Biometrika (2013), *pp.* 1–18
doi: 10.1093/biomet/ast015

Printed in Great Britain

Adaptive Bayesian multivariate density estimation with Dirichlet mixtures

BY WEINING SHEN

Department of Statistics, North Carolina State University, 5109 SAS Hall, 2311 Stinson Drive, Raleigh, North Carolina 27695, U.S.A.
wshen2@ncsu.edu

SURYA T. TOKDAR

Department of Statistical Science, Duke University, 219A Old Chemistry Building, Box 90251, Durham, North Carolina 27708, U.S.A.
tokdar@stat.duke.edu

AND SUBHASHIS GHOSAL

Department of Statistics, North Carolina State University, 5109 SAS Hall, 2311 Stinson Drive, Raleigh, North Carolina 27695, U.S.A.
sghosal@ncsu.edu

SUMMARY

We show that rate-adaptive multivariate density estimation can be performed using Bayesian methods based on Dirichlet mixtures of normal kernels with a prior distribution on the kernel's covariance matrix parameter. We derive sufficient conditions on the prior specification that guarantee convergence to a true density at a rate that is minimax optimal for the smoothness class to which the true density belongs. No prior knowledge of smoothness is assumed. The sufficient conditions are shown to hold for the Dirichlet location mixture-of-normals prior with a Gaussian base measure and an inverse Wishart prior on the covariance matrix parameter. Locally Hölder smoothness classes and their anisotropic extensions are considered. Our study involves several technical novelties, including sharp approximation of finitely differentiable multivariate densities by normal mixtures and a new sieve on the space of such densities.

Some key words: Anisotropy; Dirichlet mixture; Multivariate density estimation; Nonparametric Bayesian method; Rate adaptation.

1. INTRODUCTION

Asymptotic frequentist properties of Bayesian nonparametric methods have received much attention recently. It is now recognized that a single fully Bayesian method can offer adaptive optimal rates of convergence for large collections of true data-generating distributions ranging over several smoothness classes. Examples include: signal estimation in the presence of Gaussian white noise (Belitser & Ghosal, 2003); density estimation and regression based on a mixture model of spline or wavelet bases (Huang, 2004; Ghosal et al., 2008); regression, classification and density estimation based on a rescaled Gaussian process model (van der Vaart & van Zanten, 2009); density estimation based on a hierarchical finite mixture

model of beta densities (Rousseau, 2010); and density estimation (Kruijer et al., 2010) and regression (de Jonge & van Zanten, 2010) based on hierarchical, finite mixture models of location-scale kernels.

Results on adaptive convergence rates for nonparametric Bayesian methods are useful for at least two reasons. First, they provide frequentist justification of these methods in large samples, which can be attractive to non-Bayesian practitioners who use these methods because they are easy to implement, provide estimation and prediction intervals, do not require the adjustment of tuning parameters, and can handle multivariate data. Second, these results supply indirect validation that the spread of the underlying prior distribution is well balanced across its infinite-dimensional support. Such a prior distribution quantifies the rate at which it packs mass into a sequence of shrinking neighbourhoods around any given point in its support. When the support of the prior can be partitioned into smoothness classes in the space of continuous functions, a sharp bound on this rate can be calculated for all support points within each smoothness class. These calculations have a nearly one-to-one relationship with the asymptotic convergence rates of the resulting method.

In this article we focus on a collection of nonparametric Bayesian density estimation methods based on Dirichlet process mixture-of-normals priors. Dirichlet process mixture priors (Ferguson, 1983; Lo, 1984) form a cornerstone of nonparametric Bayesian methodology (Escobar & West, 1995; Müller et al., 1996; Müller & Quintana, 2004; Dunson, 2010), and density estimation methods based on these priors are among the first Bayesian nonparametric methods for which convergence results were obtained (Ghosal et al., 1999; Ghosal & van der Vaart, 2001; Tokdar, 2006). However, because of two major technical difficulties, rate adaptation results have not been available so far and convergence rates remain unknown beyond univariate density estimation (Ghosal & van der Vaart, 2001, 2007). The first major difficulty lies in showing adaptive prior concentration rates for mixture priors on density functions. Taylor expansions do not suffice because of the nonnegativity constraint on the densities. The second major difficulty is to construct a suitable low-entropy, high-mass sieve on the space of infinite-component mixture densities. Such sieve constructions form an integral part of the current technical machinery for deriving rates of convergence. The sieves that have been used to study Dirichlet process mixture models, e.g., in Ghosal & van der Vaart (2007), do not scale to higher dimensions and lack the ability to adapt to smoothness classes (Wu & Ghosal, 2010).

We plug these two gaps and establish rate adaptation properties of a collection of multivariate density estimation methods based on Dirichlet process mixture-of-normals priors. Our priors include the commonly used specification of mixing over multivariate normal kernels with a location parameter drawn from a Dirichlet process having a Gaussian base measure, while using an inverse Wishart prior on the common covariance matrix parameter of the kernels. Rate adaptation is established with respect to Hölder smoothness classes. In particular, when any density estimation method from our collection is applied to independent observations $X_1, \ldots, X_n \in \mathbb{R}^d$ drawn from a density f_0 which belongs to the smoothness class of locally β-Hölder functions, it is shown to produce a posterior distribution on the unknown density of the X_i that converges to f_0 at a rate of $n^{-\beta/(2\beta+d)}(\log n)^t$, where t depends on β, d and tail properties of f_0. This rate, without the $(\log n)^t$ term, is minimax optimal for the β-Hölder class (Barron et al., 1999). It is further shown that if f_0 is anisotropic with Hölder smoothness coefficients $\beta_1, \ldots, \beta_d$ along the d axes, then the posterior convergence rate is $n^{-\beta_0/(2\beta_0+d)}$ times a factor $\log n$, where β_0 is the harmonic mean of $\beta_1, \ldots, \beta_d$. Again, this rate is minimax optimal for this class of functions (Hoffmann & Lepski, 2002).

To the best of our knowledge, such rate adaptation results are new for any kernel-based multivariate density estimation method. The performance of a non-Bayesian multivariate kernel

density estimator depends heavily on the difficult choice of a bandwidth and a smoothing kernel (Scott, 1992). Optimal rates are possible only by using higher-order kernels and choices of bandwidth that require knowing the smoothness level. In contrast, our results show that a single Bayesian nonparametric method based on a single choice of Dirichlet process mixture of normal kernels achieves optimal convergence rates universally across all smoothness levels.

2. Posterior convergence rates for Dirichlet mixtures

2·1. *Notation*

For any $d \times d$ positive definite real matrix Σ, let $\phi_\Sigma(x)$ denote the d-variate normal density $(2\pi)^{-d/2}(\det \Sigma)^{-1/2}\exp(-x^{\mathrm{T}}\Sigma^{-1}x/2)$ with mean zero and covariance matrix Σ. For a probability measure F on $\mathbb{R}^d$ and a $d \times d$ positive definite real matrix Σ, the F-induced location mixture of ϕ_Σ is denoted by $p_{F,\Sigma}$; that is, $p_{F,\Sigma}(x) = \int \phi_\Sigma(x - z)F(\mathrm{d}z)$ for $x \in \mathbb{R}^d$. For a scalar $\sigma > 0$ and any function f on $\mathbb{R}^d$, we let $K_\sigma f$ denote the convolution of f and $\phi_{\sigma^2 I}$, i.e., $(K_\sigma f)(x) = \int \phi_{\sigma^2 I}(x - z)f(z)\,\mathrm{d}z$.

For any finite positive measure α on $\mathbb{R}^d$, let $\mathcal{D}_\alpha$ denote the Dirichlet process distribution with parameter α (Ferguson, 1973); that is, an $F \sim \mathcal{D}_\alpha$ is a random probability measure on $\mathbb{R}^d$ such that for any Borel-measurable partition $B_1, \ldots, B_k$ of $\mathbb{R}^d$, the joint distribution of $F(B_1), \ldots, F(B_k)$ is the k-variate Dirichlet distribution with parameters $\alpha(B_1), \ldots, \alpha(B_k)$.

Let $\mathbb{N}_0 = \{0, 1, 2, \ldots\}$, and let $\Delta_J = \{(x_1, \ldots, x_J) : x_i > 0,\ i = 1, \ldots, J;\ \sum_{i=1}^J x_i = 1\}$ be the J-dimensional probability simplex. Let the indicator function of a set A be denoted by $\mathbb{1}(A)$. We write $\lesssim$ to mean an inequality up to a constant multiple, where the underlying constant of proportionality is universal or is unimportant for our purposes. For any $x \in \mathbb{R}$, define $\lfloor x \rfloor$ to be the largest integer that is strictly smaller than x. Similarly, define $\lceil x \rceil$ to be the smallest integer strictly greater than x. For a multi-index $k = (k_1, \ldots, k_d) \in \mathbb{N}_0^d$, define $k_\cdot = k_1 + \cdots + k_d$ and $k! = k_1! \cdots k_d!$, and let D^k denote the mixed partial derivative operator $\partial^{k_\cdot}/\partial x_1^{k_1} \cdots \partial x_d^{k_d}$.

For any $\beta > 0$, $\tau_0 \geqslant 0$ and nonnegative function L on $\mathbb{R}^d$, define the locally β-Hölder class with envelope L, denoted by $\mathcal{C}^{\beta,L,\tau_0}(\mathbb{R}^d)$, to be the set of all functions $f : \mathbb{R}^d \to \mathbb{R}$ that have finite mixed partial derivatives $D^k f$ of all orders up to $k_\cdot \leqslant \lfloor \beta \rfloor$, such that for every $k \in \mathbb{N}_0^d$ with $k_\cdot = \lfloor \beta \rfloor$,

$$|(D^k f)(x + y) - (D^k f)(x)| \leqslant L(x)\exp(\tau_0\|y\|^2)\|y\|^{\beta - \lfloor\beta\rfloor} \quad (x, y \in \mathbb{R}^d).$$

In our discussion, we shall assume that the true density f lies in $\mathcal{C}^{\beta,L,\tau_0}(\mathbb{R}^d)$. This condition is essentially weaker than the one in Kruijer et al. (2010), where $\log f \in \mathcal{C}^{\beta,L,0}(\mathbb{R})$ is assumed; see Lemma B4.

For any $d \times d$ matrix A, we denote its eigenvalues by $\mathrm{eig}_1(A) \leqslant \cdots \leqslant \mathrm{eig}_d(A)$, its spectral norm by $\|A\|_2 = \sup_{x \neq 0} \|Ax\|/\|x\|$ and its max norm by $\|A\|_{\max}$, the maximum of the absolute values of the elements of A.

2·2. *Dirichlet process mixture-of-normals prior*

Consider drawing inference on an unknown probability density function f on $\mathbb{R}^d$ based on independent observations $X_1, \ldots, X_n$ from f. A nonparametric Bayesian method assigns a prior distribution Π on f and draws inference on f based on the posterior distribution $\Pi_n(\cdot \mid X_1, \ldots, X_n)$. A Dirichlet process location mixture-of-normals prior Π is the distribution of a random probability density function $p_{F,\Sigma}$ where $F \sim \mathcal{D}_\alpha$ for some finite positive measure α on $\mathbb{R}^d$ and $\Sigma \sim G$, a probability distribution on $d \times d$ positive definite real matrices.

We restrict our discussion to a collection of such prior distributions Π for which the associated $\mathcal{D}_\alpha$ and G satisfy the following conditions. Let $|\alpha| = \alpha(\mathbb{R}^d)$ and $\bar{\alpha} = \alpha/|\alpha|$. We assume that $\bar{\alpha}$ has a positive density function on the whole of $\mathbb{R}^d$ and that there exist positive constants $a_1, a_2, a_3, b_1, b_2, b_3, C_1, C_2$ such that

$$1 - \bar{\alpha}([-x, x]^d) \leqslant b_1 \exp(-C_1 x^{a_1}) \quad \text{for all sufficiently large } x > 0, \tag{1}$$

$$G\{\Sigma : \mathrm{eig}_d(\Sigma^{-1}) \geqslant x\} \leqslant b_2 \exp(-C_2 x^{a_2}) \quad \text{for all sufficiently large } x > 0, \tag{2}$$

$$G\{\Sigma : \mathrm{eig}_1(\Sigma^{-1}) < x\} \leqslant b_3 x^{a_3} \quad \text{for all sufficiently small } x > 0. \tag{3}$$

We also assume that there exist $\kappa, a_4, a_5, b_4, C_3 > 0$ such that for any $0 < s_1 \leqslant \cdots \leqslant s_d$ and $t \in (0, 1)$,

$$G\{\Sigma : s_j < \mathrm{eig}_j(\Sigma^{-1}) < s_j(1 + t),\ j = 1, \ldots, d\} \geqslant b_4 s_1^{a_4} t^{a_5} \exp(-C_3 s_d^{\kappa/2}). \tag{4}$$

Our assumption on $\bar{\alpha}$ is analogous to (11) of Kruijer et al. (2010) and holds, for example, when $\bar{\alpha}$ is a Gaussian measure on $\mathbb{R}^d$. Unlike previous treatments of Dirichlet process mixture models (Ghosal & van der Vaart, 2001, 2007), we allow a full-support prior on Σ, including the widely used inverse Wishart distribution. The following lemma shows that such a G satisfies our assumptions; see Appendix A for a proof.

