IRIS AperTO

UNIVERSITÀ DEGLI STUDI DI TORINO

AperTO - Archivio Istituzionale Open Access dell'Università di Torino

Posterior asymptotics of nonparametric location-scale mixtures for multivariate density estimation

This is the author's manuscript

Original Citation:

Availability:

This version is available http://hdl.handle.net/2318/139500 since 2017-05-15T12:38:32Z

Published version:

DOI:doi:10.3150/15-BEJ746

Terms of use:

Open Access

Anyone can freely access the full text of works made available as "Open Access". Works made available under a Creative Commons license can be used according to the terms and conditions of said license. Use of all other works requires consent of the right holder (author or publisher) if not exempted from copyright protection by the applicable law.

(Article begins on next page)

24 April 2024

Posterior asymptotics of nonparametric location-scale mixtures for multivariate density estimation

Antonio Canale*, Pierpaolo De Blasi †

Accepted version, June 29, 2015

Abstract

Density estimation represents one of the most successful applications of Bayesian nonparametrics. In particular, Dirichlet process mixtures of normals are the gold standard for density estimation and their asymptotic properties have been studied extensively, especially in the univariate case. However a gap between practitioners and the current theoretical literature is present. So far, posterior asymptotic results in the multivariate case are available only for location mixtures of Gaussian kernels with independent prior on the common covariance matrix, while in practice as well as from a conceptual point of view a location-scale mixture is often preferable. In this paper we address posterior consistency for such general mixture models by adapting a convergence rate result which combines the usual low-entropy, high-mass sieve approach with a suitable summability condition. Specifically, we establish consistency for Dirichlet process mixtures of Gaussian kernels with various prior specifications on the covariance matrix. Posterior convergence rates are also discussed.

Keywords: Bayesian Nonparametrics; density estimation; Dirichlet mixture; factor model; posterior asymptotics; sparse random eigenmatrices

1 Introduction

Multivariate density estimation is a fundamental problem in nonparametric inference being also the starting point for nonparametric regression, clustering, and robust estimation. For modeling continuous densities, standard nonparametric Bayes methods rely on Dirichlet process (DP) (Ferguson, 1973) mixture models of the form

$$f(x) = \int K(x;\theta)dP(\theta), \quad P \sim DP(\alpha\, P^*), \tag{1}$$

where $K(x;\theta)$ is a probability kernel depending on some finite-dimensional parameter θ and $DP(\alpha P^*)$ is a Dirichlet process with total mass $\alpha > 0$ and P^* a probability measure over the space of parameters θ. Model (1) has been introduced by Lo (1984) and made popular by Escobar and West (1995); Müller et al. (1996). For densities on $\mathbb{R}^d$, a common choice for the kernel $K(x;\theta)$ is the normal density $\phi_\Sigma(x-\mu)$, that is

$$\phi_\Sigma(x-\mu) = (2\pi)^{-d/2}\det(\Sigma)^{-1/2}\exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\},$$

*antonio.canale@unito.it
†pierpaolo.deblasi@unito.it

for $\mu \in \mathbb{R}^d$ a d-dimensional vector of locations and Σ a positive definite symmetric matrix of variance-covariances. Depending on whether or not the mixture involves the scale parameter Σ, we will speak of location and location-scale mixtures, respectively. In the former case, all components in the mixture share the same Σ which is typically modeled with an independent prior.

In the univariate case, the asymptotic properties of DP mixtures are well known for both location (Ghosal et al., 1999; Lijoi et al., 2005b; Ghosal and van der Vaart, 2007; Walker et al., 2007) and location-scale mixtures (Tokdar, 2006; Ghosal and van der Vaart, 2001). In the multivariate case, the only results, to our knowledge, are confined to the case of location mixtures. Posterior consistency is studied in Wu and Ghosal (2010) assuming a truncated inverse-Wishart prior for Σ. Shen et al. (2013) improve considerably on these results by deriving adaptive posterior convergence rates and remove the artificial truncation of the Wishart prior, which prevents an effective implementation. One of the main tools for achieving their result is the use of the stick-breaking representation of the DP to build a low-entropy, high-mass sieve on the space of mixed densities, a procedure whose extension to the multivariate case is a challenging task.

The lack of asymptotic results for mutivariate location-scale mixtures is somehow in contrast with their predominant use in applications (Müller et al., 1996; MacEachern and Müller, 1998; Gorur and Rasmussen, 2009; Chen et al., 2010). In this paper, we fill this gap in the current literature by establishing posterior consistency for DP location-scale mixtures of multivariate normals with ready to verify conditions on the prior parameter P^*. In particular, a distinctive condition with respect to the location mixtures case involves the existence of moments to a certain order of the ratio between the largest and smallest eigenvalues of Σ, a quantity known in random matrix literature as condition number. This moment condition is satisfied by the inverse-Wishart as well as by other prior specifications that enable scaling to higher dimension. The consistency result exploits the sieve construction suggested by Shen et al. (2013) by adapting a convergence rate theorem of Ghosal and van der Vaart (2007). Such an adaptation allows to relax the growth condition on the entropy of the sieve through a summability condition which suitably weighs entropy numbers with prior probabilities, an idea first appeared in Lijoi et al. (2005b) and Walker et al. (2007). We also discuss some of the technical issues related to the challenging task of deriving posterior convergence rates for heavy tailed densities together with some preliminary results.

The layout of the paper is as follows. In Section 2, sufficient conditions, on the true f_0 and on the prior, to obtain posterior consistency of DP location-scale mixtures are given. In Section 3 some particular prior specifications satisfying such conditions are discussed. Section 4 is about convergence rates and the paper ends with a final discussion.

2 Posterior consistency

For any $d \times d$ matrix A with real eigenvalues, let $\lambda_1(A) \geq \ldots \geq \lambda_d(A)$ denote its eigenvalues in decreasing order and $\|A\|_2 = \max_{x \neq 0} \|Ax\|/\|x\|$ be its spectral norm. We denote by $\mathcal{S}$ be the space of $d \times d$ positive definite matrices and by $\mathscr{P}$ the space of probability measures on $\mathbb{R}^d \times \mathcal{S}$. We consider DP location-scale mixtures of the type

$$f_P(x) = \int \phi_\Sigma(x - \mu) \mathrm{d}P(\mu, \Sigma), \quad P \sim DP(\alpha P^*). \tag{2}$$

where $P^* \in \mathscr{P}$ with μ and Σ independent under P^*. In marginalizing out P one can write model (2) as

$$f_P(x) = \sum_{h=1}^{\infty} \pi_h \phi_{\Sigma_h}(x - \mu_h), \quad (\mu_h, \Sigma_h) \sim \text{iid } P^*, \quad \pi_h = V_h \prod_{k<h} (1 - V_k), \tag{3}$$

where $V_h \sim$ iid beta$(1, \alpha)$. We denote by Π^* the DP prior on $\mathscr{P}$ and by Π the prior induced by (2) on the space $\mathcal{F}$ of density functions on $\mathbb{R}^d$.

As customary in Bayesian asymptotics, we take the data $X_1 \ldots, X_n$ to be i.i.d. from some "true" density $f_0 \in \mathcal{F}$ and study the behavior of the posterior as $n \to \infty$ with respect to the n-products measure F_0^n, F_0 being the probability measure associated to f_0. As metrics on $\mathcal{F}$ we consider the Hellinger $d(f,g) = \{\int(\sqrt{f} - \sqrt{g})^2\}^{1/2}$, and the L_1 $\|f-g\|_1 = \int |f-g|$, which induce equivalent topologies in view of $d^2(f,g) \leq \|f-g\|_1 \leq 2d(f,g)$. By posterior consistency at f_0 we mean that, for any $\epsilon > 0$,

$$\Pi(\{f : \ \rho(f_0, f) > \epsilon\} \mid X_1, \ldots, X_n) \to 0$$

in F_0^n-probability where ρ is either the Hellinger or the L_1-metric.

It is known that, for the posterior distribution to accumulate around f_0, one has to establish first some support condition of the prior. We say that Π satisfies the Kullback-Leibler (KL) property at f_0 if

$$\Pi\left\{f : \ \int \log(f_0/f) f_0 \leq \eta\right\} \geq 0, \quad \text{for any } \eta > 0. \tag{4}$$

The KL property has been established in Theorem 5 of Wu and Ghosal (2008) for location-scale mixtures of Gaussian kernels with scalar covariance matrices, i.e. $\Sigma = \sigma^2 I$ for $\sigma^2 > 0$ and I the $d \times d$ identity matrix. Minor adaptations are needed to extend this result to the case of non scalar Σ. Sufficient conditions for (4) involve a mild requirement on the weak support of the DP prior Π^* on $\mathscr{P}$ together with some regularity assumptions on f_0 like the existence of moments up to a certain order. As for the weak support of Π^*, we shall assume that the prior mean P^* of the DP is supported on all $\mathbb{R}^d \times \mathcal{S}$, where we take as distance the sum of the Euclidean norm on $\mathbb{R}^d$ and the spectral norm on $\mathcal{S}$. All priors considered in Section 3 satisfy this requirement. As for f_0, the same regularity conditions of Wu and Ghosal (2008, Theorem 5) apply; they are repeated, for readers' convenience, in Lemma 1 whose proof is reported in the Appendix.

Lemma 1. *Let $f_0 \in \mathcal{F}$ and Π denote the prior on $\mathcal{F}$ induced by (2). Assume that f_0 satisfies the following conditions: $0 < f_0(x) < M$ for some constant M and all $x \in \mathbb{R}^d$; $|\int f_0(x) \log f_0(x) \mathrm{d}x| < \infty$; for some $\delta > 0$, $\int f_0(x) \log \frac{f_0(x)}{\phi_\delta(x)} \mathrm{d}x < \infty$, where $\phi_\delta(x) = \inf_{||t-x||<\delta} f_0(t)$; for some $\eta > 0$, $\int ||x||^{2(1+\eta)} f_0(x) \mathrm{d}x < \infty$. Then Π satisfies (4).*

The KL property (4) plays a very important role in consistency since it provides a lower bound for the denominator of the posterior probability. However, posterior consistency in non-compact spaces, such as $\mathcal{F}$, requires an additional condition on the prior which involves the metric entropy of $\mathcal{F}$ (see below for a formal definition). A critical step is to introduce a compact subset $\mathcal{F}_n$, called sieve, which is indexed by the sample size n and eventually grows to fill the entire parameter space as $n \to \infty$. According to Theorem 2 of Ghosal et al. (1999), the metric entropy of $\mathcal{F}_n$ has to grow slower than linearly in n, while the prior probability assigned to $\mathcal{F}_n^c$, the complement of the sieve, needs to decrease exponentially fast in n. See Theorem 2.1 of Ghosal et al. (2000) for similar ideas applied to posterior convergence rates. The choice of the sieve is a delicate issue in multivariate density estimation, since the metric entropy tends to blow up with the dimension d. A novel sieve construction has been introduced in Shen et al. (2013) and it has proven successful in deriving adaptive posterior convergence rates in the case of location mixtures with independent inverse-Wishart prior on Σ. In particular, the sieve relies on the stick-breaking representation of the DP as in (3). In adapting this sieve construction to location-scale mixtures (see Lemma 2 in the Appendix), the tail behavior of P^* with respect to the condition number, i.e. the ratio of the maximum and minimum eigenvalue of Σ, plays a crucial role. A straight application

of Theorem 2 of Ghosal et al. (1999) would require a too restrictive condition on this tail behavior, ruling out common choices for the part of P^* involving Σ like the inverse-Wishart distribution. See Remark 2 in the Appendix for a detailed explanation. For this reason we resort to a different posterior consistency theorem which consists in a modification of Theorem 5 of Ghosal and van der Vaart (2007). The main idea is to relax the growth condition on the entropy of $\mathcal{F}_n$ through a summability condition of entropy numbers weighted by square roots of prior probabilities. This modification can be applied also to relax the usual exponential tail behavior of the marginal of P^* on the location parameters μ, with a weaker power tail decay. As pointed out by Ghosal and van der Vaart (2007), there is a trade off between entropy and summability which is worth exploring in Bayesian asymptotics. Our result is a step in this direction. Similar ideas had earlier appeared in Lijoi et al. (2005b) and in Walker et al. (2007).

To state the posterior convergence result, we recall the definition of entropy of $\mathcal{G} \subset \mathcal{F}$ as $\log N(\epsilon, \mathcal{G}, d)$ where $N(\epsilon, \mathcal{G}, d)$ is the minimum integer N for which there exists $f_1, \ldots, f_N \in \mathcal{F}$ such that $\mathcal{G} \subset \bigcup_{j=1}^{N} \{f : d(f, f_j) < \epsilon\}$.

Theorem 1. *Suppose $\mathcal{F}_n \subset \mathcal{F}$ can be partitioned as $\bigcup_j \mathcal{F}_{n,j}$ such that, for $\epsilon > 0$,*

$$\Pi(\mathcal{F}_n^c) \lesssim \mathrm{e}^{-bn}, \quad \text{for some } b > 0 \tag{5}$$

$$\textstyle\sum_j \sqrt{N(2\epsilon, \mathcal{F}_{n,j}, d)}\sqrt{\Pi(\mathcal{F}_{n,j})}\mathrm{e}^{-(4-c)n\epsilon^2} \to 0, \quad \text{for some } c > 0 \tag{6}$$

Then $\Pi(f : \ d(f_0, f) > 8\epsilon | X_1, \ldots, X_n) \to 0$ in F_0^n-probability for any f_0 satisfying (4).

The proof is similar to that of Theorem 5 of Ghosal and van der Vaart (2007) and is presented in the Appendix. Here we state and prove the main result on posterior consistency for DP location-scale mixtures of Gaussian kernels.

Theorem 2. *Let f_0 satisfy the conditions of Lemma 1. Consider the prior Π defined in (2) with P^* that satisfies the following tail behaviors: for some positive constants c_1, c_2, c_3, $r > (d-1)/2$ and $\kappa > d(d-1)$,*

$$P^*(\|\mu\| > x) \lesssim x^{-2(r+1)}, \tag{7}$$

$$P^*(\lambda_1(\Sigma^{-1}) > x) \lesssim \exp(-c_1 x^{c_2}), \tag{8}$$

$$P^*(\lambda_d(\Sigma^{-1}) < 1/x) \lesssim x^{-c_3}, \tag{9}$$

$$P^*(\lambda_1(\Sigma^{-1})/\lambda_d(\Sigma^{-1}) > x) \lesssim x^{-\kappa} \tag{10}$$

for all sufficiently large $x > 0$. Then the posterior is consistent at f_0.

