}, \quad |f(\cdot - y_1) - f(\cdot - y_2)| \le \frac{1}{2}|y_1 - y_2|,$$

see, *e.g.*, Lemma A.1 in Scricciolo (2011), pp. 299–300, on the set

$$\int \bar{K}^2\, \mathrm{d}\mathbb{P}_n \le T^2 \log(1/\delta_n), \tag{16}$$

where $T > 0$ is a finite constant, we find that, for $\mathrm{d}\mathbb{Q}_n := \mathrm{d}\mathbb{P}_n/(T^2 \log(1/\delta_n))$,

$$N(\delta,\ \mathscr{K}/p_0,\ ||\cdot||_{2,\mathbb{Q}_n}) \lesssim \delta^{-1} \quad \text{for } \delta > 0,$$

where $||\cdot||_{2,\mathbb{Q}_n}$ denotes the $L^2(\mathbb{Q}_n)$-norm, that is, $||g||_{2,\mathbb{Q}_n} := (\int |g|^2\, \mathrm{d}\mathbb{Q}_n)^{1/2}$. So, in view of the result of Ball and Pajor (1990), reported as Theorem 1.1 in Van de Geer (1996), p. 295, on the same set as in (16), we have

$$\log N(\delta,\ \overline{\mathrm{conv}}(\mathscr{K}/p_0),\ ||\cdot||_{2,\mathbb{Q}_n}) \lesssim \delta^{-2/3},$$

hence

$$\log N(\delta,\ \overline{\mathrm{conv}}(\mathscr{K}/p_0),\ ||\cdot||_{2,\mathbb{P}_n}) \lesssim \left(\frac{T \log^{1/2}(1/\delta_n)}{\delta}\right)^{2/3}.$$

Next, defined the class

$$\mathscr{P}_{\sigma_n}^{(\mathrm{conv})} := \left\{\frac{2p_G}{p_G + p_0}\mathbf{1}\{p_0 > \sigma_n\} :\ p_G \in \mathscr{P}\right\}$$

considered in condition (7.20) of Theorem 7.7 in Van de Geer (2000), p. 104, since

$$\log N(2\delta,\ \mathscr{P}_{\sigma_n}^{(\mathrm{conv})},\ ||\cdot||_{2,\mathbb{P}_n}) \le \log N(\delta,\ \overline{\mathrm{conv}}(\mathscr{K}/p_0),\ ||\cdot||_{2,\mathbb{P}_n}),$$

in view of (15), we have

$$\sup_{\delta > 0} \frac{\log N(\delta,\ \mathscr{P}_{\sigma_n}^{(\mathrm{conv})},\ ||\cdot||_{2,\mathbb{P}_n})}{H(\delta)} = O_{\mathbf{P}}(1)$$

for the non-increasing function of δ

$$H(\delta) := \delta^{-2/3} \log^{1/3}(1/\delta_n), \quad \delta > 0.$$

Taken $\Psi(\delta) := c_1\delta^{2/3}\log^{1/6}(1/\delta_n)$ with a suitable finite constant $c_1 > 0$, we have

$$\forall\, \delta \in (0,\ 1),\quad \Psi(\delta) \geq \left(\int_{\delta^2/c}^{\delta} H^{1/2}(u)\,\mathrm{d}u\right) \vee \delta$$

and, for some $\varepsilon > 0$, $\Psi(\delta)/\delta^{2-\varepsilon}$ is non-increasing. Then, for δ_n such that $\sqrt{n}\delta_n^2 \geq \Psi(\delta_n)$, cf. condition (7.22) of Theorem 7.7 in Van de Geer (2000), p. 104, which implies that, consistently with the initial choice, we can take $\delta_n \propto n^{-3/8}\log^{1/8} n$, we have $h(\hat{p}_n,\ p_0) = O_{\mathbf{P}}(\delta_n)$ and the proof is complete. □

Remark 3 If $p_0 > 0$ and $\mathscr{Y}$ is a compact interval $[-a,\ a]$, with $a > 0$, then $h(\hat{p}_n,\ p_0) = O_{\mathbf{P}}(n^{-3/8})$. In fact, the sequence $\sigma_n \equiv 0$, $||\bar{K}||_\infty \leq e^{2a}$ and $\int \bar{K}^2\,\mathrm{d}P_0 \leq e^{4a}$ so that, on the set $\{\int \bar{K}^2\,\mathrm{d}\mathbb{P}_n \leq T\}$, the entropy $\log N(\delta,\ \overline{\mathrm{conv}}(\mathscr{K}/p_0),\ ||\cdot||_{2,\mathbb{P}_n}) \lesssim \delta^{-2/3}$ and, reasoning as in Proposition 4, we find the rate $n^{-3/8}$.

We now derive a consequence of Proposition 4 on the rate of convergence in the L^1-Wasserstein metric for the MLE of G_0. A MLE $\hat{p}_n$ of the *mixed* density p_0 corresponds to a MLE $\hat{G}_n$ of the *mixing* distribution G_0, that is, $\hat{p}_n \equiv p_{\hat{G}_n}$, such that

$$\hat{G}_n \in \arg\max_{G\in\mathscr{G}} \frac{1}{n}\sum_{i=1}^{n} \log p_G(X_i) = \arg\max_{G\in\mathscr{G}} \int (\log p_G)\,\mathrm{d}\mathbb{P}_n.$$

Clearly, $\hat{G}_n$ is a discrete distribution, but we do not know the number of its components: Lindsay (1995) showed that the MLE $\hat{G}_n$ is a discrete distribution supported on at most $k \leq n$ support points, k being the number of distinct observed values or data points.

Corollary 2 *Suppose that the assumptions of Proposition 4 hold. If, in addition, the mixing distribution G_0 has finite moment generating function in some interval $(-s_0,\ s_0)$, with $0 < s_0 < 1$, and*

$$\begin{aligned}\forall\, 0 < s < s_0,\quad \limsup\nolimits_{n\to+\infty} P_0^n M_{\hat{G}_n}(-s) \leq M_{G_0}(-s) \quad \textit{and}\\ \limsup\nolimits_{n\to+\infty} P_0^n M_{\hat{G}_n}(s) \leq M_{G_0}(s),\end{aligned} \tag{17}$$

where $M_{\hat{G}_n}(s) := \int_{\mathscr{Y}} e^{sy}\,\mathrm{d}\hat{G}_n(y)$, $s \in \mathbb{R}$, then

$$W_1(\hat{G}_n,\ G_0) = O_{\mathbf{P}}(n^{-1/8}(\log n)^{13/24}).$$

Proof Let $k_n := n^{-1/8}(\log n)^{13/24}$ and, for a suitable finite constant $c_2 > 0$, $M_n = c_2(\log n)$. Fix numbers s and u such that $0 < u < s < s_0 < 1$. For sufficiently large constants $0 < T,\ T',\ T'' < +\infty$, reasoning as in Lemma 4, we have

$$\begin{aligned}&P_0^n(W_1(\hat{G}_n,\ G_0) > Tk_n)\\ &\quad\leq P_0^n(h(\hat{p}_n,\ p_0) > T'k_n^3(\log n)^{-3/2})\\ &\qquad+ P_0^n(M_{\hat{G}_n}(-s) + M_{\hat{G}_n}(s) > T''k_n e^{uM_n}) =: P_1 + P_2.\end{aligned}$$

The term P_1 can be made arbitrarily small because $h(\hat{p}_n,\ p_0) = O_{\mathbf{P}}(n^{-3/8}\log^{1/8} n)$ by Proposition 4. The term P_2 goes to zero as $n \to +\infty$: in fact, by Markov's inequality and assumption (17), for some real $0 < l < +\infty$,

$$\begin{aligned} P_2 &\lesssim e^{-uM_n} k_n^{-1}[P_0^n M_{\hat{G}_n}(-s) + P_0^n M_{\hat{G}_n}(s)] \\ &\lesssim \frac{1}{n^l}[P_0^n M_{\hat{G}_n}(-s) + P_0^n M_{\hat{G}_n}(s)] \to 0 \quad \text{as } n \to +\infty \end{aligned}$$

and the assertion follows. □

Remark 4 Assumption (17) essentially requires that $M_{\hat{G}_n}$ is an asymptotically unbiased estimator of M_{G_0} in some neighborhood of zero $(-s_0,\ s_0)$, with $0 < s_0 < 1$. An analysis of the asymptotic behaviour of certain linear functionals of the MLE $\hat{G}_n$ is presented in Van de Geer (1995), wherein sufficient conditions are provided so that they are $\sqrt{n}$-consistent, asymptotically normal and efficient.