LEMMA 1. *The inverse Wishart distribution* $\mathrm{IW}(\nu, \Psi)$ *with* ν *degrees of freedom and a positive definite scale matrix* Ψ *satisfies* (2), (3) *and* (4) *with* $\kappa = 2$.

From a computational point of view, another useful specification is to consider a G that supports only diagonal covariance matrices $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$, with each diagonal component independently assigned a prior distribution G_0. By choosing an inverse gamma distribution as G_0, we get a G that again satisfies (2), (3) and (4) with $\kappa = 2$. Alternatively, we could take G_0 to be the distribution of the square of an inverse gamma random variable. Such a G_0 leads to a G that satisfies (2), (3) and (4) with $\kappa = 1$. This difference in κ matters, with smaller κ leading to optimal convergence rates for a wider class of true densities.

2·3. *Convergence rates results*

Let Π be a Dirichlet process mixture prior as defined in § 2·2, and let $\Pi_n(\cdot \mid X_1, \ldots, X_n)$ denote the posterior distribution based on n observations $X_1, \ldots, X_n$ modelled as $X_i \sim f$, $f \sim \Pi$. Let $\{\epsilon_n\}_{n \geqslant 1}$ be a sequence of positive numbers with $\lim_{n \to \infty} \epsilon_n = 0$. Also, let ρ denote a suitable metric on the space of probability densities on $\mathbb{R}^d$, such as the L_1 metric $\|f - g\|_1 = \int |f(x) - g(x)| \, \mathrm{d}x$, or the Hellinger metric $d_{\mathrm{H}}(f, g) = [\int \{f^{1/2}(x) - g^{1/2}(x)\}^2 \, \mathrm{d}x]^{1/2}$. Fix any probability density f_0 on $\mathbb{R}^d$. For the density estimation method based on Π, we say that its posterior convergence rate at f_0 in the metric ρ is ϵ_n if for any $M < \infty$,

$$\lim_{n \to 0} \Pi_n [\{f : \rho(f_0, f) > M\epsilon_n\} \mid X_1, \ldots, X_n] = 0 \text{ almost surely}, \tag{5}$$

whenever $X_1, X_2, \ldots$ are independent and identically distributed with density f_0.

Although (5) only establishes $(\epsilon_n)_{n \geqslant 1}$ as a bound on the convergence rate at f_0, it serves as a useful calibration when checked against the optimal rate for the smoothness class to which f_0 belongs. It is known that the minimax rate associated with a β-Hölder class is $n^{-\beta/(2\beta+d)}$. We establish (5) for this class with ϵ_n as $n^{-\beta/(2\beta+d)}$, up to a factor that is a power of $\log n$. A formal result requires some additional conditions on f_0, as summarized in Theorem 1.

THEOREM 1. *Suppose that* $f_0 \in \mathcal{C}^{\beta, L, \tau_0}(\mathbb{R}^d)$ *is a probability density function satisfying*

$$P_0\big(|D^k f_0|/f_0\big)^{(2\beta+\epsilon)/k.} < \infty \quad (k \in \mathbb{N}_0^d,\ k. \leqslant \lfloor\beta\rfloor), \qquad P_0\big(L/f_0\big)^{(2\beta+\epsilon)/\beta} < \infty \tag{6}$$

for some $\epsilon > 0$, *where* $P_0 g = \int g(x) f(x)\,\mathrm{d}x$ *denotes the expectation of* $g(X)$ *under* $X \sim f_0$. *Further, suppose that there exist positive constants* a, b, c *and* τ *such that*

$$f_0(x) \leqslant c \exp(-b\|x\|^\tau) \quad (\|x\| > a). \tag{7}$$

For the prior Π *constructed in* § 2·2, (5) *holds in the Hellinger or the* L_1 *metric with* $\epsilon_n = n^{-\beta/(2\beta+d^*)}(\log n)^t$, *where* $t > \{d^*(1 + 1/\tau + 1/\beta) + 1\}/(2 + d^*/\beta)$ *and* $d^* = \max(d, \kappa)$.

We prove this result by verifying a set of sufficient conditions presented originally in Ghosal et al. (2000) and subsequently modified by Ghosal & van der Vaart (2007). For $\epsilon > 0$ and any subset A of a metric space equipped with a metric ρ, let $N(\epsilon, A, \rho)$ denote the ϵ-covering number of A, i.e., $N(\epsilon, A, \rho)$ is the smallest number of balls of radius ϵ needed to cover A. The logarithm of this number is referred to as the ϵ-entropy of A. Also, define $\mathcal{K}(f_0, \epsilon) = \{f : \int f_0 \log(f_0/f) < \epsilon^2,\ \int f_0 \log^2(f_0/f) < \epsilon^2\}$, the Kullback–Leibler ball around f_0 of size ϵ. Ghosal & van der Vaart (2007) showed that (5) holds whenever there exist positive constants c_1, c_2, c_3 and c_4, a sequence of positive numbers $(\tilde{\epsilon}_n)_{n\geqslant 1}$ with $\tilde{\epsilon}_n \leqslant \epsilon_n$ and $\lim_{n\to\infty} n\tilde{\epsilon}_n^2 = \infty$, and a sequence of compact subsets $(\mathcal{F}_n)_{n\geqslant 1}$ of probability densities such that

$$\log N(\epsilon_n, \mathcal{F}_n, \rho) \leqslant c_1 n\epsilon_n^2, \tag{8}$$

$$\Pi(\mathcal{F}_n^{\mathrm{c}}) \leqslant c_3 \exp\{-(c_2 + 4)n\tilde{\epsilon}_n^2\}, \tag{9}$$

$$\Pi\{\mathcal{K}(f_0, \tilde{\epsilon}_n)\} \geqslant c_4 \exp(-c_2 n\tilde{\epsilon}_n^2). \tag{10}$$

The sequence of sets $\mathcal{F}_n$ is often called a sieve, and the Kullback–Leibler ball probability in (10) is called the prior thickness at f_0. In Theorem 4 we show that (10) holds for $\Pi = \mathcal{D}_\alpha \times G$ with $\tilde{\epsilon}_n = n^{-\beta/(2\beta+d^*)}(\log n)^{t_0}$, where $t_0 = \{d^*(1 + 1/\tau + 1/\beta) + 1\}/(2 + d^*/\beta)$. In Theorem 5 we show that (8) and (9) hold with $\tilde{\epsilon}_n$ as before and $\epsilon_n = n^{-\beta/(2\beta+d^*)}(\log n)^t$ for every $t > t_0$. The following sections lay out the machinery needed to establish these two fundamental results.

When $\kappa = 1$, the rate in Theorem 1 equals the optimal rate $n^{-\beta/(2\beta+d)}$ up to a factor of $\log n$. However, the commonly used inverse Wishart specification of G leads to $\kappa = 2$, and hence Theorem 1 gives the optimal rate only for $d \geqslant 2$. We will see later that κ has a bigger impact on rates of convergence for anisotropic densities.

Our result also applies to a finite mixture prior specification Π where the density function f is represented by $f(x) = \sum_{h=1}^H \omega_h \phi_\Sigma(x - \mu_h)$ and priors are assigned on $H, \Sigma, \omega = (\omega_1, \ldots, \omega_H)$ and $\mu_1, \ldots, \mu_H$. We assume $\Sigma \sim G$, which satisfies (2), (3) and (4), and that there exist positive constants $a_4, b_4, b_5, b_6, b_7, C_4, C_5, C_6, C_7$ such that $b_4 \exp\{-C_4 x(\log x)^{\tau_1}\} \leqslant \Pi(H \geqslant x) \leqslant b_5 \exp\{-C_5 x(\log x)^{\tau_1}\}$ for sufficiently large $x > 0$, while for every fixed $H = h$,

$$\Pi(\mu_i \notin [-x, x]^d) \leqslant b_6 \exp(-C_6 x^{a_4}) \text{ for sufficiently large } x > 0 \quad (i = 1, \ldots, h),$$

$$\Pi(\|\omega - \omega_0\| \leqslant \epsilon) \geqslant b_7 \exp\{-C_7 h \log(1/\epsilon)\} \text{ for all } 0 < \epsilon < 1/h \text{ and all } \omega_0 \in \Delta_h.$$

Theorem 2 summarizes our findings for a finite mixture prior. Its proof is similar to that of Theorem 1 except that in verifying (9) we need $\exp\{-H(\log H)^{\tau_1}\} \lesssim \exp\{-n\tilde{\epsilon}_n^2\}$. Together with $H = \lfloor n\epsilon_n^2/(\log n)\rfloor$, we have $\epsilon_n^2(\log n)^{\tau_1 - 1} \geqslant \tilde{\epsilon}_n^2$, leading to $\tilde{\epsilon}_n = n^{-\beta/(2\beta+d^*)}(\log n)^{t_0}$ where

$t_0 = \{d^*(1 + 1/\tau + 1/\beta) + 1\}/(2 + d^*/\beta)$ and $\epsilon_n = n^{-\beta/(2\beta+d^*)}(\log n)^t$ with $t > t_0 + \max\{0, (1 - \tau_1)/2\}$.

THEOREM 2. *Suppose that $f_0 \in \mathcal{C}^{\beta,L,\tau_0}(\mathbb{R}^d)$ is a probability density function satisfying (6) and (7) for some positive constants a, b, c, τ and ϵ. For a finite mixture prior Π as above, (5) holds in the Hellinger or the L_1 metric with $\epsilon_n = n^{-\beta/(2\beta+d^*)}(\log n)^t$ for every $t > \{d^*(1 + 1/\tau + 1/\beta) + 1\}/(2 + d^*/\beta) + \max\{0, (1 - \tau_1)/2\}$, where $d^* = \max(d, \kappa)$.*

3. PRIOR THICKNESS RESULTS

Functions in $\mathcal{C}^{\beta,L,\tau_0}$ can be approximated by mixtures of $\phi_{\sigma^2 I}$ with an accuracy that improves with β. We establish this through the following constructions and lemma, which are adapted from Lemma 3.4 of de Jonge & van Zanten (2010) and univariate approximation results of Kruijer et al. (2010). The proofs are given in Appendix A.

For each $k \in \mathbb{N}_0^d$, let m_k denote the kth moment $m_k = \int y^k \phi_1(y)\,dy$ of the standard normal distribution on $\mathbb{R}^d$. For $n \in \mathbb{N}_0^d$, define two sequences of numbers by the following recursion. If $n_\cdot = 1$ set $c_n = 0$ and $d_n = -m_n/n!$, and for $n_\cdot \geqslant 2$ define

$$c_n = -\sum_{\substack{n=l+k \\ l_\cdot \geqslant 1,\, k_\cdot \geqslant 1}} \frac{(-1)^{k_\cdot}}{k!} m_k d_l, \quad d_n = \frac{(-1)^{n_\cdot} m_n}{n!} + c_n. \tag{11}$$

Given $\beta > 0$ and $\sigma > 0$, define a transform $T_{\beta,\sigma}$ on $f : \mathbb{R}^d \to \mathbb{R}$ with derivatives up to order $\lfloor\beta\rfloor$ by

$$T_{\beta,\sigma} f = f - \sum_{\substack{k \in \mathbb{N}_0^d \\ 1 \leqslant k_\cdot \leqslant \lfloor\beta\rfloor}} d_k \sigma^{k_\cdot} D^k f.$$

LEMMA 2. *For any $\beta, \tau_0 > 0$, there is a positive constant M_β such that any $f \in \mathcal{C}^{\beta,L,\tau_0}(\mathbb{R}^d)$ satisfies $|\{K_\sigma(T_{\beta,\sigma} f) - f\}(x)| < M_\beta L(x)\sigma^\beta$ for all $x \in \mathbb{R}^d$ and all $\sigma \in (0, 1/(2\tau_0)^{1/2})$.*

Lemma 2 applies to any function $f \in \mathcal{C}^{\beta,L,\tau_0}$, not necessarily a probability density, and the mixing function $T_{\beta,\sigma} f$ need not be a density and could be negative. Fortunately, when f is a probability density, we can derive a density h_σ from $T_{\beta,\sigma} f$ so that $K_\sigma h_\sigma$ provides an order-σ^β approximation to f. The construction of h_σ can be viewed as a multivariate extension of results in Kruijer et al. (2010, § 3). The main difference is that we establish approximation results under the Hellinger distance and employ Taylor expansions on f_0 instead of $\log f_0$, which lead to a more elegant proof.