Condition (7) is weaker than the usual exponential tail condition in Tokdar (2006) (univariate case) and in Shen et al. (2013) (location mixture case). The distinctive condition for multivariate location-scale mixtures turns out to be (10) as explained in Remark 2. It corresponds to the existence of the $d(d-1)+1$ moment of the quantity $\lambda_1(\Sigma^{-1})/\lambda_d(\Sigma^{-1})$ which is known in random matrix theory as condition number of Σ^{-1}. See Edelman and Sutton (2005) for a recent contribution on the topic. In particular, the condition number plays an important role in electronical engineering problems in the context of multiple-input multiple-output communication systems. See Matthaiou et al. (2010) for related results.

Proof of Theorem 2. The proof is an application of Theorem 1 and is based on the entropy upper bounds of Lemma 2 in the Appendix. Let $M_n = \underline{\sigma}_n^{-2c_2} = n$ and $H_n = \lfloor Cn\epsilon^2 / \log n \rfloor$ for a positive constant C to

be determined later. Also, let $\boldsymbol{j} = (j_1, \ldots, j_{H_n})$, $j_h \in \mathbb{N}^*$, and $\boldsymbol{l} = (l_1, \ldots, l_{H_n})$, $l_h \in \mathbb{N}$. Define

$$\mathcal{F}_n = \{f_P \text{ with } P = \textstyle\sum_{h\geq 1}\pi_h\delta_{(\mu_h,\Sigma_h)} : \ \sum_{h>H_n}\pi_h \leq \epsilon; \\ \underline{\sigma}_n^2 \leq \lambda_d(\Sigma_h), \lambda_1(\Sigma_h) \leq \underline{\sigma}_n^2\left(1+\epsilon/\sqrt{d}\right)^{M_n}, \text{for } h \leq H_n\}$$

$$\mathcal{F}_{n,\boldsymbol{j},\boldsymbol{l}} = \left\{f_P \in \mathcal{F}_n : \ \sqrt{n}(j_h-1) < \|\mu_h\| \leq \sqrt{n}j_h, \ n^{2^{l_h-1}\mathbb{1}_{(l_h\geq 1)}} < \frac{\lambda_1(\Sigma_h)}{\lambda_d(\Sigma_h)} \leq n^{2^{l_h}}, \text{ for } h \leq H_n\right\}$$

so that $\mathcal{F}_n \uparrow \mathcal{F}$ as $n \to \infty$ and $\mathcal{F}_n \subset \bigcup_{\boldsymbol{j},\boldsymbol{l}} \mathcal{F}_{n,\boldsymbol{j},\boldsymbol{l}}$.

As for (5) of Theorem 1, note that

$$\Pi(\mathcal{F}_n^c) \leq \Pr\left\{\textstyle\sum_{h>H_n}\pi_h > \epsilon\right\} + H_n\left[P^*(\lambda_d(\Sigma) < \underline{\sigma}_n^2) + P^*(\lambda_1(\Sigma) > \underline{\sigma}_n^2(1+\epsilon/\sqrt{d})^{M_n})\right].$$

Use the stick-breaking representation of the DP for the first term (see the proof of Proposition 2 of Shen et al. (2013)), (8) and (9) to get

$$\Pi(\mathcal{F}_n^c) \lesssim \left\{\frac{\mathrm{e}\alpha}{H_n}\log\frac{1}{\epsilon}\right\}^{H_n} + H_n\left[\mathrm{e}^{-c_1\underline{\sigma}_n^{-2c_2}} + \underline{\sigma}_n^{-2c_3}\left(1+\frac{\epsilon}{\sqrt{d}}\right)^{-c_3M_n}\right], \tag{11}$$

so that

$$\begin{aligned}\Pi(\mathcal{F}_n^c) &\lesssim (Cn\epsilon^2/\log n)^{-Cn\epsilon^2/\log n} + (Cn\epsilon^2/\log n)\left[\mathrm{e}^{-c_1 n} + n^{c_3/c_2}(1+\epsilon/\sqrt{d})^{-c_3 n}\right] \\ &\lesssim \exp\{-(Cn\epsilon^2/\log n)\log(Cn\epsilon^2/\log n)\} + \mathrm{e}^{-c_1 n} + \exp\{-c_3 n\log(1+\epsilon/\sqrt{d})\}.\end{aligned}$$

Note that $(Cn\epsilon^2/\log n)\log(Cn\epsilon^2/\log n) > Cn\epsilon^2/2$ for large enough n, therefore

$$\Pi(\mathcal{F}_n^c) \lesssim \mathrm{e}^{-Cn\epsilon^2/2} + \mathrm{e}^{-c_1 n} + \mathrm{e}^{-c_3\log(1+\epsilon/\sqrt{d})n} \lesssim \mathrm{e}^{-bn}$$

for $0 < b < \min\{C\epsilon^2/2, c_1, c_3\log(1+\epsilon/\sqrt{d})\}$. Hence (5) is satisfied.

We next show that $\mathcal{F}_{n,\boldsymbol{j},\boldsymbol{l}}$ satisfies the summability condition (6) of Theorem 1. By an application of Lemma (2) in the Appendix,

$$\begin{aligned}N(\epsilon, \mathcal{F}_{n,\boldsymbol{j},\boldsymbol{l}}, \|\cdot\|_1) \lesssim \exp\Bigg\{&dH_n\log M_n + H_n\log\tfrac{C_1}{\epsilon} \\ &+ \textstyle\sum_{h\leq H_n}\log\left[\left(\frac{\sqrt{n}j_h}{\underline{\sigma}_n\epsilon/2}+1\right)^d - \left(\frac{\sqrt{n}(j_h-1)}{\underline{\sigma}_n\epsilon/2}-1\right)^d\right] + \tfrac{d(d-1)}{2}\log\tfrac{2d\,n^{2^{l_h}}}{\epsilon^2}\Bigg\}\end{aligned}$$

for some positive C_1. Let $c_4 = 1/2 + 1/(2c_2)$ and note that

$$\begin{aligned}\left(\frac{\sqrt{n}j_h}{\underline{\sigma}_n\epsilon/2}+1\right)^d - \left(\frac{\sqrt{n}(j_h-1)}{\underline{\sigma}_n\epsilon/2}-1\right)^d &= \left(\frac{2n^{c_4}j_h}{\epsilon}+1\right)^d - \left(\frac{2n^{c_4}j_h}{\epsilon}+1-\frac{2n^{c_4}}{\epsilon}-2\right)^d \\ &\lesssim d\left(\frac{2n^{c_4}}{\epsilon}+2\right)\left(\frac{2n^{c_4}j_h}{\epsilon}+1\right)^{d-1} \lesssim \frac{n^{c_4 d}j_h^{d-1}}{\epsilon^d}\end{aligned}$$

where inequality sign "$\lesssim$" is for both large n and j_h. Hence we have

$$\begin{aligned}N(\epsilon, \mathcal{F}_{n,\boldsymbol{j},\boldsymbol{l}}, \|\cdot\|_1) \lesssim \exp\Bigg\{&dH_n\,\log n + H_n\log\tfrac{C_1}{\epsilon} \\ &+ \textstyle\sum_{h\leq H_n}\log\frac{n^{c_4 d}j_h^{d-1}}{\epsilon^d} + \tfrac{d(d-1)}{2}\log\tfrac{2d\,n^{2^{l_h}}}{\epsilon^2}\Bigg\}\end{aligned} \tag{12}$$

Moreover, by using tail conditions (7) and (10),

$$\begin{aligned}\Pi(\mathcal{F}_{n,\boldsymbol{j},\boldsymbol{l}}) &\leq \textstyle\prod_{h\leq H_n} P^*(\|\mu\| > \sqrt{n}(j_h-1), \lambda_1(\Sigma)/\lambda_d(\Sigma) > n^{2^{l_h-1}\mathbb{1}_{(l_h\geq 1)}}) \\ &\lesssim \textstyle\prod_{h\leq H_n} [\sqrt{n}(j_h-1)]^{-\mathbb{1}_{(j_h\geq 2)}2(r+1)} n^{-\mathbb{1}_{(l_h\geq 1)}2^{l_h-1}\kappa} \end{aligned} \tag{13}$$

with the convention $0^0 = 1$. A combination of (12), (13) and $d^2(f,g) \leq \|f-g\|_1$ imply that $\sqrt{N(2\epsilon, \mathcal{F}_{n,\boldsymbol{j},\boldsymbol{l}}, d)}\sqrt{\Pi(\mathcal{F}_{n,\boldsymbol{j},\boldsymbol{l}})}$ is bounded by a multiple of

$$\exp\left\{\tfrac{d+c_4 d}{2} C n\epsilon^2\right\} \textstyle\prod_{h\leq H_n} j_h^{\frac{d-1}{2}} [\sqrt{n}(j_h-1)]^{-\mathbb{1}_{(j_h\geq 2)}(r+1)}\ n^{\frac{d(d-1)}{4}2^{l_h}} n^{-\mathbb{1}_{(l_h\geq 1)}2^{l_h-1}\frac{\kappa}{2}} \tag{14}$$

By hypothesis, $r > (d-1)/2$ so that $K := \sum_{j\geq 2} j^{(d-1)/2}(j-1)^{-(r+1)} < \infty$. Hence by summing (14) with respect to $\boldsymbol{j}$ we get

$$\exp\left\{\tfrac{d+c_4 d}{2} C n\epsilon^2\right\} (1+n^{-\frac{r+1}{2}}K)^{H_n} \textstyle\prod_{h\leq H_n} n^{\frac{d(d-1)}{4}2^{l_h}} n^{-\mathbb{1}_{(l_h\geq 1)}2^{l_h-1}\frac{\kappa}{2}} \tag{15}$$

Moreover, note that

$$\begin{aligned}\sum_{l\geq 0} n^{\frac{d(d-1)}{4}2^l} n^{-\mathbb{1}_{(l\geq 1)}2^{l-1}\frac{\kappa}{2}} &= n^{\frac{d(d-1)}{4}} + \sum_{l\geq 1} \exp\left\{-\frac{d(d-1)}{2}\log n\left(\frac{\kappa}{d(d-1)}-1\right)2^{l-1}\right\} \\ &\leq 2n^{\frac{d(d-1)}{4}}\end{aligned}$$

for n large enough since, by hypothesis, $\kappa/[d(d-1)] - 1 > 0$. Hence, by summing (15) with respect to $\boldsymbol{l}$ we get the upper bound

$$\begin{aligned}&\exp\left\{\frac{d+c_4 d}{2} C n\epsilon^2\right\} (1+n^{-\frac{r+1}{2}}K)^{H_n} \left[2n^{\frac{d(d-1)}{4}}\right]^{H_n} \\ &\leq \exp\left\{\frac{d+c_4 d}{2} C n\epsilon^2\right\} 4^{H_n} \exp\left\{C n\epsilon^2 \frac{d(d-1)}{4}\right\} \lesssim \exp\left\{\frac{1}{2}\left[d + c_4 d + \frac{d(d-1)}{2}\right] C\epsilon^2 n\right\}\end{aligned}$$

where in the first inequality we have used $n^{-\frac{r+1}{2}}K \leq 1$ for n large enough. By taking C sufficiently small to satisfy $C < 2(4-c)/[d + c_4 d + d(d-1)/2]$ for some $c > 0$, (6) is satisfied and the proof is complete. □

3 Illustration

Theorem 2 holds for DP location-scale mixtures of multivariate Gaussian kernels with minimal requirements on the prior parameter P^*. The power tail decay (7) for μ is indeed important as it covers the prior specification of Müller et al. (1996):

$$P^*(\mu, \Sigma) = N(\mu; m, B)\, IW(\Sigma; \Sigma_0, \nu),$$

where an additional hyperprior on B is given by $IW(B; B_0, \nu_B)$, with $IW(B_0, \nu_B)$ denoting the inverse-Wishart distribution with scale parameter B_0 and ν_B degrees of freedom. In this case μ under P^* has multivariate Student's t-distribution with ν_B degree of freedom so that $P^*(\|\mu\|^2 > x) \sim x^{-(\nu_B - d + 1)/2}$ for x large enough. Hence $\nu_B > 2d$ is sufficient for (7) to hold with $r > (d-1)/2$.

In what follows, we focus on the prior specification of the scale parameter Σ and show that conditions (8)-(10) are verified for some important choices which make a substantial practical difference in applications, particularly in high-dimensional settings. Henceforth, we refer to the prior specification for Σ as $\Sigma \sim \mathcal{L}$ meaning that $\mathcal{L}$ is the marginal of Σ with respect to the prior mean P^*. All proofs are reported in the Appendix.

The first result is about $\mathcal{L}$ being the inverse-Wishart distribution $IW(\Sigma_0, \nu)$. While conditions (8) and (9) are always satisfied, see e.g. the proof of Lemma 1 of Shen et al. (2013), ν needs to be sufficiently large for tail condition (10) on the condition number to hold. Corollary 1 makes this statement precise.

Corollary 1. *Assume f_0 satisfies the conditions of Lemma 1. Consider a prior Π induced by (2) with P^* satisfying (7) and $\mathcal{L} = IW(\Sigma_0, \nu)$ with $\nu > 2d(d-1) + d - 1$. Then conditions (8)-(10) are satisfied and the posterior is consistent at f_0.*

There is a rich literature providing alternatives to the inverse-Wishart prior when the dimension of the data is large. For example a commonly used and successful approach consists on analytic factorizations (West, 2003; Carvalho et al., 2008) where

$$\Sigma = \Gamma\Gamma^T + \Omega, \ \Gamma \sim \mathcal{L}_\Gamma, \ \Omega \sim \mathcal{L}_\Omega, \tag{16}$$

where Γ is a $d \times r$ matrix with $r < d$, independent from the $d \times d$ diagonal matrix Ω. Let γ_{jh} be the (j,h)-th element of Γ (factor loading) and σ_j^2 be the j-th diagonal element of Ω (residual variance). The specification of $\mathcal{L}_\Gamma$ and $\mathcal{L}_\Omega$ corresponds then to a distribution for γ_{jh} and σ_j^2. The next corollary addresses the case of normal factor loadings with inverse gamma distributed residual variances. It turns out that conditions (8)-(9) are automatically satisfied, while a constraint on the shape parameter of the inverse gamma prior is needed for the verification of condition (10).