4 Merging of Bayes and ML for L^1-Wasserstein deconvolution of Laplace mixtures

In this section, we show that the Bayes' estimator and the MLE of G_0 merge in the L^1-Wasserstein metric, their discrepancy vanishing, at worst, at rate $n^{-1/8}(\log n)^{2/3}$ because they both consistently estimate G_0 at a speed which is within a $(\log n)$-factor of $n^{-1/8}$, cf. Proposition 3 and Corollary 2.

Proposition 5 *Under the assumptions of Proposition 3 and Corollary 2, we have*

$$W_1(\hat{G}_n^{\mathrm{B}},\ \hat{G}_n) = O_{\mathbf{P}}(n^{-1/8}(\log n)^{2/3}). \tag{18}$$

Proof By the triangle inequality,

$$W_1(\hat{G}_n^{\mathrm{B}},\ \hat{G}_n) \le W_1(\hat{G}_n^{\mathrm{B}},\ G_0) + W_1(G_0,\ \hat{G}_n),$$

where $W_1(\hat{G}_n^{\mathrm{B}},\ G_0) = O_{\mathbf{P}}(n^{-1/8}(\log n)^{2/3})$ and $W_1(G_0,\ \hat{G}_n) = O_{\mathbf{P}}(n^{-1/8}(\log n)^{13/24})$ by Proposition 3 and Corollary 2, respectively. Relationship (18) follows. □

Proposition 5 states that the Bayes' estimator and the MLE of G_0 will eventually be indistinguishable and (an upper bound on) the speed of convergence for their L^1-Wasserstein discrepancy is determined by the stochastic orders of their errors in recovering G_0. The crucial question that remains open is whether the Bayes' estimator and the MLE are rate-optimal. Concerning this issue, we note that, on the one hand, other deconvolution estimators for the distribution function attain the rate $n^{-1/8}$ when the error distribution is the standard Laplace, with the proviso, however, that the L^1-Wasserstein metric is not linked to the integrated quadratic risk between the c.d.f.'s used in the result we are going to mention, so that the rates are not comparable. For instance, the estimator $G_n^K(h_n)(y) := \int_{-\infty}^{y} p_n^K(h_n)(u)\,\mathrm{d}u$, $y \in \mathbb{R}$, of the c.d.f. G_0 based on the standard deconvolution kernel density estimator is such that $\{\int_{-\infty}^{+\infty} E[G_n^K(h_n)(y) - G_0(y)]^2 \, \mathrm{d}y\}^{1/2} = O(n^{-1/8})$ when no assumptions on G_0 are postulated, except for the existence of the first absolute moment, see (3.12) in Corollary 3.3 of Hall and Lahiri (2008), p. 2117. On the other hand, a recent lower bound result, due to Dedecker and Michel (2015), Theorem 4.1, pp. 246–248, suggests that better rates are possible. For $M > 0$ and $r \geq 1$, let $\mathscr{D}(M, r)$ be the class of all probability measures G on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ such that $\int_{-\infty}^{+\infty} |y|^r \, \mathrm{d}G(y) \leq M$. Let f be the error density. Assume that there exist $\beta > 0$ and $c > 0$ such that, for every $\ell \in \{0, 1, 2\}$, it holds $|\hat{f}^{(\ell)}(t)| \leq c(1 + |t|)^{-\beta}$, $t \in \mathbb{R}$. Then, there exists a finite constant $C > 0$ such that, for *any* estimator $\hat{G}_n$ (we warn the reader of the clash of notation with the symbol $\hat{G}_n$ previously used to denote the MLE of G_0),

$$\liminf_{n \to +\infty} n^{p/(2\beta+1)} \sup_{G \in \mathscr{D}(M, r)} E W_p^p(\hat{G}_n, G) > C.$$

For $p = 1$ and the (standard) Laplace error distribution, this renders the lower bound $n^{-1/5}$, which is better than the leading term $n^{-1/8}$ of the upper bounds we have found, even if it is not said that either the Bayes' estimator or the MLE attains it.

Finally, a remark on the use of the term "merging". Even if this term is herein declined with a different meaning from that considered in Barron (1988), where merging is intended as the convergence to one of the ratio of the marginal likelihood to the joint density of the first n observations, or from that in Diaconis and Freedman (1986), where merging refers to the "intersubjective agreement", as more and more data become available, between two Bayesians with different prior opinions, the underlying idea is, in a broad sense, the same: different inferential procedures become essentially indistinguishable for large sample sizes.

5 Final remarks

In this note, we have studied rates of convergence for Bayes and maximum likelihood estimation of Laplace mixtures and for their L^1-Wasserstein deconvolution. The result on the convergence rate in the Hellinger metric for the MLE of Laplace mixtures is achieved taking a different approach from that adopted in Ghosal and van der Vaart (2001), which is based on the L^1-metric entropy with bracketing of the set of densities under consideration and is difficult to apply in the present context, due to the non-analyticity of the Laplace density. Posterior contraction rates for Dirichlet–Laplace mixtures have been previously studied by Gao and van der Vaart (2016) in the case of compactly supported mixing distributions and have been here extended to mixing distributions with a possibly unbounded set of locations, this accounting for the derivation of more general entropy estimates, cf. Appendix B. An interesting extension to pursue would be that of considering general kernel densities with polynomially decaying Fourier transforms in the sense of Definition 1: indeed, in the proof of Proposition 2, which gives an assessement of the posterior contraction rate in the L^2-metric for Dirichlet–Laplace mixtures, all conditions, except for the Kullback–Leibler prior mass requirement, hold for any kernel density as in Definition 1, provided that $\beta > 1$. The missing piece is an extension of Lemma 2 in Gao and van der Vaart

(2016), pp. 615–616, which is preliminary for checking the Kullback–Leibler prior mass condition and guarantees that a Laplace mixture, with mixing distribution that is the re-normalized restriction of G_0 to a compact interval, can be approximated in the Hellinger metric by a Laplace mixture with a discrete mixing distribution having a sufficiently restricted number of support points. We believe that, as for the Laplace kernel, the number of support points of the approximating mixing distribution will ultimately depend only on the decay rate of the Fourier transform of the kernel density, even though, in a general proof, the explicit expression of the kernel density cannot be exploited as in the Laplace case. Extending the result on posterior contraction rates to general kernel mixtures would be of interest in itself and for extending the L^1-Wasserstein deconvolution result, even though this would pose in more general terms the rate-optimality question, as it happens for the $n^{-1/8}$-rate in the Laplace case, see the remarks at the end of Sect. 4. We hope to report on these issues in a follow-up contribution.