THEOREM 3. *Let $f_0 \in \mathcal{C}^{\beta,L,\tau_0}(\mathbb{R}^d)$ be a probability density function and write $f_\sigma = T_{\beta,\sigma} f_0$. Suppose that f_0 satisfies (6) for some $\epsilon > 0$. Then there exist $s_0 > 0$ and $K > 0$ such that for any $0 < \sigma < s_0$, $g_\sigma = f_\sigma + (1/2) f_0\, 1\!\!1\{f_\sigma < (1/2) f_0\}$ is a nonnegative function with $\int g_\sigma(x)\,dx < \infty$ and the density $h_\sigma = g_\sigma / \int g_\sigma(x)\,dx$ satisfies $d_{\mathrm{H}}^2(f_0, K_\sigma h_\sigma) \leqslant K\sigma^{2\beta}$.*

The next result trades g_σ for a compactly supported density h_σ whose convolution with $\phi_{\sigma^2 I}$ inherits the same order-σ^β approximation to f_0. We need the tail condition (7) on f_0 to obtain a suitable compact support.

PROPOSITION 1. *Let $f_0 \in \mathcal{C}^{\beta,L,\tau_0}(\mathbb{R}^d)$ be a probability density function satisfying (6) and (7) for some positive constants ϵ, a, b, c and τ. For any $\sigma > 0$, define $E_\sigma = \{x \in \mathbb{R}^d : f_0(x) \geqslant \sigma^{(4\beta+2\epsilon+8)/\delta}\}$. Then there exist $s_0, a_0, B_0, K_0 > 0$ such that for every $0 < \sigma < s_0$, $P_0(E_\sigma^c) \leqslant B_0\sigma^{4\beta+2\epsilon+8}$, $E_\sigma \subset \{x \in \mathbb{R}^d : \|x\| \leqslant a_\sigma\}$ where $a_\sigma = a_0\{\log(1/\sigma)\}^\tau$, and there is a probability density $\tilde{h}_\sigma$ with support inside $\{x \in \mathbb{R}^d : \|x\| \leqslant a_\sigma\}$ satisfying $d_{\mathrm{H}}(f_0, K_\sigma\tilde{h}_\sigma) \leqslant K_0\sigma^\beta$.*

Proposition 1 paves the way to calculating prior thickness around f_0, because the probability density $K_\sigma\tilde{h}_\sigma$ can be well approximated by densities $p_{F,\Sigma}$ with (F, Σ) chosen from a suitable set. Towards this, we present the final theorem of this section and a proof of it that overlaps with § 9 of Ghosal & van der Vaart (2007). However, our proof requires new calculations to handle a non-compactly supported f_0 and a matrix-valued Σ.

THEOREM 4. *Let $f_0 \in \mathcal{C}^{\beta,L,\tau_0}(\mathbb{R}^d)$ be a bounded probability density function satisfying (6) and (7) for some positive constants ϵ, a, b, c and τ. Then, for some $A, C > 0$ and all sufficiently large n,*

$$(\mathcal{D}_\alpha \times G)\left\{(F, \Sigma) : P_0 \log\frac{f_0}{p_{F,\Sigma}} \leqslant A\tilde{\epsilon}_n^2,\ P_0\left(\log\frac{f_0}{p_{F,\Sigma}}\right)^2 \leqslant A\tilde{\epsilon}_n^2\right\} \geqslant \exp(-Cn\tilde{\epsilon}_n^2) \qquad (12)$$

where $\tilde{\epsilon}_n = n^{-\beta/(2\beta+d^*)}(\log n)^t$ with any $t \geqslant \{d^*(1 + 1/\tau + 1/\beta) + 1\}/(2 + d^*/\beta)$.

Proof. Let δ, s_0, a_0 and K_0 be as in Proposition 1. Take n large enough so that $\tilde{\epsilon}_n < s_0^\beta$. Fix $\sigma^\beta = \tilde{\epsilon}_n\{\log(1/\tilde{\epsilon}_n)\}^{-1}$ and, as in Proposition 1, define $E_\sigma = \{x \in \mathbb{R}^d : f_0(x) \geqslant \sigma^{(4\beta+2\epsilon+8)/\delta}\}$ and $a_\sigma = a_0\{\log(1/\sigma)\}^{1/\tau}$. Recall that $P_0(E_\sigma^c) \leqslant B_0\sigma^{4\beta+2\epsilon+8}$ for some constant B_0 and that $E_\sigma \subset \{x \in \mathbb{R}^d : \|x\| \leqslant a_\sigma\}$. Apply Proposition 1 to find $\tilde{h}_\sigma$ with support E_σ such that $d_{\mathrm{H}}(f_0, K_\sigma\tilde{h}_\sigma) \leqslant K_0\sigma^\beta$. Find $b_1 > \max\{1, 1/(2\beta)\}$ such that $\tilde{\epsilon}_n^{b_1}\{\log(1/\tilde{\epsilon}_n)\}^{5/4} \leqslant \tilde{\epsilon}_n$.

By Corollary B1, there is a discrete probability measure $F_\sigma = \sum_{j=1}^N p_j\delta_{z_j}$ with at most $N \leqslant D_0\sigma^{-d}\{\log(1/\sigma)\}^{d/\tau}\{\log(1/\tilde{\epsilon}_n)\}^d \leqslant D_1\sigma^{-d}\{\log(1/\tilde{\epsilon}_n)\}^{d+d/\tau}$ support points inside $\{x \in \mathbb{R}^d : \|x\| \leqslant a_\sigma\}$, and with at least $\sigma\tilde{\epsilon}_n^{2b_1}$ separation between any $z_i \neq z_j$ such that $d_{\mathrm{H}}(K_\sigma\tilde{h}_\sigma, K_\sigma F_\sigma) \leqslant A_1\tilde{\epsilon}_n^{b_1}\{\log(1/\tilde{\epsilon}_n)\}^{1/4}$ for some constants A_1 and D_1.

Place disjoint balls U_j centred at $z_1, \ldots, z_N$ with diameter $\sigma\tilde{\epsilon}_n^{2b_1}$ each. Extend $\{U_1, \ldots, U_N\}$ to a partition $\{U_1, \ldots, U_K\}$ of $\{x \in \mathbb{R}^d : \|x\| \leqslant a_\sigma\}$ such that each U_j $(j = N + 1, \ldots, K)$ has a diameter of at most σ. This can be done with $K \leqslant D_2\sigma^{-d}\{\log(1/\tilde{\epsilon}_n)\}^{d+d/\tau}$ for some constant D_2. Further extend this to a partition $U_1, \ldots, U_M$ of $\mathbb{R}^d$ such that $a_1(\sigma\tilde{\epsilon}_n^{2b_1})^d \leqslant \alpha(U_j) \leqslant 1$ for all $j = 1, \ldots, M$, for some constant a_1. We can still have $M \leqslant D_3\sigma^{-d}\{\log(1/\tilde{\epsilon}_n)\}^{d+d/\tau} \leqslant D_4\tilde{\epsilon}_n^{-d/\beta}\{\log(1/\tilde{\epsilon}_n)\}^{sd}$ with $s = 1 + 1/\beta + 1/\tau$, for some constants D_3 and D_4. None of these constants depends on n or σ.

Define $p_j = 0$ $(j = N + 1, \ldots, M)$. Let $\mathcal{P}_\sigma$ denote the set of probability measures F on $\mathbb{R}^d$ with $\sum_{j=1}^M |F(U_j) - p_j| \leqslant 2\tilde{\epsilon}_n^{2db_1}$ and $\min_{1\leqslant j\leqslant M} F(U_j) \geqslant \tilde{\epsilon}_n^{4db_1}/2$. Observe that

$$M\tilde{\epsilon}_n^{2db_1} \leqslant D_4\left[\tilde{\epsilon}_n^{b_1-1/(2\beta)}\{\log(1/\tilde{\epsilon}_n)\}^{s/2}\right]^{2d} \leqslant 1,$$

$$\min_{1\leqslant j\leqslant M} \alpha(U_j)^{1/2} \geqslant a_1^{1/2}\tilde{\epsilon}_n^{2db_1}\{\tilde{\epsilon}_n^{b_1-1/(2\beta)}\log(1/\tilde{\epsilon}_n)\}^{-d} \geqslant (a_1/D_4)^{1/2}\,\tilde{\epsilon}_n^{2db_1},$$

provided n has been chosen large enough. By Lemma 10 of Ghosal & van der Vaart (2007), $\mathcal{D}_\alpha(\mathcal{P}_\sigma) \geqslant C_1\exp\{-c_1M\log(1/\tilde{\epsilon}_n)\} \geqslant C_1\exp[-c_2\tilde{\epsilon}_n^{-d/\beta}\{\log(1/\tilde{\epsilon}_n)\}^{sd+1}]$ for some constants C_1 and c_2 that depend on $\alpha(\mathbb{R}^d)$, a_1, D_4, d and b_1. Also, let $\mathcal{S}_\sigma$ denote the set of all $d \times d$ nonsingular matrices Σ such that all eigenvalues of Σ^{-1} lie between σ^{-2} and $\sigma^{-2}(1 + \sigma^{2\beta})$.

By (4), $G(\mathcal{S}_\sigma) \geqslant \sigma^{D_5} \exp(-D_6/\sigma^\kappa) \geqslant C_3 \exp[-c_3 \tilde{\epsilon}_n^{-\kappa/\beta} \{\log(1/\tilde{\epsilon}_n)\}^{s\kappa+1}]$ for some constants C_3 and c_3. Any $\Sigma \in \mathcal{S}_\sigma$ satisfies $\det(\Sigma^{-1}) \geqslant \sigma^{-2d}$, $y^\mathrm{T}\Sigma^{-1}y \leqslant 2\|y\|^2/\sigma^2$ for any $y \in \mathbb{R}^d$ and $|\mathrm{tr}(\sigma^2\Sigma^{-1}) - d - \log\det(\sigma^2\Sigma^{-1})| < d\sigma^{2\beta}$.

Apply Lemma B1 with $V_i = U_i$ $(i = 1, \ldots, N)$ and $V_0 = \bigcup_{j>N} U_j$ to conclude that for any $F \in \mathcal{P}_\sigma$, $d_\mathrm{H}(K_\sigma F_\sigma, K_\sigma F) \leqslant A_2 \tilde{\epsilon}_n^{b_1}$ for some universal constant A_2, and hence

$$\begin{aligned} d_\mathrm{H}(f_0, K_\sigma F) &\leqslant d_\mathrm{H}(f_0, K_\sigma \tilde{h}_\sigma) + d_\mathrm{H}(K_\sigma \tilde{h}_\sigma, K_\sigma F_\sigma) + d_\mathrm{H}(K_\sigma F_\sigma, \phi_{\sigma^2 I} * F) \\ &\leqslant K_0\sigma^\beta + A_1 \tilde{\epsilon}_n^{b_1} \{\log(1/\tilde{\epsilon}_n)\}^{1/4} + A_2 \tilde{\epsilon}_n^{b_1} \leqslant A_3 \sigma^\beta \end{aligned}$$

for some constant A_3. Therefore, for any $F \in \mathcal{P}_\sigma$ and $\Sigma \in \mathcal{S}_\sigma$, $d_\mathrm{H}(f_0, p_{F,\Sigma}) \leqslant d_\mathrm{H}(f_0, K_\sigma F) + d_\mathrm{H}(p_{F,\sigma^2 I}, p_{F,\Sigma}) \leqslant A_4\sigma^\beta$ for some constant A_4, because $d_\mathrm{H}(p_{F,\sigma^2 I}, p_{F,\Sigma}) \leqslant |\mathrm{tr}(\sigma^2\Sigma^{-1}) - d - \log\det(\sigma^2\Sigma^{-1})|^{1/2}$ for any F. Moreover, for every $x \in \mathbb{R}^d$ with $\|x\| < a_\sigma$,

$$\frac{p_{F,\Sigma}(x)}{f_0(x)} \geqslant \frac{K_1}{\sigma^d} \int_{\|x-z\| \leqslant \sigma} \exp\left(-\frac{\|x-z\|^2}{\sigma^2}\right) F(\mathrm{d}z) \geqslant \frac{K_2}{\sigma^d} F(U_{J(x)}) \geqslant K_3 \frac{\tilde{\epsilon}_n^{4db_1}}{\sigma^d}$$

for some constants K_1, K_2 and K_3, where $J(x)$ denotes the index $j \in \{1, \ldots, K\}$ for which $x \in U_j$. The penultimate inequality holds because $U_{J(x)}$ with diameter no larger than σ must be a subset of a ball of radius σ around x. Also, for any $x \in \mathbb{R}^d$ with $\|x\| > a_\sigma$,

$$\frac{p_{F,\Sigma}(x)}{f_0(x)} \geqslant \frac{K_1}{\sigma^d} \int_{\|z\| \leqslant a_\sigma} \exp\left(-\frac{\|x-z\|^2}{\sigma^2}\right) F(\mathrm{d}z) \geqslant \frac{K_4}{\sigma^d} \exp(-4\|x\|^2/\sigma^2)$$

for some constant K_4, because $\|x-z\|^2 \leqslant 2\|x\|^2 + 2\|z\|^2 \leqslant 4\|x\|^2$ and $F(\{x \in \mathbb{R}^d : \|x\| \leqslant a_\sigma\}) \geqslant 1 - 2\tilde{\epsilon}_n^{2db_1}$. Set $\lambda = K_3 \tilde{\epsilon}_n^{4db_1}/\sigma^d$, and notice that $\log(1/\lambda) \leqslant K_5 \log(1/\tilde{\epsilon}_n)$ for some constant K_5. For any $F \in \mathcal{P}_\sigma$ and $\Sigma \in \mathcal{S}_\sigma$,

