Electronic Journal of Statistics
Vol. 11 (2017) 4723–4745
ISSN: 1935-7524
DOI: 10.1214/17-EJS1369

A novel approach to Bayesian consistency

Minwoo Chae and Stephen G. Walker

Department of Mathematics, The University of Texas at Austin
e-mail: minwooo.chae@gmail.com; s.g.walker@math.utexas.edu

Abstract: It is well-known that the Kullback–Leibler support condition implies posterior consistency in the weak topology, but is not sufficient for consistency in the total variation distance. There is a counter–example. Since then many authors have proposed sufficient conditions for strong consistency; and the aim of the present paper is to introduce new conditions with specific application to nonparametric mixture models with heavy–tailed components, such as the Student-t. The key is a more focused result on sets of densities where if strong consistency fails then it fails on such densities. This allows us to move away from the traditional types of sieves currently employed.

MSC 2010 subject classifications: Primary 62G05, 62G20; secondary 62G07.
Keywords and phrases: Kullback–Leibler divergence, Lévy–Prokhorov metric, mixture of Student's t distributions, posterior consistency, total variation.

Received May 2017.

1. Introduction

In this paper we consider a novel approach to Bayesian consistency in nonparametric problems, specifically concentrating on mixture models, which are the usual type of nonparametric model used in practice. The first formulation is given by Doob [9]; but this approach has a drawback in infinite dimensional models, see [7, 8]. Instead it is commonly assumed that observations are i.i.d. from some fixed but unknown density function, and a general sufficient condition for weak consistency is given in Schwartz [25].

To set the scene, assume that the observations $X_1, \ldots, X_n$ are i.i.d. real-valued random variables from a true density p_0. Let the model $\mathscr{L}$ be the space of all Lebesgue densities on $(\mathbb{R}, \mathcal{R})$ equipped with the total variation metric, and Π be a prior on $(\mathscr{L}, \mathcal{L})$, where $\mathcal{R}$ and $\mathcal{L}$ are Borel σ-algebras. Formally, for a (pseudo-)metric d on $\mathscr{L}$, the posterior distribution $\Pi(\cdot|X_1, \ldots, X_n)$ is called to be d-consistent at p_0 if $\Pi(d(p_0, p) > \eta | X_1, \ldots, X_n)$ converges to zero in probability for every $\eta > 0$. When d is the total variation (Lévy–Prokhorov, resp.) metric, it is often called strongly (weakly, resp.) consistent.

Let $K(p, q) = \int p \log(p/q)\, d\mu$ be the Kullback–Leibler (KL) divergence, where μ is the Lebesgue measure. Schwartz [25] has shown that if p_0 lies in the KL support of Π, *i.e.*

$$\Pi\Big(p \in \mathscr{L} : K(p_0, p) < \delta\Big) > 0 \quad \text{for every } \delta > 0, \tag{1.1}$$

then the posterior distribution is weakly consistent at p_0. Along with the KL support condition (1.1), various sufficient conditions for strong consistency have been studied in infinite–dimensional models, see [3, 29, 28, 6] for general conditions. Some important references concerning specific models and priors are [1, 11, 13, 5]. Further work incorporating convergence rates can be found in [16, 12, 31, 21], for example.

Since the total variation is a stronger metric than Lévy–Prokhorov, see [20, p. 34], the KL support condition (1.1) is often insufficient for strong consistency. In this regard, Barron, Schervish and Wasserman [3] constructed a prior satisfying the KL support condition (1.1) but the corresponding posterior distribution is not strongly consistent. Walker, Lijoi and Prünster [30] explained this phenomenon with the notion of data tracking.

In this paper we present a new sufficient condition for strong consistency and apply it to nonparametric mixture models. Since the convergence in Lévy–Prokhorov metric is equivalent to weak convergence, once the prior satisfies the KL support condition (1.1), Schwartz's theorem implies that there exists some sequence $\epsilon_n \downarrow 0$ such that $\Pi(d_P(p_0, p) > \epsilon_n | X_1, \ldots, X_n)$ converges to zero in probability. For strong consistency, therefore, it suffices to show that $\Pi(A_{n,\eta} | X_1, \ldots, X_n) \to 0$ in probability for every $\eta > 0$, where

$$A_{n,\eta} = \Big\{p \in \mathscr{L} : d_P(p_0, p) \le \epsilon_n, d_V(p_0, p) > \eta\Big\}. \tag{1.2}$$

The new approach is based on the fact that $A_{n,\eta}$ is a collection of "weird" densities in the sense that it consists of highly fluctuating densities with a centering around p_0. With a reasonable prior, therefore, prior mass imposed on $A_{n,\eta}$ is negligible, which in turn implies strong consistency. The focus on $A_{n,\eta}$ allows us to move away from the typical uses of sieves.

Our approach is very different from [30], relying on a special property of densities in $A_{n,\eta}$. The new approach entails different kinds of sieves avoiding the calculation of Hellinger entropy or prior probabilities of small Hellinger balls. Instead, we require a Lévy–Prokhorov convergence rate (ϵ_n) for which we provide a general sufficient condition. Our new approach significantly simplifies conditions required on the hyperparameter of a Dirichlet process in a mixture model, for example. In particular, a mean parameter can have an arbitrarily heavy tail. We also consider a mixture of Student's t distributions which can be used to model heavy-tailed distributions; the consistency of which is yet to been done in the literature.

Notation

For $p \in \mathscr{L}$, the corresponding probability measure is denoted as P, and vice versa. The expectation of a function f with respect to P is denoted Pf, *i.e.* $Pf =$

$\int f(x)dP(x)$. The expectation under the true distribution is denoted $\mathbb{E}$. Let $d_V(p_1,p_2) = 2\sup_{A\in\mathcal{R}} |P_1(A) - P_2(A)|$ and $d_H(p_1,p_2) = \{\int(\sqrt{p_1} - \sqrt{p_2})^2 d\mu\}^{1/2}$ be the total variation and Hellinger metrics. The indicator function for a set A is denoted 1_A. For two positive sequence (a_n) and (b_n), $a_n \ll b_n$ represents $a_n/b_n \to 0$. The maximum of two numbers a and b are denoted $a \vee b$. The inequality $\lesssim$ represents "less than up to a constant multiplication," where the constant is universal (such as $2, \pi, e$) unless specified explicitly.

2. Main results

For $p \in \mathscr{L}$ and $\gamma > 0$, define a non–negative function p_γ on $\mathbb{R}$ as

$$p_\gamma(x) = \frac{P(B_\gamma(x))}{\mu(B_\gamma(x))} = (2\gamma)^{-1}\int_{\{y:|y-x|<\gamma\}} p(y)\,dy,$$

where $B_\gamma(x) = \{y \in \mathbb{R} : |y-x| < \gamma\}$. Note that $p_\gamma = p * U_\gamma$, where $*$ denotes the convolution and U_γ is the uniform distribution on the interval $[-\gamma,\gamma]$. Therefore, p_γ is also a probability density which can be understood as a smoothed version of p, where γ controls the degree of smoothness. For example, suppose

$$p(y) = \begin{cases} 2 & 0 < y < 1/4 \ \text{ or } \ 1/2 < y < 3/4 \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$p_{1/4}(y) = \begin{cases} 1 & 0 \le y \le 3/4 \\ 0 & y \le -1/4 \ \text{ or } \ y \ge 1 \\ \text{linear} & -1/4 \le y \le 0 \ \text{ or } \ 3/4 < y < 1. \end{cases}$$

See Figure 1. For simplicity, $(p_0)_\gamma$ is written as $p_{0,\gamma}$. For two probability measures P and Q, let

$$d_P(P,Q) = \inf\Big\{\epsilon > 0 : P(A) \le Q(A^\epsilon) + \epsilon \text{ and } Q(A) \le P(A^\epsilon) + \epsilon,\ \forall A \in \mathcal{R}\Big\}$$

be the Lévy–Prokhorov metric, where $A^\epsilon = \cup_{x\in A} B_\epsilon(x)$. Note that the convergence in d_P is equivalent to weak convergence, and one inequality in the definition of d_P can be omitted; see [17].

For a given density p_0, suppose that a density p is close to p_0 in d_P but far away from p_0 in d_V. This is only possible when p is a "weird" density in the sense that it highly fluctuates with a centering around p_0; as illustrated in Figure 2. It is an important property of such a density that $d_V(p, p_\gamma)$ is large even for small γ. Note that for every fixed $p \in \mathscr{L}$, $d_V(p, p_\gamma)$ converges to zero as γ goes to 0 by Lebesgue differentiation theorem and Scheffé's lemma, but never converges uniformly over $\mathscr{L}$ due to highly fluctuating densities. Therefore, if the prior probability for large $d_V(p, p_\gamma)$ is sufficiently small, the posterior distribution would be strongly consistent. The key point here is that after excluding weird densities from $\mathscr{L}$, $d_V(p, p_\gamma)$ can be shown to converge uniformly.

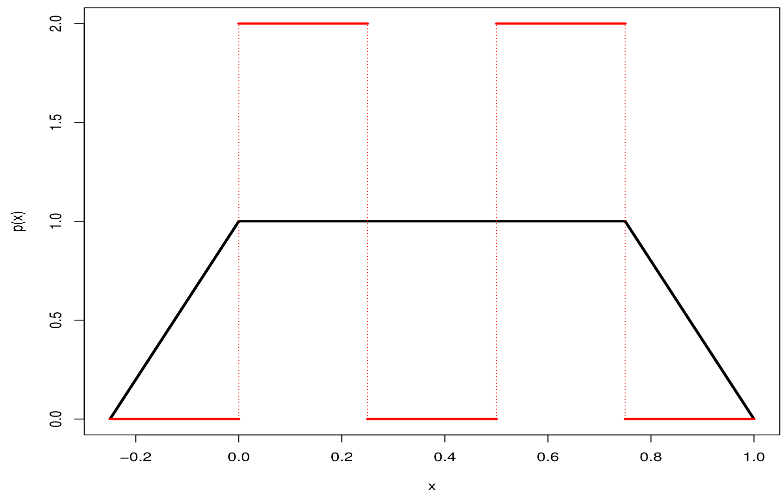

FIG 1. *An illustration of* p *(red) and* p_γ *(black) with* $\gamma = 1/4$.