Acknowledgements The author would like to thank the Editor and an anonymous Referee for their careful reading of the manuscript and helpful comments that have led to an improved presentation of the results. She gratefully acknowledges financial support from MIUR, Grant n° 2015SNS29B "Modern Bayesian nonparametric methods".

Appendix A: Auxiliary results

In this section, a sufficient condition on a convolution kernel $K \in L^1(\mathbb{R})$ is stated in terms of its Fourier transform $\hat{K}$ so that the exact order of the L^2-norm error for approximating any probability density f, with polynomially decaying characteristic function $\hat{f}$ of degree $\beta > 1/2$ (see Definition 1 below) by its convolution with $K_h := h^{-1}K(\cdot/h)$, that is, by $f * K_h$, is assessed in terms of the bandwidth h. The result is instrumental to the proof of Proposition 2 to show that any mixture density $p_G = G * f$, irrespective of the mixing distribution $G \in \mathscr{G}$, verifies the *bias* condition $||p_G * K_h - p_G||_2 = O(h^{\beta-1/2})$, which is involved in the definition of the sieve set in (15) of Theorem 2 in Giné and Nickl (2011), p. 2891. We refer to the difference $(p_G * K_h - p_G)$ as the *bias* because it is indeed the bias of the kernel density estimator $p_n^K(h) := \mathbb{P}_n * K_h$, when the observations are sampled from p_G: in fact, the bias $b[p_n^K(h)] := E[p_n^K(h)] - p_G = p_G * K_h - p_G$. The condition in (20) below, which traces back to Watson and Leadbetter (1963), see the first Theorem of Sect. 3B, pp. 486–487, is verified for any kernel K of order r greater than or equal to β, as later on spelled out in Remark 5.

Definition 1 Let f be a probability density function on $\mathbb{R}$. The Fourier transform of f or the characteristic function of the corresponding probability measure on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$, denoted by $\hat{f}$, is said to decrease algebraically of degree $\beta > 0$ if there exists a constant $0 < B_f < +\infty$ such that

$$\lim_{|t|\to+\infty} |t|^{\beta}|\hat{f}(t)| = B_f. \tag{19}$$

Relationship (19) describes the tail behaviour of $|\hat{f}|$ by stating that it decays polynomially as $|t|^{-\beta}$. The class of probability measures on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ that have characteristic functions satisfying condition (19) includes

- any gamma distribution with shape and scale parameters $\nu > 0$ and $\lambda > 0$, respectively, whose characteristic function has expression $(1 + it/\lambda)^{-\nu}$, the role of β in (19) being played by ν;
- any distribution with characteristic function $(1 + |t|^{\alpha})^{-1}$, $t \in \mathbb{R}$, for $0 < \alpha \leq 2$, which is called an *α-Laplace distribution* or *Linnik's distribution*, cf. Devroye (1990); the case $\alpha = 2$ renders the characteristic function of a standard Laplace distribution. The role of β in (19) is played by α;
- any distribution with characteristic function $(1 + |t|^{\alpha})^{-1/\beta}$, which, for $\beta = 1$, reduces to that of an α-Laplace distribution. The exponent α/β plays the role of the polynomial's degree β in (19). Devroye (1990) observes that, if S_α is any symmetric stable r.v. with characteristic function $e^{-|t|^{\alpha}}$, $0 < \alpha \leq 2$, and V_β is an independent r.v. with density $e^{-v^{\beta}}/\Gamma(1 + 1/\beta)$, $v > 0$, then the r.v. $S_\alpha V_\beta^{\beta/\alpha}$ has characteristic function $(1 + |t|^{\alpha})^{-1/\beta}$.

Lemma 1 *Let $f \in L^2(\mathbb{R})$ be a probability density function with Fourier transform $\hat{f}$ satisfying condition (19) for some $\beta > 1/2$ and a constant $0 < B_f < +\infty$. If $K \in L^1(\mathbb{R})$ has Fourier transform $\hat{K}$ such that $\hat{K}(0) = 1$ and*

$$I_\beta^2[\hat{K}] := \int_{\{t \neq 0\}} \frac{|1 - \hat{K}(t)|^2}{|t|^{2\beta}}\, \mathrm{d}t < +\infty, \tag{20}$$

then

$$h^{-2(\beta-1/2)} \| f - f * K_h \|_2^2 \to \frac{1}{2\pi} \times B_f^2 \times I_\beta^2[\hat{K}] \quad \textit{as } h \to 0.$$

Proof Since it is assumed that $f \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$, then $\hat{f} \in L^2(\mathbb{R})$ and necessarily $\beta > 1/2$. Also, as $K \in L^1(\mathbb{R})$, then $\|f * K_h\|_p \leq \|f\|_p \|K_h\|_1 < +\infty$ for $p = 1, 2$. Thus, $(f - f * K_h) \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ and, by Plancherel's Theorem, $\|f - f * K_h\|_2^2 = (2\pi)^{-1} \|\hat{f} - \hat{f} \times \hat{K}_h\|_2^2$. By the change of variable $z = ht$,

$$\begin{aligned} \|f - f * K_h\|_2^2 &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} |\hat{f}(t)|^2 |1 - \hat{K}(ht)|^2\, \mathrm{d}t \\ &= \frac{1}{2\pi} h^{2(\beta-1/2)} \left\{ B_f^2 \times I_\beta^2[\hat{K}] + \int_{\{z \neq 0\}} \frac{|1 - \hat{K}(z)|^2}{|z|^{2\beta}} \left[|z/h|^{2\beta} |\hat{f}(z/h)|^2 - B_f^2 \right] \mathrm{d}z \right\}, \end{aligned}$$

where, for every sequence of positive real numbers $h_n \to 0$, the integral on the right-hand side of the last display tends to zero by the dominated convergence theorem due to assumption (20). The assertion follows. □

In the following remark, which is essentially due to Davis (1977), cf. Sect. 3, pp. 532–533, sufficient conditions on a kernel $K \in L^1(\mathbb{R})$ are given so that $\hat{K}(0) = 1$ and the requirement in (20) is satisfied. The conditions in (21) below require that K is a *kernel of order* $r \geq \beta > 1/2$, the order of a kernel being the first non-zero "moment" of the kernel, cf. Definition 1.3 in Tsybakov (2004), p. 5.

Remark 5 For $K \in L^1(\mathbb{R})$, the Fourier transform $\hat{K}$ is continuous and bounded so that the integral $\int_{-\infty}^{+\infty} |t|^{-2\beta}|1-\hat{K}(t)|^2\mathbf{1}_{[1,\,+\infty)}(|t|)\,\mathrm{d}t < +\infty$ for $\beta > 1/2$. The problem with condition (20) is therefore the integrability of the function $t \mapsto |t|^{-2\beta}|1-\hat{K}(t)|^2$ for $|t| \in (0,\ 1)$. Suppose that

$$\begin{aligned}
&\int_{-\infty}^{+\infty} K(x)\,\mathrm{d}x = 1,\\
&\exists\, r \in \mathbb{N}, r \geq \beta > \tfrac{1}{2}: \int_{-\infty}^{+\infty} x^j K(x)\,\mathrm{d}x = 0 \ \text{ for } j = 1,\ \ldots,\ r-1 \ \text{ only if } \ r \geq 2,\\
&\text{and} \int_{-\infty}^{+\infty} x^r K(x)\,\mathrm{d}x \neq 0
\end{aligned} \tag{21}$$

and

$$\int_{-\infty}^{+\infty} |x|^r |K(x)|\,\mathrm{d}x < +\infty, \tag{22}$$

(the value r being called the *characteristic exponent* of $\hat{K}$, see Parzen (1962), pp. 1072–1073), then

$$\hat{K}(0) = 1 \ \text{ and } \ \int_{-\infty}^{+\infty} |t|^{-2\beta}|1-\hat{K}(t)|^2\mathbf{1}_{(0,\,1)}(|t|)\,\mathrm{d}t < +\infty.$$