$$\begin{aligned} P_0\left\{\left(\log\frac{f_0}{p_{F,\Sigma}}\right)^2 \mathbb{1}\mathbb{1}\left(\frac{p_{F,\Sigma}}{f_0} < \lambda\right)\right\} &\leqslant \frac{K_6}{\sigma^4} \int_{\|x\|>a_\sigma} \|x\|^4 f_0(x)\,\mathrm{d}x \\ &\leqslant \frac{K_6}{\sigma^4} (P_0\|X\|^8)^{1/2} P_0(E_\sigma^\mathrm{c})^{1/2} \leqslant K_7 \sigma^{2\beta+\epsilon} \end{aligned}$$

for some constant K_7, since $P_0\|X\|^m < \infty$ for all $m > 0$ because of the tail condition (7). Given n sufficiently large, we have $\lambda < e^{-1}$ and hence $\log(f_0/p_{F,\Sigma})\mathbb{1}\mathbb{1}(p_{F,\Sigma}/f_0 < \lambda) \leqslant \{\log(f_0/p_{F,\Sigma})\}^2 \mathbb{1}\mathbb{1}(p_{F,\Sigma}/f_0 < \lambda)$. Therefore $P_0\{\log(f_0/p_{F,\Sigma})\mathbb{1}\mathbb{1}(p_{F,\Sigma}/f_0 < \lambda)\} \leqslant K_7\sigma^{2\beta+\epsilon}$. Now apply Lemma B2 to conclude that both $P_0\{\log(f_0/p_{F,\Sigma})\}$ and $P_0\{\log(f_0/p_{F,\Sigma})\}^2$ are bounded by $K_8 \log(1/\lambda)^2\sigma^{2\beta} \leqslant K_9\sigma^{2\beta}\{\log(1/\tilde{\epsilon}_n)\}^2 \leqslant A\tilde{\epsilon}_n^2$ for some positive constant A. Therefore

$$\begin{aligned} &(\mathcal{D}_\alpha \times G)\left[P_0 \log\frac{f_0}{p_{F,\Sigma}} \leqslant A\tilde{\epsilon}_n^2,\ P_0\left(\log\frac{f_0}{p_{F,\Sigma}}\right)^2 \leqslant A\tilde{\epsilon}_n^2\right] \\ &\quad\geqslant \mathcal{D}_\alpha(\mathcal{P}_\sigma) G(\mathcal{S}_\sigma) \\ &\quad\geqslant C_4 \exp\left[-c_4 \tilde{\epsilon}_n^{-d^*/\beta} \left\{\log(1/\tilde{\epsilon}_n)\right\}^{sd^*+1}\right]. \end{aligned}$$

This gives (12), provided that $\tilde{\epsilon}_n^{-d^*/\beta}\{\log(1/\tilde{\epsilon}_n)\}^{sd^*+1} \leqslant n\tilde{\epsilon}_n^2$. With $\tilde{\epsilon}_n = n^{-\beta/(2\beta+d^*)}(\log n)^t$, the condition is satisfied if $t \geqslant (sd^* + 1)/(2 + d^*/\beta)$. □

4. Sieve construction

In the following proposition, based on the stick-breaking representation of a Dirichlet process, we give an explicit definition of the sieve and derive upper bounds for its entropy and the prior probability of its complement. This result serves as the main tool in obtaining adaptive posterior convergence rates; a proof is given in Appendix A.

Proposition 2. *Fix $\epsilon, a, \sigma_0 > 0$ and integers $M, H \geqslant d$. Define*

$$\mathcal{Q} = \left\{ p_{F,\Sigma} \text{ with } F = \sum_{h=1}^{\infty} \pi_h \delta_{z_h} : \begin{array}{l} z_h \in [-a, a]^d,\ h \leqslant H;\ \sum_{h>H} \pi_h < \epsilon; \\ \sigma_0^2 \leqslant \mathrm{eig}_j(\Sigma) < \sigma_0^2 \left(1 + \epsilon^2/d\right)^M,\ j = 1, \ldots, d \end{array} \right\}. \quad (13)$$

Then:

(i) $\log N(\epsilon, \mathcal{Q}, \rho) \leqslant K[dH \log\{a/(\sigma_0\epsilon)\} - H \log \epsilon + \log M + M\epsilon^2]$ *for some constant K, where ρ is either the Hellinger or the L_1 metric;*

(ii) $(\mathcal{D}_\alpha \times G)(\mathcal{Q}^{\mathrm{c}}) \leqslant b_1 H \exp\{-C_1 a^{a_1}\} + \{(e|\alpha|/H)\log(1/\epsilon)\}^H + b_2 \exp\{-C_2 \sigma_0^{-2a_2}\} + b_3 \sigma_0^{-2a_3}(1 + \epsilon^2/d)^{-2Ma_3}$, *with the constants as defined in (1)–(4).*

The sieve defined here can easily adapt to different rates of convergence of the form $\epsilon_n = n^{-\gamma}(\log n)^{(d+1+s)/2}$ for $0 < \gamma \leqslant 1/2$ and $s > 0$. The extreme case of $\gamma = 1/2$ corresponds to the class of Gaussian mixtures (Ghosal & van der Vaart, 2001). For a β-Hölder-class convergence rate we need to work with $\gamma = \beta/(2\beta + d^*)$. The following theorem makes this precise.

Theorem 5. *Fix $\gamma \in (0, 1/2)$ and a pair of numbers t and t_0 such that $t > t_0 \geqslant (d+1)/2$. For $n \geqslant 1$, take $\epsilon_n = n^{-\gamma}(\log n)^t$ and $\tilde{\epsilon}_n = n^{-\gamma}(\log n)^{t_0}$, and define $\mathcal{F}_n$ as $\mathcal{Q}$ in (13) with $\epsilon = \epsilon_n$, $H = \lfloor n\epsilon_n^2/(\log n)\rfloor$ and $M = a^{a_1} = \sigma_0^{-2a_2} = n$. Then $\mathcal{F}_n$ satisfies (8) and (9) for all large n, for some $c_1, c_3 > 0$ and every $c_2 > 0$.*

Proof. By Proposition 2,

$$\begin{aligned} \log N(\bar{\epsilon}_n, \mathcal{F}_n, \rho) &\leqslant K\{dn^{1-2\gamma}(\log n)^{2t} + n^{1-2\gamma}(\log n)^{2t} + \log n + n^{1-2\gamma}(\log n)^{2t}\} \\ &\leqslant c_1 n^{1-2\gamma}(\log n)^{2t} = c_1 n\epsilon_n^2 \end{aligned}$$

for some $c_1 > 0$, and hence (8) holds. By the second assertion of the same proposition,

$$\begin{aligned} (\mathcal{D}_\alpha \times G)(\mathcal{F}_n^{\mathrm{c}}) &\leqslant b_1 n^{1-2\gamma}(\log n)^{2t-1} \exp(-b_1 n) + n^{-(1-2\gamma)n^{1-2\gamma}(\log n)^{2t-1}} \\ &\quad + b_2 \exp(-C_2 n) + b_3 n^{a_3/a_2} \exp\{-2a_3 n \log(1 + \bar{\epsilon}_n^2/d)\} \\ &\leqslant c_3 \exp\{-(1-2\gamma)n^{1-2\gamma}(\log n)^{2t}\} \leqslant c_3 \exp\{-(c_2+4)n^{1-2\gamma}(\log n)^{2t_0}\} \end{aligned}$$

for all large n, some $c_3 > 0$ and every $c_2 > 0$. □

5. Anisotropic Hölder functions

Anisotropic functions are those that have different orders of smoothness along different axes. The isotropic result presented earlier gives adaptive rates corresponding to the least smooth direction. Sharper results can be obtained by explicitly factoring in the anisotropy. For any $a = (a_1, \ldots, a_d)$ and $b = (b_1, \ldots, b_d)$, let $\langle a, b\rangle$ denote $a_1 b_1 + \cdots + a_d b_d$; for $y = (y_1, \ldots, y_d)$, let $\|y\|_1$ denote the L_1-norm $|y_1| + \cdots + |y_d|$. For a $\beta > 0$, an $\alpha = (\alpha_1, \ldots, \alpha_d) \in (0, \infty)^d$ with

$\alpha_\cdot = d$, and an $L : \mathbb{R}^d \to (0, \infty)$ satisfying $L(x + y) \leqslant L(x) \exp(\tau_0 \|y\|_1^2)$ for all $x, y \in \mathbb{R}^d$ and some $\tau_0 > 0$, the α-anisotropic β-Hölder class with envelope L is defined to be the set of all functions $f : \mathbb{R}^d \to \mathbb{R}$ that have continuous mixed partial derivatives $D^k f$ of all orders $k \in \mathbb{N}_0^d$ with $\beta - \alpha_{\max} \leqslant \langle k, \alpha \rangle < \beta$ where $\alpha_{\max} = \max(\alpha_1, \ldots, \alpha_d)$, such that

$$|D^k f(x + y) - D^k f(x)| \leqslant L(x) \exp(\tau_0 \|y\|_1^2) \sum_{j=1}^{d} |y_j|^{\min(\beta/\alpha_j - k_j, 1)} \quad (x, y \in \mathbb{R}^d).$$

We denote this set of functions by $\mathcal{C}^{\alpha,\beta,L,\tau_0}(\mathbb{R}^d)$. Here β refers to the mean smoothness and α the anisotropy index. An $f \in \mathcal{C}^{\alpha,\beta,L,\tau_0}$ has partial derivatives of all orders up to $\lfloor \beta_j \rfloor$ along axis j, where $\beta_j = \beta/\alpha_j$, and β is the harmonic mean $d/(\beta_1^{-1} + \cdots + \beta_d^{-1})$ of these axial smoothness coefficients. In the special case of $\alpha = (1, \ldots, 1)$, the anisotropic set $\mathcal{C}^{\alpha,\beta,L,\tau_0}(\mathbb{R}^d)$ equals the isotropic set $\mathcal{C}^{\beta,L,\tau_0}(\mathbb{R}^d)$.

THEOREM 6. *Suppose that $f_0 \in \mathcal{C}^{\alpha,\beta,L,\tau_0}(\mathbb{R}^d)$ is a probability density function satisfying*

$$P_0\big(|D^k f_0|/f_0\big)^{(2\beta+\epsilon)/\langle k,\alpha\rangle} < \infty \quad (k \in \mathbb{N}_0^d,\ \langle k, \alpha \rangle < \beta), \qquad P_0\big(L/f_0\big)^{(2\beta+\epsilon)/\beta} < \infty$$

for some $\epsilon > 0$ and that (7) holds for some constants $a, b, c, \tau > 0$. If Π is as in § 2·2, then the posterior convergence rate at f_0 in the Hellinger or the L_1 metric is $\epsilon_n = n^{-\beta/(2\beta+d^*)}(\log n)^t$, where $t \geqslant \{d^*(1 + \tau^{-1} + \beta^{-1}) + 1\}/(2 + d^*/\beta)$ and $d^* = \max(d, \kappa\alpha_{\max})$.

A proof, given in Appendix A, is similar to the proofs of the results presented in § 3, except that to obtain an approximation to f_0, we replace the single bandwidth σ with bandwith σ^{α_j} along the jth axis. An f_0 satisfying the conditions of the above theorem also satisfies the conditions of Theorem 1 with smoothness index $\beta/\alpha_{\max}$, which is strictly smaller than β as long as not all of the α_j are equal to 1. Therefore, when the true density is anisotropic, Theorem 6 indeed leads to a sharper convergence rate result.

With the standard inverse Wishart prior G, we have $\kappa = 2$, and consequently the optimal rate $n^{-\beta/(2\beta+d)}$ is recovered up to a $\log n$ factor only when $\alpha_{\max} \leqslant d/2$. Therefore, in a two-dimensional case, only the isotropic case is addressed, and for higher dimensions we get optimal results for a limited amount of anisotropy. But, when $\kappa \leqslant 1$, as in the case of a diagonal Σ with squared inverse gamma diagonal components, Theorem 6 provides optimal rates for any dimension and any degree of anisotropy, because $\alpha_{\max}$ can never exceed d.

ACKNOWLEDGEMENT

The authors thank the editor, the associate editor and the reviewers for comments that have helped to improve the paper substantially. The research of the first and third authors was partially supported by the U.S. National Science Foundation.