Corollary 2. *Assume f_0 satisfies the conditions of Lemma 1. Consider a prior Π induced by (2) with P^* satisfying (7) and $\mathcal{L}$ induced by (16). Assume that $\gamma_{ij} \sim$ iid $N(0,1)$ and $\sigma_j^{-2} \sim$ iid $Ga(a,b)$ with $a > d(d-1)$. Then conditions (8)-(10) are satisfied and the posterior is consistent at f_0.*

Remark 1. *Motivated by the need of a method that scales for increasing dimension, Bhattacharya and Dunson (2011) introduce a sparse Bayesian factor model for the estimation of high-dimensional covariance matrices according to (16). The model consists of a multiplicative gamma process shrinkage prior on the factor loadings, i.e.*

$$\gamma_{jh}|\phi_{jh}\tau_h \sim N(0, \phi_{jh}^{-1}\tau_h^{-1}), \ \phi_{jh} \sim Ga(3/2, 3/2), \ \tau_h = \prod_{l=1}^{h} \delta_l,$$
$$\delta_1 \sim Ga(a_1, 1), \ \delta_l \sim Ga(a_2, 1), l > 1, \ \sigma_j^{-2} \sim Ga(a,b), j = 1, \ldots, d.$$

Using similar arguments to Corollary 2 it can be proved that conditions (8)–(10) are satisfied for $a > d(d-1)$.

One can easily build $\mathcal{L}$ which satisfy (8)–(10) by modeling directly the distribution of the eigenvalues. The idea is to use the spectral decomposition of Σ:

$$\Sigma = O\Lambda O^T, \ O \sim \mathcal{L}_O, \ \Lambda \sim \mathcal{L}_\Lambda, \tag{17}$$

where O is a $d \times d$ orthogonal matrix independent from $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_d)$ with $\lambda_i > 0$. It is clear that the verification of (8)–(10) involves only the distribution $\mathcal{L}_\Lambda$. Motivated by high dimensional sparse

random matrices modeling, Cron and West (2015) propose a similar approach. The authors set $O_{i,j}$ to be the rotation matrix for the rotator angle $\omega_{i,j}$ and $O = \prod_{i<j} O_{i,j}(\omega_{i,j})$ and define prior on O throught a prior on the rotator angles ω which naturally accommodates sparsity, namely

$$p(\omega) = \beta_{\pi/2}\mathbb{I}(\omega = \pi/2) + (1-\beta_{\pi/2})\beta_0\mathbb{I}(\omega = 0) + (1-\beta_{\pi/2})(1-\beta_0)p_c(\omega),$$
$$p_c(\omega) = c(\kappa)\exp\{\kappa\cos^2\omega\}\mathbb{I}(|\omega| < \pi/2).$$

This class of models is particularly useful to induce sparsity without the assumption of a reduced dimensional latent factor and hence can be appealing in many practical situations. The next corollary discusses sufficient conditions on $\mathcal{L}_\Lambda$ to obtain posterior consistency under formulation (17).

Corollary 3. *Assume f_0 satisfies the conditions of Lemma 1. Consider a prior Π induced by (2) with P^* satisfying (7) and $\mathcal{L}$ following (17) with $\lambda_i^{-1} \sim Ga(a,b)$, $a > d(d-1)$. Then conditions (8)-(10) are satisfied and the posterior is consistent at f_0.*

4 Posterior convergence rates

In this section we discuss some relevant issues related to the derivation of posterior convergence rates for location-scale mixtures. Under the prior conditions (7)–(10), the sieve construction laid down in the proof of Theorem 2 would adapt to any rate $\epsilon_n = n^{-\gamma}$ (for any $\gamma \in (0, 1/2)$ and up to a logarithmic term) determined by the prior concentration rate. This statement is made precise in the following proposition, whose proof is reported in the Appendix.

Proposition 1. *Let $\tilde{\epsilon}_n = n^{-\gamma}(\log n)^t$ for some $\gamma \in (0,1/2)$ and $t \geq 0$ and suppose that*

$$\Pi\left(f : \int f_0 \log(f_0/f) \leq \tilde{\epsilon}_n^2, \int f_0 \log^2(f_0/f) \leq \tilde{\epsilon}_n^2\right) \geq \mathrm{e}^{-n\tilde{\epsilon}_n^2} \tag{18}$$

Then, under the hypothesis of Theorem 2, $\Pi_n\{f : d(f_0, f) > (\log n)^s \tilde{\epsilon}_n\} \to 0$ in F_0^n-probability for any $s > 0$.

As for the prior contraction rate (18), in the case of location mixtures the derivation of adaptive $\tilde{\epsilon}_n$ for β-Hölder f_0 consists of three main steps: (i) construct a density h_σ (depending on f_0 and β) such that $d(f_0, \phi_\sigma \star h_\sigma) \lesssim \sigma^\beta$ as $\sigma \to 0$; (ii) modify h_σ to a compactly supported density $\tilde{h}_\sigma$ with the same approximation properties; (iii) approximate the continuous mixture $\phi_\sigma \star \tilde{h}_\sigma$ by a discrete mixture $\phi_\sigma \star F_\sigma$ with a convenient lower bound of the ratio $\phi_\sigma \star F_\sigma / f_0$. Here $\phi_\sigma \star h$ denotes the convolution of h and $\phi_{\sigma^2 I}$. See Shen et al. (2013) and Kruijer et al. (2010). Steps (ii) and (iii) rely on the assumption of exponential tail of f_0, which makes these techniques not suitable for location-scale mixtures, since the latter are more flexible than location mixtures in modeling densities with heavy tails. We come back to this point later. Before, we present a result of the type of convergence rates which are attainable by adapting these techniques to our setting. To this aim, refer to the definition of the locally β-Hölder class with envelope L, denoted $\mathcal{C}^{\beta,L,\tau_0}(\mathbb{R}^d)$, from Shen et al. (2013). Moreover, for a multi-index $k = (k_1, \ldots, k_d)$, $k_i \in \mathbb{N}^*$, define $k_\cdot = k_1 + \cdots + k_d$ and let D^k denote the mixed partial derivative operator $\partial^{k_\cdot}/\partial^{k_1}\cdots\partial^{k_d}$. Finally, we make the following two assumptions on the prior mean P^*. As for the scale parameter, we resort to an assumption analogous to (4) in Shen et al. (2013) about the mass in neighborhood of the eigenvalues of Σ: there exist $\kappa^*, a_4, a_5, b_4, C_3$ such that for any $s_1 \geq \cdots \geq s_d \geq 0$ and $t \in (0,1)$,

$$P^*\left(s_i < \lambda_i(\Sigma^{-1}) < s_i(1+t),\ i = 1, \ldots, d\right) \gtrsim b_4 s_d^{a_4} t^{a_5} \exp\{-C_3 s_1^{\kappa^*/2}\}. \tag{19}$$

As for the mean parameter, we assume that the marginal of P^* has density f^* and that f^* has tails no lighter than the Gaussian, i.e. for $b^*, \tau^* > 0$ and $\|x\|$ sufficiently large,

$$f^*(x) \gtrsim \mathrm{e}^{-b^* \|x\|^{\tau^*}} \tag{20}$$

The proof of the following proposition is deferred to the Appendix.

Proposition 2. *Let $f_0 \in \mathcal{C}^{\beta, L, \tau_0}(\mathbb{R}^d)$ be a bounded probability density function satisfying $F_0(|D^k f_0|/f_0)^{(2\beta+\epsilon)/k_\cdot} < \infty$, $k_\cdot \leq \lfloor \beta \rfloor$, $F_0(L/f_0)^{(2\beta+\epsilon)/\beta} < \infty$ for some $\epsilon > 0$ and*

$$f_0(x) \leq c \exp\{-b\|x\|^\tau\}, \tag{21}$$

for some $b, c > 0$, $\tau \geq 2$ and $\|x\|$ sufficiently large. Assume that conditions (19) and (20) hold for $\tau^* \leq \tau$. Then (18) holds for

$$\tilde{\epsilon}_n = n^{-\beta/(2\beta+d+\kappa^*)} (\log n)^t, \quad t \geq \frac{d(1 + (\kappa^* + 1)/\beta + 1/\tau)}{2 + (d + \kappa^*)/\beta)} \tag{22}$$

Proposition 2 yields, via Proposition 1, a posterior convergence rate which is suboptimal with respect to the minimax rate $n^{-\beta/(2\beta+d)}$ by a term which depends on the constant κ^* appearing in (19). For illustration, $\kappa^* = 2$ if, under P^*, Σ has Inverse Wishart distribution $\mathrm{IW}(\Psi, \nu)$ with ν degrees of freedom and a positive definite scalar matrix Ψ, see Lemma 1 in Shen et al. (2013). Another useful specification is to consider that, under P^*, each $\lambda_i(\Sigma^{-1})$ have been independently assigned the distribution of the square of an inverse gamma random variable. Then $\kappa^* = 1$, which leads to a better convergence rate, still suboptimal. The technical reason for which the minimax rate is not achieved is to be found in the prior probability of L_1-balls around the frequencies of the approximating mixture of the true density. Specifically, the approximating mixture has all covariance matrices equal to $\sigma_n^2 I$ for a scaling factor σ_n which goes to zero as a function of $\tilde{\epsilon}_n$. In location-scale mixtures P^* put mass proportional to $\exp\{-C_3 \sigma_n^{-\kappa^*}\}$ in neighborhoods of $\sigma_n^2 I$ and this accounts for an extra factor in the radius of the L_1-ball with respect to the location mixture case. This, in turns, determines worse probability estimates of Kullback-Leibler neighborhoods. See the proof of Proposition 2 for details. It can be proved that the minimax rate $n^{-\beta/(2\beta+d)}$ is recovered upon setting the marginal of P^* on Σ depending on n by the scaling factor equal to $n^{-1/(2\beta+d)} (\log n)^{-1/\beta}$, however this has clearly a limited relevance since the rate would be non adaptive and it would rule out the Inverse-Wishart distribution as well as any other commonly used prior specification for the covariance matrix. The suboptimal convergence rate might be related to the fact that location-scale mixtures are more robust to tails than location mixtures. This should emerge through posterior convergence rates which are slower still adaptive to the minimax rate of a suitably defined class of densities with heavy tails. To this aim, an approximation scheme different than the one used for location mixture is in order since the latter relies on the exponential tail condition (21). This remains an open problem and is left as an argument for future research.

We discuss next how minimax rates for density estimation depend on the tail. It is well known that smoothness alone is not sufficient in order to guarantee consistency of density estimators in the L_1-norm. In fact, there is a vast literature on minimax and adaptive minimax density estimation with L_p norm which indicates the existence of a tail zone, i.e. a range of values of p, for which the minimax rate depends on p and deteriorates to 1 as p decreases to 1. Such phenomenon does not appear in density estimation on a compact domain, see Goldenshluger and Lepski (2014) for an up to date literature review. Goldenshluger and Lepski (2014) have also determined a tail dominance condition which illustrates how

the tail zone shrinks to an empty set according to the tails of the density. The result which is relevant to our study is about the minmax rate under L_1-norm for β-Holder classes of densities which is, up to $\log n$ factors, given by

$$\max\left\{n^{-\frac{\beta}{2\beta+d}},\ n^{-\frac{1-\theta}{1+d/\beta}}\right\} \tag{23}$$

where θ is a positive parameter in $(0,1]$ which determines the heaviness of the tails. See Remark 4.3 in Goldenshluger and Lepski (2014). The fastest rate $n^{-\beta/(2\beta+d)}$ is recovered for $\theta < \beta/(2\beta+d)$ (light tail), while for $\theta \geq \beta/(2\beta+d)$ (heavy tail) a wide range of slower minimax rates is obtained. See Theorem 13 of Devroye and Gyorfi (1985) for a closely related result in the light tail case. Clearly the exponential tail assumption (21) corresponds to small value of θ and this explain why in Shen et al. (2013) the usual $n^{-\beta/(2\beta+d)}$ rate is achieved.

5 Discussion

In this paper we have discussed asymptotic properties of DP location-scale mixtures of Gaussian kernels for multivariate density estimation. To our knowledge, this is the first contribution to posterior asymptotics in the context of multivariate location-scale mixtures, a modelling approach which is a common practice in many applications. We have given sufficient conditions on the DP prior measure on the space of means and covariance matrices in order to achieve posterior consistency. While showing that these conditions are satisfied for widely used inverse-Wishart distribution, with the same practical motivation of providing theoretical justification for models used in practice, we showed that the conditions hold if one uses priors that parsimoniously model the covariance in high dimensional settings, such as a factor model or spectral decomposition having both computational tractability and better fit in finite samples. Future work will deal with mixtures of more general random probability measures, like the two-parameter Poisson-Dirichlet process (Perman et al., 1992), the normalized inverse Gaussian process (Lijoi et al., 2005a) and, in general, Gibbs-type priors (Gnedin and Pitman, 2005). A characteristic feature of Gibbs-type priors is given by their heavy-tailedness, which has certain advantages in terms of statistical inference. See De Blasi et al. (2015). From the perspective of frequentist asymptotics this implies that the sieve based on the truncated stick-breaking construction does not carry over in a straightforward way and hence this problem deserves further investigations.

Acknowledgements

Comments on an earlier version of this work from David B. Dunson are greatly appreciated. The authors also thank the editor, the associate editor and the reviewers for comments that have helped to improve the paper substantially, Alberto Zanella for providing useful references about the condition number distribution and Bertrand Lods, Bernardo Nipoti, and Igor Prünster for fruitful discussions. The present work is supported by Regione Piemonte and by the European Research Council (ERC) through StG "N-BNP" 306406.

Appendix

Proof of Lemma 1. It is well known that $P \in \mathscr{P}$ belongs to the weak support of Π^* (meaning that any weak neighborhood of P has positive probability under Π^*) if the support of P is contained in the

support of P^*. We recall that the support of P is the smallest closed set of P-measure 1. Hence, since the support of P^* is the whole space $\mathbb{R}^d \times \mathcal{S}$, the weak support of Π^* is the whole space $\mathscr{P}$ and it contains, in particular, any compactly supported P.