In fact, $\hat{K}(0) = \int_{-\infty}^{+\infty} K(x)\,\mathrm{d}x = 1$. Also, for every real number $t \neq 0$,

$$\begin{aligned}
\frac{1-\hat{K}(t)}{t^r} = -\frac{\hat{K}(t)-1}{t^r} &= -\frac{1}{t^r}\int_{-\infty}^{+\infty} (e^{itx}-1)K(x)\,\mathrm{d}x\\
&= -\frac{1}{t^r}\int_{-\infty}^{+\infty} \left[e^{itx} - \sum_{j=0}^{r-1}\frac{(itx)^j}{j!}\right]K(x)\,\mathrm{d}x\\
&= -\frac{i^r}{(r-1)!}\int_{-\infty}^{+\infty} x^r K(x)\int_0^1 (1-u)^{r-1}e^{itux}\,\mathrm{d}u\,\mathrm{d}x.
\end{aligned}$$

By the dominated convergence theorem, condition (22) implies that

$$\frac{1-\hat{K}(t)}{t^r} \to -\frac{i^r}{r!}\int_{-\infty}^{+\infty} x^r K(x)\,\mathrm{d}x \quad \text{as } t \to 0,$$

where the limit is non-zero in virtue of the last condition on the right-hand side of (21). It is seen by comparison that, since $r \geq \beta$, the integral $\int_{-\infty}^{+\infty} |t|^{-2\beta}|1-\hat{K}(t)|^2\mathbf{1}_{(0,\,1)}(|t|)\,\mathrm{d}t < +\infty$ and condition (20) is satisfied. If, for instance, $1/2 <$

$\beta \le 2$, then any symmetric probability density K on $\mathbb{R}$, with finite, non-zero second moment $\mu_2 := \int_{-\infty}^{+\infty} x^2 K(x)\,\mathrm{d}x \neq 0$ is such that $I_\beta^2[\hat{K}] < +\infty$.

Appendix B: Entropy estimates

In this section, Hellinger and L^1-metric entropy estimates for a class of Laplace mixture densities, with mixing distributions having tails dominated by a given decreasing function, are provided. The result of Lemma 2 extends, along the lines of Theorem 7 in Ghosal and van der Vaart (2007a), pp. 708–709, Proposition 2 of Gao and van der Vaart (2016), p. 617, which deals with Laplace mixtures having compactly supported mixing distributions. Lemma 2 is invoked in the proof of Proposition 1, reported in Appendix C, to verify that the entropy condition is satisfied.

Lemma 2 *For a given decreasing function* $A : (0,\ +\infty) \to [0,\ 1]$, *with inverse* A^{-1}, *define the class of Laplace mixture densities*

$$\mathscr{P}_A := \{p_G :\ G([-a,\ a]^c) \le A(a)\ \textit{for all}\ a > 0\}.$$

Then, for every $0 < \varepsilon < 1$,

- *taking* $a \equiv a_\varepsilon := A^{-1}(\varepsilon)$ *in the definition of* $\mathscr{P}_A$, *we have*

$$\log N(3\varepsilon,\ \mathscr{P}_A,\ ||\cdot||_1) \lesssim \varepsilon^{-2/3} \log \frac{A^{-1}(\varepsilon)}{\varepsilon^2}, \tag{23}$$

- *taking* $a \equiv a_{\varepsilon^2} := A^{-1}(\varepsilon^2)$ *in the definition of* $\mathscr{P}_A$, *we have*

$$\log N((\sqrt{2}+1)\varepsilon,\ \mathscr{P}_A,\ h) \lesssim \varepsilon^{-2/3} \log \frac{A^{-1}(\varepsilon^2)}{\varepsilon^2}. \tag{24}$$

Proof Concerning the L^1-metric entropy in (23), since $a \equiv a_\varepsilon := A^{-1}(\varepsilon)$ satisfies $G([-a_\varepsilon,\ a_\varepsilon]^c) \le A(a_\varepsilon) = \varepsilon$ for all G as in the definition of $\mathscr{P}_A$, Lemma A.3 of Ghosal and van der Vaart (2001), p. 1261, implies that the L^1-distance between any density $p_G \in \mathscr{P}_A$ and the corresponding density p_{G^*}, with mixing distribution G^* defined as the re-normalized restriction of G to $[-a_\varepsilon,\ a_\varepsilon]$, is bounded above by 2ε. Then, in virtue of the inequality in (2), a Hellinger $(\varepsilon/2)$-net over the class of densities $\mathscr{P}_{a_\varepsilon} := \{p_G :\ G([-a_\varepsilon,\ a_\varepsilon]) = 1\}$ is an L^1-metric 3ε-net over $\mathscr{P}_A$, where

$$\log N\big(\varepsilon/2,\ \mathscr{P}_{a_\varepsilon},\ h\big) \lesssim \varepsilon^{-2/3} \log \frac{a_\varepsilon}{\varepsilon^2}$$

by Proposition 2 of Gao and van der Vaart (2016), p. 617. The inequality in (23) follows.

Concerning the Hellinger-metric entropy in (24), by taking $a \equiv a_{\varepsilon^2} := A^{-1}(\varepsilon^2)$, for every $p_G \in \mathscr{P}_A$ and the corresponding p_{G^*}, with mixing distribution G^* defined as the re-normalized restriction of G to $[-a_{\varepsilon^2},\ a_{\varepsilon^2}]$, by the inequality in (1), we have $h^2(p_G,\ p_{G^*}) \le ||p_G - p_{G^*}||_1 \le 2G([-a_{\varepsilon^2},\ a_{\varepsilon^2}]^c) \le 2\varepsilon^2$, which implies that

$h(p_G,\ p_{G^*}) \le \sqrt{2}\varepsilon$. Thus, a Hellinger ε-net over $\mathscr{P}_{a_{\varepsilon^2}} := \{p_G:\ G([-a_{\varepsilon^2},\ a_{\varepsilon^2}]) = 1\}$ is a $(\sqrt{2}+1)\varepsilon$-net over $\mathscr{P}_A$, where

$$\log N\big(\varepsilon,\ \mathscr{P}_{a_{\varepsilon^2}},\ h\big) \lesssim \varepsilon^{-2/3}\log\frac{a_{\varepsilon^2}}{\varepsilon^2}$$

again by Proposition 2 of Gao and van der Vaart (2016), p. 617. The inequality in (24) follows. □

Appendix C: Posterior contraction rates in L^r-metrics, $1 \le r \le 2$, for Dirichlet–Laplace mixtures

In this section, we prove Proposition 1 and Proposition 2 of Sect. 2 on contraction rates in the L^1 and L^2-metrics, respectively, for the posterior distribution corresponding to a Dirichlet process mixture of Laplace densities.