APPENDIX A

Proof of Lemma 1. Let $\Sigma \sim \text{IW}(\nu, \Psi)$ and suppose $\Psi = I$. It is well known that $\text{tr}(\Sigma^{-1}) \sim \chi^2_{\nu d}$, the chi-squared distribution with νd degrees of freedom. The cumulative distribution function $F(x; k)$ of χ^2_k satisfies $1 - F(zk; k) \leqslant \{z \exp(1 - z)\}^{k/2}$ for all $z > 1$. Therefore, for all $x > \nu d$,

$$\text{pr}\{\text{eig}_d(\Sigma^{-1}) > x\} \leqslant \text{pr}\{\text{tr}(\Sigma^{-1}) > x\} \leqslant \Big(\frac{x}{\nu d}\Big)^{\nu d/2} \exp\{(\nu d - x)/2\} \leqslant b_2 \exp(-C_2 x)$$

for some constants b_2 and C_2. Furthermore, the joint probability density of $\mathrm{eig}_1(\Sigma^{-1}), \ldots, \mathrm{eig}_d(\Sigma^{-1})$ is

$$f(x_1, \ldots, x_d) = c_{d,\nu} \exp\left(-\sum_j x_j/2\right) \prod_{j=1}^d x_j^{(\nu+1-d)/2} \prod_{j<k} (x_k - x_j)$$

over the set $\{(x_1, \ldots, x_d) \in (0, \infty)^d : x_1 \leqslant \cdots \leqslant x_d\}$, for a known constant $c_{d,\nu}$. Since $\prod_{j<k}(x_k - x_j) \leqslant \prod_{j<k} x_k = \prod_{k=2}^d x_k^{k-1}$, the probability density of $\mathrm{eig}_1(\Sigma^{-1})$ satisfies

$$\begin{aligned} f(x_1) &\leqslant c_{d,\nu} x_1^{(\nu+1-d)/2} \exp(-x_1/2) \prod_{k=2}^d \left\{ \int_0^\infty x_k^{(\nu+1-d)/2+k-1} \exp(-x_k/2)\, \mathrm{d}x_k \right\} \\ &= \tilde{c}_{d,\nu} x_1^{(\nu+1-d)/2} \exp(-x_1/2) \end{aligned}$$

for all $x_1 > 0$ and some positive constant $\tilde{c}_{d,\nu}$. Therefore, for any $x > 0$,

$$\mathrm{pr}\left\{\mathrm{eig}_1(\Sigma^{-1}) < x\right\} \leqslant \tilde{c}_{d,r} \int_0^x x_1^{(\nu+1-d)/2}\, \mathrm{d}x_1 \leqslant b_3 x^{a_3}$$

for some positive constants a_3 and b_3.

Next, notice that the set on the left-hand side of (4) contains all Σ which have $\mathrm{eig}_j(\Sigma^{-1}) \in I_j = \big(s_j\{1 + (j - 1/2)t/d\},\ s_j(1 + jt/d)\big)$ $(j = 1, \ldots, d)$ and that for any positive integers $k > j$, $x_j \in I_j$ and $x_k \in I_k$ implies that $x_k - x_j > s_k\{1 + (k - 1/2)t/d\} - s_j(1 + jt/d) \geqslant s_1 t/(2d)$. Therefore

$$\begin{aligned} &\mathrm{pr}\{s_j < \mathrm{eig}_j(\Sigma^{-1}) < s_j(1 + t),\ j = 1, \ldots, d\} \\ &\quad \geqslant \int_{I_d} \cdots \int_{I_1} c_{d,\nu} \exp\left(-\sum_j x_j/2\right) \prod_{j=1}^d x_j^{(\nu+1-d)/2} \prod_{j<k} (x_k - x_j)\, \mathrm{d}x_1 \cdots \mathrm{d}x_d \\ &\quad \geqslant c_{d,\nu} \exp(-ds_d)\, s_1^{d(\nu+1-d)/2} \{t/(2d)\}^{d(d-1)/2} \int_{I_d} \cdots \int_{I_1} \mathrm{d}x_1 \cdots \mathrm{d}x_d \\ &\quad = c_{d,\nu} \exp(-ds_d)\, s_1^{d(\nu+1-d)/2} \{t/(2d)\}^{d(d-1)/2} \{s_1 t/(2d)\}^d, \end{aligned}$$

which gives (4) for some positive constants a_4, a_5, b_4 and C_3.

If $\Psi \neq I$, by applying the above results for $\Psi^{-1}\Sigma \sim \mathrm{IW}(\nu, I)$ one sees that the conclusion holds for a different set of constants. □

Proof of Lemma 2. From multivariate Taylor expansion of any $f \in \mathcal{C}^{\beta, L, \tau_0}(\mathbb{R}^d)$,

$$f(x - y) - f(x) = \sum_{1 \leqslant k_\cdot \leqslant \lfloor\beta\rfloor} \frac{(-y)^{k_\cdot}}{k!} (D^k f)(x) + R(x, y),$$

with the residual satisfying $|R(x, y)| \leqslant K_1 L(x) \exp(\tau_0 \|y\|^2) \|y\|^\beta$ for every $x, y \in \mathbb{R}^d$ and for a universal constant K_1. Therefore, for any $\sigma \in (0, 1/(2\tau_0)^{1/2})$,

$$\begin{aligned} \{K_\sigma(T_{\beta,\sigma} f) - f\}(x) &= \int \phi_{\sigma^2 I}(y)\{f(x - y) - f(x)\}\, \mathrm{d}y - \sum_{2 \leqslant k_\cdot \leqslant \lfloor\beta\rfloor} d_k \sigma^{k_\cdot} \{K_\sigma(D^k f)\}(x) \\ &= \int \phi_{\sigma^2 I}(y) R(x, y)\, \mathrm{d}y + \sum_{2 \leqslant k_\cdot \leqslant \lfloor\beta\rfloor} \sigma^{k_\cdot} \left[\frac{(-1)^{k_\cdot} m_k}{k!} (D^k f)(x) - d_k \{K_\sigma(D^k f)\}(x)\right]. \end{aligned} \tag{A1}$$

The first term of (A1) is bounded by $K_2 L(x) \sigma^\beta$ for some universal constant K_2. If $\beta \leqslant 2$, then the second term of (A1) does not exist and we get a proof with $M_\beta = K_2$. For $\beta > 2$ we use induction on $\lfloor\beta\rfloor$.

From (11) we can rewrite the second term of (A1) as

$$\sum_{2\leqslant k_\cdot\leqslant\lfloor\beta\rfloor}\left[\frac{(-1)^{k_\cdot}m_k\sigma^{k_\cdot}}{k!}\{D^kf-K_\sigma(D^kf)\}(x)-c_k\sigma^{k_\cdot}\{K_\sigma(D^kf)\}(x)\right].$$

For each $1\leqslant k_\cdot\leqslant\lfloor\beta\rfloor$, the induction hypothesis implies that $D^kf\in\mathcal{C}^{\beta-k_\cdot,L,\tau_0}(\mathbb{R}^d)$ and

$$D^kf-K_\sigma(D^kf)=\{D^kf-K_\sigma T_{\beta-k_\cdot,\sigma}(D^kf)\}+K_\sigma\{T_{\beta-k_\cdot,\sigma}(D^kf)-D^kf\}$$

with $|\{D^kf-K_\sigma T_{\beta-k_\cdot,\sigma}(D^kf)\}(x)|\leqslant M_{\beta-k_\cdot}L(x)\sigma^{\beta-k_\cdot}$ for all $x\in\mathbb{R}^d$. This establishes the claim with $M_\beta=K_2+\sum_{2\leqslant k_\cdot\leqslant\lfloor\beta\rfloor}(m_k/k!)M_{\beta-k_\cdot}$, because

$$\begin{aligned}&\sum_{2\leqslant k_\cdot\leqslant\lfloor\beta\rfloor}\left[\frac{(-1)^{k_\cdot}m_k\sigma^{k_\cdot}}{k!}\{T_{\beta-k_\cdot,\sigma}(D^kf)-D^kf\}-c_k\sigma^{k_\cdot}D^kf\right]\\&\quad=\sum_{2\leqslant k_\cdot\leqslant\lfloor\beta\rfloor}\left\{\frac{(-1)^{k_\cdot}m_k\sigma^{k_\cdot}}{k!}\sum_{1\leqslant j_\cdot\leqslant\lfloor\beta\rfloor-k_\cdot}d_j\sigma^{j_\cdot}D^{k+j}f-c_k\sigma^{k_\cdot}D^kf\right\}\\&\quad=\sum_{3\leqslant n_\cdot\leqslant\lfloor\beta\rfloor}\left\{\sum_{\substack{n=l+k\\l_\cdot\geqslant1,k_\cdot\geqslant2}}\frac{(-1)^{k_\cdot}}{k!}m_kd_l-c_n\right\}\sigma^{n_\cdot}D^nf=0\end{aligned}$$

identically by the definitions of c_n and d_n. □

Proof of Theorem 3. Fix $s_0\in(0,1/(2\tau_0)^{1/2})$ such that $\sum_{1\leqslant k_\cdot\leqslant\lfloor\beta\rfloor}|d_k|\,|\log\sigma|^{-k_\cdot/2}<1/2$ and $\sigma^\epsilon|\log\sigma|^{(2\beta+\epsilon)/2}<1$ for all $0<\sigma<s_0$. For any $\sigma\in(0,s_0)$, define

$$A_\sigma=\left\{x:\frac{|D^kf_0(x)|}{f_0(x)}\leqslant\sigma^{-k_\cdot}|\log\sigma|^{-k_\cdot/2},\ k_\cdot\leqslant\lfloor\beta\rfloor;\ \frac{L(x)}{f_0(x)}\leqslant\sigma^{-\beta}|\log\sigma|^{-\beta/2}\right\}$$

and notice that, by Markov's inequality,

$$\begin{aligned}P_0(A_\sigma^{\mathrm{c}})&\leqslant\sum_{k_\cdot\leqslant\lfloor\beta\rfloor}P_0\left\{\frac{|D^kf_0(X)|}{f_0(X)}>\sigma^{-k_\cdot}|\log\sigma|^{-k_\cdot/2}\right\}+P_0\left\{\frac{L(X)}{f_0(X)}>\sigma^{-\beta}|\log\sigma|^{-\beta/2}\right\}\\&=\sum_{k_\cdot\leqslant\lfloor\beta\rfloor}P_0\left\{\left(|D^kf_0|/f_0\right)^{(2\beta+\epsilon)/k_\cdot}>\sigma^{-(2\beta+\epsilon)}|\log\sigma|^{-(2\beta+\epsilon)/2}\right\}\\&\quad+P_0\{(L/f_0)^{(2\beta+\epsilon)/\beta}>\sigma^{-(2\beta+\epsilon)}|\log\sigma|^{-(2\beta+\epsilon)/2}\}\\&\leqslant\sigma^{2\beta+\epsilon}|\log\sigma|^{(2\beta+\epsilon)/2}\left\{\sum_{k_\cdot\leqslant\lfloor\beta\rfloor}P_0\left(|D^kf_0|/f_0\right)^{(2\beta+\epsilon)/k_\cdot}+P_0\,(L/f_0)^{(2\beta+\epsilon)/\beta}\right\},\end{aligned}$$

which is bounded by $K_1\sigma^{2\beta}$ for some constant K_1. Also, for any $x\in A_\sigma$,

$$|(f_\sigma-f_0)(x)|\leqslant\sum_{1\leqslant k_\cdot\leqslant\lfloor\beta\rfloor}|d_k|\sigma^{k_\cdot}|D^kf_0(x)|\leqslant f_0(x)\sum_{1\leqslant k_\cdot\leqslant\lfloor\beta\rfloor}|d_k|\,|\log\sigma|^{-k_\cdot/2}\leqslant\frac{1}{2}f_0(x).$$

Consequently, $f_\sigma\geqslant f_0/2$ on A_σ. Because of integrability conditions on D^kf_0/f_0, it turns out that in calculating $\int D^kf_0(x)\,\mathrm{d}x$ for any $1\leqslant k_\cdot\leqslant\lfloor\beta\rfloor$, one can integrate under the derivative and conclude that

$\int D^k f_0(x)\,\mathrm{d}x = 0$ as f_0 is a density. So $\int f_\sigma(x)\,\mathrm{d}x = 1$, and for some constant K_2 and all $\sigma < s_0$,

$$1 \leqslant \int g_\sigma(x)\,\mathrm{d}x \leqslant 1 + \frac{1}{2}\int f_0(x)\,\mathrm{ll}\{f_\sigma(x) < f_0(x)/2\}\,\mathrm{d}x \leqslant 1 + \frac{1}{2}P_0(A_\sigma^{\mathrm{c}}) \leqslant 1 + K_2\sigma^{2\beta}.$$

Thus $\int g_\sigma(x)\,\mathrm{d}x < \infty$ and h_σ is a well-defined probability density function on $\mathbb{R}^d$.