To be more specific, note that by Scwartz's theorem [25], the KL support condition (1.1) guarantees the existence of a sequence $\epsilon_n \downarrow 0$ such that

$$\mathbb{E}\,\Pi\Big(d_P(p_0, p) \geq \epsilon_n \big| X_1, \ldots, X_n\Big) \to 0. \tag{2.1}$$

If $d_P(p, p_0) \leq \epsilon_n$, then for any sequence (γ_n), with $\gamma_n \to 0$ and $\epsilon_n/\gamma_n \to 0$, we have

$$\begin{aligned} d_V(p, p_0) &\leq d_V(p, p_{\gamma_n}) + d_V(p_{\gamma_n}, p_{0,\gamma_n}) + d_V(p_{0,\gamma_n}, p_0) \\ &\leq d_V(p, p_{\gamma_n}) + o(1) \end{aligned} \tag{2.2}$$

as $n \to \infty$, where the $o(1)$ term depends on ϵ_n, γ_n and p_0 only, see the proof of Theorem 2.1. Thus, strong consistency holds if there exists a sequence $(\mathscr{L}_n)$ of subsets of $\mathscr{L}$ such that

$$\mathbb{E}\,\Pi(\mathscr{L}_n^c | X_1, \ldots, X_n) = o(1) \quad \text{and} \quad \sup_{p \in \mathscr{L}_n} d_V(p, p_{\gamma_n}) = o(1). \tag{2.3}$$

This is summarized in Theorem 2.1 with more details.

Before stating the main theorem, we assume for the true density p_0 that

$$\begin{aligned} &\text{(i)} \ \sup_{x \in \mathbb{R}} p_0(x) \text{ is bounded by a universal constant, and} \\ &\text{(ii)} \ P_0\left(x : |x| > \epsilon^{-1}\right) \lesssim \epsilon \text{ for every } \epsilon > 0, \end{aligned} \tag{2.4}$$

which is not essential but simplifies the proof. Note that condition (ii) holds if the tail of p_0 is not heavier than that of the Cauchy distribution.

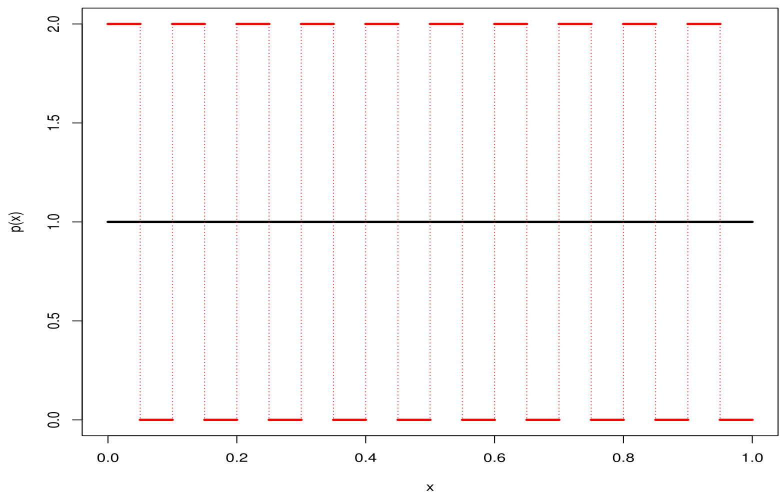

FIG 2. *An example that $d_P(p_0, p)$ is small but $d_V(p_0, p)$ is large, where p_0 (black) is the standard uniform density and p (red) fluctuates with a centering around p_0.*

Theorem 2.1. *Suppose that p_0 satisfies (2.4) and that (2.1) holds for some $\epsilon_n \downarrow 0$. Furthermore, assume that there exists $\mathscr{L}_n \subset \mathscr{L}$ satisfying (2.3) for some (γ_n) with $\epsilon_n \ll \gamma_n \ll 1$. Then, $\Pi(\cdot|X_1, \ldots, X_n)$ is strongly consistent.*

Since $d_V(p, p_{\gamma_n}) = o(1)$ for every fixed $p \in \mathscr{L}$, $(\mathscr{L}_n)$ can typically be chosen to increase to $\mathscr{L}$, constituting new sieves. In the existing Bayesian literature, such sieves are required to have bounded entropy [11, 3] or satisfy a certain prior summability condition [29]. Instead of these conditions, our requirement is (2.3), which eventually gives $A_{n,\eta} \cap \mathscr{L}_n = \emptyset$, where $A_{n,\eta}$ is defined as (1.2). Note that $A_{n,\eta}$ decreases to the singleton $\{p_0\}$, while $\mathscr{L}_n$ grows to the whole set $\mathscr{L}$. As illustrated in the next section, we can easily find (γ_n) and $(\mathscr{L}_n)$ satisfying (2.3) in nonparametric mixture models.

Note that in Barron's counter–example [3], the prior puts large mass on a set of weird densities such as the one in Figure 2. As a consequence, we cannot choose a sequence of sets $(\mathscr{L}_n)$ satisfying (2.3), resulting in posterior inconsistency.

It should be emphasized that to prove (2.3), we need to know a Lévy–Prokhorov rate (ϵ_n), which can be interpreted as the "price" for avoiding the construction of complicated sieves. Note that the KL support condition guarantees the existence of "some" rate sequence (ϵ_n). If we do not know what ϵ_n is, we only know that there exists a sequence γ_n such that $\gamma_n \to 0$ and $\epsilon_n/\gamma_n \to 0$. If γ_n converges too slowly, however, $\mathscr{L}_n$ satisfying the second assumption of (2.3) cannot contain sufficiently many densities. As a consequence the posterior probability of $\mathscr{L}_n^c$ might not be sufficiently small.

For a given sequence (δ_n), define a specialized KL ball around p_0 as

$$\mathcal{K}_n = \left\{ p \in \mathscr{L} : K(p_0, p) \le \delta_n^2,\ P_0\Big(\log\frac{p_0}{p}\Big)^2 \le \delta_n^2 \right\}. \tag{2.5}$$

Note that $\Pi(\mathcal{K}_n) \ge e^{-n\delta_n^2}$ is a standard assumption to achieve the posterior convergence rate of at least (δ_n); see for example [12, 31]. Let $B_n = \{p \in \mathscr{L} : d_P(p, p_0) \le \epsilon_n\}$. Since d_P induces the weak topology, there exist a number $r > 0$ and finite number of bounded continuous functions $g_1, \ldots, g_k$ such that

$$B_n \supset \bigcap_{j=1}^k \left\{ p \in \mathscr{L} : \int g_j dP < \int g_j dP_0 + r \right\}. \tag{2.6}$$

Note that the number k of sub–bases and radius r may depend on p_0 and ϵ_n. The key idea for obtaining the Lévy–Prokhorov rate is to find these numbers. In this context, it is shown in Lemma 5.2 that for every $\epsilon_n \downarrow 0$ with $n\epsilon_n^4 \to \infty$, there exists a sequence of tests (φ_n) such that

$$P_0^n \varphi_n \le e^{-Kn\epsilon_n^2} \quad \text{and} \quad \sup_{\{p \in \mathscr{L} : d_P(p,p_0) \ge \epsilon_n\}} P^n(1 - \varphi_n) \le e^{-Kn\epsilon_n^2}, \tag{2.7}$$

where $K > 0$ is a universal constant. As a consequence, $\Pi(\mathcal{K}_n) \ge e^{-n\delta_n^2}$ implies (2.1) for every $\epsilon_n \gg \delta_n \vee n^{-1/4}$. Although $n^{-1/4}$ might be far away from the optimal rate, it is sufficient for strong consistency in many examples.

Theorem 2.2. *Suppose that p_0 satisfies (2.4) and $\Pi(\mathcal{K}_n) \ge e^{-c_1 n\delta_n^2}$ for a constant $c_1 > 0$ and a sequence $\delta_n \downarrow 0$. Then, (2.1) holds for every sequence (ϵ_n) with $\delta_n \vee n^{-1/4} \ll \epsilon_n$.*

Note that $\Pi(\mathcal{K}_n) \ge e^{-c_1 n\delta_n^2}$ and $\Pi(\mathscr{L}_n^c) \le e^{-(c_1+4)n\delta_n^2}$ implies

$$\mathbb{E}\Pi(\mathscr{L}_n^c | X_1, \ldots, X_n) = o(1),$$

see Lemma 8.1 of [12]. Combining this with the previous two theorems, we have the following corollary.

Corollary 2.1. *Under the assumption of Theorem 2.2, let (ϵ_n) be a sequence satisfying (2.1). Also, suppose that there exist $\mathscr{L}_n \subset \mathscr{L}$ such that $\Pi(\mathscr{L}_n^c) \le e^{-(c_1+4)n\delta_n^2}$ and $\sup_{p \in \mathscr{L}_n} d_V(p_{\gamma_n}, p) = o(1)$ for some (γ_n) with $\epsilon_n \ll \gamma_n \ll 1$. Then, $\Pi(\cdot | X_1, \ldots, X_n)$ is strongly consistent.*

3. Illustrations

3.1. Mixture of normal distributions

Consider a location mixture of normal distributions

$$p_{F,\sigma}(x) = (\phi_\sigma * F)(x) = \int \phi_\sigma(x - z) dF(z),$$

where $\phi_\sigma(x) = \phi(x/\sigma)/\sigma$, ϕ is the standard normal density and F is a probability measure. A prior Π on $\mathscr{L}$ can be constructed by putting independent priors for σ and F. With a slight abuse of notation, we use the notation Π for denoting both a prior for (σ, F) and a prior for p.