Proof of Proposition 1 In order to derive the Hellinger or the L^1-metric posterior contraction rate, we can appeal to Theorem 2.1 of Ghosal et al. (2000), p. 503, or Theorem 2.1 of Ghosal and van der Vaart (2001), p. 1239. We define a sieve set for which conditions (2.2) or (2.8) and (2.3) or (2.9), postulated in the aforementioned theorems, are satisfied. To the aim, once recalled that $\alpha(\mathbb{R}) < +\infty$, let $\bar{\alpha} := \alpha/\alpha(\mathbb{R})$ be the probability measure corresponding to the baseline measure α of the Dirichlet process. Consistently with the notation adopted throughout, $\bar{\alpha}$ is also used to denote the corresponding cumulative distribution function. By a result of Doss and Sellke (1982), p. 1304, which concerns the tails of probability measures chosen from a Dirichlet prior, we have that, for almost every sample distribution G, if $a > 0$ is large enough so that $\bar{\alpha}(-a) = 1 - \bar{\alpha}(a)$ is sufficiently small, then

$$\begin{aligned} G([-a,\ a]^c) &\le G(-a) + 1 - G(a) \\ &\le \exp\left\{-\frac{1}{\bar{\alpha}(-a)|\log\bar{\alpha}(-a)|^2}\right\} \\ &\quad + \exp\left\{-\frac{1}{[1-\bar{\alpha}(a)]|\log[1-\bar{\alpha}(a)]|^2}\right\} \\ &= 2\exp\left\{-\frac{1}{\bar{\alpha}(-a)\,|\log\bar{\alpha}(-a)|^2}\right\} \\ &< A_\eta(a), \end{aligned}$$

having set the position $A_\eta(a) := 2\exp\{-[\bar{\alpha}(-a)]^{-\eta}\}$ for some fixed $0 < \eta < 1$. The inverse function $A_\eta^{-1}:\ (0,\ 1) \to (0,\ +\infty)$ is defined as $A_\eta^{-1}:\ u \mapsto -\bar{\alpha}^{-1}(\log^{-1/\eta}(2/u))$, where the function $\bar{\alpha}^{-1}(\cdot)$ is the left-continuous inverse of $\bar{\alpha}(\cdot)$, that is, $\bar{\alpha}^{-1}(u) := \inf\{y \in \mathbb{R}:\ \bar{\alpha}(y) \ge u\}$, $u \in (0,\ 1)$. Considered the class of densities $\mathscr{P}_{A_\eta} := \{p_G:\ G([-a,\ a]^c) \le A_\eta(a)\ \text{ for all } a > 0\}$, we have $\Pi(\mathscr{P}_{A_\eta}) = 1$. For any sequence of positive real numbers $\bar{\varepsilon}_n \downarrow 0$, set the position $a \equiv a_{\bar{\varepsilon}_n} := A_\eta^{-1}(\bar{\varepsilon}_n)$ and defined the sieve set $\mathscr{P}_n := \{p_G:\ G([-a_{\bar{\varepsilon}_n},\ a_{\bar{\varepsilon}_n}]^c) \le A_\eta(a_{\bar{\varepsilon}_n}) = \bar{\varepsilon}_n\}$, we have

$$\Pi(\mathscr{P} \setminus \mathscr{P}_n) = 0$$

and condition (2.3) or (2.9) is satisfied. As for condition (2.2) or (2.8), taking $\bar{\varepsilon}_n = n^{-3/8} \log^{3/8} n$, by Lemma 2, we have

$$\begin{aligned}\log D(\bar{\varepsilon}_n,\ \mathscr{P}_n,\ ||\cdot||_1) &\leq \log N(\bar{\varepsilon}_n/2,\ \mathscr{P}_n,\ ||\cdot||_1)\\ &\lesssim (\bar{\varepsilon}_n)^{-2/3} \log \frac{A_\eta^{-1}(\bar{\varepsilon}_n/6)}{\bar{\varepsilon}_n^2}\\ &\lesssim n\bar{\varepsilon}_n^2. \end{aligned} \tag{25}$$

The same bound as in (25) also holds for the Hellinger metric entropy. The Kullback-Leibler prior mass condition (2.4) of Theorem 2.1 of Ghosal et al. (2000), p. 503, or, equivalently, condition (2.10) of Theorem 2.1 of Ghosal and van der Vaart (2001), p. 1239, can be seen to be satisfied for $\tilde{\varepsilon}_n := n^{-3/8} \log^{5/8} n$. For the verification of this condition, we refer the reader to condition (2) of Proposition 2 below, whose requirement (5) is satisfied under assumption (4) of Proposition 1. The proof is completed by taking $\varepsilon_n := \max\{\bar{\varepsilon}_n,\ \tilde{\varepsilon}_n\} = n^{-3/8} \log^{5/8} n$. For the sake of clarity, we remark that the role of $\tilde{\varepsilon}_n$ is played by ε_n in the proof of Proposition 2.

We now prove Proposition 2 on the posterior contraction rate in the L^2-metric. The result relies on Theorem 3 of Giné and Nickl (2011), p. 2892, which gives sufficient conditions for deriving posterior contraction rates in L^r-metrics, $1 < r < +\infty$. All assumptions of Theorem 3, except for condition (2), are shown to be satisfied for any kernel density f as in Definition 1 with $\beta > 1$. This includes the (standard) Laplace kernel density as a special case when $\beta = 2$. Condition (2), which requires the prior mass in Kullback–Leibler type neighborhoods of the sampling density $p_0 \equiv p_{G_0} = G_0 * f$ to be not exponentially small, relies on a preliminary approximation result of the density $p_{G_0^*} = G_0^* * f$, with mixing distribution G_0^* obtained as the re-normalized restriction of G_0 to a compact interval, by a mixture density that has a discrete mixing distribution with a sufficiently restricted number of support points. This result is known to hold for the Laplace kernel density in virtue of Lemma 2 of Gao and van der Vaart (2016), pp. 615–616.

Proof of Proposition 2 We apply Theorem 3 of Giné and Nickl (2011), p. 2892, with $r = 2$. We refer to the conditions of this theorem using the same letters/numbers as in the original article. Let $\gamma_n \equiv 1$ and $\delta_n \equiv \varepsilon_n := n^{-3/8} \log^{5/8} n$, $n \in \mathbb{N}$.

- *Verification of condition* (b) Condition (b), which requires that $\varepsilon_n^2 = O(n^{-1/2})$, is satisfied in the general case for $\varepsilon_n = n^{-(\beta-1/2)/2\beta} \log^\kappa n$, with some $\kappa > 0$ and $\beta > 1$.
- *Verification of condition* (1) Condition (1) requires that the prior probability of the complement of a sieve set $\mathscr{P}_n$ is exponentially small. We show that, in the present setting, the prior probability of a sieve set $\mathscr{P}_n$, chosen as prescribed by (15) in Theorem 2 of Giné and Nickl (2011), p. 2891, is equal to zero. Let J_n be any sequence of positive real numbers satisfying $2^{J_n} \leq cn\varepsilon_n^2$ for some fixed constant $0 < c < +\infty$. Let K be a convolution kernel such that it is of bounded p-variation for some finite real number $p \geq 1$, right (or left) continuous and satisfies

$||K||_\infty < +\infty$, $\int_{-\infty}^{+\infty}(1+|z|)^w|K(z)|\,\mathrm{d}z < +\infty$ for some $w > 2$, $\hat{K}(0) = 1$ and $I_\beta^2[\hat{K}] < +\infty$, cf. condition (20) in Lemma 1. Defined the sieve set

$$\mathscr{P}_n := \big\{p_G \in \mathscr{P} : \, ||p_G * K_{2^{-J_n}} - p_G||_2 \le C\delta_n\big\},$$

where $K_{2^{-J_n}}(\cdot) := 2^{J_n}K(\cdot 2^{J_n})$ and $C > 0$ is a finite constant depending only on K and f, we have

$$\Pi(\mathscr{P} \setminus \mathscr{P}_n) = 0 \quad \text{for all } n \in \mathbb{N}.$$