The proofs of Theorem 5 and Theorem 2 of Wu and Ghosal (2008) imply that, under the conditions of Lemma 1, for any $\epsilon > 0$, there exists $P_\epsilon \in \mathscr{P}$ such that $\int f_0 \log(f_0/f_{P_\epsilon}) \leq \epsilon$. In particular, P_ϵ can be taken as compactly supported, that is $P_\epsilon(D) = 1$ where $D = [-a, a]^p \times \{\Sigma \in \mathcal{S} : \underline{\sigma}^2 \leq \lambda_i(\Sigma) \leq \bar{\sigma}^2, i = 1, \ldots, d\})$ for some constants $a > 0$, $0 < \underline{\sigma} < \bar{\sigma}$. Since

$$\int f_0(x) \log \frac{f_0(x)}{f(x)} \mathrm{d}x = \int f_0(x) \log \frac{f_0(x)}{f_{P_\epsilon}(x)} \mathrm{d}x + \int f_0(x) \log \frac{f_{P_\epsilon}(x)}{f(x)} \mathrm{d}x$$

in order to prove (4), we next show that there exists $W \subset \mathscr{P}$ with $\Pi^*(W) > 0$ such that, for any $P \in W$, $\int f_0 \log(f_{P_\epsilon}/f_P) \leq \epsilon$. To this aim, we verify the hypotheses of Lemma 3 in Wu and Ghosal (2008). It is clear that P_ϵ belongs to the weak support of Π^*. Next, condition (A7) of Lemma 3 in Wu and Ghosal (2008) requires $\log f_{P_\epsilon}$ and $\log \inf_{(\mu,\Sigma)\in D} \phi_\Sigma(x - \mu)$ to be f_0-integrable. Note that, for $\|x\| < a$, $\inf_{(\mu,\Sigma)\in D} \phi_\Sigma(x - \mu)$ is bounded, while, for $\|x\| \geq a$,

$$\inf_{(\mu,\Sigma)\in D} \phi_\Sigma(x - \mu) = \bar{\sigma}^{-d} \exp\left\{-\frac{4\|x\|^2}{2\underline{\sigma}^2}\right\}.$$

Hence, under the hypotheses made, $\log \inf_{(\mu,\Sigma)\in D} \phi_\Sigma(x - \mu)$ is f_0-integrable. Also

$$|\log f_{P_\epsilon}| \leq \left|\log\left\{\bar{\sigma}^{-d} \exp\left\{-\frac{4\|x\|^2}{2\underline{\sigma}^2}\right\} P_\epsilon(D)\right\}\right|$$

for $\|x\| \geq a$, so that $\log f_{P_\epsilon}$ is f_0-integrable by a similar argument.

As for condition (A8) of Lemma 3 in Wu and Ghosal (2008), it is obviously satisfied since the multivariate normal kernel $\phi_\Sigma(x - \mu)$ is bounded away from zero for x in a compact set of $\mathbb{R}^d$ and $(\mu, \Sigma) \in D$. Finally, condition (A9) of Lemma 3 in Wu and Ghosal (2008) requires that, for $C \subset \mathbb{R}^d$ a given compact set, $\{\phi_\Sigma(x - \mu), x \in C\}$ is uniformly equicontinuous as a family of functions of (μ, Σ) on D. This can be shown by adapting to the present context the arguments given in the last part of the proof of Theorem 2 of Wu and Ghosal (2008). □

Proof of Theorem 1. According to Corollary 1 of Ghosal and van der Vaart (2007), for any set of probability measures $\mathcal{Q}$ with $\inf\{d(F_0, Q) : Q \in \mathcal{Q}\} \geq 4\epsilon$, any $\alpha, \beta > 0$ and all $n \geq 1$, there exists a test ϕ_n such that

$$F_0^n \phi_n \leq \sqrt{\frac{\beta}{\alpha}} N(\epsilon, \mathcal{Q}, d) \mathrm{e}^{-n\epsilon^2}, \quad \sup_{Q\in\mathcal{Q}} Q^n(1 - \phi_n) \leq \sqrt{\frac{\alpha}{\beta}} \mathrm{e}^{-n\epsilon^2}. \tag{24}$$

Let $A_\epsilon = \{f : d(f_0, f) \leq 8\epsilon\}$. Write

$$\Pi(A_\epsilon^c | X_1, \ldots X_n) = \Pi(A_\epsilon^c \cap \mathcal{F}_n | X_1, \ldots X_n) + \Pi(A_\epsilon^c \cap \mathcal{F}_n^c | X_1, \ldots X_n). \tag{25}$$

From Lemma 4 of Barron et al. (1999), (4) implies that, for every $\eta > 0$

$$F_0^\infty \left\{\int \prod_{i\leq n} (f(x_i)/f_0(x_i)) \, \mathrm{d}\Pi(f) \leq \exp(-n\eta), \ i.o.\right\} = 0, \tag{26}$$

so that, under (5), the second expression on the r.h.s. of (25) goes to 0 in F_0^n-probability (actually in F_0^∞-almost surely). So it is sufficient to prove that the first expression in the r.h.s. of (25) goes

to 0 in F_0^n-probability. Let $R_n(f) = \prod_{i\le n} f(X_i)/f_0(X_i)$ and $I_n = \int R_n(f)\mathrm{d}\Pi(f)$. The event B_n that $I_n \ge \mathrm{e}^{-cn\epsilon^2}$ for c in (6) satisfies $F_0^n(B_n) \to 1$, see (26). Therefore

$$F_0^n\left[\Pi(A_\epsilon^c \cap \mathcal{F}_n|X_1,\dots,X_n)\right] = F_0^n\left[\Pi(A_\epsilon^c \cap \mathcal{F}_n|X_1,\dots,X_n)\mathbb{1}_{B_n}\right] + o_p(1).$$

For arbitrary tests $\phi_{n,j}$, we have

$$\begin{aligned} F_0^n\left[\Pi(A_\epsilon^c \cap \mathcal{F}_{n,j}|X_1,\dots,X_n)\mathbb{1}_{B_n}\right] &\le F_0^n\phi_{n,j} + F_0^n\left((1-\phi_{n,j})\int_{A_\epsilon^c\cap\mathcal{F}_{n,j}} R_n(f)\mathrm{d}\Pi(f)\right)\mathrm{e}^{cn\epsilon^2} \\ &\le F_0^n\phi_{n,j} + \sup_{f\in\mathcal{F}_{n,j}:d(f_0,f)\ge 8\epsilon} F^n(1-\phi_{n,j})\Pi(\mathcal{F}_{n,j})\mathrm{e}^{cn\epsilon^2} \end{aligned}$$

which can be bounded by a multiple of

$$\sqrt{\frac{\beta_j}{\alpha_j}}N(2\epsilon,\mathcal{F}_{n,j},d)\mathrm{e}^{-4n\epsilon^2} + \sqrt{\frac{\alpha_j}{\beta_j}}\mathrm{e}^{-(4-c)n\epsilon^2}\Pi(\mathcal{F}_{n,j})$$

for the choice of $\phi_{n,j}$ from Corollary 1 with 2ϵ in place of ϵ, $\mathcal{Q} = \{F \in \mathcal{F}_{n,j} : d(f_0,f) \ge 8\epsilon\}$ and any $\alpha_j, \beta_j > 0$. Put $\alpha_j = N(2\epsilon,\mathcal{F}_{n,j},d), \beta_j = \Pi(\mathcal{F}_{n,j})$ to obtain

$$\begin{aligned} F_0^n\left[\Pi(A_\epsilon^c \cap \mathcal{F}_n|X_1,\dots,X_n)\mathbb{1}_{B_n}\right] \le{}& \textstyle\sum_j \sqrt{N(2\epsilon,\mathcal{F}_{n,j},d)}\sqrt{\Pi(\mathcal{F}_{n,j})}\mathrm{e}^{-4n\epsilon^2} \\ &+ \textstyle\sum_j \sqrt{N(2\epsilon,\mathcal{F}_{n,j},d)}\sqrt{\Pi(\mathcal{F}_{n,j})}\mathrm{e}^{-(4-c)n\epsilon^2} \end{aligned}$$

so that (6) yields the result. □

Lemma 2. *Let $H, M \in \mathbb{N}$, $\underline{\sigma} > 0$ and, for $h = 1,\dots,H$, $0 \le \underline{a}_h < \bar{a}_h$, $1 \le u_h$. Define*

$$\begin{aligned} \mathcal{G} = \{f_P \text{ with } P = \textstyle\sum_{h\ge1}\pi_h\delta_{(\mu_h,\Sigma_h)} :\ & \textstyle\sum_{h>H}\pi_h \le \epsilon;\ \text{ for } h\le H,\ \underline{a}_h < \|\mu_h\| \le \bar{a}_h, \\ & \underline{\sigma}^2 \le \lambda_d(\Sigma_h), \lambda_1(\Sigma_h) \le \underline{\sigma}^2(1+\epsilon/\sqrt{d})^M,\ \lambda_1(\Sigma_h)/\lambda_d(\Sigma_h) \le u_h\} \end{aligned}$$

Then, for some positive constant C_1,

$$\begin{aligned} N(\epsilon,\mathcal{G},\|\cdot\|_1) \lesssim \exp\Big\{ & dH\log M + H\log\tfrac{C_1}{\epsilon} \\ & + \textstyle\sum_{h\le H}\log\left[\left(\frac{\bar{a}_h}{\underline{\sigma}\epsilon/2}+1\right)^d - \left(\frac{\underline{a}_h}{\underline{\sigma}\epsilon/2}-1\right)^d\right] + \frac{d(d-1)}{2}\log\frac{2d\,u_h}{\epsilon^2}\Big\}. \end{aligned} \tag{27}$$

Proof. The proof of (27) is based on a modification of the arguments of Proposition 2 in Shen et al. (2013) to the location-scale mixture case. We recall here that a set $\hat{\mathcal{G}} \subset \mathcal{G}$ with the property that any element of $\mathcal{G}$ is within ϵ-distance from an element of $\hat{\mathcal{G}}$ is called an *ϵ-net* over $\mathcal{G}$. Since $N(\epsilon,\mathcal{G},d)$ is the minimal cardinality of an ϵ-net over $\mathcal{G}$, $N(\epsilon,\mathcal{G},d) \le \#(\hat{\mathcal{G}})$.

Let $P_1 = \sum_{h\ge1}\pi_h^{(1)}\delta_{(\mu_h^{(1)},\Sigma_h^{(1)})}$ and $P_2 = \sum_{h\ge1}\pi_h^{(2)}\delta_{(\mu_h^{(2)},\Sigma_h^{(2)})}$. Then

$$\|f_{P_1} - f_{P_2}\|_1 \le \sum_{h\le H}\pi_h^{(1)}\left\|\phi_{\Sigma_h^{(1)}}(\cdot-\mu_h^{(1)}) - \phi_{\Sigma_h^{(2)}}(\cdot-\mu_h^{(2)})\right\|_1 + \sum_{h\le H}\left|\pi_h^{(1)}-\pi_h^{(2)}\right| + \sum_{h>H}\left\{\pi_h^{(1)}+\pi_h^{(2)}\right\}.$$

Note that

$$\left\|\phi_{\Sigma_h^{(1)}}(\cdot-\mu_h^{(1)})-\phi_{\Sigma_h^{(2)}}(\cdot-\mu_h^{(2)})\right\|_1 \leq \left\|\phi_{\Sigma_h^{(2)}}(\cdot-\mu_h^{(1)})-\phi_{\Sigma_h^{(2)}}(\cdot-\mu_h^{(2)})\right\|_1 + \left\|\phi_{\Sigma_h^{(1)}}-\phi_{\Sigma_h^{(2)}}\right\|_1 .$$

The first term in the r.h.s. is smaller than $\sqrt{2/\pi}\|\mu_h^{(1)}-\mu_h^{(2)}\|/\sqrt{\lambda_d(\Sigma_h^{(2)})}$. As for the second term, use spectral decompositions $\Sigma_h^{(j)} = [O_h^{(j)}\Lambda_h^{(j)}(O_h^{(j)})^T]^{-1}$, $j=1,2$, where $O_h^{(1)}, O_h^{(2)}$ are orthogonal matrices and $\Lambda_h^{(1)}, \Lambda_h^{(2)}$ are diagonal matrices. Drop the index h so to ease the notation and write $\Sigma_h^{(j)} = \Sigma_j$, $O_h^{(j)} = O_j$ and $\Lambda_h^{(j)} = \Lambda_j$. Moreover, let $\lambda_{j,1} \geq \ldots \geq \lambda_{j,d}$ be the diagonal elements of Λ_j. Finally, define $\widetilde{\Sigma} = (O_2\Lambda_1 O_2^T)^{-1}$. By triangular inequality

$$\|\phi_{\Sigma_1}-\phi_{\Sigma_2}\|_1 \leq \left\|\phi_{\widetilde{\Sigma}}-\phi_{\Sigma_2}\right\|_1 + \left\|\phi_{\widetilde{\Sigma}}-\phi_{\Sigma_1}\right\|_1 . \tag{28}$$

For each term in the r.h.s., use Csiszár's inequality, $\|f-g\|_1^2 \leq 2\int \log(f/g)f$, and the exact expression of the Kullback-Leibler divergence between zero mean multivariate Gaussians to get

$$\int \log(\phi_{\Sigma_2}/\phi_{\Sigma_1})\phi_{\Sigma_2} = \frac{1}{2}\left(\mathrm{tr}(\Sigma_1^{-1}\Sigma_2) - \log\det(\Sigma_1^{-1}\Sigma_2) - d\right).$$

As for the first term in the r.h.s. of (28),

$$\begin{aligned}\left\|\phi_{\widetilde{\Sigma}}-\phi_{\Sigma_2}\right\|_1 &\leq \left\{\mathrm{tr}(\Sigma_2^{-1}\widetilde{\Sigma}) - \log\det(\Sigma_2^{-1}\widetilde{\Sigma}) - d\right\}^{1/2} \\ &= \left\{\textstyle\sum_{i=1}^d (\lambda_{2,i}/\lambda_{1,i} - \log\lambda_{2,i}/\lambda_{1,i} - 1)\right\}^{1/2}\end{aligned}$$

where $\lambda_{2,i}/\lambda_{1,i}$ corresponds to $\lambda_{d-i+1}(\Sigma_h^{(1)})/\lambda_{d-i+1}(\Sigma_h^{(2)})$ in the original notation. As for the second term in the r.h.s. of (28),

$$\begin{aligned}\left\|\phi_{\widetilde{\Sigma}}-\phi_{\Sigma_1}\right\|_1 &\leq \left\{\mathrm{tr}(\Sigma_1^{-1}\widetilde{\Sigma}) - \log\det(\Sigma_1^{-1}\widetilde{\Sigma}) - d\right\}^{1/2} \\ &= \left\{\mathrm{tr}(Q\Lambda_1 Q^T\Lambda_1^{-1}) - d\right\}^{1/2} = \left\{\mathrm{tr}(Q\Lambda_1 Q^T\Lambda_1^{-1} - I)\right\}^{1/2}\end{aligned}$$

where $Q = O_2^T O_1$. Let $B = Q - I$ and $\|B\|_{\max} = \max|b_{i,j}|$. Then

$$\|B\|_{\max} \leq \|B\|_2 = \|O_2^T O_1 - I\|_2 \leq \|O_2^T\|_2\|O_1 - O_2\|_2 = \|O_1 - O_2\|_2.$$