For $p = p_{F,\sigma}$, it can be shown that

$$d_V(p, p_\gamma) \lesssim \frac{\gamma}{\sigma}\left(1 + \frac{\gamma}{\sigma} + \frac{\gamma^2}{\sigma^2}\right), \tag{3.1}$$

see Lemma 5.3. Note that the right hand side of (3.1) depends on p through σ only. Therefore, sieves $(\mathscr{L}_n)$ can be constructed independent of F.

Theorem 3.1. *Let Π be a mixture of normal prior described above. Suppose that p_0 satisfies (2.4) and that (2.1) holds with a rate ϵ_n. Furthermore, if $\mathbb{E}\Pi(\mathscr{L}_n^c|X_1,\dots,X_n) = o(1)$, where*

$$\mathscr{L}_n = \{p_{F,\sigma} : \sigma \geq \sigma_n\} \quad \textit{for some } \sigma_n \gg \epsilon_n,$$

then $\Pi(\cdot|X_1,\dots,X_n)$ is strongly consistent.

For a concrete example, we consider an inverse gamma $\Gamma^{-1}(a_1, a_2)$ prior for σ^2, which is standard in both theory and practice, where $a_1, a_2 > 0$ are hyperparameters and $\Gamma^{-1}(a_1, a_2)$ denotes the inverse gamma distribution whose density is proportional to $x \mapsto x^{-a_1-1}e^{-a_2/x}$. Note that the prior on σ^2 puts little mass around zero implying that prior mass for large $d_V(p, p_\gamma)$ with small γ is nearly zero, *c.f.* (3.1).

Theorem 3.2. *Put a $\Gamma^{-1}(a_1, a_2)$ prior on σ^2. Suppose that p_0 satisfies (2.4) and $\Pi(\mathcal{K}_n) \geq e^{-cn\delta_n^2}$ for a sequence $\delta_n \ll n^{-1/4}$. Then, $\Pi(\cdot|X_1,\dots,X_n)$ is strongly consistent.*

In most examples $\Pi(\mathcal{K}_n) \geq e^{-n\delta_n^2}$ with δ_n much smaller than $n^{-1/4}$, so the condition given in Theorem 3.2 is very mild. A natural choice for the prior on F is $\mathrm{DP}(a_3, G)$, where $\mathrm{DP}(a_3, G)$ denotes the Dirichlet process with precision $a_3 > 0$ and mean G. For the Dirichlet process mixture of normal prior, the prior concentration condition has been extensively studied in literature, see for example [15, 26, 22, 4]. In most existing papers, the true density p_0 is firstly approximated by a finite mixture $p^*(x) = \sum_{j=1}^N w_j \phi_\sigma(x - z_j)$ with a sufficiently small number N, and then prove that a DP mixture prior puts sufficiently large mass around p^*. It should be noted that in the above mentioned papers, the tail of G must be exponentially thin to construct suitable sieves. Lijoi, Prünster and Walker [23] partly resolved this problem using the martingale approach of [29], but it is still required that G has a finite mean. With our approach, however, the only requirement is the prior concentration on $\mathcal{K}_n$ which holds if the tail of G is not extremely thin, see Proposition 2 in [4] for the most recent result. Therefore, conditions on G can be significantly weakened. For example, the Cauchy and heavier-tailed distributions can be taken for G which are not allowed with any other methods.

3.2. Mixture of Student's t distributions

If the true density p_0 is heavy-tailed, *e.g.* the tail is of a polynomial order, then it is theoretically unknown that Bayesian procedures based on normal mixtures work well. Practically, there are two possible methods to utilize a Dirichlet process mixture of normal for fitting data generated from a heavy-tailed distribution. The first one is to use a location-scale mixture. In this regard, Tokdar [27] proved the posterior consistency with a location-scale mixture under mild conditions. His result allows a heavy-tailed distribution such as Cauchy for the true density. Secondly, one may use a heavy-tailed mean parameter G. Unfortunately for both methods, it is challenging to generalize the theoretical results beyond consistency. In particular, existing mathematical tools for getting convergence rates might be difficult to apply because it is rarely possible to find (δ_n) satisfying $\Pi(\mathcal{K}_n) \geq e^{-n\delta_n^2}$ with a heavy-tailed p_0. We are not aware of whether this is due to the mathematical difficulty or the intrinsic limit of normal mixtures.

As an another alternative, we consider a mixture of Student's t distributions. While a mixture of Student's t distributions has been considered in some application, see for example [24, 10, 18], its asymptotic behavior has not been studied in the literature. In Bayesian analysis, this is due to the technical challenge for constructing suitable sieves with heavy-tailed components. Since the approach given in the present paper avoids the construction of complicated sieves, it can also be applied to Student's t mixtures.

Let h be the density of Student's t distribution with $v > 0$ degrees of freedom, and $h_\sigma(x) = \sigma^{-1} h(x/\sigma)$. For a fixed v, consider a location mixture of the form

$$p_{F,\sigma}(x) = \int h_\sigma(x-z) dF(z).$$

Similarly as in normal mixtures, for $p = p_{F,\sigma}$, we have

$$d_V(p, p_\gamma) \lesssim \frac{\gamma}{\sigma}\left(1 + \frac{\gamma}{\sigma}\right) \tag{3.2}$$

by Lemma 5.4, where the constant in the inequality depends only on v. A prior Π can be constructed by putting independent priors for σ and F. As in the case of normal mixtures, we can put an inverse gamma prior on σ^2. We abbreviate the proof of the following two theorems because after replacing (3.1) by (3.2), it is identical to the normal mixture case.

Theorem 3.3. *Let Π be a mixture of Student's t prior described above. Suppose that p_0 satisfies (2.4) and that (2.1) holds with a rate ϵ_n. Furthermore, if $\mathbb{E}\Pi(\mathscr{L}_n^c | X_1, \ldots, X_n) = o(1)$, where*

$$\mathscr{L}_n = \{p_{F,\sigma} : \sigma \geq \sigma_n\} \quad \textit{for some } \sigma_n \gg \epsilon_n,$$

then $\Pi(\cdot | X_1, \ldots, X_n)$ is strongly consistent.

Theorem 3.4. *Put a $\Gamma^{-1}(a_1, a_2)$ prior on σ^2. Suppose that p_0 satisfies (2.4) and $\Pi(\mathcal{K}_n) \geq e^{-n\delta_n^2}$ for a sequence $\delta_n \ll n^{-1/4}$. Then, $\Pi(\cdot|X_1, \ldots, X_n)$ is strongly consistent.*

We put $\mathrm{DP}(a_3, G)$ prior on F. Although the required condition for the prior concentration is mild, it is technically demanding to prove $\Pi(\mathcal{K}_n) \geq e^{-cn\delta_n^2}$. We imitate techniques known for normal mixtures. As mentioned earlier, the key part of the proof is the approximation of p_0, which can be approximated by $p_{F,\sigma}$ for some (F, σ), a finite mixture of normal distributions. To be a bit more specific, for any probability measure F on a compact interval $[-a, a]$, the total variation between $\phi_\sigma * F$ and $\phi_\sigma * F'$ is small if the first few, say N, moments of F and F' are the same, see Lemma 3.1 of [14]. Also, there exists discrete measures F' at most N components such that this moment condition is satisfied, see Lemma A.1 of [14]. Since Student's t distribution is a scale mixture of normal distributions [2], we have, see (5.3) for details,

$$d_V(p_{F,\sigma}, p_{F',\sigma}) \leq \int d_V(\phi_{\sigma\tau} * F, \phi_{\sigma\tau} * F') dH(\tau^{-2}),$$

where H is $\Gamma(v/2, v/2)$ distribution. Therefore, by applying the finite approximation technique of continuous normal mixtures, a mixture of Student's t distribution can also be approximated by a finite mixture. Combining with known concentration results for the Dirichlet distribution, we have the following theorem. Although the proof is long and quite similar to [15], we provide full details for the reader's convenience. We note that the main difference from normal mixtures is that a discrete measure F' should be constructed independent of the scale parameter, see Lemma 5.9.

Theorem 3.5. *Put independent $\Gamma^{-1}(a_1, a_2)$ and $DP(a_3, G)$ priors for σ^2 and F, respectively, where $v > 4$ and G is the standard Cauchy. Suppose that p_0 satisfies (2.4) and twice continuously differentiable with first and second order derivatives p_0' and p_0''. Furthermore, assume that $\int (p_0''/p_0)^2 p_0 d\mu < \infty$, $\int (p_0'/p_0)^4 p_0 d\mu < \infty$ and $P_0([-x, x]) \geq 1 - x^{-\beta}$ for some $\beta > 4/3$ and every large enough x. Then, $\Pi(\mathcal{K}_n) \geq e^{-n\delta_n^2}$ for some $\delta_n \ll n^{-1/4}$.*

Note that the mean parameter G of the Dirichlet process is assumed to be the standard Cauchy, but it can be replaced by other distribution whose tail is of a polynomial order. Although Theorem 3.5 cannot allow the Cauchy distribution as p_0 due to the tail assumption required for p_0, it is not difficult to extend the result further with more elaborate proof. For example, if p_0 is smoother than the twice differentiablity condition in Theorem 3.5, refined approximation techniques can be applied to obtain better rates as in [22, 4, 26].