In fact, for every $G \in \mathscr{G}$, by Plancherel's Theorem, $||p_G * K_{2^{-J_n}} - p_G||_2^2 = ||p_G - p_G * K_{2^{-J_n}}||_2^2 = (2\pi)^{-1}||\hat{p}_G - \hat{p}_G \times \hat{K}_{2^{-J_n}}||_2^2 \le (2\pi)^{-1}||\hat{f} - \hat{f} \times \hat{K}_{2^{-J_n}}||_2^2$ and, by Lemma 1, $||\hat{f} - \hat{f} \times \hat{K}_{2^{-J_n}}||_2^2 \sim (2^{-J_n})^{2\beta-1} \times B_f^2 \times I_\beta^2[\hat{K}]$, where, for $\beta = 2$, we have $(2^{-J_n})^{2\beta-1} = (2^{-J_n})^3 = O(\delta_n^2)$. Thus,

$$\forall\, G \in \mathscr{G}, \quad ||p_G * K_{2^{-J_n}} - p_G||_2 = O(\delta_n) \tag{26}$$

and condition (1) is verified. Relationship (26) holds, in particular, for $p_0 \equiv p_{G_0} = G_0 * f$. Furthermore, $p_0 \in L^2(\mathbb{R})$ if $f \in L^2(\mathbb{R})$, which is the case for the (standard) Laplace kernel density, because $||p_0||_2^2 = (2\pi)^{-1}||\hat{p}_0||_2^2 \le (2\pi)^{-1}||\hat{f}||_2^2 = ||f||_2^2 < +\infty$.

- *Verification of condition* (2)
  Condition (2) requires that, for some finite constant $C_1 > 0$, the prior probability of Kullback–Leibler type neighborhoods of P_0 of radius ε_n^2 is at least $\exp(-C_1 n\varepsilon_n^2)$, that is, $\Pi(B_{\mathrm{KL}}(P_0;\, \varepsilon_n^2)) \gtrsim \exp(-C_1 n\varepsilon_n^2)$. Fix $0 < \varepsilon \le (1 - e^{-1})/\sqrt{2}$ and let $a_\varepsilon := A_0^{-1}(\varepsilon^2)$, where A_0^{-1} is the inverse of the function A_0 in condition (3). Define G_0^* as the re-normalized restriction of G_0 to $[-a_\varepsilon,\, a_\varepsilon]$. By Lemma A.3 of Ghosal and van der Vaart (2001), p. 1261, and assumption (3), we have $||p_{G_0} - p_{G_0^*}||_1 \le 2G_0([-a_\varepsilon,\, a_\varepsilon]^c) \lesssim \varepsilon^2$. From the inequality in (1), $h^2(p_{G_0},\, p_{G_0^*}) \le ||p_{G_0} - p_{G_0^*}||_1 \lesssim \varepsilon^2$, whence $h(p_{G_0},\, p_{G_0^*}) \lesssim \varepsilon$. It is known from Lemma 2 of Gao and van der Vaart (2016), pp. 615–616, that there exists a discrete distribution G_0' such that $h(p_{G_0'},\, p_{G_0^*}) \lesssim \varepsilon$. The distribution G_0' has at most $N \asymp \varepsilon^{-2/3}$ support points $y_1,\, \ldots,\, y_N$ in $[-a_\varepsilon,\, a_\varepsilon]$, which we may assume to be at least $2\varepsilon^2$-separated. If not, we can take a maximal $2\varepsilon^2$-separated set in the support points of G_0' and replace G_0' with the discrete distribution G_0'' obtained by relocating the masses of G_0' to the nearest points of the $2\varepsilon^2$-net. Then, $h^2(p_{G_0'},\, p_{G_0''}) \lesssim \max_{1\le j\le N}|y_j' - y_j''| \lesssim \varepsilon^2$, as shown in Proposition 2 of Gao and van der Vaart (2016), p. 617. Let $G_0' = \sum_{j=1}^N p_j\delta_{y_j}$, with $|y_j - y_k| \ge 2\varepsilon^2$ for all $1 \le j \ne k \le N$. For any distribution G such that

$$\sum_{j=1}^{N} |G([y_j - \varepsilon^2,\, y_j + \varepsilon^2]) - p_j| \le \varepsilon^2, \tag{27}$$

  we have $||p_G - p_{G_0'}||_1 \lesssim \varepsilon^2$ by Lemma 5 of Gao and van der Vaart (2016), p. 620. Thus,

$$h^2(p_G,\ p_{G_0}) \lesssim h^2(p_G,\ p_{G_0'}) + h^2(p_{G_0'},\ p_{G_0^*}) + h^2(p_{G_0^*},\ p_{G_0})$$
$$\lesssim ||p_G - p_{G_0'}||_1 + \varepsilon^2 + ||p_{G_0^*} - p_{G_0}||_1 \lesssim \varepsilon^2.$$

We can now invoke Lemma A.10 in Scricciolo (2011), p. 305, taking into account Remark A.3 of the same article. To this aim, note that, if G satisfies (27), then $G([-(a_\varepsilon + 1),\ (a_\varepsilon + 1)]) > 1/2$. The reader may also refer to Scricciolo (2014), p. 305. For any $G \in \mathscr{G}$, let P_G stand for the probability measure with density $p_G \in \mathscr{P}$. The inclusion

$$\left\{P_G : \sum_{j=1}^{N} |G([y_j - \varepsilon^2,\ y_j + \varepsilon^2]) - p_j| \leq \varepsilon^2\right\} \subseteq B_{\mathrm{KL}}\big(P_0;\ \varepsilon^2 \log^2(1/\varepsilon)\big)$$

holds. To apply Lemma A.2 of Ghosal and van der Vaart (2001), p. 1260, note that, for every y_j, $1 \leq j \leq N$, we have $\alpha([y_j - \varepsilon^2,\ y_j + \varepsilon^2]) \gtrsim \varepsilon^{b'}$ for some finite constant $b' > 0$. Thus,

$$\log \Pi(B_{\mathrm{KL}}(P_0;\ \varepsilon^2 \log^2(1/\varepsilon))) \gtrsim -N \log(1/\varepsilon) \asymp -\varepsilon^{-2/3} \log(1/\varepsilon).$$

Taking $\varepsilon_n := \varepsilon \log(1/\varepsilon)$, we have $\Pi(B_{\mathrm{KL}}(P_0;\ \varepsilon_n^2)) \gtrsim \exp(-C_1 n \varepsilon_n^2)$ and condition (2) is satisfied.

- *Verification of condition* (3)
  Condition (3) requires that there exists a finite constant $B > 0$ such that $\Pi(||p_G||_\infty > B \mid X^{(n)}) = o_{\mathbf{P}}(1)$. If $||f||_\infty < +\infty$, then $||p_G||_\infty \leq ||f||_\infty < +\infty$ for all $G \in \mathscr{G}$, see Lemma 3. In particular, $||p_0||_\infty = ||p_{G_0}||_\infty \leq ||f||_\infty < +\infty$. Taking $B := ||f||_\infty$, we have

  $$\forall\, n \in \mathbb{N},\quad \Pi(||p_G||_\infty > B \mid X^{(n)}) = 0 \quad P_0^n\text{-almost surely,}$$

  and condition (3) is satisfied. For the (standard) Laplace kernel density, $||f||_\infty = 1/2$.