Let b_{ij} be the element (i,j) of B. Note that

$$\begin{aligned}\mathrm{tr}(Q\Lambda_1 Q^T\Lambda_1^{-1} - I) &= \mathrm{tr}(B + \Lambda_1 B^T\Lambda_1^{-1} + B\Lambda_1 B^T\Lambda_1^{-1}) \\ &= \mathrm{tr}(B) + \mathrm{tr}(\Lambda_1 B^T\Lambda_1^{-1}) + \mathrm{tr}(B\Lambda_1 B^T\Lambda_1^{-1}) \\ &= \mathrm{tr}(B) + \mathrm{tr}(B^T) + \mathrm{tr}(B\Lambda_1 B^T\Lambda_1^{-1}) \\ &\leq 2\mathrm{tr}(B) + \tfrac{\lambda_{11}}{\lambda_{1d}}\mathrm{tr}(BB^T) \\ &= 2\mathrm{tr}(B) + 2\tfrac{\lambda_{11}}{\lambda_{1d}}\mathrm{tr}(I-Q) = 2\mathrm{tr}(B)\left[1 - \tfrac{\lambda_{11}}{\lambda_{1d}}\right] \\ &\leq 2d\|B\|_{\max}\left[\tfrac{\lambda_{1,1}}{\lambda_{1,d}} - 1\right] \leq 2d\|B\|_{\max}\tfrac{\lambda_{1,1}}{\lambda_{1,d}}\end{aligned}$$

where in the second last inequality we use $\mathrm{tr}(B) \geq -d\|B\|_{\max}$ together with $\lambda_{1,1}/\lambda_{1,d} \geq 1$. Hence,

$$\left\|\phi_{\widetilde{\Sigma}} - \phi_{\Sigma_1}\right\|_1 \leq \left\{2d\|O_1 - O_2\|_2 \tfrac{\lambda_{1,1}}{\lambda_{1,d}}\right\}^{1/2}$$

where $\lambda_{1,1}/\lambda_{1,d}$ corresponds to $\lambda_1(\Sigma_h^{(1)})/\lambda_d(\Sigma_h^{(1)})$ in the original notation. In summary, back to the original notation,

$$\begin{aligned}\|f_{P_1} - f_{P_2}\|_1 \leq \sum_{h\leq H} \pi_h^{(1)} \Bigg[\sqrt{\frac{2}{\pi}} \frac{\|\mu_h^{(1)} - \mu_h^{(2)}\|}{\sqrt{\lambda_d(\Sigma_h^{(2)})}} + \left\{\sum_{i=1}^d \left(\frac{\lambda_i(\Sigma_h^{(1)})}{\lambda_i(\Sigma_h^{(2)})} - \log \frac{\lambda_i(\Sigma_h^{(1)})}{\lambda_i(\Sigma_h^{(2)})} - 1\right)\right\}^{1/2} \\ + \left\{2d\|O_h^{(1)} - O_h^{(2)}\|_2 \frac{\lambda_1(\Sigma_h^{(1)})}{\lambda_d(\Sigma_h^{(1)})}\right\}^{1/2}\Bigg] + \sum_{h\leq H}\left|\pi_h^{(1)} - \pi_h^{(2)}\right| + \sum_{h>H}\left\{\pi_h^{(1)} + \pi_h^{(2)}\right\}.\end{aligned}$$

Now pick $f_P \in \mathcal{G}$ with $P = \sum_{h\geq 1} \pi_h \delta_{(\mu_h,\Sigma_h)}$, $\Sigma_h = (O_h \Lambda_h O_h^T)^{-1}$ and $\Lambda_h = \mathrm{diag}(\lambda_{h,1}, \ldots, \lambda_{h,d})$. Then find

• $\hat{\mu}_h \in \hat{R}_h$, $h = 1, \ldots, H$, where $\hat{R}_h$ is a $\underline{\sigma}\epsilon$-net of $R_h = \{\mu \in \mathbb{R}^d : \ \underline{a}_h < \|\mu\| \leq \bar{a}_h\}$ such that $\|\mu_h - \hat{\mu}_h\| < \underline{\sigma}\epsilon$;

• $\hat{\pi}_1, \ldots, \hat{\pi}_H \in \hat{\Delta}$, where $\hat{\Delta}$ is a ϵ-net of the H-dimensional probability simplex Δ such that $\sum_{h\leq H} |\tilde{\pi}_h - \hat{\pi}_h| \leq \epsilon$, and $\tilde{\pi}_h = \pi_h \Big/ \sum_{l\leq H} \pi_l$, $h \leq H$;

• $\hat{O}_h \in \hat{\mathcal{O}}_h$, $h = 1, \ldots, H$, where $\hat{\mathcal{O}}_h$ is a δ_h-net of the set $\mathcal{O}$ of $d \times d$ orthogonal matrices with respect to the spectral norm $\|\cdot\|_2$, with $\delta_h = \epsilon^2/(2d\, u_h)$ such that $\|O_h - \hat{O}_h\|_2 \leq \delta_h$;

• $(m_{h,1}, \ldots, m_{h,d}) \in \{1, \ldots, M\}^d$, $h = 1, \ldots, H$, such that $\hat{\lambda}_{h,i} = \{\underline{\sigma}^2(1 + \epsilon/\sqrt{d})^{m_{h,i}-1}\}^{-1}$ satisfies $1 \leq \hat{\lambda}_{h,i}/\lambda_{h,i} < (1 + \epsilon/\sqrt{d})$.

Take $\hat{P} = \sum_{h\leq H} \hat{\pi}_h \delta_{(\hat{\mu}_h, \hat{\Sigma}_h)}$ and $\hat{\Sigma}_h = (\hat{O}_h \hat{\Lambda}_h \hat{O}_h^T)^{-1}$, where $\hat{\Lambda}_h = \mathrm{diag}(\hat{\lambda}_{h,1}, \ldots, \hat{\lambda}_{h,d})$. Then

$$\begin{aligned}\|f_P - f_{\hat{P}}\|_1 \leq & \max_{h\leq H}\left[\sqrt{\frac{2}{\pi}}\frac{\|\mu_h - \hat{\mu}_h\|}{\sqrt{\lambda_d(\hat{\Sigma}_h)}} + \left\{\sum_{i=1}^d \left(\frac{\lambda_i(\Sigma_h)}{\lambda_i(\hat{\Sigma}_h)} - \log\frac{\lambda_i(\Sigma_h)}{\lambda_i(\hat{\Sigma}_h)} - 1\right)\right\}^{1/2}\right] \\ & + \sum_{h\leq H} \pi_h \left\{2d\|O_h - \hat{O}_h\|_2 \frac{\lambda_1(\Sigma_h)}{\lambda_d(\hat{\Sigma}_h)}\right\}^{1/2} + \sum_{h\leq H}|\pi_h - \hat{\pi}_h| + \sum_{h>H}\pi_h \\ \leq & \max_{h\leq H}\left[\sqrt{\frac{2}{\pi}}\frac{\|\mu_h - \hat{\mu}_h\|}{\underline{\sigma}} + \left\{\sum_{i=1}^d \left(\frac{\hat{\lambda}_{h,d-i+1}}{\lambda_{h,d-i+1}} - 1\right)^2\right\}^{1/2}\right] \\ & + \sum_{h\leq H} \pi_h \left\{2d\|O_h - \hat{O}_h\|_2 \frac{\lambda_{h,1}}{\lambda_{h,d}}\right\}^{1/2} + \sum_{h\leq H}|\pi_h - \hat{\pi}_h| + \sum_{h>H}\pi_h,\end{aligned}$$

where in the second inequality $x - \log x - 1 \leq (x-1)^2$ for $x \geq 1$ together with $\hat{\lambda}_{h,d-i+1}/\lambda_{h,d-i+1} \geq 1$

has been used to bound the second term. Moreover,

$$\begin{aligned}\sum_{h\le H}|\pi_h-\hat\pi_h| &\le \sum_{h\le H}|\pi_h-(1-\textstyle\sum_{h>H}\pi_h)\hat\pi_h| + \sum_{h\le H}|(1-\textstyle\sum_{h>H}\pi_h)\hat\pi_h-\hat\pi_h| \\ &= (1-\textstyle\sum_{h>H}\pi_h)\sum_{h\le H}|\tilde\pi_h-\hat\pi_h| + (\textstyle\sum_{h>H}\pi_h)\sum_{h\le H}\hat\pi_h \\ &\le \sum_{h\le H}|\tilde\pi_h-\hat\pi_h| + \sum_{h>H}\pi_h \le 2\epsilon.\end{aligned}$$

Hence,

$$\begin{aligned}\|f_P-f_{\hat P}\|_1 &\le \sqrt{\frac{2}{\pi}}\epsilon + \left\{\sum_{i=1}^d\left(1+\frac{\epsilon}{\sqrt{d}}-1\right)^2\right\}^{1/2} + \sum_{h\le H}\pi_h\left\{2d\frac{\epsilon^2}{2d\,u_h}\frac{\lambda_{h,1}}{\lambda_{h,d}}\right\}^{1/2} + 2\epsilon + \epsilon \\ &\le \epsilon + \{\epsilon^2\}^{1/2} + \{\epsilon^2\}^{1/2} + 3\epsilon = 6\epsilon.\end{aligned}$$

Thus a 6ϵ-net of $\mathcal{G}$, in the L_1 distance, can be constructed with $f_{\hat P}$ as above. Recalling that $\#(\hat\Delta)\lesssim\epsilon^{-H}$, $\#(\hat{\mathcal{O}}_h)\lesssim\delta_h^{-\frac{d(d-1)}{2}}$ and $\#(\hat R_h)\le(\bar a_h/(\underline\sigma\epsilon/2)+1)^d-(\underline a_h/(\underline\sigma\epsilon/2)-1)^d$, the total number is

$$\lesssim (M)^{dH}\epsilon^{-H}\prod_{h\le H}\left[\left(\frac{\bar a_h}{\underline\sigma\epsilon/2}+1\right)^d-\left(\frac{\underline a_h}{\underline\sigma\epsilon/2}-1\right)^d\right]C_1\left(\frac{2d\,u_h}{\epsilon^2}\right)^{\frac{d(d-1)}{2}}$$

for some positive constant C_1. Finally, the constant factor by 6 can be absorbed in the bound, hence (27) follows. □

Remark 2. *The usual low-entropy, high mass sieve approach of Theorem 2 by Ghosal et al. (1999) is not suitable for multivariate DP location-scale mixture. This theorem, indeed, requires a sieve $\mathcal{F}_n$ such that: (i) $\Pi(\mathcal{F}_n^c)$ decreases exponentially fast in n, and (ii) the L_1 metric entropy of $\mathcal{F}_n$ grows linearly with n. As for (i), let $\mathcal{F}_n$ to be defined as $\mathcal{G}$ of Lemma 2 with $M=\underline\sigma^{-2c_2}=\bar a_h=u_h=n$, $H_n=\lfloor Cn\epsilon^2/\log n\rfloor$, $\underline a_h$=0 and $l_h=1$. Clearly $\mathcal{F}_n\uparrow\mathcal{F}$ as $n\to\infty$. However, in order to have $\mathcal{F}_n$ satisfying condition (i), stronger tail conditions on P^* are necessary than those in (7) and (10), namely exponential tail behaviors. While one can be still satisfied with a Gaussian specification for μ, exponential tail for the condition number turns out to be too restrictive ruling out all the models discussed in Section 3 as well as any other reasonable prior specification for non diagonal covariance matrices. However, as an inspection of the proof of Lemma 2 reveals, with diagonal covariance matrices the metric entropy is considerably smaller and one can prove consistency by applying Theorem 2 of Ghosal et al. (1999) under essentially the same tail requirements of DP location mixture of Shen et al. (2013). We omit the details here.*

Proof of Corollary 1. That conditions (8) and (9) are satisfied when $\mathcal{L}=IW(\Sigma_0,\nu)$ has been proved in Lemma 1 of Shen et al. (2013). Therefore we focus here on condition (10) on the condition number. Let for the moment $\Sigma_0=I$. We directly work with the joint distribution of the ordered eigenvalues of a Wishart-distributed matrix:

$$f_{\lambda_1,\dots,\lambda_d}(x_1,\dots,x_d)=K\exp\left\{-\frac{1}{2}\sum_{i=1}^d x_i\right\}\prod_{j=1}^d(x_j)^{(\nu-d-1)/2}\prod_{i<j}(x_i-x_j),$$

over the set $\{(x_1,\ldots,x_d)\in(0,\infty)^d:\ x_1\geq\ldots\geq x_d\}$, where K is a known normalizing constant. See Corollary 3.2.19, page 107, of Muirhead (1982). To show condition (10), let $z=\lambda_1/\lambda_d$ and apply the Jacobi's transformation formula for $g(\lambda_1,\ldots,\lambda_d)=(z,\lambda_2,\ldots,\lambda_d)$. For $J_{g^{-1}}$ denoting the Jacobian of g^{-1}, we have

$$f_{z,\lambda_2,\ldots,\lambda_d}(z,\ldots,y)=|J_{g^{-1}}|f_{\lambda_1,\ldots,\lambda_d}(x_dz,x_2,\ldots,x_d),$$

and the marginal of z can be obtained integrating out $x_2,\ldots,x_d$. Using $|J_{g^{-1}}|=x_d$, $x_d\geq 0$ and $x_d\leq x_i\leq zx_d$ for $i=2,\ldots,d-1$, after some algebra one gets

$$f_z(z)\leq K''z^{-\frac{\nu-d+3}{2}}$$

for some constant K'', so that $P^*(z>x)\lesssim x^{-(\nu-d+1)/2}$ as $x\to\infty$. Cfr. Theorem 3.2 of Edelman and Sutton (2005). If $\Sigma_0\neq I$, the conclusion holds for a different set of constants, see Theorem 4 of Matthaiou et al. (2010) for the complex Wishart case. Condition (10) is then satisfied under the hypothesis on ν. □