4. Discussion

The key idea for the proof of Theorem 2.1 lies in the inequality (2.2). This can be extended to the consistency incorporating a rate (η_n). Assume for the moment that the support of p_0 is bounded. To find an upper bound of (2.2), we applied

$d_V(p, p_{\gamma_n}) \lesssim \gamma_n/\sigma_n$, $d_V(p_0, p_{0,\gamma_n}) \lesssim \gamma_n$ and $d_V(p_{0,\gamma_n}, p_{\gamma_n}) \lesssim \epsilon_n/\gamma_n$. By taking $\gamma_n^2 \asymp \epsilon_n\sigma_n$, a rate sequence (η_n) can be chosen as $\eta_n \asymp \sqrt{\epsilon_n/\sigma_n}$. However, this rate is far from optimal rate even when $\epsilon_n \asymp n^{-1/2}$ and $\sigma_n \asymp 1/\log n$. Better rates can be obtained if we have a better bounds for $d_V(p, p_{\gamma_n})$, $d_V(p_0, p_{0,\gamma_n})$ and $d_V(p_{0,\gamma_n}, p_{\gamma_n})$ (or similar quantities). For example, if p_0 belongs to a β-Hölder class, the bound for $d_V(p_0, p_{0,\gamma_n})$ might be improved to $d_V(p_0, p_{0,\gamma_n}) \lesssim \gamma_n^\beta$, as with normal mixtures [22]. We leave this more delicate analysis of rates as future work; and since our approach does not require entropy calculations, we believe that it can eliminate additional $\log n$ terms in the existing literature.

5. Proofs

5.1. *Proof of Theorem 2.1*

Lemma 5.1. *Let $\epsilon \equiv d_P(p, p_0)$. Then*

$$p_{0,\gamma}(x) \le p_\gamma(x) + \frac{\epsilon + 2\xi(\epsilon)}{2\gamma}$$

for every $\gamma > \epsilon$ and $x \in \mathbb{R}$, where $\xi(\epsilon) = \sup_{x\in\mathbb{R}} P_0(B_\epsilon(x))$.

Proof. For $\gamma > \epsilon$, note that $B_{\gamma-\epsilon}(x)$ and $B_\gamma(x)$ are equal to intervals $(x - \gamma + \epsilon, x + \gamma - \epsilon)$ and $(x - \gamma, x + \gamma)$, respectively. Thus,

$$\begin{aligned} P_0(B_{\gamma-\epsilon}(x)) &= P_0(B_\gamma(x)) - \int_{x-\gamma}^{x-\gamma+\epsilon} p_0(x)dx - \int_{x+\gamma-\epsilon}^{x+\gamma} p_0(x)dx \\ &\ge P_0(B_\gamma(x)) - 2\xi(\epsilon). \end{aligned}$$

From the definition of d_P, we have

$$\epsilon + P(B_\gamma(x)) \ge P_0(B_{\gamma-\epsilon}(x)).$$

Combining the last two display, we have $P_0(B_\gamma(x)) \le P(B_\gamma(x)) + \epsilon + 2\xi(\epsilon)$. By dividing both sides by 2γ, the proof is complete. □

Proof of Theorem 2.1. It suffices to prove that for every $\eta > 0$, $A_{n,\eta} \cap \mathscr{L}_n$ is an empty set for large enough n, where $A_{n,\eta}$ is defined as (1.2). For every $p \in A_{n,\eta}$, we have

$$\begin{aligned} d_V(p_{\gamma_n}, p) &\ge d_V(p, p_0) - d_V(p_0, p_{0,\gamma_n}) - d_V(p_{0,\gamma_n}, p_{\gamma_n}) \\ &\ge \eta - d_V(p_0, p_{0,\gamma_n}) - d_V(p_{0,\gamma_n}, p_{\gamma_n}). \end{aligned} \tag{5.1}$$

Note that $\gamma_n = o(1)$ implies $d_V(p_0, p_{0,\gamma_n}) = o(1)$. Also, for any $M > 0$ and $p \in A_{n,\eta}$, we have

$$\begin{aligned} d_V(p_{0,\gamma_n}, p_{\gamma_n}) &= 2 \sup_{B\in\mathcal{R}} |P_{0,\gamma_n}(B) - P_{\gamma_n}(B)| \\ &= 2\int \{p_{0,\gamma_n}(x) - p_{\gamma_n}(x)\}1_{\{p_{0,\gamma_n}(x) > p_{\gamma_n}(x)\}}dx \end{aligned}$$

$$\leq 2\int_{[-M,M]}\{p_{0,\gamma_n}(x)-p_{\gamma_n}(x)\}1_{\{p_{0,\gamma_n}(x)>p_{\gamma_n}(x)\}}dx$$
$$+2P_{0,\gamma_n}([-M,M]^c).$$

The last integral is bounded by $2M\gamma_n^{-1}(\epsilon_n+\xi(\epsilon_n))$ by Lemma 5.1. Note that if X and U_{γ_n} are independent random variables following P_0 and Unif$[\gamma_n,\gamma_n]$, respectively, the law of $X+U_{\gamma_n}$ is equal to P_{0,γ_n}. This implies that

$$\begin{aligned} P_{0,\gamma_n}([-M,M]^c) &= \Pr(|X+U_{\gamma_n}|>M)\leq \Pr(|X|>M-\gamma_n) \\ &= P_0([-M+\gamma_n,M-\gamma_n]^c). \end{aligned}$$

Therefore, we have

$$d_V(p_{0,\gamma_n},p_{\gamma_n})\leq 4M\frac{\epsilon_n+2\xi(\epsilon_n)}{\gamma_n}+2P_0([-M+\gamma_n,M-\gamma_n]^c)$$

Since $\xi(\epsilon_n)\lesssim\epsilon_n$ by (2.4), M can be arbitrarily large and $\epsilon_n\ll\gamma_n=o(1)$, the right hand side of (5.1) is bounded below by $\eta/2$ for every $p\in A_{n,\eta}$ and large enough n. Since $\sup_{p\in\mathscr{L}_n}d_V(p,p_{\gamma_n})=o(1)$, we conclude that $A_{n,\eta}\cap\mathscr{L}_n$ is an empty set for large enough n. ◻

5.2. *Proof of Theorem 2.2*

Lemma 5.2. *Assume that (2.4) holds. Then, for every $\epsilon_n\downarrow 0$ with $n\epsilon_n^4\to\infty$, there exists a sequence of tests φ_n such that (2.7) holds for a universal constant $K>0$ and every large enough n.*

Proof. For a given $\epsilon>0$, let $U=\{p\in\mathscr{L}:d_P(p,p_0)<\epsilon\}$. Let $M_\epsilon>0$ be a number such that $P_0(\{x:|x|>M_\epsilon\})\leq\epsilon/4$. By condition (2.4), M_ϵ can be chosen so that $M_\epsilon\lesssim\epsilon^{-1}$ for every small enough ϵ. Let N be the smallest positive integer greater than or equal to $2M_\epsilon/\epsilon$, and choose a real sequence $(a_j)_{j=0}^N$ such that $a_0\leq -M_\epsilon$ and $a_N\geq M_\epsilon$, where $a_j=a_{j-1}+\epsilon$ for $j=1,\ldots,N$. Let $B_{-\infty}=(-\infty,a_0]$, $B_\infty=(a_N,\infty)$ and $B_j=(a_{j-1},a_j]$ for $j=1,\ldots,N$. For $\delta>0$, define bounded continuous functions ψ_j, for $j=1,\ldots,N$, such that $\psi_j(x)=1$ for $x\in[a_{j-1}+\delta,a_j-\delta]$, $\psi_j(x)=0$ for $x\leq a_{j-1}$ or $x\geq a_j$ and ψ_j is linear on the intervals $[a_{j-1},a_{j-1}+\delta]$ and $[a_j-\delta,a_j]$. We can choose δ sufficiently small so that

$$P_0(B_j)\leq P_0\psi_j+\frac{\epsilon}{4N}\quad\text{for } j=1,\ldots,N.$$

For $k=(k_1,\ldots,k_N)\in\{0,1\}^N$ with $k\neq(0,\ldots,0)$, let $g_k=\sum_{k_j=1}\psi_j$ and

$$U_k=\left\{p\in\mathscr{L}:P_0g_k\leq Pg_k+\frac{\epsilon}{2}\right\}.$$

For $A\in\mathcal{R}$, let $J=\{j:1\leq j\leq N,A\cap B_j\neq\emptyset\}$. Then, for $p\in\cap_k U_k$, where the intersection is taken over 2^N-1 indices, we have

$$P_0(A) \le \frac{\epsilon}{4} + \sum_{j\in J} P_0(A\cap B_j) \le \frac{\epsilon}{4} + \sum_{j\in J} P_0(B_j) \le \frac{\epsilon}{4} + \sum_{j\in J}\left(P_0\psi_j + \frac{\epsilon}{4N}\right)$$
$$\le \frac{\epsilon}{2} + \sum_{j\in J} P_0\psi_j \le \epsilon + \sum_{j\in J} P\psi_j \le \epsilon + \sum_{j\in J} P(B_j) \le \epsilon + P(A^\epsilon).$$

It follows that $\cap_k U_k \subset U$.