To prove the final result of Theorem 3, write $r_\sigma = (1/2)\,f_0\,\mathrm{ll}\{f_\sigma < (1/2)\,f_0\}$ and $c_\sigma = \int g_\sigma(x)\,\mathrm{d}x$; note that for $a, b > 0$ we have $(a^{1/2} - b^{1/2})^2 = (a-b)^2/(a^{1/2} + b^{1/2})^2 \leqslant (a-b)^2/(a+b)$ and hence

$$\begin{aligned}
d_{\mathrm{H}}^2(f_0, K_\sigma h_\sigma) &\leqslant \int \frac{(f_0 - K_\sigma h_\sigma)^2(x)}{f_0(x) + (K_\sigma h_\sigma)(x)}\,\mathrm{d}x \\
&= \frac{1}{c_\sigma}\int \frac{(c_\sigma f_0 - K_\sigma g_\sigma)^2(x)}{c_\sigma f_0(x) + (K_\sigma g_\sigma)(x)}\,\mathrm{d}x \\
&\leqslant 3\int \frac{(c_\sigma - 1)^2 f_0^2(x) + (f_0 - K_\sigma f_\sigma)^2(x) + (K_\sigma r_\sigma)^2(x)}{c_\sigma f_0(x) + (K_\sigma g_\sigma)(x)}\,\mathrm{d}x \\
&\leqslant 3\left\{\int (c_\sigma - 1)^2 f_0(x)\,\mathrm{d}x + \int \frac{(f_0 - K_\sigma f_\sigma)^2(x)}{f_0(x)}\,\mathrm{d}x + \int \frac{(K_\sigma r_\sigma)^2(x)}{(K_\sigma g_\sigma)(x)}\,\mathrm{d}x\right\} \\
&\leqslant 3\left\{K_2^2\sigma^{4\beta} + M_\beta^2\sigma^{2\beta}P_0\,(L/f_0)^2 + \int (K_\sigma r_\sigma)(x)\,\mathrm{d}x\right\},
\end{aligned}$$

because $1 \leqslant c_\sigma \leqslant 1 + K_2\sigma^{2\beta}$, $|(f_0 - K_\sigma f_\sigma)(x)| < M_\beta L(x)\sigma^\beta$ and $K_\sigma r_\sigma \leqslant K_\sigma g_\sigma$ since $r_\sigma \leqslant g_\sigma$. By Jensen's inequality, $P_0(L/f_0)^2 \leqslant \{P_0(L/f_0)^{(2\beta+\epsilon)/\beta}\}^{\beta/(\beta+\epsilon/2)} < \infty$. Also, $\int (K_\sigma r_\sigma)(x)\,\mathrm{d}x$ is

$$\frac{1}{2}\int\int \phi_{\sigma^2 I}(x-y) f_0(y)\,\mathrm{ll}\{f_\sigma(y) < f_0(y)/2\}\,\mathrm{d}x\,\mathrm{d}y = \frac{1}{2}\int f_0(y)\,\mathrm{ll}\{f_\sigma(y) < f_0(y)/2\}\,\mathrm{d}y,$$

which is bounded by $P_0(A_\sigma^{\mathrm{c}}) \leqslant K_1\sigma^{2\beta}$. □

Proof of Proposition 1. Define g_σ and h_σ as in the statement of Theorem 3. This theorem implies that there are $s_1, K > 0$ such that $d_{\mathrm{H}}^2(f_0, K_\sigma h_\sigma) \leqslant K\sigma^{2\beta}$ for all $0 < \sigma < s_1$. The tail condition on f_0 implies existence of a small $\delta > 0$ such that B_0, which is defined as $P_0(f_0^{-\delta})$, satisfies $B_0 < \infty$. Let $s_2 \in (0, 1/(2\tau_0)^{1/2})$ be such that $\{(4\beta + 2\epsilon + 8)/(b\delta)\}\log(1/s_2) > \max\{(1/b)\log c,\ a^\tau/2\}$. Set $s_0 = \min(s_1, s_2)$ and pick any $\sigma \in (0, s_0)$. Define $E_\sigma = \{x \in \mathbb{R}^d : f_0(x) \geqslant \sigma^{(4\beta+2\epsilon+8)/\delta}\}$ and $a_\sigma = a_0\log(1/\sigma)^{1/\tau}$ with $a_0 = \{(8\beta + 4\epsilon + 16)/(b\delta)\}^{1/\tau}$. Then $a_\sigma > a$ and $E_\sigma \subset \{x \in \mathbb{R}^d : \|x\| \leqslant a_\sigma\}$.

By Markov's inequality, $P_0(E_\sigma^{\mathrm{c}}) = P_0\{f_0(X)^{-\delta} > \sigma^{-(4\beta+2\epsilon+8)}\} \leqslant B_0\sigma^{4\beta+2\epsilon+8} \leqslant B_0\sigma^{2\beta+\epsilon}$; consequently, by (6) and applications of Hölder's inequality,

$$\begin{aligned}
\int_{E_\sigma^{\mathrm{c}}} g_\sigma(x)\,\mathrm{d}x &\leqslant \frac{3}{2}\int_{E_\sigma^{\mathrm{c}}} f_0(x)\,\mathrm{d}x + \sum_{k.=1}^{\lfloor\beta\rfloor}\sigma^{k.}|d_k|\int_{E_\sigma^{\mathrm{c}}} |D^k f_0(x)|\,\mathrm{d}x \\
&\leqslant \frac{3}{2}P_0(E_\sigma^{\mathrm{c}}) + \sum_{k.=1}^{\lfloor\beta\rfloor}\sigma^{k.}|d_k|\left\{P_0\left(|D^k f_0|/f_0\right)^{(2\beta+\epsilon)/k.}\right\}^{k./2\beta+\epsilon} P_0(E_\sigma^{\mathrm{c}})^{(2\beta+\epsilon-k.)/(2\beta+\epsilon)},
\end{aligned}$$

which is bounded by $B_1\sigma^{2\beta+\epsilon}$ for some constant B_1 that does not depend on σ. Hence $\int_{E_\sigma^{\mathrm{c}}} h_\sigma(x)\,\mathrm{d}x \leqslant \int_{E_\sigma^{\mathrm{c}}} g_\sigma(x)\,\mathrm{d}x \leqslant B_1\sigma^{2\beta+\epsilon}$.

Define $\tilde{h}_\sigma$ to be the restriction of h_σ to E_σ, that is, $\tilde{h}_\sigma(x) = h_\sigma(x)\,\mathrm{ll}(x \in E_\sigma)/\int_{E_\sigma} h(x)\,\mathrm{d}x$. Then $d_{\mathrm{H}}(K_\sigma h_\sigma, K_\sigma\tilde{h}_\sigma) \leqslant d_{\mathrm{H}}(h_\sigma, \tilde{h}_\sigma) = [2 - 2\{\int_{E_\sigma} h_\sigma(x)\,\mathrm{d}x\}^{1/2}]^{1/2} = O(\sigma^{\beta+\epsilon/2})$. This completes the proof, because $d_{\mathrm{H}}(f_0, K_\sigma\tilde{h}_\sigma) \leqslant d_{\mathrm{H}}(f_0, K_\sigma h_\sigma) + d_{\mathrm{H}}(K_\sigma h_\sigma, K_\sigma\tilde{h}_\sigma)$. □

Proof of Proposition 2. Let $\hat{R}$ be a $(\sigma_0\epsilon)$-net of $[-a, a]^d$, $\hat{S}$ an ϵ-net of the H-simplex $\mathcal{S}_H = \{p = (p_1, \ldots, p_H) : p_h \geqslant 0,\ \sum_{h=1}^H p_h = 1\}$, and $\hat{\mathcal{O}}$ an δ-net of $\mathcal{O}_d$, the group of $d \times d$ orthogonal matrices

equipped with the spectral norm $\|\cdot\|_2$, where $\delta = \epsilon^2/\{3d(1+\epsilon^2/d)^M\}$. It is well known that the cardinalities of these nets are such that $\text{card}(\hat{R}) \lesssim \{a/(\sigma_0\epsilon)\}^d$, $\text{card}(\hat{S}) \lesssim \epsilon^{-H}$ and $\text{card}(\hat{\mathcal{O}}) \lesssim \delta^{-d(d-1)/2}$.

Pick any $p_{F,\Sigma} \in \mathcal{Q}$ with $F = \sum_{h=1}^{\infty} z_h \delta_{z_h}$, and let the spectral decomposition of Σ^{-1} be $P\Lambda P^{\mathrm{T}}$ where $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_d)$ and P is an orthogonal matrix. Find $\hat{z}_1, \ldots, \hat{z}_H \in \hat{R}$, $\hat{\pi} = (\hat{\pi}_1, \ldots, \hat{\pi}_H) \in \hat{S}$, $\hat{P} \in \hat{\mathcal{O}}$ and $\hat{m}_1, \ldots, \hat{m}_d \in \{1, \ldots, M\}$ such that

$$\max_{1 \leqslant h \leqslant H} \|z_h - \hat{z}_h\| < \sigma_0\epsilon,$$

$$\sum_{h=1}^{H} |\tilde{\pi}_h - \hat{\pi}_h| < \epsilon \quad \text{where } \tilde{\pi}_h = \frac{\pi_h}{1 - \sum_{l>H} \pi_l} \quad (1 \leqslant h \leqslant H),$$

$$\|P - \hat{P}\|_2 \leqslant \epsilon^2,$$

$$\hat{\lambda}_j = \{\sigma_0^2(1+\epsilon^2/d)^{\hat{m}_j - 1}\}^{-1} \text{ satisfies } 1 \leqslant \hat{\lambda}_j/\lambda_j < 1 + \epsilon^2/d \quad (j = 1, \ldots, d).$$

Take $\hat{F} = \sum_{h=1}^{H} \hat{\pi}_h \delta_{\hat{z}_h}$ and $\hat{\Sigma} = (\hat{P}\hat{\Lambda}\hat{P}^{\mathrm{T}})^{-1}$ where $\hat{\Lambda} = \text{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_d)$. Also define $\tilde{\Sigma} = (\hat{P}\Lambda\hat{P}^{\mathrm{T}})^{-1}$ and $Q = \hat{P}^{\mathrm{T}}P$. By the triangle inequality,

$$\|p_{F,\Sigma} - p_{\hat{F},\hat{\Sigma}}\|_1 \leqslant \|p_{F,\Sigma} - p_{F,\hat{\Sigma}}\|_1 + \|p_{F,\hat{\Sigma}} - p_{\hat{F},\hat{\Sigma}}\|_1. \tag{A2}$$

The first term on the right-hand side can be bounded by

$$\int \|\phi_\Sigma(\cdot - z) - \phi_{\hat{\Sigma}}(\cdot - z)\|_1 \, \mathrm{d}F(z) = \|\phi_\Sigma - \phi_{\hat{\Sigma}}\|_1 \leqslant \|\phi_\Sigma - \phi_{\tilde{\Sigma}}\|_1 + \|\phi_{\tilde{\Sigma}} - \phi_{\hat{\Sigma}}\|_1.$$

Since the total variation distance is bounded by $2^{1/2}$ times the square root of the Kullback–Leibler divergence, we have $\|\phi_{\tilde{\Sigma}} - \phi_{\hat{\Sigma}}\|_1 \leqslant \{\text{tr}(\hat{\Sigma}^{-1}\tilde{\Sigma}) - \log\det(\hat{\Sigma}^{-1}\tilde{\Sigma}) - d\}^{1/2}$. But $\text{tr}(\hat{\Sigma}^{-1}\tilde{\Sigma}) = \text{tr}(\hat{\Lambda}\Lambda^{-1}) = \sum_{j=1}^{d} \hat{\lambda}_j/\lambda_j < d + \epsilon^2$ and $\det(\hat{\Sigma}^{-1}\tilde{\Sigma}) = \prod_{j=1}^{d}(\hat{\lambda}_j/\lambda_j) > 1$. Thus $\|\phi_{\tilde{\Sigma}} - \phi_{\hat{\Sigma}}\|_1 \leqslant \epsilon$. For the other term, we have $\|\phi_\Sigma - \phi_{\tilde{\Sigma}}\|_1 \leqslant \{\text{tr}(\Sigma^{-1}\hat{\Sigma}) - \log\det(\Sigma^{-1}\hat{\Sigma}) - d\}^{1/2} = \{\text{tr}(Q\Lambda Q^{\mathrm{T}}\Lambda^{-1} - I)\}^{1/2}$ because $\Sigma^{-1}\hat{\Sigma} = P\Lambda P^{\mathrm{T}}\hat{P}\Lambda^{-1}\hat{P}^{\mathrm{T}}$ has determinant 1 and trace equal to that of $Q\Lambda Q^{\mathrm{T}}\Lambda^{-1}$. Write $Q = I + B$. Then $\|B\|_{\max} \leqslant \|B\|_2 = \|\hat{P}^{\mathrm{T}}P - I\|_2 = \|P - \hat{P}\|_2 \leqslant \delta$ and hence

$$\text{tr}(Q\Lambda Q^{\mathrm{T}}\Lambda^{-1} - I) = \text{tr}(B + \Lambda B^{\mathrm{T}}\Lambda^{-1} + B\Lambda B^{\mathrm{T}}\Lambda^{-1}) \leqslant 3d\,\|B\|_{\max}\frac{\max(\lambda_1, \ldots, \lambda_d)}{\min(\lambda_1, \ldots, \lambda_d)} \leqslant \epsilon^2.$$

Hence the first term on the right-hand side of (A2) is bounded by 2ϵ. The last term of (A2) equals

$$\left\| \sum_{h>H} \pi_h \phi_{\hat{\Sigma}}(\cdot - z_h) + \sum_{h=1}^{H} \pi_h \{\phi_{\hat{\Sigma}}(\cdot - z_h) - \phi_{\hat{\Sigma}}(\cdot - \hat{z}_h)\} + \sum_{h=1}^{H} (\pi_h - \hat{\pi}_h)\phi_{\hat{\Sigma}}(\cdot - \hat{z}_h) \right\|_1$$

$$\leqslant \sum_{h>H} \pi_h + \sum_{h=1}^{H} \pi_h \|\phi_{\hat{\Sigma}}(\cdot - z_h) - \phi_{\hat{\Sigma}}(\cdot - \hat{z}_h)\|_1 + \sum_{h=1}^{H} |\pi_h - \hat{\pi}_h|.$$

The first term above is smaller than ϵ, and so is the second term because

$$\|\phi(\cdot - z_h) - \phi(\cdot - \hat{z}_h)\|_1 \leqslant \left(\frac{2}{\pi}\right)^{1/2} \|\hat{\Sigma}^{-1/2}(z_h - \hat{z}_h)\| \leqslant \epsilon.$$

The last term is less than or equal to $(1 - \sum_{h>H}\pi_h)\sum_{h=1}^{H}|\tilde{\pi}_h - \hat{\pi}_h| + \sum_{h>H}\pi_h \sum_{h=1}^{H}\hat{\pi}_h \leqslant 2\epsilon$. Thus a (6ϵ)-net of $\mathcal{Q}$, in the L_1 topology, can be constructed with $\hat{p} = p_{\hat{F},\hat{\Sigma}}$ as above. The total number of such $\hat{p}$ is bounded by a multiple of $\{a/(\sigma_0\epsilon)\}^{dH}\epsilon^{-H}\delta^{-d(d-1)/2}M^d$. This proves the first assertion with $\rho = \|\cdot\|_1$, because $M\log(1+\epsilon^2/d) \lesssim M\epsilon^2$ and the constant factor 6 can be absorbed into the bound. The same holds when ρ is the Hellinger metric, because it is bounded by the square root of the L_1 metric.