Proof of Corollary 2. To prove condition (8), consider

$$P^*\left\{\lambda_1(\Sigma^{-1})\geq x\right\}\leq P^*\left\{\mathrm{tr}(\Sigma^{-1})\geq x\right\}\leq P^*\left\{\mathrm{tr}(\Omega^{-1})\geq x\right\},$$

where the last inequality is verified since $\mathrm{tr}(\Sigma^{-1})\leq\mathrm{tr}(\Omega^{-1})$ by an application of the Woodbury's identity. Since the trace of Ω^{-1} is the sum of i.i.d. gamma random variable, and the gamma has exponential tail, (8) is verified. As for (9), consider

$$P^*\left\{\lambda_d(\Sigma^{-1})<1/x\right\}\leq x^{-1}E_{P^*}\left\{\lambda_1(\Sigma)\right\}\leq x^{-1}\left[E_{P^*}\left\{\lambda_1(\Omega)\right\}+E_{P^*}\left\{\lambda_1(\Gamma\Gamma^T)\right\}\right]$$

by Markov's and Weyl's inequalities. The expectation of $\lambda_1(\Omega)$ is finite and since $\lambda_1(\Gamma\Gamma^T)\leq\mathrm{tr}(\Gamma\Gamma^T)\sim\chi^2_{rd}$, also the expectation of $\lambda_1(\Gamma\Gamma^T)$ is finite. Finally, to prove condition (10), first use Markov's inequality. For $k>d(d-1)$, we have

$$\begin{aligned}P^*\left\{\lambda_1(\Sigma^{-1})/\lambda_d(\Sigma^{-1})\geq x\right\}&=P^*\left\{\left(\lambda_1(\Sigma)/\lambda_d(\Sigma)\right)^k\geq x^k\right\}\\&\leq x^{-k}E\left\{\left(\lambda_1(\Sigma)/\lambda_d(\Sigma)\right)^k\right\}.\end{aligned}$$

Using again Weyl's inequalities and noting that $\lambda_d(\Gamma\Gamma^T)=0$ when $r<d$, we have

$$\frac{\lambda_1(\Sigma)}{\lambda_d(\Sigma)}\leq\frac{\lambda_1(\Omega)+\lambda_1(\Gamma\Gamma^T)}{\lambda_d(\Omega)+\lambda_d(\Gamma\Gamma^T)}=\frac{\lambda_1(\Omega)+\lambda_1(\Gamma\Gamma^T)}{\lambda_d(\Omega)}\leq\frac{\lambda_1(\Omega^{-1})}{\lambda_d(\Omega^{-1})}+\frac{\mathrm{tr}(\Gamma\Gamma^T)}{\lambda_d(\Omega)}.$$

By convexity of $g(x)=x^k$, for any $x,y\in\mathbb{R}^+$ we have $(x+y)^k\leq 2^{k-1}(x^k+y^k)$. Hence $\lambda_1(\Sigma)/\lambda_d(\Sigma)$ has finite k-th moment if both summand have finite k-th moment. First consider the distribution of $\lambda_1(\Omega^{-1})/\lambda_d(\Omega^{-1})$. Since Ω^{-1} is a diagonal matrix with iid gamma distributed diagonal elements, the joint distribution of the ordered eigenvalues of Ω^{-1} is the joint distribution of the ordered statistics of an iid sample from $\mathrm{Ga}(a,b)$, i.e.

$$f_{\lambda_1(\Omega^{-1})\ldots\lambda_d(\Omega^{-1})}(x_1,\ldots,x_d)=d!f_{\sigma_1^{-2}}(x_1)\times\ldots\times f_{\sigma_d^{-2}}(x_d)=K\prod_{j=1}^d x_j^{a-1}\exp\left\{-b\sum_{j=1}^d x_j\right\},$$

over the set $\{(x_1,\dots,x_d)\in(0,\infty)^d:\ x_1\geq\ldots\geq x_d\}$, where K is a known normalizing constant. Now we use the Jacobi's transformation formula as in the proof of Corollary 1 to get, for $z=\lambda_1(\Omega^{-1})/\lambda_d(\Omega^{-1})$,

$$f_z(z)=K\int_{x_3}^{\infty}\dots\int_{x_{d-2}}^{\infty}\int_{x_d}^{\infty}\prod_{j=2}^{d-1}x_j^{a-1}\exp\left\{-b\sum_{j=1}^{d}x_j\right\}\mathrm{d}x_2\dots\mathrm{d}x_{d-1}\times$$
$$\int_0^{\infty}x_d(zx_d^2)^{a-1}\exp\{-bx_d(z+1)\}\mathrm{d}x_d\leq\ldots\leq K'\frac{z^{a-1}}{(z+1)^{2a}}.$$

Hence $\lambda_1(\Omega^{-1})/\lambda_d(\Omega^{-1})$ has finite k-th moment as long as $a>k$. Now, since

$$\frac{\mathrm{tr}(\Gamma\Gamma^T)}{\lambda_d(\Omega)}=\mathrm{tr}(\Gamma\Gamma^T)\lambda_1(\Omega^{-1})$$

and the k-th moment of the product of independent random variables is the product of the k-th moments, it remains to show that both $\mathrm{tr}(\Gamma\Gamma^T)$ and $\lambda_1(\Omega^{-1})$ have finite k-th moment. To this aim, it is sufficient to note that $\mathrm{tr}(\Gamma\Gamma^T)\sim\chi^2_{dr}$ and $\lambda_1(\Omega^{-1})$ has the distribution of the first order statistics of a sample of d independent gamma random variables. The proof is then complete. □

Proof of Corollary 3. Since each eigenvalues has independent inverse-gamma distribution, conditions (8)-(9)-(10) are satisfied following part of the proof of Corollary 2. □

Proof of Proposition 1. Let $M_n=\underline{\sigma}_n^{-2c_2}=n$ with $\epsilon_n=n^{-\gamma}(\log n)^t$ for $t_0=t+s$ and $H_n=\lfloor Cn\epsilon_n^2/\log n\rfloor$ for a positive constant to be determined later. Also define $\mathcal{F}_n$ and $\mathcal{F}_{n,\boldsymbol{j},\boldsymbol{l}}$ as in the proof of Theorem 2 with $\epsilon=\epsilon_n$. We prove next that

$$\Pi(\mathcal{F}_n^c)\lesssim\mathrm{e}^{-4n\tilde{\epsilon}_n^2} \tag{29}$$
$$\sum\nolimits_{\boldsymbol{j},\boldsymbol{l}}\sqrt{N(\epsilon_n,\mathcal{F}_{n,\boldsymbol{j},\boldsymbol{l}},d)}\sqrt{\Pi(\mathcal{F}_{n,\boldsymbol{j},\boldsymbol{l}})}\mathrm{e}^{-n\epsilon_n^2}\to 0. \tag{30}$$

The thesis then follows by an application of Theorem 5 by Ghosal and van der Vaart (2007), cfr. version in Theorem 3 of Kruijer et al. (2010).

As for (29), refer to equation (11). Note that $n\epsilon_n^2=n^{1-2\gamma}(\log n)^{2t_0}$ and $H_n=Cn^{1-2\gamma}(\log n)^{2t_0-1}$. We have

$$\begin{aligned}\left\{\frac{\mathrm{e}\alpha}{H_n}\log\frac{1}{\epsilon_n}\right\}^{H_n}&\lesssim\exp\{-H_n\log H_n\}=\exp\left\{-Cn^{1-2\gamma}(\log n)^{2t_0-1}\log(Cn^{1-2\gamma}(\log n)^{2t_0-1})\right\}\\&\lesssim\exp\left\{-Cn^{1-2\gamma}(\log n)^{2t_0-1}(1-2\gamma)\log n\right\}=\exp\left\{-(1-2\gamma)Cn\epsilon_n^2\right\}\\H_n\mathrm{e}^{-c_1\underline{\sigma}_n^{-2c_2}}&=Cn^{1-2\gamma}(\log n)^{2t_0-1}\mathrm{e}^{-nc_1}=o\left(\exp\left\{-n\epsilon_n^2\right\}\right)\\H_n\underline{\sigma}_n^{-2c_3}\left(1+\epsilon_n/\sqrt{d}\right)^{-c_3M_n}&\leq Cn^{c_3/c_2+1-2\gamma}(\log n)^{2t_0-1}\exp\{-c_3n\epsilon_n/(2\sqrt{d})\}=o\left(\exp\left\{-n\epsilon_n^2\right\}\right)\end{aligned}$$

Therefore $\Pi(\mathcal{F}_n^c)\lesssim\exp\left\{-(1-2\gamma)Cn\epsilon_n^2\right\}\lesssim\mathrm{e}^{-4n\tilde{\epsilon}_n^2}$ for all large n since $t<t_0$, Note that the rate at which the prior probability of the complement of the sieve vanishes is determined by the truncation level of the stick-breaking weights.

As for (30), refer to equation (12). Based on the inequality $d(f,g)^2\leq\|f-g\|_1$,

$$N(\epsilon_n,\mathcal{F}_{n,\boldsymbol{j},\boldsymbol{l}},d)\lesssim\exp\left\{dH_n\log n+H_n\log\tfrac{C_1}{\epsilon_n^2}+\textstyle\sum_{h\leq H_n}\log\frac{n^{c_4d}j_h^{d-1}}{\epsilon_n^{2d}}+\frac{d(d-1)}{2}\log\frac{2dn2^{l_h}}{\epsilon_n^4}\right\}$$

Since $H_n \log(1/\epsilon_n) = \gamma C n\epsilon_n^2 + o\big(n\epsilon_n^2\big)$, we also have

$$N(\epsilon_n, \mathcal{F}_{n,\boldsymbol{j},\boldsymbol{l}}, d) \lesssim \exp\left\{[d + c_4 d + 2\gamma + 2\gamma d + 2\gamma d(d-1)]\, C n\epsilon_n^2\right\} \prod_{h\le H_n} j_h^{d-1} n^{\frac{d(d-1)}{2} 2^{l_h}}$$
$$< \exp\left\{d^* C n\epsilon_n^2\right\} \prod_{h\le H_n} j_h^{d-1} n^{\frac{d(d-1)}{2} 2^{l_h}}$$

where $d^* = d + c_4 d + 1 + d + d(d-1)$ and $\gamma < 1/2$ has been used in the last inequality. By using (13), $\sqrt{N(\epsilon_n, \mathcal{F}_{n,\boldsymbol{j},\boldsymbol{l}}, d)}\sqrt{\Pi(\mathcal{F}_{n,\boldsymbol{j},\boldsymbol{l}})}$ can be bounded by a multiple of

$$\exp\left\{(d^*/2) C n\epsilon_n^2\right\} \prod_{h\le H_n} j_h^{\frac{d-1}{2}} [\sqrt{n}(j_h - 1)]^{-\mathbb{1}_{(j_h\ge 2)}(r+1)} n^{\frac{d(d-1)}{4} 2^{l_h} - \mathbb{1}_{(l_h\ge 1)} \frac{\kappa}{2} 2^{l_h - 1}}$$

Summing with respect to $\boldsymbol{j}$ and $\boldsymbol{l}$, by arguments similar to those used in the proof of Theorem 2, we obtain, for n large enough and when $r > (d-1)/2$ and $\kappa > d(d-1)$, the upper bound

$$\sum\nolimits_{\boldsymbol{j},\boldsymbol{l}} \sqrt{N(\epsilon_n, \mathcal{F}_{n,\boldsymbol{j},\boldsymbol{l}}, d)}\sqrt{\Pi(\mathcal{F}_{n,\boldsymbol{j},\boldsymbol{l}})} \le \exp\left\{(d^*/2) C n\epsilon_n^2\right\} (1 + n^{-\frac{r+1}{2}} K)^{H_n} \left[2n^{\frac{d(d-1)}{4}}\right]^{H_n}$$
$$\lesssim \exp\left\{(d^*/2 + d(d-1)/4)\, C n\epsilon_n^2\right\}$$

where $K := \sum_{j\ge 2} j^{(d-1)/2}(j-1)^{-(r+1)} < \infty$. Hence (30) is satisfied for C sufficiently small to satisfy $C \le 4/[2d^* + d(d-1)]$. The proof is complete. □

Proof of Proposition 2. The proof follows the arguments presented in Theorem 4 of Shen et al. (2013) adapted to the location-scale case. Denote by $F_0\, g = \int g(x) f_0(x) \mathrm{d}x$ the expectation of $g(X)$ under $X \sim f_0$. Consider n large enough such that $\tilde\epsilon_n < s_0^\beta$ and fix $\sigma_n^\beta = \tilde\epsilon_n \{\log(1/\tilde\epsilon_n)\}^{-1}$. As in Proposition 1 of Shen et al. (2013), define $E_{\sigma_n} = \{\|x\| \in \mathbb{R}^d : \ f_0(x) \ge \sigma_n^{(4\beta+2\epsilon+8)/\delta}\}$ for some $\delta \in (0,1)$ and $a_{\sigma_n} = a_0\{\log(1/\sigma_n)\}^{1/\tau}$. Then, $F_0(E_{\sigma_n}^c) \le B_0 \sigma_n^{4\beta+2\epsilon+8}$ for some constant B_0, $E_{\sigma_n} \subset \{\|x\| \in \mathbb{R}^d : \|x\| \le a_{\sigma_n}\}$ and, there exists a density $\tilde h_{\sigma_n}$ with support E_{σ_n} such that $d(f_0, \phi_{\sigma_n} \star \tilde h_{\sigma_n}) \lesssim \sigma_n^\beta$. Find $b_1 > \max\{1, 1/(2\beta)\}$ such that $\tilde\epsilon_n^{b_1}\{\log(1/\epsilon_n)\}^{5/4} \le \tilde\epsilon_n$. Apply Theorem B1 of Shen et al. (2013) for $\sigma = \sigma_n$, $a = a_{\sigma_n}$, $F = \tilde h_{\sigma_n}$ and $\epsilon = \tilde\epsilon_n^{2b_1}$ to get $F_{\sigma_n} = \sum_{j=1}^N p_j \delta_{\mu_j}$ with

$$N \le D[(a_{\sigma_n}\sigma_n^{-1} \vee 1)\log(1/\tilde\epsilon_n^{2b_1})]^d \lesssim \sigma_n^{-d}\{\log(1/\tilde\epsilon_n)\}^{d+d/\tau}$$

many support points inside $\{x \in \mathbb{R} : \ \|x\| \le a_{\sigma_n}\}$ such that

$$\|\phi_{\sigma_n} \star \tilde h_{\sigma_n} - \phi_{\sigma_n} \star F_{\sigma_n}\|_1 \lesssim \tilde\epsilon_n^{2b_1}\{\log(1/\tilde\epsilon_n^{2b_1})\}^{1/2}.$$