Let

$$\varphi_{n,k} = 1_{\{n^{-1}\sum_{i=1}^n g_k(X_i) < P_0 g_k - \epsilon/4\}}$$

and $\varphi_n = \max_k \varphi_{n,k}$. Since g_k is bounded by 1, Hoeffding's inequality [19] implies

$$P_0^n\varphi_{n,k} = P_0^n\left(\frac{1}{n}\sum_{i=1}^n g_k(X_i) < P_0 g_k - \frac{\epsilon}{4}\right) \le \exp\left(-\frac{n\epsilon^2}{8}\right)$$

and for $p \in U_k^c$,

$$\begin{aligned} P^n(1-\varphi_{n,k}) &= P^n\left(\frac{1}{n}\sum_{i=1}^n g_k(X_i) \ge P_0 g_k - \frac{\epsilon}{4}\right) \\ &= P^n\left(\frac{1}{n}\sum_{i=1}^n g_k(X_i) \ge P g_k + (P_0 - P)g_k - \frac{\epsilon}{4}\right) \\ &\le P^n\left(\frac{1}{n}\sum_{i=1}^n g_k(X_i) \ge P g_k + \frac{\epsilon}{4}\right) \le \exp\left(-\frac{n\epsilon^2}{8}\right). \end{aligned}$$

The last display implies

$$\sup_{p\in U^c} P^n(1-\varphi_n) \le \exp\left(-\frac{n\epsilon^2}{8}\right).$$

If $N \le n\epsilon^2/(16\log 2)$, we have

$$P_0^n\varphi_n \le \sum_k P_0^n\varphi_{n,k} \le 2^N \exp\left(-\frac{n\epsilon^2}{8}\right) \le \exp\left(-\frac{n\epsilon^2}{16}\right).$$

Since $N \lesssim M_\epsilon/\epsilon \lesssim \epsilon^{-2}$, the desired sequence of tests exists provided that $n\epsilon_n^4$ is bigger than a universal constant. ☐

Proof of Theorem 2.2. Let ϵ_n be a sequence such that $\delta_n \vee n^{-1/4} \ll \epsilon_n \ll 1$ and $A_n = \{p \in \mathscr{L} : d_P(p_0, p) \ge \epsilon_n\}$. By Lemma 8.1 of [12] and Lemma 5.2, there exists a constant $c_2 > 0$ such that $P_0^n(\Omega_n) \to 1$, where Ω_n is the event that $\int R_n(p)d\Pi(p) > e^{-c_2 n\delta_n^2}$. Also, by Lemma 5.2, there exist a constant $K > 0$ and a sequence of tests (φ_n) satisfying (2.7) for every large enough n. It follows that $\Pi(A_n|X_1,\ldots,X_n) = \Pi(A_n|X_1,\ldots,X_n)1_{\Omega_n}(1-\varphi_n) + o_{P_0}(1)$. Since

$$\begin{aligned} \mathbb{E}\Pi(A_n|X_1,\ldots,X_n)1_{\Omega_n}(1-\varphi_n) &\le e^{c_2 n\delta_n^2}\mathbb{E}\int_{A_n}(1-\varphi_n)R_n(p)d\Pi(p) \\ &\le e^{c_2 n\delta_n^2}\sup_{p\in A_n} P^n(1-\varphi_n) \le e^{c_2 n\delta_n^2 - Kn\epsilon_n^2} = o(1), \end{aligned}$$

the proof is complete. ☐

5.3. Proof of Theorem 3.1

Lemma 5.3. *For $p(x) = \int \phi_\sigma(x-z)dF(z,\sigma)$, where F is a probability measure on $\mathbb{R} \times (0,\infty)$, we have*

$$d_V(p, p_\gamma) \lesssim \int \frac{\gamma}{\sigma}\left(1 + \frac{\gamma}{\sigma} + \frac{\gamma^2}{\sigma^2}\right) dF(z,\sigma).$$

Proof. Let $\gamma > 0$ be given, $\phi'_\sigma(x) = \partial\phi_\sigma(x)/\partial x$ and $g_\sigma(x) = \sup_{|y-x|<\gamma} |\phi'_\sigma(y)|$. Since

$$\phi'_\sigma(x) = -\frac{x}{\sigma^2}\phi_\sigma(x)$$

and

$$\sup_{|y-x|<\gamma} \phi_\sigma(x) = \begin{cases} \phi_\sigma(x-\gamma) & \text{if } x > \gamma \\ \phi_\sigma(x+\gamma) & \text{if } x < -\gamma \\ \phi_\sigma(0) & \text{if } |x| \le \gamma \end{cases}$$

we have

$$\begin{aligned} &\int_{-\infty}^{\infty} g_\sigma(x)dx \\ \le\ & \int_\gamma^\infty \frac{x+\gamma}{\sigma^2}\phi_\sigma(x-\gamma)dx + \int_{-\infty}^{-\gamma} \frac{-x+\gamma}{\sigma^2}\phi_\sigma(x+\gamma)dx + \int_{-\gamma}^{\gamma} \frac{2\gamma}{\sigma^2}\phi_\sigma(0)dx \\ \le\ & \int_0^\infty \frac{x+2\gamma}{\sigma^2}\phi_\sigma(x)dx + \int_{-\infty}^0 \frac{-x+2\gamma}{\sigma^2}\phi_\sigma(x)dx + \frac{4\gamma^2}{\sigma^2}\phi_\sigma(0) \\ =\ & \int_{-\infty}^\infty \frac{|x|}{\sigma^2}\phi_\sigma(x)dx + \int_{-\infty}^\infty \frac{2\gamma}{\sigma^2}\phi_\sigma(x)dx + \frac{4\gamma^2}{\sigma^2}\phi_\sigma(0) \\ \lesssim\ & \frac{1}{\sigma}\left(1 + \frac{\gamma}{\sigma} + \frac{\gamma^2}{\sigma^2}\right). \end{aligned}$$

Note that

$$\begin{aligned} |p(x) - p_\gamma(x)| \le& \frac{1}{2\gamma}\int |p(x) - p(y)| 1_{\{|y-x|<\gamma\}} dy \\ \le& \frac{1}{2\gamma}\int\int |\phi_\sigma(x-z) - \phi_\sigma(y-z)| 1_{\{|y-x|<\gamma\}} dy dF(z,\sigma) \\ \le& \frac{1}{2\gamma}\int\int |x-y| g_\sigma(x-z) 1_{\{|y-x|<\gamma\}} dy dF(z,\sigma) \\ \le& \gamma \int g_\sigma(x-z) dF(z,\sigma), \end{aligned} \tag{5.2}$$

where the third inequality holds by the Taylor expansion. It follows that

$$\begin{aligned} d_V(p, p_\gamma) &= \int |p(x) - p_\gamma(x)| dx \le \gamma \int\int g_\sigma(x-z) dx dF(z,\sigma) \\ &\lesssim \int \frac{\gamma}{\sigma}\left(1 + \frac{\gamma}{\sigma} + \frac{\gamma^2}{\sigma^2}\right) dF(z,\sigma). \end{aligned}$$

This completes the proof. ∎

Proof of Theorem 3.1. Take a sequence (γ_n) such that $\epsilon_n \ll \gamma_n \ll \sigma_n$. Then,

$$\sup_{p\in\mathscr{L}_n} d_V(p, p_{\gamma_n}) \lesssim \frac{\gamma_n}{\sigma_n} = o(1)$$

by Lemma 5.3. Therefore, strong consistency holds by Theorem 2.1. □

5.4. Proof of Theorem 3.2

For a sufficiently slowly diverging (will be described below) sequence $(M_n) \to \infty$, let $\epsilon_n = n^{-1/4} M_n$ and $\gamma_n = \epsilon_n M_n$. If M_n grows sufficiently slowly, we can choose a seqeunce (β_n) such that $n\delta_n^4 M_n^8 \ll \beta_n^4$ and $\gamma_n \ll \beta_n \ll 1$.

Let $\mathscr{L}_n = \{p_{F,\sigma} : \gamma_n/\sigma \le \beta_n\}$ Then,

$$\Pi\left(\mathscr{L}_n^c\right) = \Pi\left(\sigma^2 < \frac{\gamma_n^2}{\beta_n^2}\right) = C\int_0^{\gamma_n^2/\beta_n^2} x^{-a_1-1}e^{-a_2/x}dx$$
$$\le \int_0^{\gamma_n^2/\beta_n^2} e^{-a_2/(2x)}dx \le \frac{\gamma_n^2}{\beta_n^2}e^{-a_2\beta_n^2/(2\gamma_n^2)} \le e^{-a_2\beta_n^2/(2\gamma_n^2)}$$

for large enough n, where C is a constant depending only on a_1 and a_2. By the construction of (β_n), we have

$$\frac{\beta_n^2}{\gamma_n^2} \gg \frac{\sqrt{n}\delta_n^2 M_n^4}{\gamma_n^2} = n\delta_n^2.$$

It follows that $\Pi(\mathscr{L}_n^c) \le e^{-5n\delta_n^2}$. Also, for any $p = p_{F,\sigma}$ in $\mathscr{L}_n$, we have

$$d_V(p_{\gamma_n}, p) \lesssim \frac{\gamma_n}{\sigma} \le \beta_n = o(1),$$

where the first inequality holds by Lemma 5.3. Therefore, the strong consistency holds by Corollary 2.1. □

5.5. Proof of Theorem 3.5

Throughout this subsection, h is the density of Student's t distribution with v degrees of freedom, $h_\sigma(x) = \sigma^{-1}h(x/\sigma)$, and constants in $\lesssim$ may depend on v.