For the second assertion, we know that a Dirichlet process $F \sim \mathcal{D}_\alpha$ can be represented by Sethuraman's stick-breaking process as

$$F = \sum_{h=1}^{\infty} \pi_h \delta_{Z_h}, \quad \pi_h = V_h \prod_{j<h} (1 - V_j), \tag{A3}$$

where δ_x is the Dirac measure at x, $\{V_h, h \geqslant 1\}$ are independent beta-distributed random variables with parameters 1 and $|\alpha| = \alpha(\mathbb{R}^d)$, $\{Z_h, h \geqslant 1\}$ are independently distributed according to the probability measure $\bar{\alpha} = \alpha/|\alpha|$, and these two sets of random variables are mutually independent. Hence $p_{F,\Sigma} = \sum_{h=1}^{\infty} \pi_h \phi_\Sigma(\cdot - Z_h)$ with π_h and Z_h as described in (A3). Therefore, with Π denoting the Dirichlet mixture prior of § 2·2, we have

$$\Pi(\mathcal{Q}^{\mathrm{c}}) \leqslant H\bar{\alpha}([-a, a]^d)^{\mathrm{c}} + \mathrm{pr}\left(\sum_{h>H} \pi_h > \epsilon\right) + \mathrm{pr}\{\mathrm{eig}_d(\Sigma^{-1}) > \sigma_0^{-2}\}$$
$$+ \mathrm{pr}\left\{\mathrm{eig}_1(\Sigma^{-1}) \leqslant \sigma_0^{-2}\left(1 + \frac{\epsilon^2}{d}\right)^{-M}\right\}.$$

The first term is bounded by $b_1 H \exp(-C_1 a^{a_1})$ by the assumption on α. Because $W = -\sum_{h=1}^{H} \log(1 - V_h)$ is gamma-distributed with parameters H and $|\alpha|$, we have

$$\mathrm{pr}\left(\sum_{h>H} \pi_h > \epsilon\right) = \mathrm{pr}\left(W < \log\frac{1}{\epsilon}\right) \leqslant \frac{(-|\alpha| \log \epsilon)^H}{\Gamma(H+1)} \leqslant \left(\frac{e|\alpha|}{H} \log\frac{1}{\epsilon}\right)^H$$

by Stirling's formula. The last two terms are bounded by a multiple of $b_2 \exp\{-C_2 \sigma_0^{-2a_2}\} + b_3 \sigma_0^{-2a_3} \left(1 + \epsilon^2/d\right)^{-Ma_3}$. This proves the second assertion. □

Proof of Theorem 6. For any $\sigma > 0$, define the transformation $T_{\alpha,\beta,\sigma}$ on $\mathcal{C}^{\alpha,\beta,L,\tau_0}(\mathbb{R}^d)$ as

$$T_{\alpha,\beta,\sigma} f = f - \sum_{k \in \mathbb{N}_0^d:\, 1 \leqslant \langle k,\alpha\rangle < \beta} d_k \sigma^{\langle k,\alpha\rangle} f.$$

Also, define $K_{\alpha,\sigma} f$ to be the convolution of f and the normal density with mean zero and variance $\mathrm{diag}(\sigma^{2\alpha_1}, \ldots, \sigma^{2\alpha_d})$. The anisotropic analogue of Lemma 2 is that there exists a constant $M_{\alpha,\beta}$ such that for any $f \in \mathcal{C}^{\alpha,\beta,L,\tau_0}$ and any $\sigma \in (0, 1/(2\tau_0)^{1/2\alpha_{\max}})$, $|\{K_{\alpha,\sigma}(T_{\alpha,\beta,\sigma} f) - f\}(x)| < M_\beta L(x) \sigma^\beta$ for all $x \in \mathbb{R}^d$. This follows the lines of our proof of Lemma 2, starting from the anisotropic Taylor approximation

$$f(x + y) - f(x) = \sum_{1 \leqslant \langle k,\alpha\rangle < \beta} \frac{(-y)^k}{k!} (D^k f)(x) + R(x, y),$$

where the residual $R(x, y)$ is bounded in absolute value by a sum over terms of the form

$$\frac{|y|^k}{k!} \Big| (D^k f)(x_1, \ldots, x_{j-1}, x_j + \xi_j, x_{j+1} + y_{j+1}, \ldots, x_d + y_d)$$
$$- (D^k f)(x_1, \ldots, x_{j-1}, x_j, x_{j+1} + y_{j+1}, \ldots, x_d + y_d) \Big|$$
$$\leqslant L(x) \exp(\tau_0 \|y\|_1^2)\, |y|^k |y_j|^{\min(\beta/\alpha_j - k_j, 1)}/k!,$$

with j such that $\beta > \langle k, \alpha\rangle > \beta - \alpha_j$. Consequently, $\int |R(x, y)| \phi_{\mathrm{diag}(\sigma^{2\alpha})}(y)\, \mathrm{d}y \leqslant K_1 L(x) \sigma^\beta$ for some constant K_1. Applying the rest of the induction argument in our proof of Lemma 2, we obtain the pointwise error bound between f_0 and $K_{\alpha,\sigma}(T_{\alpha,\beta,\sigma} f)$. Then it leads to exact analogues of Theorem 3 and Proposition 1, giving a $\tilde{h}_\sigma$ with support inside $\{x \in \mathbb{R}^d : \|x\| \leqslant a_0\{\log(1/\sigma)\}^\tau\}$ satisfying $d_{\mathrm{H}}(f_0, K_{\alpha,\sigma}\tilde{h}_\sigma) \leqslant K_0 \sigma^\beta$ for some constant K_0. Next, the arguments in the proof of Theorem 4 can be replicated, with $\mathcal{P}_\sigma$ built around a discrete $F_\sigma = \sum_{j=1}^{N} p_j \delta_{z_j}$ with $N \leqslant D_1 \sigma^{-d} \{\log(1/\tilde{\epsilon}_n)\}^{d+d/\tau}$ support points such that

$d_{\mathrm{H}}(K_{\alpha,\sigma}\tilde{h}_\sigma, K_{\alpha,\sigma}F_\sigma) \leqslant A_1\tilde{\epsilon}_n^{b_1}\{\log(1/\tilde{\epsilon}_n)\}^{1/4}$. We also need to define $\mathcal{S}_\sigma$ as the set of Σ such that $\mathrm{eig}_j(\Sigma^{-1})$ lies between $\sigma^{-2\alpha_j}$ and $\sigma^{-2\alpha_j}(1+\sigma^{2\beta})$ for each $j = 1, \ldots, d$. The prior probability of this set under G is bounded from below by $C_3\exp[-c_3\tilde{\epsilon}_n^{-\kappa\alpha_{\max}/\beta}\{\log(1/\tilde{\epsilon}_n)\}^{s\kappa+1}]$, which contributes the $\kappa\alpha_{\max}$ term in $d^* = \max(d, \kappa\alpha_{\max})$. □

APPENDIX B

Supplementary results

THEOREM B1. *Let P_0 be a probability measure on $\{x \in \mathbb{R}^d : \|x\| \leqslant a\} \subset \mathbb{R}^d$. For any $\varepsilon > 0$ and $\sigma > 0$, there is a discrete probability measure F_σ on $\{x \in \mathbb{R}^d : \|x\| \leqslant a\}$ with at most $N_{\sigma,\varepsilon} = D[\{(a/\sigma) \vee 1\}\log(1/\varepsilon)]^d$ support points, such that $\|p_{P_0,\sigma} - p_{F_\sigma,\sigma}\|_\infty \lesssim \varepsilon/\sigma^d$ and $\|p_{P_0,\sigma} - p_{F_\sigma,\sigma}\|_1 \lesssim \varepsilon\{\log(1/\varepsilon)\}^{1/2}$ for some universal constant D.*

Proof. The proof is a straightforward extension to d dimensions of Lemma 2 of Ghosal & van der Vaart (2007) and Lemma 3.1 of Ghosal & van der Vaart (2001). For any probability distribution F on $\mathbb{R}^d$, there exists a discrete distribution F' with at most $\{(2k-2)^d + 1\}$ support points such that the mixed moments $z_1^{l_1}z_2^{l_2}\cdots z_d^{l_d}$ are matched up for every $1 \leqslant l_i \leqslant 2k-2$ $(i = 1, \ldots, d)$. This power of d propagates all the way through the required extensions and appears in $N_{\sigma,\epsilon}$ in the statement of the current theorem. □

COROLLARY B1. *Let P_0 be a probability measure on $\{x \in \mathbb{R}^d : \|x\| \leqslant a\}$. For any $\varepsilon > 0$ and $\sigma > 0$, there is a discrete probability measure F_σ^* on $\{x \in \mathbb{R}^d : \|x\| \leqslant a\}$ with at most $N_{\sigma,\varepsilon} = D[\{(a/\sigma) \vee 1\}\log(1/\varepsilon)]^d$ support points from the set $\{(n_1, \ldots, n_p)\sigma\varepsilon : n_i \in \mathbb{Z},\ |n_i| < \lceil a/(\sigma\varepsilon)\rceil,\ i = 1, \ldots, p\}$, such that $\|p_{P_0,\sigma} - p_{F_\sigma^*,\sigma}\|_\infty \lesssim \varepsilon/\sigma^d$ and $\|p_{P_0,\sigma} - p_{F_\sigma^*,\sigma}\|_1 \lesssim \varepsilon\{\log(1/\varepsilon)\}^{1/2}$.*

Proof. First, obtain F_σ as in Theorem B1, and then move each of its support points to the nearest point on the grid $\{(n_1, \ldots, n_p)\sigma\varepsilon : n_i \in \mathbb{Z},\ |n_i| < \lceil a/(\sigma\varepsilon)\rceil,\ i = 1, \ldots, p\}$ to get F_σ^*. These moves cost at most a constant times ϵ^2/σ^d to the supremum-norm distance and at most a constant times ϵ to the L_1 distance. □

LEMMA B1. *Let $V_0, V_1, \ldots, V_N$ be a partition of $\mathbb{R}^d$ and let $F' = \sum_{j=1}^N p_j\delta_{z_j}$ be a probability measure on $\mathbb{R}^d$ with $z_j \in V_j$ $(j = 1, \ldots, N)$. Then, for any probability measure F on $\mathbb{R}^d$ and any $\sigma > 0$,*

$$\|p_{F,\sigma} - p_{F',\sigma}\|_\infty \lesssim \frac{1}{\sigma^{d+1}}\max_{1\leqslant j\leqslant N}\mathrm{diam}(V_j) + \frac{1}{\sigma^d}\sum_{j=1}^N |F(V_j) - p_j|,$$

$$\|p_{F,\sigma} - p_{F',\sigma}\|_1 \lesssim \frac{1}{\sigma}\max_{1\leqslant j\leqslant N}\mathrm{diam}(V_j) + \sum_{j=1}^N |F(V_j) - p_j|,$$

where $\mathrm{diam}(A) = \sup\{\|z_1 - z_2\| : z_1, z_2 \in A\}$ denotes the diameter of a set A.