Move each of the support point of F_{σ_n} to the nearest point on the grid $\{(n_1, \ldots, n_d)\sigma_n\tilde\epsilon_n^{2b_1} : n_i \in \mathbb{Z}, |n_i| < \lceil a_{\sigma_n}/(\sigma_n\tilde\epsilon_n^{2b_1})\rceil, i = 1, \ldots, d\}$, so that $\min_{j\ne l}\|\mu_j - \mu_l\| > \sigma_n\tilde\epsilon_n^{2b_1}$. These moves cost at most a constant times $\tilde\epsilon_n^{2b_1}$ to the L_1 distance. Modify further F_{σ_n} so that $(\mu_1, \ldots, \mu_N)$ form a $\sigma_n a_{\sigma_n}^{\tau/2}$-net of $\{x \in \mathbb{R}^d : \|x\| \le a_{\sigma_n}\}$. This can be achieved by adding at most $[a_{\sigma_n}/(\sigma_n a_{\sigma_n}^{\tau/2}) + 1]^d \lesssim \sigma_n^{-d} a_{\sigma_n}^{d - d\tau/2} \lesssim \sigma_n^{-d}\{\log(1/\sigma_n)\}^{d/\tau - d/2}$ support points so that we still have $N \lesssim \sigma_n^{-d}\{\log(1/\tilde\epsilon_n)\}^{d+d/\tau}$. Finally, modify the probability masses $p_1, \ldots, p_N$ to a version for which $p_j \ge \tilde\epsilon_n^{4db_1}$ for $j = 1, \ldots, N$. By a suitable extension of Lemma 3 of Kruijer et al. (2010) to d dimensions, the L_1 distance changes by the L_1 distance between the original and the new probability weights, and the latter can be shown to be smaller

than $2(N-1)\tilde{\epsilon}_n^{4db_1}$. In conclusion, we have

$$\begin{aligned}\|\phi_{\sigma_n} \star \tilde{h}_{\sigma_n} - \phi_{\sigma_n} \star F_{\sigma_n}\|_1 &\lesssim \tilde{\epsilon}_n^{2b_1}\{\log(1/\tilde{\epsilon}_n^{2b_1})\}^{1/2} + \tilde{\epsilon}_n^{2b_1} + N\tilde{\epsilon}_n^{4db_1} \\ &\lesssim \tilde{\epsilon}_n^{2b_1}\{\log(1/\tilde{\epsilon}_n)\}^{1/2} + \sigma_n^{-d}\{\log(1/\tilde{\epsilon}_n)\}^{d+d/\tau}\tilde{\epsilon}_n^{4db_1} \\ &= \tilde{\epsilon}_n^{2b_1}\{\log(1/\tilde{\epsilon}_n)\}^{1/2} + \tilde{\epsilon}_n^{d(4b_1-1/\beta)}\{\log(1/\tilde{\epsilon}_n)\}^{1+1/\tau+1/\beta} \\ &\lesssim \tilde{\epsilon}_n^{2b_1}\{\log(1/\tilde{\epsilon}_n)\}^{1/2} \end{aligned} \tag{31}$$

since $d(4b_1 - 1/\beta) > 2b_1$ for $b_1 > 1/(4d-2)\beta$ and the latter follows by the assumption made on b_1.

Place disjoint balls U_j with centers at $\mu_1, \ldots, \mu_N$ with diameter $\sigma_n\tilde{\epsilon}_n^{2b_1}$ each. Because of assumption (20) on f^*, for any $\|x\| \le a_{\sigma_n}$,

$$f^*(x) \ge c^* \mathrm{e}^{-b^* a_{\sigma_n}^{\tau^*}} \ge c^* \exp\{-b^* a_0^{\tau^*} \log(1/\sigma_n)^{\tau^*/\tau}\} \ge c^* \sigma_n^{b^* a_0^{\tau^*} \log(1/\sigma_n)^{\tau^*/\tau - 1}} \gtrsim \sigma_n^{b^* a_0^{\tau^*}},$$

where the second inequality follows from the assumption that $\tau^* \le \tau$. This implies that for n sufficiently large and some constant a_1, $a_1\sigma_n^{b^* a_0^{\tau^*}}(\sigma_n\tilde{\epsilon}_n^{2b_1})^d \le P^*(U_j) \le 1$. Further extend this to a partition $U_1, \ldots, U_K$ of $\mathbb{R}^d$ such that $a_1\sigma_n^{b^* a_0^{\tau^*}}(\sigma_n\tilde{\epsilon}_n^{2b_1})^d \le P^*(U_j) \le 1$ still holds. We can still have $K \lesssim \sigma_n^{-d}\{\log(1/\tilde{\epsilon}_n)\}^{d+d/\tau} = \tilde{\epsilon}_n^{-d/\beta}\{\log(1/\tilde{\epsilon}_n)\}^{sd}$ for $s = 1 + 1/\beta + 1/\tau$. The next step is to construct a partition of $\mathbb{R}^d \times \mathcal{S}$. To this aim, let

$$S_{\sigma_n} = \{\Sigma \in \mathcal{S} : \ \sigma_n^{-2} < \lambda_i(\Sigma^{-1}) < \sigma_n^{-2}(1 + \sigma_n^{2\beta}), \ i = 1, \ldots, d\}$$

which satisfies $P^*(S_{\sigma_n}) \gtrsim \mathrm{e}^{-C_3\sigma_n^{-\kappa^*}}$ by hypothesis (19). Define $V_j = U_j \times S_{\sigma_n}$ for $j = 1, \ldots, K$, and extend $\{V_1, \ldots, V_K\}$ to a partition $\{V_1, \ldots, V_M\}$ of $\mathbb{R}^d \times \mathcal{S}$ such that

$$a_2\sigma_n^{b^* a_0^{\tau^*} + d}\sigma_n\tilde{\epsilon}_n^{2db_1}\mathrm{e}^{-C_3\sigma_n^{-\kappa^*}} \le \alpha P^*(V_j) \le 1$$

for all $j = 1, \ldots, M$, for some positive constant a_2 and n sufficiently large. We can still have $M \lesssim \tilde{\epsilon}_n^{-d/\beta}\{\log(1/\tilde{\epsilon}_n)\}^{sd}$.

Set $p_j = 0$ for $j > N$ and consider the set $\mathcal{P}_{\sigma_n} \subset \mathscr{P}$ of probability measures P on $\mathbb{R}^d \times \mathcal{S}$ with

$$\sum\nolimits_{j=1}^{M} |P(V_j) - p_j| \le 2\tilde{\epsilon}_n^{2db_1}\mathrm{e}^{-C_3\sigma_n^{-\kappa^*}}.$$

Note that

$$\begin{aligned} M\,\tilde{\epsilon}_n^{2db_1}\mathrm{e}^{-C_3\sigma_n^{-\kappa^*}} &\lesssim \tilde{\epsilon}_n^{-d/\beta}\{\log(1/\tilde{\epsilon}_n)\}^{sd}\tilde{\epsilon}_n^{2db_1}\mathrm{e}^{-C_3\sigma_n^{-\kappa^*}} \le 1 \\ \min_{1\le j\le M}[\alpha P^*(V_j)]^{1/3} &\ge \left\{a_2\sigma_n^{b^* a_0^{\tau^*}+d}\tilde{\epsilon}_n^{2db_1}\mathrm{e}^{-C_3\sigma_n^{-\kappa^*}}\right\}^{1/3} \\ &= a_2^{1/3}\tilde{\epsilon}_n^{2db_1}\mathrm{e}^{-C_3\sigma_n^{-\kappa^*}}\left\{\sigma_n^{b^* a_0^{\tau^*}+d}\tilde{\epsilon}_n^{-4db_1}\mathrm{e}^{2C_3\sigma_n^{-\kappa^*}}\right\}^{1/3} \ge a_2^{1/3}\tilde{\epsilon}_n^{2db_1}\mathrm{e}^{-C_3\sigma_n^{-\kappa^*}} \end{aligned}$$

for n sufficiently large. Hence, by Lemma 10 of Ghosal and van der Vaart (2007),

$$\begin{aligned}\mathcal{D}_{\alpha,P^*}(\mathcal{P}_{\sigma_n}) &\ge C'\exp\left\{-c\,M\log\left[1/\tilde{\epsilon}_n^{2db_1}\mathrm{e}^{-C_3\sigma_n^{-\kappa^*}}\right]\right\} \ge C''\exp\{-c'M\sigma_n^{-\kappa^*}\} \\ &\ge C''\exp\{-c''\tilde{\epsilon}_n^{-d/\beta}\{\log(1/\tilde{\epsilon}_n)\}^{sd}\sigma_n^{-\kappa^*}\} \\ &= C''\exp\{-c'''\tilde{\epsilon}_n^{-(d+\kappa^*)/\beta}\{\log(1/\tilde{\epsilon}_n)\}^{d(s+\kappa^*/\beta)}\} \ge \exp\{-Cn\tilde{\epsilon}_n^2\}\end{aligned} \tag{32}$$

for some constants $C, C', C'', c, c', c'', c'''$ and for $t \geq d(s+\kappa^*/\beta)/(2+(d+\kappa^*)/\beta)$ since

$$\begin{aligned}\tilde{\epsilon}_n^{-(d+\kappa^*)/\beta}\{\log(1/\tilde{\epsilon}_n)\}^{d(s+\kappa^*/\beta)} &= n^{\frac{d+\kappa^*}{2\beta+d+\kappa^*}}(\log n)^{-(d+\kappa^*)t/\beta}\{\log[n^{\frac{\beta}{2\beta+d+\kappa^*}}(\log n)^{-t}]\}^{d(s+\kappa^*/\beta)} \\ &\lesssim n^{\frac{d+\kappa^*}{2\beta+d+\kappa^*}}(\log n)^{d(s+\kappa^*/\beta)-(d+\kappa^*)t/\beta} \leq n\tilde{\epsilon}_n^2\end{aligned}$$

when $d(s+\kappa^*/\beta)-(d+\kappa^*)t/\beta \leq 2t$, cfr. (22).

We show next that, for any $P \in \mathcal{P}_{\sigma_n}$, $d(f_0, f_P) \lesssim \sigma_n^\beta$. By triangle inequality,

$$d(f_0, f_P) \leq d(f_0, \phi_{\sigma_n} \star \tilde{h}_{\sigma_n}) + d(\phi_{\sigma_n} \star \tilde{h}_{\sigma_n}, \phi_{\sigma_n} \star F_{\sigma_n}) + d(f_P, \phi_{\sigma_n} \star F_{\sigma_n})$$

and $d(f_0, \phi_{\sigma_n} \star \tilde{h}_{\sigma_n}) \lesssim \sigma_n^\beta$ by Proposition 1 of Shen et al. (2013), while, by (31),

$$d(\phi_{\sigma_n} \star \tilde{h}_{\sigma_n}, \phi_{\sigma_n} \star F_{\sigma_n}) \lesssim \tilde{\epsilon}_n^{b_1}\{\log(1/\tilde{\epsilon}_n)\}^{1/4} \leq \tilde{\epsilon}_n\{\log(1/\tilde{\epsilon}_n)\}^{-1} = \sigma_n^\beta$$

since, by assumption, $\tilde{\epsilon}_n^{b_1}\{\log(1/\epsilon_n)\}^{5/4} \leq \tilde{\epsilon}_n$. In order to show that also $d(f_P, \phi_{\sigma_n} \star F_{\sigma_n}) \lesssim \sigma_n^\beta$, we use arguments presented in Lemma 5 of Ghosal and van der Vaart (2007). Let $V_0 = \bigcup_{j>N} V_j$. After some algebra,

$$\begin{aligned}f_P(x) - \phi_{\sigma_n} \star F_{\sigma_n}(x) = &\int_{V_0} \phi_\Sigma(x-\mu)\mathrm{d}P(\mu,\Sigma) + \sum_{j=1}^N \int_{V_j} (\phi_\Sigma(x-\mu) - \phi_\Sigma(x-\mu_j)\mathrm{d}P(\mu,\Sigma) \\ &+ \sum_{j=1}^N \int_{V_j} (\phi_\Sigma(x-\mu_j) - \phi_{\sigma_n}(x-\mu_j)\mathrm{d}P(\mu,\Sigma) + \sum_{j=1}^N \phi_{\sigma_n}(x-\mu_j)\left[P(V_j) - p_j\right]\end{aligned}$$

Note that $P(V_0) = 1 - \sum_{j=1}^N P(V_j) \leq \sum_{j=1}^N |P(V_j) - p_j|$. Hence, by triangle inequality,

$$\begin{aligned}\|f_P - \phi_{\sigma_n} \star F_{\sigma_n}\|_1 \leq &\sum_{j=1}^N \int_{V_j} \|\phi_\Sigma(\cdot-\mu) - \phi_\Sigma(\cdot-\mu_j)\|_1 \mathrm{d}P(\mu,\Sigma) \\ &+ \sum_{j=1}^N \int_{V_j} \|\phi_\Sigma - \phi_{\sigma_n}\|_1 \mathrm{d}P(\mu,\Sigma) + 2\sum_{j=1}^N |P(V_j) - p_j|\end{aligned}$$

Recall now that $V_j = U_j \times S_{\sigma_n}$ for $j = 1, \ldots, N$ so that, by construction, $\|\mu - \mu'\| \leq \sigma_n\tilde{\epsilon}_n^{2b_1}$ for any $\mu, \mu' \in U_j$. Also any $\Sigma \in S_{\sigma_n}$ satisfies $\det(\Sigma^{-1}) \geq 2^{-d}\sigma_n^{-2d}$, $y^T\Sigma^{-1}y \leq \|y\|\sigma_n^{-2}$ for any $y \in \mathbb{R}^d$ and $|\mathrm{tr}(\sigma_n^2\Sigma^{-1}) - d - \log\det(\sigma_n^2\Sigma^{-1})| \leq d\sigma_n^{2\beta}$. Hence, by using arguments similar to the proof of Lemma 2,

$$\begin{aligned}&\|\phi_\Sigma(\cdot-\mu) - \phi_\Sigma(\cdot-\mu_j)\|_1 \leq \|\mu-\mu_j\|\lambda_1(\Sigma^{-1})^{1/2} \leq \sigma_n\tilde{\epsilon}_n^{2b_1}\sigma_n^{-1} = \tilde{\epsilon}_n^{2b_1} \\ &\|\phi_\Sigma - \phi_{\sigma_n}\|_1 \leq \left|\mathrm{tr}(\sigma_n^{-2}\Sigma^{-1}) - d - \log\det(\sigma_n\Sigma^{-1})\right| \leq d\sigma_n^{2\beta}\end{aligned}$$

for any $(\mu, \Sigma) \in V_j$ and any $j = 1, \ldots, N$. Hence, for $P \in \mathcal{P}_{\sigma_n}$,

$$\|f_P - \phi_{\sigma_n} \star F_{\sigma_n}\|_1 \leq \tilde{\epsilon}_n^{2b_1} + d\sigma_n^{2\beta} + 4\tilde{\epsilon}_n^{2db_1}\mathrm{e}^{-C_3\sigma_n^{-\kappa^*}} \lesssim \sigma_n^{2\beta}$$

since $\tilde{\epsilon}_n^{2b_1} \leq \sigma_n^{2\beta}$ by the assumption on b_1. It follows that $d(f_P, \phi_{\sigma_n} \star F_{\sigma_n}) \lesssim \sigma_n^\beta$ as desired.