Lemma 5.4. *Let F be a probability measure on $\mathbb{R}$ and $p(x) = \int h_\sigma(x-z)dF(z)$. Then, (3.2) holds, where the constant in the inequality depends only on the degree of freedom v.*

Proof. Let $\gamma > 0$ be given, $h'_\sigma(x) = \partial h_\sigma(x)/\partial x$ and $g_\sigma(x) = \sup_{|y-x|<\gamma}|h'_\sigma(y)|$. Since

$$|h'_\sigma(x)| = (v+1)h_\sigma(x)\frac{|x|}{v\sigma^2 + x^2} \le \frac{v+1}{2\sqrt{v}\sigma}h_\sigma(x)$$

and

$$\sup_{|y-x|<\gamma} h_\sigma(x) = \begin{cases} h_\sigma(x-\gamma) & \text{if } x > \gamma \\ h_\sigma(x+\gamma) & \text{if } x < -\gamma \\ h_\sigma(0) & \text{if } |x| \leq \gamma \end{cases}$$

we have

$$\int_{-\infty}^{\infty} g_\sigma(x)dx \lesssim \int_{\{x:|x|>\gamma\}} \frac{h_\sigma(|x|-\gamma)}{\sigma}dx + \int_{\{x:|x|\leq\gamma\}} \frac{h_\sigma(0)}{\sigma}dx \lesssim \frac{1}{\sigma}\left(1+\frac{\gamma}{\sigma}\right).$$

As (5.2), we have

$$|p(x) - p_\gamma(x)| \leq \gamma \int g_\sigma(x-z)dF(z).$$

Therefore, the desired bound for $d_V(p, p_\gamma)$ can be obtained by the Fubini's theorem. ☐

Lemma 5.5. *Let $p(x) = \int h_\sigma(x-z)dF(z)$ with $\sigma \leq 1$. Suppose that p_0 satisfies (2.4) and $\int |z|^{\delta(v+1)}dF(z) \vee \int |x|^{\delta(v+1)}dP_0(x) \leq M$ for some positive $\delta \leq (v+1)^{-1}$ and $M \geq 1$. Then,*

$$\int \left(\frac{p_0}{p}\right)^\delta dP_0 \lesssim M\sigma^{-v\delta},$$

where the constant in $\lesssim$ depends only on v.

Proof. Note that

$$\begin{aligned}
&\int \{h_\sigma(x-z)\}^\delta dF(z) \gtrsim \int \frac{\sigma^{-\delta}}{\{1+(x-z)^2/(v\sigma^2)\}^{\delta(v+1)/2}}dF(z) \\
&\geq \frac{\sigma^{-\delta}}{\int\{1+(x-z)^2/(v\sigma^2)\}^{\delta(v+1)/2}dF(z)} \\
&\gtrsim \frac{\sigma^{-\delta}}{1+\{x^2/(v\sigma^2)\}^{\delta(v+1)/2}+\int\{z^2/(v\sigma^2)\}^{\delta(v+1)/2}dF(z)} \\
&\geq \frac{\sigma^{-\delta}}{1+(v\sigma^2)^{-\delta(v+1)/2}(|x|^{\delta(v+1)}+M)} \gtrsim \frac{1}{\sigma^\delta + \sigma^{-\delta v}(|x|^{\delta(v+1)}+M)},
\end{aligned}$$

where the second inequality holds by Jensen, the third inequality holds by that $(a+b)^2 \leq 2(a^2+b^2)$ and $(a+b)^\zeta \leq a^\zeta + b^\zeta$ for any positive numbers a, b and $\zeta \leq 1$, and the last inequality holds by treating v as a constant. Therefore,

$$\begin{aligned}
\int \left(\frac{p_0}{p}\right)^\delta dP_0 &= \int \left[\frac{p_0(x)}{\int h_\sigma(x-z)dF(z)}\right]^\delta dP_0(x) \\
&\lesssim \int \frac{1}{\int\{h_\sigma(x-z)\}^\delta dF(z)}dP_0(x) \\
&\lesssim \sigma^\delta + \sigma^{-\delta v}\left(\int |x|^{\delta(v+1)}dP_0(x) + M\right) \lesssim \sigma^{-\delta v}M,
\end{aligned}$$

where the first inequality holds by Jensen. ☐

Lemma 5.6. *Let $(T_z h_\sigma)(x) = h_\sigma(x-z)$. For $\sigma, \sigma_1, \sigma_2 > 0$ and $z \in \mathbb{R}$,*

$$d_V(T_z h_\sigma, h_\sigma) \lesssim \frac{|z|}{\sigma} \quad \textit{and} \quad d_V(h_{\sigma_1}, h_{\sigma_2}) \lesssim \frac{|\sigma_1 - \sigma_2|}{\min\{\sigma_1, \sigma_2\}},$$

where constants in $\lesssim$ depends only on v.

Proof. Let

$$h'_\sigma(x) = \frac{\partial h_\sigma(x)}{\partial x} = -\frac{(v+1)x}{v\sigma^2 + x^2} h_\sigma(x)$$

and

$$\dot{h}_\sigma(x) = \frac{\partial h_\sigma(x)}{\partial \sigma} = \frac{1}{\sigma} h_\sigma(x) \left(\frac{(v+1)x^2}{v\sigma^2 + x^2} - 1 \right).$$

Let g be the density proportional to $(1+x^2/v)^{-(v+1)/2-1}$ and $g_\sigma(x) = \sigma^{-1} g(x/\sigma)$. Then, by the Taylor expansion, we have

$$\begin{aligned} & d_V(T_z h_\sigma, h_\sigma) = \int |h_\sigma(x-z) - h_\sigma(x)| dx \leq \int |z| \int_0^1 |h'_\sigma(x - tz)| dt dx \\ & = |z| \int |h'_\sigma(x)| dx \lesssim \frac{|z|}{\sigma^2} \int \frac{|x|}{1 + x^2/(v\sigma^2)} h_\sigma(x) dx \lesssim \frac{|z|}{\sigma^2} \int |x| g_\sigma(x) dx \\ & \leq \frac{|z|}{\sigma} \int \frac{|x|}{\sigma} g_\sigma(x) dx = \frac{|z|}{\sigma}. \end{aligned}$$

This proves the first inequality.

For the second inequality, assume that $\sigma_1 < \sigma_2$ without loss of generality, and let $\sigma_t = \sigma_1 + t(\sigma_2 - \sigma_1)$ for $t \in (0,1)$. Then,

$$d_V(h_{\sigma_1}, h_{\sigma_2}) = \int |h_{\sigma_1}(x) - h_{\sigma_2}(x)| dx \leq \int |\sigma_1 - \sigma_2| \int_0^1 |\dot{h}_{\sigma_t}(x)| dt dx.$$

Since

$$\int \frac{1}{\sigma} \frac{(v+1)x^2}{v\sigma^2 + x^2} h_\sigma(x) dx \lesssim \int \frac{x^2}{\sigma^3} g_\sigma(x) dx \lesssim \frac{1}{\sigma},$$

$d_V(h_{\sigma_1}, h_{\sigma_2})$ is bounded by a constant multiple of $|\sigma_1 - \sigma_2|/\sigma_1$. □

Lemma 5.7. *Let $F^* = \sum_{j=1}^N p_j \delta_{z_j}$ such that $\sum_{j=1}^N p_j = 1$ and $|z_j - z_k| \geq 2\epsilon$ for $j \neq k$, where δ_z denotes the Dirac measure at z. Then,*

$$d_V(p_{F,\sigma}, p_{F^*,\sigma^*}) \lesssim \frac{\epsilon + |\sigma - \sigma^*|}{\min\{\sigma, \sigma^*\}} + \sum_{j=1}^N \big| F[z_j - \epsilon, z_j + \epsilon] - p_j \big|,$$

where the constant in $\lesssim$ depends only on v.

Proof. Note that

$$d_V(p_{F,\sigma}, p_{F^*,\sigma^*}) \leq d_V(p_{F,\sigma}, p_{F^*,\sigma}) + d_V(p_{F^*,\sigma}, p_{F^*,\sigma^*})$$

and

$$d_V(p_{F^*,\sigma}, p_{F^*,\sigma^*}) \le d_V(h_\sigma, h_{\sigma^*}) \lesssim \frac{|\sigma - \sigma^*|}{\min\{\sigma, \sigma^*\}},$$

where the last inequality holds by Lemma 5.6. Also,

$$|p_{F,\sigma}(x) - p_{F^*,\sigma}(x)| = \left|\int h_\sigma(x-z)d(F-F^*)(z)\right|$$

$$\le \int_{\{|z-z_j|>\epsilon,\forall j\}} h_\sigma(x-z)dF(z) + \sum_{j=1}^N \left|\int_{\{|z-z_j|\le\epsilon\}} h_\sigma(x-z)d(F-F^*)(z)\right|$$

and the summand in the right hand side is bounded by

$$\int_{\{|z-z_j|\le\epsilon\}} |h_\sigma(x-z) - h_\sigma(x-z_j)|dF(z) + h_\sigma(x-z_j)\big|F[z_j-\epsilon, z_j+\epsilon] - p_j\big|.$$

Combining the last two displays, we have

$$d_V(p_{F,\sigma}, p_{F^*,\sigma}) \le F\Big(z : |z-z_j| > \epsilon, \forall j\Big)$$
$$+ \sum_{j=1}^N \int_{\{|z-z_j|<\epsilon\}} d_V(T_z h_\sigma, T_{z_j} h_\sigma)dF(z) + \sum_{j=1}^N \big|F[z_j-\epsilon, z_j+\epsilon] - p_j\big|,$$

where $T_z h_\sigma$ is defined as in Lemma 5.6. Since

$$F\Big(z : |z-z_j| > \epsilon, \forall j\Big) = 1 - \sum_{j=1}^N F[z_j-\epsilon, z_j+\epsilon] \le \sum_{j=1}^N \big|F[z_j-\epsilon, z_j+\epsilon] - p_j\big|,$$

$d_V(p_{F,\sigma}, p_{F^*,\sigma})$ is bounded by a constant multiple of

$$\frac{\epsilon}{\sigma} + \sum_{j=1}^N \big|F[z_j-\epsilon, z_j+\epsilon] - p_j\big|$$

by Lemma 5.6. □

Lemma 5.8. *Let $p(x) = \int h_\sigma(x-z)dP_0(z)$. Assume that p_0 is twice continuously differentiable with first and second order derivatives p_0' and p_0''.*

- *If $v > 2$ and p_0 is bounded with $\int |p_0''|d\mu < \infty$, then $d_V(p_0, p) \le c_1\sigma^2$.*
- *If $v > 4$, $\int (p_0''/p_0)^2 p_0 d\mu < \infty$ and $\int (p_0'/p_0)^4 p_0 d\mu < \infty$, then $d_H(p_0, p) \le c_2\sigma^2$.*

In both cases, constants c_1, c_2 depend on v and given integrals only.