Proof. The proof is an extension to d dimensions of Lemma 5 of Ghosal & van der Vaart (2007). □

LEMMA B2. *There is a $\lambda_0 \in (0, 1)$ such that for any $\lambda \in (0, \lambda_0)$ and any two probability measures P and Q with respective densities p and q,*

$$P\log\frac{p}{q} \leqslant d_{\mathrm{H}}^2(p,q)\left(1 + 2\log\frac{1}{\lambda}\right) + 2P\left\{\left(\log\frac{p}{q}\right)\mathrm{ll}\left(\frac{q}{p} \leqslant \lambda\right)\right\},$$

$$P\left(\log\frac{p}{q}\right)^2 \leqslant d_{\mathrm{H}}^2(p,q)\left\{12 + 2\left(\log\frac{1}{\lambda}\right)^2\right\} + 8P\left\{\left(\log\frac{p}{q}\right)^2\mathrm{ll}\left(\frac{q}{p} \leqslant \lambda\right)\right\}.$$

Proof. Our proof follows the argument presented in the proof of Lemma 7 of Ghosal & van der Vaart (2007). The function $r : (0, \infty) \to \mathbb{R}$ defined implicitly by $\log x = 2(x^{1/2} - 1) - r(x)(x^{1/2} - 1)^2$ is

nonnegative and decreasing, and there exists a $\lambda_0 > 0$ such that $r(x) \leqslant 2\log(1/x)$ for all $x \in (0, \lambda_0)$. Using these properties and the fact that $d_{\mathrm{H}}^2(p, q) = -2P\{(q/p)^{1/2} - 1\}$, we obtain

$$\begin{aligned} P\log\frac{p}{q} &= d_{\mathrm{H}}^2(p,q) + P\left\{r\left(\frac{q}{p}\right)\left(\frac{q^{1/2}}{p^{1/2}} - 1\right)^2\right\} \\ &\leqslant d_{\mathrm{H}}^2(p,q) + r(\lambda)d_{\mathrm{H}}^2(p,q) + P\left\{r\left(\frac{q}{p}\right)\mathrm{ll}\left(\frac{q}{p} < \lambda\right)\right\} \\ &\leqslant d_{\mathrm{H}}^2(p,q) + 2\left(\log\frac{1}{\lambda}\right)d_{\mathrm{H}}^2(p,q) + 2P\left\{\left(\log\frac{p}{q}\right)\mathrm{ll}\left(\frac{q}{p} < \lambda\right)\right\} \end{aligned}$$

for any $\lambda < \lambda_0$, proving the first inequality of the lemma.

To prove the other inequality, note that $|\log x| \leqslant 2|x^{1/2} - 1|$ for $x \geqslant 1$ and so

$$P\left\{\left(\log\frac{p}{q}\right)^2 \mathrm{ll}\left(\frac{q}{p} \geqslant 1\right)\right\} \leqslant 4P\left(\frac{q^{1/2}}{p^{1/2}} - 1\right)^2 = 4d_{\mathrm{H}}^2(p,q).$$

On the other hand,

$$\begin{aligned} &P\left\{\left(\log\frac{p}{q}\right)^2 \mathrm{ll}\left(\frac{q}{p} \leqslant 1\right)\right\} \\ &\leqslant 8P\left(\frac{q^{1/2}}{p^{1/2}} - 1\right)^2 + 2P\left\{r^2\left(\frac{q}{p}\right)\left(\frac{q^{1/2}}{p^{1/2}} - 1\right)^4 \mathrm{ll}\left(\frac{q}{p} \leqslant 1\right)\right\} \\ &\leqslant 8d_{\mathrm{H}}^2(p,q) + 2r^2(\lambda)P\left(\frac{q^{1/2}}{p^{1/2}} - 1\right)^2 + 2P\left\{r^2\left(\frac{q}{p}\right)\mathrm{ll}\left(\frac{q}{p} \leqslant \lambda\right)\right\} \\ &\leqslant 8d_{\mathrm{H}}^2(p,q) + 2\left(\log\frac{1}{\lambda}\right)^2 d_{\mathrm{H}}^2(p,q) + 8P\left\{\left(\log\frac{p}{q}\right)^2 \mathrm{ll}\left(\frac{q}{p} \leqslant \lambda\right)\right\}. \end{aligned}$$

This completes the proof. □

LEMMA B3. *Let $\mathcal{A}$ and $\mathcal{X}$ be metric spaces and suppose that $\{p_\alpha\}_{\alpha\in\mathcal{A}}$ and $\{q_\alpha\}_{\alpha\in\mathcal{A}}$ are collections of probability density functions on $\mathcal{X}$ with respect to a dominating measure ν. Then, for any probability measure G on $\mathcal{A}$, $d_{\mathrm{H}}^2(\int p_\alpha\,\mathrm{d}G, \int q_\alpha\,\mathrm{d}G) \leqslant \int d_{\mathrm{H}}^2(p_\alpha, q_\alpha)\,\mathrm{d}G$. In particular, for any three densities p, q and ϕ on $\mathbb{R}^d$, $d_{\mathrm{H}}(\phi * p, \phi * q) \leqslant d_{\mathrm{H}}(p, q)$.*

Proof. By the Cauchy–Schwartz inequality, $1 - \int d_{\mathrm{H}}^2(p_\alpha, q_\alpha)\,\mathrm{d}G/2$ equals

$$\int\int\{p_\alpha(x)q_\alpha(x)\}^{1/2}\,\nu(\mathrm{d}x)\,G(\mathrm{d}\alpha) \leqslant \int\left\{\int p_\alpha(x)\,G(\mathrm{d}\alpha)\int q_\alpha(x)\,G(\mathrm{d}\alpha)\right\}^{1/2}\nu(\mathrm{d}x),$$

which is the same as $1 - \frac{1}{2}d_{\mathrm{H}}^2(\int p_\alpha\,\mathrm{d}G, \int q_\alpha\,\mathrm{d}G)$. This gives the first result. The second assertion follows from choosing $\mathcal{A} = \mathcal{X} = \mathbb{R}^d$, $p_\alpha(x) = p(x-\alpha)$, $q_\alpha(x) = q(x-\alpha)$ and $G(\mathrm{d}\alpha) = \phi(\alpha)\mathrm{d}\alpha$. □

LEMMA B4. *Suppose that a probability density function f_0 satisfies the tail condition (7) and is such that $\log f_0 \in \mathcal{C}^{\beta,Q_1,0}(\mathbb{R}^d)$ for some polynomial Q_1, with $P_0|D^k \log f_0|^{(2\beta+\epsilon)/k.} < \infty$ for $k \in \mathbb{N}_0^d$, $k. \leqslant \lfloor\beta\rfloor$, and $P_0 Q_1^{(2\beta+\epsilon)/\beta} < \infty$. Additionally, suppose that*

$$\left|\frac{f_0(x+y)}{f_0(x)} - 1\right| \leqslant Q(x)\exp(\tau_0\|y\|^2)\|y\|^{\beta-\lfloor\beta\rfloor} \quad (x, y \in \mathbb{R}^d) \tag{A4}$$

for some $\tau_1 > 0$ and a function Q satisfying $P_0 Q^2 < \infty$. Then, there exist a $\tau_0 > 0$ and a positive function $L(x)$ such that $f_0 \in \mathcal{C}^{\beta,L,\tau_0}(\mathbb{R}^d)$ and (6) holds.

Without (A4), the assumptions made on f_0 in Lemma B4 match one to one with conditions C1–C3 of Kruijer et al. (2010). The additional assumption (A4) is a mild one and is satisfied by densities with tails exactly as in the bound (7) with $\tau \leqslant 2$, as well as by finite mixtures of such densities.

Proof of Lemma B4. For a multi-index $k \in \mathbb{N}_0^d$, let $\mathcal{P}$ denote the set of all solutions $\{m^{(1)}, \ldots, m^{(q)}\}$ to $k = m^{(1)} + \cdots + m^{(q)}$, $q \geqslant 1$, $m^{(j)} \in \mathbb{N}_0^d$ with $m_\cdot^{(j)} \geqslant 1$ $(j = 1, \ldots, q)$. Existence of $D^k f_0$ of all orders $k_\cdot \leqslant \lfloor \beta \rfloor$ follows from the same property of $\log f_0$. In fact, by the chain rule, $D^k f_0(x) = f_0(x) \sum_{P \in \mathcal{P}(k)} \prod_{m \in P} D^m \log f_0(x)$ and so $P_0 |(D^k f_0)/f_0|^{(2\beta+\epsilon)/k_\cdot} < \infty$ by an application of the Hölder inequality. Also, because $\log f_0 \in \mathcal{C}^{\beta, Q_1, 0}(\mathbb{R}^d)$ with Q_1 being a polynomial, for every $k \in \mathbb{N}_0^d$ with $k_\cdot < \beta$ we can find polynomials $Q_{k,1}$ and $Q_{k,2}$ such that $|D^k \log f_0(x)| < Q_{k,1}(x)$ and $|D^k \log f_0(x + y) - D^k \log f_0(x)| < Q_{k,2}(x) \exp(\|y\|^2) \|y\|^{\beta - \lfloor \beta \rfloor}$. Hence, for $k_\cdot = \lfloor \beta \rfloor$, $|D^k f_0(x + y) - D^k f_0(x)|$ can be bounded by $|f_0(x + y) - f_0(x)| Q_3(x) + f_0(x) Q_4(x) \exp(\tau_2 \|y\|^2) \|y\|^{\beta - \lfloor \beta \rfloor}$ for some polynomials Q_3 and Q_4 and a $\tau_2 > 0$. Therefore $f_0 \in \mathcal{C}^{\beta, L, \tau_0}$ for $\tau_0 = \max(\tau_1, \tau_2)$ and $L(x) = f_0(x)\{Q(x) Q_3(x) + Q_4(x)\}$. Because of the tail condition on f_0, for any polynomial $\tilde{Q}$ and $a > 0$ we have $P_0 |\tilde{Q}|^a < \infty$. Thus $P_0 (L/f_0)^{2+\epsilon/\beta} < \infty$ by Hölder's inequality and the assumption on Q. □

REFERENCES

BARRON, A., BIRGÈ, L. & MASSART, P. (1999). Risk bounds for model selection via penalization. *Prob. Theory Rel. Fields* **113**, 301–413.

BELITSER, E. & GHOSAL, S. (2003). Adaptive Bayesian inference on the mean of an infinite-dimensional normal distribution. *Ann. Statist.* **31**, 536–59.

DE JONGE, R. & VAN ZANTEN, H. (2010). Adaptive nonparametric Bayesian inference using location-scale mixture priors. *Ann. Statist.* **38**, 3300–20.

DUNSON, D. B. (2010). Nonparametric Bayes applications to biostatistics. In *Bayesian Nonparametrics*, N. L. Hjort, C. Holmes, P. Müller & S. G. Walker, eds. Cambridge: Cambridge University Press, pp. 223–73.

ESCOBAR, M. D. & WEST, M. (1995). Bayesian density estimation and inference using mixtures. *J. Am. Statist. Assoc.* **90**, 577–88.

FERGUSON, T. S. (1973). Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1**, 209–30.

FERGUSON, T. S. (1983). Bayesian density estimation by mixtures of normal distributions. In *Recent Advances in Statistics*, H. Rizvi & J. Rustagi, eds. New York: Academic Press, pp. 287–302.

GHOSAL, S., GHOSH, J. K. & RAMAMOORTHI, R. V. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *Ann. Statist.* **27**, 143–58.

GHOSAL, S., GHOSH, J. K. & VAN DER VAART, A. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28**, 500–31.

GHOSAL, S., LEMBER, J. & VAN DER VAART, A. (2008). Nonparametric Bayesian model selection and averaging. *Electron. J. Statist.* **2**, 63–89.

GHOSAL, S. & VAN DER VAART, A. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Statist.* **29**, 1233–63.

GHOSAL, S. & VAN DER VAART, A. (2007). Posterior convergence rates of Dirichlet mixtures at smooth densities. *Ann. Statist.* **35**, 697–723.

HOFFMANN, M. & LEPSKI, O. (2002). Random rates in anisotropic regression. *Ann. Statist.* **30**, 325–58.

HUANG, T.-M. (2004). Convergence rates for posterior distributions and adaptive estimation. *Ann. Statist.* **32**, 1556–93.

KRUIJER, W., ROUSSEAU, J. & VAN DER VAART, A. (2010). Adaptive Bayesian density estimation with location-scale mixtures. *Electron. J. Statist.* **4**, 1225–57.

LO, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *Ann. Statist.* **12**, 351–7.

MÜLLER, P., ERKANLI, A. & WEST, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika* **83**, 67–79.

MÜLLER, P. & QUINTANA, F. A. (2004). Nonparametric Bayesian data analysis. *Statist. Sci.* **19**, 95–111.

ROUSSEAU, J. (2010). Rates of convergence for the posterior distributions of mixtures of betas and adaptive nonparametric estimation of the density. *Ann. Statist.* **38**, 146–80.

SCOTT, D. W. (1992). *Multivariate Density Estimation: Theory, Practice and Visualization*. New York: Wiley.

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

VAN DER VAART, A. & VAN ZANTEN, H. (2009). Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *Ann. Statist.* **37**, 2655–75.

WU, Y. & GHOSAL, S. (2010). The L_1-consistency of Dirichlet mixtures in multivariate Bayesian density estimation. *J. Mult. Anal.* **101**, 2411–9.

[*Received November* 2011. *Revised February* 2013]