The proof is thus complete and assertion (6) follows. □

Appendix D: Inversion inequalities

In this section, we state a result relating, for every real number $p \geq 1$, the L^p-Wasserstein distance between any pair of mixing distributions $G,\ G' \in \mathscr{G}$ to the L^2-distance between the corresponding mixed densities $p_G = G{*}f$ and $p_{G'} = G'{*}f$, with a kernel density f that is ordinary smooth in the sense of condition (29) stated below. Lemma 4 extends Lemma 7 of Gao and van der Vaart (2016), pp. 621–622, beyond the case of compactly supported mixing distributions to mixing distributions with finite moment generating functions on some neighborhood of zero $(-s_0,\ s_0)$, with $0 < s_0 < 1$. If, furthermore, the kernel density is bounded, $||f||_\infty < +\infty$, then the inversion inequality in (30) below also holds for the Hellinger metric in virtue of the following known result, which is reported for the reader's convenience.

Lemma 3 *For a given kernel density f, let $p_G = G * f$, with $G \in \mathscr{G}$. If $||f||_\infty < +\infty$, then*

$$\forall\, G \in \mathscr{G}, \quad p_G(x) \leq ||f||_\infty \ \textit{for all } x \in \mathbb{R},$$

and

$$\forall\, G,\ G' \in \mathscr{G}, \quad ||p_G - p_{G'}||_2^2 \leq 4||f||_\infty h^2(p_G,\ p_{G'}). \tag{28}$$

We now state and prove an inequality translating the L^2-norm and the Hellinger distance between mixed densities into any L^p-Wasserstein distance, $p \geq 1$, between the corresponding mixing distributions.

Lemma 4 *Let G and G' be probability measures on some Borel-measurable space $(\mathscr{Y},\ \mathscr{B}(\mathscr{Y}))$, $\mathscr{Y} \subseteq \mathbb{R}$, such that the associated moment generating functions $M_G(s)$ and $M_{G'}(s)$ are finite for all $|s| < s_0$, with $0 < s_0 < 1$. Let f be a probability density function on $\mathbb{R}$, with Fourier transform $\hat{f}$ satisfying, for some real number $\beta > 0$, the condition*

$$\inf_{t \in \mathbb{R}} (1 + |t|^\beta)|\hat{f}(t)| > 0. \tag{29}$$

Let d stand for the L^2-distance between the mixed densities $p_G = G * f$ and $p_{G'} = G' * f$. Then, for any real number $p \geq 1$,

$$W_p(G,\ G') \lesssim d^{1/(p+\beta)} \left(\log \frac{1}{d}\right)^{(p+1/2)/(p+\beta)}$$
$$\textit{for } d = ||p_G - p_{G'}||_2 \ \textit{small enough.} \tag{30}$$

If, in addition, $||f||_\infty < +\infty$, then the upper bound in (30) also holds for d being the Hellinger distance, $d = h(p_G,\ p_{G'})$.

Proof For any real number $h > 0$, by the triangle inequality, we have

$$W_p^p(G,\ G') \leq W_p^p(G,\ G * \Phi_h) + W_p^p(G * \Phi_h,\ G' * \Phi_h) + W_p^p(G' * \Phi_h,\ G'), \tag{31}$$

where Φ_h stands for a zero-mean Gaussian probability measure with variance h^2, whose density is denoted by $\phi_h(\cdot) := h^{-1}\phi(\cdot/h)$, for ϕ the density of a standard normal r.v. W. The first and third terms on the right-hand side of (31) can be bounded above as follows. By standard arguments, see, for instance, the proof of Theorem 2 in Nguyen (2013), pp. 389–391,

$$\max\{W_p^p(G,\ G * \Phi_h),\ W_p^p(G' * \Phi_h,\ G')\} \leq E[|hW|^p] \lesssim h^p \tag{32}$$

because $E[|W|^p] < +\infty$ for every real number $p > 0$, hence, *a fortiori*, for every real $p \geq 1$. Concerning the second term on the right-hand side of (31), reasoning as in Lemma 7 of Gao and van der Vaart (2016), pp. 621–622, for any real number $M > 0$,

$$W_p^p(G * \Phi_h,\ G' * \Phi_h) \lesssim \left(\int_{|x| \leq M} + \int_{|x| > M}\right) |x|^p |(G - G') * \phi_h(x)|\,\mathrm{d}x =: T_1 + T_2,$$

where, for every $0 < h \leq 1$,

$$T_1 \lesssim M^{p+1/2} ||(G - G') * \phi_h||_2 \lesssim M^{p+1/2} h^{-\beta} ||p_G - p_{G'}||_2 \tag{33}$$

because $\sup_{t \in \mathbb{R}} |\hat{\phi}(ht)|/|\hat{f}(t)| \lesssim h^{-\beta}$ in virtue of assumption (29). To see it, note that assumption (29) implies the existence of a finite constant $L_f > 0$ such that $(1 + |t|^\beta)|\hat{f}(t)| \geq L_f$ for all $t \in \mathbb{R}$. Therefore, if $0 < h \leq 1$,

$$\sup_{t \in \mathbb{R}} \frac{|\hat{\phi}(ht)|}{|\hat{f}(t)|} \leq \frac{1}{L_f} \sup_{t \in \mathbb{R}} [(1 + |ht|^\beta)|\hat{\phi}(ht)|] \times \sup_{t \in \mathbb{R}} \left(\frac{1 + |t|^\beta}{1 + |ht|^\beta}\right) \lesssim h^{-\beta}.$$

If $||f||_\infty < +\infty$, then the L^2-distance between p_G and $p_{G'}$ in (33) can be replaced with the Hellinger distance (see Lemma 3), so that

$$T_1 \lesssim M^{p+1/2} h^{-\beta} h(p_G,\ p_{G'}).$$

We now deal with the term T_2. We preliminarily derive an instrumental inequality. For every $x \in \mathbb{R}$ and real numbers $p,\ u > 0$,

$$\frac{p}{u} e^{u|x|/p} = \frac{p}{u} \sum_{j=0}^{+\infty} \frac{(u|x|/p)^j}{j!} \geq |x|,$$

whence

$$|x|^p \leq (p/u)^p e^{u|x|} < (p/u)^p (e^{-ux} + e^{ux}). \tag{34}$$

Now fix any number $0 < u < s_0 < 1$. Applying the inequalities in (34) and taking into account the expression of the moment generating function of a standard Gaussian distribution $M_\Phi(s) = e^{s^2/2}$, $s \in \mathbb{R}$, we get

$$\begin{aligned}\int_{-\infty}^{+\infty} \max\{1,\ |x|^p\} e^{u|x|} \phi_h(x)\,\mathrm{d}x &\leq \int_{-\infty}^{+\infty} \max\left\{e^{u|x|},\ (p/u)^p e^{2u|x|}\right\} \phi_h(x)\,\mathrm{d}x \\ &< 2 \max\left\{e^{(uh)^2/2},\ (p/u)^p e^{2(uh)^2}\right\} \\ &< 2 \max\left\{e^{s_0^2/2},\ (p/u)^p e^{2s_0^2}\right\},\end{aligned}$$

namely, for fixed u, the above integral can be bounded above by a constant that is fixed throughout and can therefore be neglected when bounding T_2. Hence,