Proof. By the Taylor expansion with the integral form of the remainder, we have

$$p_0(x+y) - p_0(x) = yp_0'(x) + y^2 \int_0^1 (1-t)p_0''(x+ty)dt$$

for every x and y. Since $p(x) = \int p_0(x + \sigma y)h(y)dy$, we have

$$\begin{aligned} p(x) - p_0(x) &= \int \{p_0(x + \sigma y) - p_0(x)\}h(y)dy \\ &= \int \left\{ \sigma y p_0'(x) + \sigma^2 y^2 \int_0^1 (1-t)p_0''(x + t\sigma y)dt \right\} h(y)dy \\ &= \int \int_0^1 \sigma^2 y^2 (1-t) p_0''(x + t\sigma y)h(y)\, dt\, dy. \end{aligned}$$

Thus, for some constant $c_1 > 0$,

$$d_V(p, p_0) \le \tfrac{1}{2} c_1 \sigma^2 \int y^2 h(y)dy,$$

where the last integral is finite for $v > 2$.

The proof for the second inequality is identical to Lemma 4 of [15], for which $\int y^4 h(y)dy < \infty$ is required. □

The following lemma is an extension of Lemma 2 in [15] in the sense that a discrete probability measure F' can be taken independent of $\sigma \ge \sigma_0$.

Lemma 5.9. *Let $a, \sigma_0, \epsilon > 0$ be given numbers such that $a/\sigma_0 \ge 1$. For any probability measure F on $[-a, a]$, there exists a discrete probability measure F' on $[-a, a]$ with fewer than $Da\sigma_0^{-1} \log \epsilon^{-1}$ support points, such that*

$$d_V(\phi_\sigma * F, \phi_\sigma * F') \lesssim \epsilon(\log \epsilon^{-1})^{1/2}$$

for every $\sigma \ge \sigma_0$, where D is a universal constant.

Proof. Throughout this proof, $p_{F,\sigma}$ denotes $\phi_\sigma * F$, not $h_\sigma * F$. Partition the interval $[-a, a]$ into k disjoint, consecutive subintervals $I_1, \ldots, I_k$ of length σ_0 and a final interval I_{k+1} of length l_{k+1} smaller than σ, where k is the largest integer less than or equal to $2a/\sigma_0$. Write $F = \sum_{i=1}^{k+1} F(I_i)F_i$, where each F_i is a probability measure concentrated on I_i, then $p_{F,\sigma} = \sum_{i=1}^{k+1} F(I_i)p_{F_i,\sigma}$. Let Z_i be a random variable distributed according to F_i, and for a_i the left endpoint of I_i, let G_i be the law of $W_i = (Z_i - a_i)/\sigma_0$. For $\sigma \ge \sigma_0$, let $G_{i,\sigma}$ be the law of $W_{i,\sigma} = W_i\sigma_0/\sigma$. Thus, G_i and $G_{i,\sigma}$ are supported on $[0, 1]$ and $[0, \sigma_0/\sigma] \subset [0, 1]$, respectively.

As the proof of Lemma 2 in [15], it can be shown that for each i, there exists a discrete probability measure G_i' with fewer than $N_i \lesssim \log \epsilon^{-1}$ support points such that $d_V(p_{G_i,1}, p_{G_i',1}) \lesssim \epsilon(\log \epsilon^{-1})^{1/2}$. Note that, from the construction, the first N_i moments of G_i and G_i' are identical, see the proof of Lemma 3.1 in [14]. Let $G_{i,\sigma}'$ be the law of $W_{i,\sigma}' = W_i'\sigma_0/\sigma$, where W_i' is a random variable distributed according to G_i'. Then, the first N_i moments of $G_{i,\sigma}$ and $G_{i,\sigma}'$ are also identical, so $d_V(p_{G_{i,\sigma},1}, p_{G_{i,\sigma}',1}) \lesssim \epsilon(\log \epsilon^{-1})^{1/2}$ by Lemmas 3.1 and 3.2 in [14]. Let F_i' be the law of $a_i + \sigma_0 W_i'$ and set $F' = \sum_{i=1}^{k+1} F(I_i)F_i'$. Note that

$$p_{F_i,\sigma}(x) = \mathbb{E}\phi_\sigma(x - Z_i) = \sigma^{-1}\mathbb{E}\phi\big((x - a_i)/\sigma - W_{i,\sigma}\big) = \sigma^{-1} p_{G_{i,\sigma},1}\big((x - a_i)/\sigma\big)$$

and similarly for F_i' and $G_{i,\sigma}'$. It follows that

$$d_V(p_{F_i,\sigma}, p_{F_i',\sigma}) = d_V(p_{G_{i,\sigma},1}, p_{G_{i,\sigma}',1}) \lesssim \epsilon(\log \epsilon^{-1})^{1/2}.$$

Since $d_V(p_{F,\sigma}, p_{F',\sigma}) \leq \sum_{i=1}^{k+1} F(I_i) d_V(p_{F_i,\sigma}, p_{F_i',\sigma}) \lesssim \epsilon(\log \epsilon^{-1})^{1/2}$ and the number of support points of F' is bounded by $\sum_{i=1}^{k+1} N_i \lesssim a\sigma_0^{-1}\log \epsilon^{-1}$, the proof is complete. □

Lemma 5.10. *Let $a, \sigma, \epsilon > 0$ be given numbers such that $a/\sigma \geq 1$. For any probability measure F on $[-a, a]$, there exists a discrete probability measure F' on $[-a, a]$ with fewer than $Da\sigma^{-1}(\log \epsilon^{-1})^{3/2}$ support points, such that*

$$d_V(p_{F,\sigma}, p_{F',\sigma}) \lesssim \epsilon(\log \epsilon^{-1})^{1/2},$$

where D and the constant in $\lesssim$ depend only on v.

Proof. Since d_V is bounded by 2, we may assume that $\epsilon > 0$ is sufficiently small. Note that $h(x) = \int \phi_\tau(x) dH(\tau^{-2})$, where H is $\Gamma(v/2, v/2)$ distribution (mean 1 and variance $2/v$), see [2]. Thus, $h_\sigma(x) = \int \phi_{\sigma\tau}(x) dH(\tau^{-2})$. Let F be a given probability measure on $[-a, a]$. Then, for any probability measure F' on $[-a, a]$, we have

$$\begin{aligned}
&d_V(p_{F,\sigma}, p_{F',\sigma}) = \int \left| \int h_\sigma(x-z) dF(z) - \int h_\sigma(x-z) dF'(z) \right| dx \\
&= \int \left| \int \int \phi_{\sigma\tau}(x-z) dH(\tau^{-2}) dF(z) - \int \int \phi_{\sigma\tau}(x-z) dH(\tau^{-2}) dF'(z) \right| dx \\
&\leq \int \int \left| \int \phi_{\sigma\tau}(x-z) dF(z) - \int \phi_{\sigma\tau}(x-z) dF'(z) \right| dx dH(\tau^{-2}) \\
&= \int d_V(\phi_{\sigma\tau} * F, \phi_{\sigma\tau} * F') dH(\tau^{-2}).
\end{aligned} \tag{5.3}$$

Since H is a Gamma distribution, we have

$$H([t, \infty)) \lesssim \int_t^\infty x^{v/2-1} e^{-vx/2} dx \leq \int_t^\infty e^{-vx/3} dx \lesssim e^{-vt/3}$$

for every large enough t. The right hand side of the last display is bounded by ϵ provided that $t \geq 3v^{-1}\log \epsilon^{-1}$. Thus, the right hand side of (5.3) is bounded by

$$C\epsilon + \int_{\{\tau^2 > \frac{v}{3}(\log \epsilon^{-1})^{-1}\}} d_V(\phi_{\sigma\tau} * F, \phi_{\sigma\tau} * F') dH(\tau^{-2}), \tag{5.4}$$

where $C > 0$ is a constant depending only on v. By Lemma 5.9, there exists a discrete probability measure F', with fewer than

$$D\sqrt{\frac{3}{v}} a\sigma^{-1}(\log \epsilon^{-1})^{3/2} \lesssim \frac{a}{\sigma}(\log \epsilon^{-1})^{3/2}$$

support points, such that $d_V(\phi_{\sigma\tau} * F, \phi_{\sigma\tau} * F') \lesssim \epsilon(\log \epsilon^{-1})^{1/2}$ for every τ with $\tau^2 > \frac{v}{3}(\log \epsilon^{-1})^{-1}$, where D is a universal constant. Therefore, the right hand side of (5.4) is bounded by a multiple of $\epsilon(\log \epsilon^{-1})^{1/2}$. □

Proof of Theorem 3.5. Throughout this proof, constants in the notation $\lesssim$ may depend on hyperparameters a_1, a_2, a_3, v, G and the true density p_0. For a given sufficiently small positive number σ^*, let P_0^* be a restricted renormalization of P_0 on $[-a, a]$, where a is the minimal number satisfying $P_0[-a, a] \geq 1 - (\sigma^*)^4$. The tail condition of p_0 implies that $a \leq (\sigma^*)^{-4/\beta}$. By Lemma A.3 of [14], $d_V(p_{P_0,\sigma^*}, p_{P_0^*,\sigma^*}) \lesssim (\sigma^*)^4$. Also, $d_H(p_0, p_{P_0,\sigma^*}) \lesssim (\sigma^*)^2$ by Lemma 5.8. Furthermore, by Lemma 5.10, there exists a discrete probability measure $F^* = \sum_{j=1}^N p_j^* \delta_{z_j^*}$ such that $N \lesssim a(\sigma^*)^{-1}(\log(\sigma^*)^{-1})^{3/2}$ and $d_V(p_{P_0^*,\sigma^*}, p_{F^*,\sigma^*}) \lesssim (\sigma^*)^4$. Therefore, for a probability measure F and $\sigma > 0$, we have

$$\begin{aligned}
& d_H^2(p_0, p_{F,\sigma}) \\
& \lesssim d_H^2(p_0, p_{P_0,\sigma^*}) + d_H^2(p_{P_0,\sigma^*}, p_{P_0^*,\sigma^*}) + d_H^2(p_{P_0^*,\sigma^*}, p_{F^*,\sigma^*}) + d_H^2(p_{F^*,\sigma^*}, p_{F,\sigma}) \\
& \lesssim (\sigma^*)^4 + d_V(p_{P_0,\sigma^*}, p_{P_0^*,\sigma^*}) + d_V(p_{P_0^*,\sigma^*}, p_{F^*,\sigma^*}) + d_V(p_{F^*,\sigma^*}, p_{F,\sigma}) \\
& \lesssim (\sigma^*)^4 + d_V(p_{F^*,\sigma^*}, p_{F,\sigma}).
\end{aligned}$$