The last step is to control the ratio f_P/f_0 for any $P \in \mathcal{P}_{\sigma_n}$ in view of an application of Lemma B2 of Shen et al. (2013). Since $(\mu_1, \ldots, \mu_N)$ forms a $\sigma_n a_{\sigma_n}$-net of $\{x \in \mathbb{R}^d : \|x\| \leq a_{\sigma_n}\}$, the set

$\{\mu \in \mathbb{R}^d : \|\mu - x\| \leq \sigma_n(a_{\sigma_n} + \tilde{\epsilon}_n^{2b_1}/2)\}$ contains at least one U_j for $j = 1, \ldots, N$. Call $J(x)$ such index j and recall that, for $P \in \mathcal{P}_\sigma$,

$$P(V_{J(x)}) \geq p_{J(x)} - 2\tilde{\epsilon}_n^{2db_1} \mathrm{e}^{-C_3 \sigma_n^{-\kappa^*}} \geq \tilde{\epsilon}_n^{4db_1} - 2\tilde{\epsilon}_n^{2db_1} \mathrm{e}^{-C_3 \sigma_n^{-\kappa^*}} \gtrsim \tilde{\epsilon}_n^{4db_1}$$

Recall that f_0 is bounded. For $\|x\| \leq a_{\sigma_n}$ we have

$$\begin{aligned}
\frac{f_P(x)}{f_0(x)} &\geq K_1 \int_{S_{\sigma_n}} \int_{\|\mu - x\| \leq \sigma_n(a_{\sigma_n}^{\tau/2} + \tilde{\epsilon}_n^{2b_1}/2)} \phi_\Sigma(x - z) \mathrm{d}P(z, \Sigma) \\
&\geq K_2 \sigma_n^{-d} \exp\left\{ -\frac{1}{2\sigma_n^2} [\sigma_n(a_{\sigma_n}^{\tau/2} + \tilde{\epsilon}_n^{2b_1}/2)]^2 \right\} P(V_{J(x)}) \\
&\geq K_2 \sigma_n^{-d} \mathrm{e}^{-2a_{\sigma_n}^{\tau}} P(V_{J(x)}) \geq K_3 \sigma_n^{-d} \mathrm{e}^{-2a_{\sigma_n}^{\tau}} \tilde{\epsilon}_n^{4db_1}
\end{aligned}$$

for some constants K_1, K_2, K_3. Also, for every $\|x\| > a_{\sigma_n}$,

$$\begin{aligned}
\frac{f_P(x)}{f_0(x)} &\geq K_1 \int_{S_{\sigma_n}} \int_{\|\mu\| \leq a_{\sigma_n}} \phi_\Sigma(x - \mu) \mathrm{d}P(\mu, \Sigma) \\
&\geq K_2 \sigma_n^{-d} \int_{S_{\sigma_n}} \int_{\|\mu\| \leq a_{\sigma_n}} \exp\left\{ -\frac{1}{2\sigma^2} \|x - \mu\|^2 \right\} \mathrm{d}P(\mu, \sigma) \\
&\geq K_2 \sigma_n^{-d} \exp\{-2\|x\|^2/\sigma_n^2\} P\left(\{\mu \in \mathbb{R} : \|\mu\| \leq a_{\sigma_n}\} \times S_{\sigma_n} \right) \\
&\geq K_2 \sigma_n^{-d} \exp\{-2\|x\|^2/\sigma_n^2\} \sum_{j=1}^N P(V_j) \geq K_4 \sigma_n^{-d} \exp\{-2\|x\|^2/\sigma_n^2\}
\end{aligned}$$

for some constant K_4 because $\|x - \mu\|^2 \leq 2\|x\|^2 + 2\|\mu\|^2 \leq 4\|x\|^2$ (since $\|\mu\| \leq a_{\sigma_n} < \|x\|$) and

$$\begin{aligned}
\sum_{j=1}^N P(V_j) &= 1 - \sum_{j>N}^M P(V_j) = 1 - \sum_{j>N}^M |P(V_j) - p_j| \\
&\geq 1 - \sum_{j=1}^M |P(V_j) - p_j| \geq 1 - 2\tilde{\epsilon}_n^{2db_1} \mathrm{e}^{-C_3 \sigma_n^{-\kappa^*}}
\end{aligned}$$

and the last inequality follows from the definition of $\mathcal{P}_\sigma$. Set $\lambda = K_3 \sigma_n^{-d} \mathrm{e}^{-2a_{\sigma_n}^{\tau}} \tilde{\epsilon}_n^{4db_1}$ and notice that

$$\begin{aligned}
\log(1/\lambda) &= \log(1/\tilde{\epsilon}_n^{4db_1} \sigma_n^{-d}) + \log\left(\exp\{2a_0^2 \log(1/\sigma_n)\} \right) \\
&= \log\left(1/\tilde{\epsilon}_n^{d(4b_1 - 1/\beta)} \{\log(1/\tilde{\epsilon}_n)\}^{d/\beta} \right) + 2a_0^2 \log(1/\sigma_n) \lesssim \log(1/\tilde{\epsilon}_n)
\end{aligned}$$

since $4b_1 > 1/\beta$. Moreover, for any $P \in \mathcal{P}_\sigma$, $\{x \in \mathbb{R} : f_P(x)/f_0(x) < \lambda\} \subset \{x \in \mathbb{R} : \|x\| > a_{\sigma_n}\}$ so that

$$
\begin{aligned}
F_0\left\{\left(\log\frac{f_0}{f_P}\right)^2 \mathbb{1}\left(\frac{f_P}{f_0} < \lambda\right)\right\} &\leq \int_{\|x\|>a_{\sigma_n}} \left(\log\frac{f_P(x)}{f_0(x)}\right)^2 f_0(x)\mathrm{d}x \\
&\leq \int_{\|x\|>a_{\sigma_n}} \left(\log\left[\frac{\sigma_n^d}{K_4}\exp\{-2\|x\|^2/\sigma_n^2\}\right]\right)^2 f_0(x)\mathrm{d}x \\
&\leq \frac{K_5}{\sigma_n^4}\int_{\|x\|>a_{\sigma_n}} \|x\|^4 f_0(x)\mathrm{d}x = \frac{K_5}{\sigma_n^4}\int_{\|x\|>a_{\sigma_n}} \|x\|^4 f_0(x)^{1/2} f_0(x)^{1/2}\mathrm{d}x \\
&\leq \frac{K_5}{\sigma_n^4}\left\{\int_{\|x\|>a_{\sigma_n}} \|x\|^8 f_0(x)\mathrm{d}x \int_{\|x\|>a_{\sigma_n}} f_0(x)\mathrm{d}x\right\}^{1/2} \\
&\leq \frac{K_5}{\sigma_n^4}(F_0\|X\|^8)^{1/2}(F_0\{x : \|x\| > a_{\sigma_n}\})^{1/2} \leq \frac{K_6}{\sigma_n^4}(F_0\{x : \|x\| > a_{\sigma_n}\})^{1/2} \\
&\leq \frac{K_6}{\sigma_n^4}F_0(E_{\sigma_n}^c)^{1/2} \leq \frac{K_6}{\sigma_n^4}B_0\sigma_n^{2\beta+\epsilon+4} \leq K_7\sigma_n^{2\beta+\epsilon}
\end{aligned}
$$

for some constants K_5, K_6, K_7. The forth inequality follows from Cauchy-Schwartz, the sixth inequality follows from $F_0\|X\|^m < \infty$ for all $m > 0$ because of the tail condition (21) on f_0, the seventh and eighth inequalities follows from Proposition 1 of Shen et al. (2013). Since $\log x \leq (\log x)^2$ for $x > \mathrm{e}^1$ and $\lambda < \mathrm{e}^{-1}$ for n sufficiently large, we also have that $F_0\{\log(f_0/f_P)\mathbb{1}(f_P/f_0 < \lambda)\} \leq K_8\sigma_n^{2\beta+\epsilon}$. Now apply Lemma B2 of Shen et al. (2013) to conclude that both $F_0\{\log f_0/f_P\}$ and $F_0\{(\log f_0/f_P)^2\}$ are bounded by

$$K_9\{\log(1/\lambda)\}^2\sigma_n^{2\beta} \leq A\sigma_n^{2\beta}\{\log(1/\tilde{\epsilon}_n)\}^2 = A\tilde{\epsilon}_n^2$$

for some positive constant A. Together with (32), this completes the proof. □

References

Barron, A., Schervish, M. J., and Wasserman, L. (1999), "The Consistency of Posterior Distributions in Nonparametric Problems," *The Annals of Statistics*, 27, pp. 536–561.

Bhattacharya, A. and Dunson, D. B. (2011), "Sparse Bayesian infinite factor models," *Biometrika*, 98, 291–306.

Carvalho, C., Lucas, J., Wand, Q., Nevins, J., and West, M. (2008), "High-dimensional sparse factor modelling: Applications in gene expression genomic," *Journal of Amer. Stat. Association*, 103, 1438–1456.

Chen, M., Silva, J., Paisley, J., Wang, C., Dunson, D. B., and Carin, L. (2010), "Compressive sensing on manifolds using a nonparametric mixture of factor analyzers: algorithm and performance bounds," *IEEE Transaction in Signal Processes*, 58, 6140–6155.

Cron, A. J. and West, M. (2015), "Models of random sparse eigenmatrices & Bayesian analysis of multivariate structure," Tech. rep., Department of Statistical Science, Duke University, submitted for publication.

De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Prünster, I., and Ruggiero, M. (2015), "Are Gibbs-type priors the most natural generalization of the Dirichlet process?" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37, 212–229.

Devroye, L. and Gyorfi, L. (1985), *Nonparametric Density Estimation. The L_1 View*, New York: Wiley & Sons.

Edelman, A. and Sutton, B. D. (2005), "Tails of condition number distributions," *SIAM Journal on Matrix Analysis and Applications*, 27, 547–560.

Escobar, M. D. and West, M. (1995), "Bayesian density estimation and inference using mixtures," *Journal of Amer. Stat. Association*, 90, 577–588.

Ferguson, T. S. (1973), "A Bayesian analysis of some nonparametric problems," *The Annals of Statistics*, 1, 209–230.

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

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

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.

— (2007), "Posterior convergence rates of Dirichlet mixtures at smooth densities," *The Annals of Statistics*, 35, 697–723.

Gnedin, A. and Pitman, J. (2005), "Exchangeable Gibbs partitions and Stirling triangles," *Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI)*, 325, 83–102.

Goldenshluger, A. and Lepski, O. (2014), "On adaptive minimax density estimation on R^d," *Probability Theory and Related Fields*, 159, 479–543.

Gorur, D. and Rasmussen, C. (2009), "Nonparametric mixtures of factor analyzers," in *Signal Processing and Communications Applications Conference, 2009. SIU 2009. IEEE 17th*, pp. 708–711.

Kruijer, W., Rousseau, J., and van der Vaart, A. W. (2010), "Adaptive Bayesian density estimation with location-scale mixtures," *Electronic Journal of Statistics*, 4, 1225–1257.

Lijoi, A., Mena, R. H., and Prünster, I. (2005a), "Hierarchical mixture modelling with normalized inverse Gaussian priors," *Journal of Amer. Stat. Association*, 100, 1278–1291.

Lijoi, A., Prünster, I., and Walker, S. G. (2005b), "On consistency of nonparametric normal mixtures for Bayesian density estimation," *Journal of Amer. Stat. Association*, 100, 1292–1296.

Lo, A. Y. (1984), "On a class of Bayesian nonparametric estimates: I. Density estimates," *The Annals of Statistics*, 12, 351–357.

MacEachern, S. N. and Müller, P. (1998), "Estimating mixture of Dirichlet process models," *Journal of Computational and Graphical Statistics*, 7, 223–238.

Matthaiou, M., McKay, M. R., Smith, P. J., and Nossek, J. A. (2010), "On the condition number distribution of complex Wishart matrices," *IEEE Transactions on Communications*, 58, 1705–1717.

Muirhead, R. J. (1982), *Aspects of Multivariate Statistical Theory*, Hoboken: Wiley & Sons.

Müller, P., Erkanli, A., and West, M. (1996), "Bayesian curve fitting using multivariate normal mixtures," *Biometrika*, 83, 67–79.

Perman, M., Pitman, J., and Yor, M. (1992), "Size–biased sampling of Poisson point processes and excursions," *Probab. Theory Related Fields*, 92, 21–39.

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

Tokdar, S. T. (2006), "Posterior consistency of Dirichlet location-scale mixture of normals in density estimation and regression," *Sankhya*, 68, 90–110.

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.

West, M. (2003), "Bayesian Factor Regression Models in the "Large p, Small n" Paradigm," in *Bayesian Statistics*, Oxford University Press, pp. 723–732.

Wu, Y. and Ghosal, S. (2008), "Kullback Leibler property of kernel mixture priors in Bayesian density estimation," *Electronic Journal of Statistics*, 2, 298–331.

— (2010), "The L_1-consistency of Dirichlet mixtures in multivariate Bayesian density estimation," *Journal of Multivariate Analysis*, 101, 2411–2419.