$$\begin{aligned}
T_2 &\lesssim e^{-uM} \int_{|x|>M} |x|^p e^{u|x|}[(G+G') * \phi_h(x)]\,\mathrm{d}x \\
&\lesssim e^{-uM} \int_{\mathscr{Y}} (1+|y|^p)e^{u|y|} \left(\int_{-\infty}^{+\infty} \max\{1,\ |x|^p\} e^{u|x|}\phi_h(x)\,\mathrm{d}x \right) \mathrm{d}(G+G')(y) \\
&\lesssim e^{-uM} \int_{\mathscr{Y}} (1+|y|^p)e^{u|y|}\,\mathrm{d}(G+G')(y) \lesssim e^{-uM}
\end{aligned}$$

because

$$\begin{aligned}
\int_{\mathscr{Y}} e^{u|y|}\,\mathrm{d}(G+G')(y) &< \int_{\mathscr{Y}} (e^{-uy}+e^{uy})\,\mathrm{d}(G+G')(y) \\
&= (M_G+M_{G'})(-u) + (M_G+M_{G'})(u) < +\infty
\end{aligned}$$

and, for any fixed real number $0 < \xi < 1$ such that $0 < s := (\xi + u) < s_0$, by the inequalities in (34),

$$\begin{aligned}
\int_{\mathscr{Y}} |y|^p e^{u|y|}\,\mathrm{d}(G+G')(y) &< (p/\xi)^p \int_{\mathscr{Y}} e^{(\xi+u)|y|}\,\mathrm{d}(G+G')(y) \\
&= (p/\xi)^p \int_{\mathscr{Y}} e^{s|y|}\,\mathrm{d}(G+G')(y) \\
&< (p/\xi)^p \int_{\mathscr{Y}} (e^{-sy}+e^{sy})\,\mathrm{d}(G+G')(y) \\
&= (p/\xi)^p [(M_G+M_{G'})(-s) + (M_G+M_{G'})(s)] < +\infty
\end{aligned}$$

by the assumption that both G and G' have finite moment generating functions on $(-s_0,\ s_0)$, for $0 < s_0 < 1$. Thus,

$$T_2 \lesssim e^{-uM}. \tag{35}$$

Combining partial results in (32), (33) and (35), we get

$$W_p^p(G,\ G') \lesssim h^p + M^{p+1/2}h^{-\beta}d + e^{-uM} \tag{36}$$

and the conclusion follows by minimizing the expression in (36) with respect to h and M, which, for sufficiently small d, implies taking $M = O(\log(1/d))$ and $h^{p+\beta} = O(d \log^{p+1/2}(1/d))$. □

Remark 6 The standard Laplace kernel density is bounded, with $||f||_\infty = 1/2$, and satisfies condition (29) for $\beta = 2$.

References

Ball K, Pajor A (1990) The entropy of convex bodies with “few” extreme points. In: Müller PFX, Schachermayer W (eds) Proceedings of the Conference in Geometry of Banach Spaces at Strobl, Austria, 1989. London Mathematical Society Lecture Note Series, vol 158, pp 25–32

Barron AR (1988) The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions. Technical Report #7, University of Illinois at Urbana-Champaign

Birgé L, Massart P (1993) Rates of convergence for minimum contrast estimators. Probab Theory Rel 97:113–150

Dall’Aglio G (1956) Sugli estremi dei momenti delle funzioni di ripartizione doppia. Ann Scuola Norm Sup Pisa 10:35–74 (Italian)

Davis KB (1977) Mean integrated square error properties of density estimates. Ann Stat 5:530–535

Dedecker J, Fischer A, Michel B (2015) Improved rates for Wasserstein deconvolution with ordinary smooth error in dimension one. Electron J Stat 9:234–265

Devroye L (1990) A note on linnik’s distribution. Stat Probab Lett 9:305–306

Diaconis P, Freedman D (1986) On the consistency of Bayes estimates. Ann Stat 14:1–26

Donnet S, Rivoirard V, Rousseau J, Scricciolo C (2018) Posterior concentration rates for empirical Bayes procedures with applications to Dirichlet process mixtures. Bernoulli 24:231–256

Doss H, Sellke T (1982) The tails of probabilities chosen from a Dirichlet prior. Ann Stat 10:1302–1305

Fan J (1991a) Global behavior of deconvolution kernel estimates. Stat Sin 1:541–551

Fan J (1991b) On the optimal rates of convergence for nonparametric deconvolution problems. Ann Stat 19:1257–1272

Gao F, van der Vaart A (2016) Posterior contraction rates for deconvolution of Dirichlet–Laplace mixtures. Electron J Stat 10:608–627

Ghosal S, Ghosh JK, van der Vaart AW (2000) Convergence rates of posterior distributions. Ann Stat 28:500–531

Ghosal S, van der Vaart AW (2001) Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. Ann Stat 29:1233–1263

Ghosal S, van der Vaart A (2007a) Posterior convergence rates of Dirichlet mixtures at smooth densities. Ann Stat 35:697–723

Ghosal S, van der Vaart A (2007b) Convergence rates of posterior distributions for noniid observations. Ann Stat 35:192–223

Ghosh JK, Ramamoorthi RV (2003) Bayesian nonparametrics, Springer Series in Statistics. Springer, New York

Giné E, Nickl R (2011) Rates of contraction for posterior distributions in L^r-metrics, $1 \leq r \leq \infty$. Ann Stat 39:2883–2911

Hall P, Lahiri SN (2008) Estimation of distributions, moments and quantiles in deconvolution problems. Ann Stat 36:2110–2134

LeCam L (1973) Convergence of estimates under dimensionality restrictions. Ann Stat 1:38–53

Lindsay BG (1995) Mixture models: theory, geometry and applications. In: In: NSF-CBMS Regional Conference Series in Probability and Statistics, vol 5. Institute of Mathematical Statistics, Hayward, CA

Lo AY (1984) On a class of Bayesian nonparametric estimates: I. Density estimates. Ann Stat 12:351–357

Meister A (2009) Deconvolution problems in nonparametric statistics, vol 193. Springer, Berlin Lecture Notes in Statistics

Nguyen X (2013) Convergence of latent mixing measures in finite and infinite mixture models. Ann Stat 41:370–400

Parzen E (1962) On estimation of a probability density function and mode. Ann Math Stat 33:1065–1076

Scricciolo C (2006) Convergence rates for Bayesian density estimation of infinite-dimensional exponential families. Ann Stat 34:2897–2920

Scricciolo C (2007) On rates of convergence for Bayesian density estimation. Scand J Stat 34:626–642

Scricciolo C (2011) Posterior rates of convergence for Dirichlet mixtures of exponential power densities. Electron J Stat 5:270–308

Scricciolo C (2014) Adaptive Bayesian density estimation in L^p-metrics with Pitman–Yor or normalized inverse-Gaussian process kernel mixtures. Bayesian Anal 9:475–520

Shorack GR, Wellner JA (1986) Empirical processes with applications to statistics. Wiley, New York

Tsybakov AB (2004) Introduction à l’estimation non-paramétrique. Springer, Berlin

Van de Geer S (1993) Hellinger-consistency of certain nonparametric maximum likelihood estimators. Ann Stat 21:14–44

Van de Geer S (1995) Asymptotic normality in mixture models. ESAIM Probab Stat 1:17–33

Van de Geer S (1996) Rates of convergence for the maximum likelihood estimator in mixture models. J Nonparametr Stat 6:293–310

Van de Geer SA (2000) Empirical processes in M-estimation. Cambridge University Press, New York

Watson GS, Leadbetter MR (1963) On the estimation of the probability density, I. Ann Math Stat 34:480–491

Wong WH, Shen X (1995) Probability inequalities for likelihood ratios and convergence rates of sieve MLES. Ann Stat 23:339–362