Without loss of generality, we may assume that the support points of F^* are at least $2(\sigma^*)^5$ separated. Otherwise, take a maximal $2(\sigma^*)^5$-separated set in the support points of F^*, and let F^{**} be the discrete measure on this $2(\sigma^*)^5$-net with weights obtained by moving the masses in F^* to the closest point in the support of F^{**}. Then, $d_V(p_{F^*,\sigma^*}, p_{F^{**},\sigma^*}) \lesssim (\sigma^*)^4$ by Lemma 5.7, and hence we can replace F^* by F^{**}.

From the last display, we have

$$\begin{aligned}
& \Pi\Big(p_{F,\sigma} : d_H^2(p_0, p_{F,\sigma}) \lesssim (\sigma^*)^4\Big) \geq \Pi\Big((\sigma, F) : d_V(p_{F^*,\sigma^*}, p_{F,\sigma}) \leq (\sigma^*)^4\Big) \\
& \geq \Pi\left((\sigma, F) : |\sigma - \sigma^*| \lesssim (\sigma^*)^5, \sum_{j=1}^N |F[z_j^* - (\sigma^*)^5, z_j^* + (\sigma^*)^5] - p_j^*| \lesssim (\sigma^*)^4\right),
\end{aligned}$$

where the last inequality holds by Lemma 5.7. For F with $\sum_{j=1}^N |F[z_j^* - (\sigma^*)^5, z_j^* + (\sigma^*)^5] - p_j^*| \leq (\sigma^*)^4$, it holds that $F[-a, a] \geq 1/2$, and thus $\int (p_0/p_{F,\sigma})^\delta dP_0 \lesssim a\sigma^{-v\delta}$ for every small enough σ and δ by Lemma 5.5. Let (δ_n) be a sequence decreasing to zero and set $(\sigma^*)^4 = \delta_n^{2+\epsilon}$, where ϵ is sufficiently small as described below. Then, by Lemma 5.5 and Theorem 5 of [32], $|\sigma - \sigma^*| \lesssim (\sigma^*)^5$ and $\sum_{j=1}^N |F[z_j^* - (\sigma^*)^5, z_j^* + (\sigma^*)^5] - p_j^*| \lesssim (\sigma^*)^4$ imply $p_{F,\sigma} \in \mathcal{K}_n$ for large enough n.

Since σ^2 follows an inverse gamma distribution,

$$\Pi\Big(\sigma : |\sigma - \sigma^*| \lesssim (\sigma^*)^5\Big) \geq e^{-c_1/(\sigma^*)^2}$$

for some constant $c_1 > 0$. Combining this with Lemma 10 of [15], there exist constants $c_2 > 0$ such that

$$\Pi(\mathcal{K}_n) \gtrsim e^{-c_2 N \log(\sigma^*)^{-1} - c_1/(\sigma^*)^2}.$$

Since $\beta > 4/3$ and $N \lesssim (\sigma^*)^{-4/\beta-1}(\log(\sigma^*)^{-1})^{3/2}$, there exists small $\zeta > 0$ depending only on β such that

$$\Pi(\mathcal{K}_n) \gtrsim \exp\left[-c_3(\sigma^*)^{-(4-\zeta)}\right] = \exp\left[-c_3\delta_n^{-(4-\zeta)(1/2+\epsilon/4)}\right],$$

where $c_3 > 0$ is a constant. If $\epsilon < \zeta/4$ then $(4-\zeta)(1/2+\epsilon/4) < 2-\zeta/4$. It follows that $\Pi(\mathcal{K}_n) \gtrsim \exp[-c_3\delta_n^{-(2-\zeta/4)}]$. Therefore, we can choose (δ_n) such that $\delta_n \ll n^{-1/4}$ and $\Pi(\mathcal{K}_n) \geq e^{-n\delta_n^2}$. □

References

[1] Amewou-Atisso, M., Ghosal, S., Ghosh, J. K. and Ramamoorthi, R. (2003). Posterior consistency for semi-parametric regression problems. *Bernoulli* **9** 291–312. MR1997031

[2] Andrews, D. F. and Mallows, C. L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **36** 99–102. MR0359122

[3] Barron, A., Schervish, M. J. and Wasserman, L. (1999). The consistency of posterior distributions in nonparametric problems. *The Annals of Statistics* **27** 536–561. MR1714718

[4] Canale, A. and De Blasi, P. (2017). Posterior consistency of nonparametric location-scale mixtures for multivariate density estimation. *Bernoulli* **23** 379–404. MR3556776

[5] Choi, T. and Schervish, M. J. (2007). On posterior consistency in nonparametric regression problems. *Journal of Multivariate Analysis* **98** 1969–1987. MR2396949

[6] De Blasi, P. and Walker, S. G. (2013). Bayesian asymptotics with misspecified models. *Statistica Sinica* **23** 169–187. MR3076163

[7] Diaconis, P. and Freedman, D. (1986a). On the consistency of Bayes estimates. *The Annals of Statistics* **14** 1–26. MR0829555

[8] Diaconis, P. and Freedman, D. (1986b). On inconsistent Bayes estimates of location. *The Annals of Statistics* **14** 68–87. MR0829556

[9] Doob, J. L. (1949). Application of the theory of martingales. In *Le Calcul des Probabilités et ses Applications* 23–27. Colloques Internationaux du Centre National de la Recherche Scientifique Paris. MR0033460

[10] Gerogiannis, D., Nikou, C. and Likas, A. (2009). The mixtures of Student's t-distributions as a robust framework for rigid registration. *Image and Vision Computing* **27** 1285–1294.

[11] Ghosal, S., Ghosh, J. K. and Ramamoorthi, R. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *The Annals of Statistics* **27** 143–158. MR1701105

[12] Ghosal, S., Ghosh, J. K. and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *The Annals of Statistics* **28** 500–531. MR1790007

[13] Ghosal, S. and Roy, A. (2006). Posterior consistency of Gaussian process prior for nonparametric binary regression. *The Annals of Statistics* **34** 2413–2429. MR2291505

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

[15] Ghosal, S. and van der Vaart, A. W. (2007). Posterior convergence rates of Dirichlet mixtures at smooth densities. *The Annals of Statistics* **35** 697–723. MR2336864

[16] Ghosal, S., Van Der Vaart, A. et al. (2007). Convergence rates of posterior distributions for noniid observations. *The Annals of Statistics* **35** 192–223. MR2332274

[17] Gibbs, A. L. and Su, F. E. (2002). On choosing and bounding probability metrics. *International Statistical Review* **70** 419–435.

[18] Hejblum, B. P., Alkhassim, C., Gottardo, R., Caron, F. and Thiébaut, R. (2017). Sequential Dirichlet Process Mixtures of Multivariate Skew t-distributions for Model-based Clustering of Flow Cytometry Data. *ArXiv:1702.04407*.

[19] Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association* **58** 13–30. MR0144363

[20] Huber, P. J. (1981). *Robust statistics*. John Wiley & Sons. MR0606374

[21] Kleijn, B. J. and van der Vaart, A. W. (2006). Misspecification in infinite-dimensional Bayesian statistics. *The Annals of Statistics* **34** 837–877. MR2283395

[22] Kruijer, W., Rousseau, J. and Van Der Vaart, A. (2010). Adaptive Bayesian density estimation with location-scale mixtures. *Electronic Journal of Statistics* **4** 1225–1257. MR2735885

[23] Lijoi, A., Prünster, I. and Walker, S. G. (2005). On consistency of nonparametric normal mixtures for Bayesian density estimation. *Journal of the American Statistical Association* **100** 1292–1296. MR2236442

[24] Peel, D. and McLachlan, G. J. (2000). Robust mixture modelling using the t distribution. *Statistics and Computing* **10** 339–348.

[25] Schwartz, L. (1965). On Bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **4** 10–26.

[26] Shen, W., Tokdar, S. T. and Ghosal, S. (2013). Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *Biometrika* **100** 623–640. MR3094441

[27] Tokdar, S. T. (2006). Posterior consistency of Dirichlet location-scale mixture of normals in density estimation and regression. *Sankhyā: The Indian Journal of Statistics* **68** 90–110. MR2301566

[28] Walker, S. (2003). On sufficient conditions for Bayesian consistency. *Biometrika* **90** 482–488. MR1986664

[29] Walker, S. (2004). New approaches to Bayesian consistency. *The Annals of Statistics* **32** 2028–2043. MR2102501

[30] Walker, S. G., Lijoi, A. and Prünster, I. (2005). Data tracking and the understanding of Bayesian consistency. *Biometrika* **92** 765–778. MR2234184

[31] Walker, S. G., Lijoi, A. and Prünster, I. (2007). On rates of convergence for posterior distributions in infinite-dimensional models. *The Annals of Statistics* **35** 738–746. MR2336866

[32] Wong, W. H. and Shen, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *The Annals of Statistics* **23** 339–362. MR1332570