Bernoulli **27**(2), 2021, 946–975
https://doi.org/10.3150/20-BEJ1261

Asymptotic confidence regions for density ridges

WANLI QIAO

Department of Statistics, George Mason University, Fairfax, VA 22030, USA. E-mail: wqiao@gmu.edu

We develop large sample theory including nonparametric confidence regions for r-dimensional ridges of probability density functions on $\mathbb{R}^d$, where $1 \leq r < d$. We view ridges as the intersections of level sets of some special functions. The vertical variation of the plug-in kernel estimators for these functions constrained on the ridges is used as the measure of maximal deviation for ridge estimation. Our confidence regions for the ridges are determined by the asymptotic distribution of this maximal deviation, which is established by utilizing the extreme value distribution of nonstationary χ-fields indexed by manifolds.

Keywords: Ridges; intersections; level sets; extreme value distribution; kernel density estimation

1. Introduction

A ridge in a data cloud is a low-dimensional geometric feature that generalizes the concept of local modes, in the sense that the density values on ridge points are local maxima constrained in some subspace. In the literature ridges are also called filaments, or filamentary structures, which usually exhibit a network-like pattern. They are widely used to model objects such as fingerprints, fault lines, road systems, and blood vessel networks. The vast amount of modern cosmological data displays a spatial structure called Cosmic Web, and ridges have been used as a mathematical model for galaxy filaments [45].

The statistical study on ridge estimation has recently attracted much attention. See [8,17–19,40]. One of the fundamental notions under ridge estimation is that ridges are sets, and most of the above statistical inference work focuses on the maximal (or global) deviation in ridge estimation, that is, how the estimated ridge approximates the ground truth as a whole. This requires an appropriately chosen measure of global deviation. For example, the Hausdorff distance is used in [8,17–19], while [40] uses the supremum of "trajectory-wise" Euclidean distance between the true and estimated ridge points, where trajectories are driven by the second eigenvectors of Hessian. Both distances measure the deviation of ridge estimation in the space where the sets live in, which we call horizontal variation (HV).

In this paper, we develop large sample theory for the nonparametric estimation of density ridges, which in particular includes the construction of confidence regions for density ridges. Our methodology is based on the measure of global deviation in ridge estimation from a different perspective. Briefly speaking, we treat ridges as intersections of special level sets, and use the measure of maximal deviation in levels, which we call vertical variation (VV).

We first give the mathematical definition of ridges. Let $\nabla f(x)$ and $\nabla^2 f(x)$ be the gradient and Hessian of a twice differentiable probability density function f at $x \in \mathbb{R}^d$ with $d \geq 2$. Let $v_1(x), \ldots, v_d(x)$ be orthonormal eigenvectors of $\nabla^2 f(x)$, with corresponding eigenvalues $\lambda_1(x) \geq \lambda_2(x) \geq \cdots \geq \lambda_d(x)$. For $r = 1, 2, \ldots, d-1$, write $V(x) = (v_{r+1}(x), \ldots, v_d(x))$. The r-ridge $\mathcal{M}^r$ induced by f is defined

1350-7265 © 2021 ISI/BS

as the collection of points x that satisfy the following two conditions:

$$V(x)^T \nabla f(x) = 0, \tag{1.1}$$

$$\lambda_{r+1}(x) < 0. \tag{1.2}$$

We fix $r \geq 1$ in this paper and denote the ridge by $\mathcal{M}$. This definition has been widely used in the literature (see, e.g., [14]). A ridge point x is a local maximum point of f in a $(d-r)$-dimensional subspace spanned by $v_{r+1}(x), \ldots, v_d(x)$. This geometric interpretation can be seen from the fact that $v_i^T \nabla f$ and λ_i are the first and second order directional derivatives of f along v_i, respectively. In fact, if we take $r = 0$, then conditions (1.1) and (1.2) just define the set of local maxima, which is the 0-ridge. Condition (1.1) indicates that an r-ridge is contained in the intersection of $(d-r)$ level sets of the functions $v_i^T \nabla f$, $i = r+1, \ldots, d$, and is an r-dimensional manifold with co-dimension $(d-r)$ under some mild assumptions (e.g., see assumption (F4) below). The confidence regions for the set of modes (0-ridge) can also be constructed by using the VV idea presented in this paper. They need to be treated in a slightly different way because of the discreteness of the sets. The study of confidence regions for modes is not included in this paper for convenience.

Given an i.i.d. sample $X_1, \ldots, X_n$ of f, the ridge $\mathcal{M}$ can be estimated using a plug-in approach based on kernel density estimators (KDE). Let $\widehat{f} \equiv \widehat{f}_{n,h}$ be the KDE of f with bandwidth $h > 0$ (see (2.1)), and let $\widehat{v}_1(x), \ldots, \widehat{v}_d(x)$ be orthonormal eigenvectors of $\nabla^2 \widehat{f}(x)$, with corresponding eigenvalues $\widehat{\lambda}_1(x) \geq \widehat{\lambda}_2(x) \geq \cdots \geq \widehat{\lambda}_d(x)$. Also write $\widehat{V}(x) = (\widehat{v}_{r+1}(x), \ldots, \widehat{v}_d(x))$. Then a plug-in estimator for $\mathcal{M}$ is $\widehat{\mathcal{M}}$, which is the set of points defined by plugging in these kernel estimators into their counterparts in conditions (1.1) and (1.2). See Figure 1 for example. [8,17–19] focus on the estimation of ridges induced by the smoothed kernel density function $f_h \equiv \mathbb{E}\widehat{f}$, instead of the true density f. Such ridges, denoted by $\mathcal{M}_h$, depend on the bandwidth h and are called surrogates. Focusing on $\mathcal{M}_h$ instead of $\mathcal{M}$ avoids the well-known bias issue in nonparametric function and set estimation.

In this paper, we consider confidence regions for both $\mathcal{M}$ and $\mathcal{M}_h$ in the form of

$$\widehat{C}_{n,h}(a_n, b_n) = \left\{x : \sqrt{nh^{d+4}} \left\| Q_n(x) \widehat{V}(x)^T \nabla \widehat{f}(x) \right\| \leq a_n, \text{ and } \widehat{\lambda}_{r+1}(x) < b_n \right\}, \tag{1.3}$$

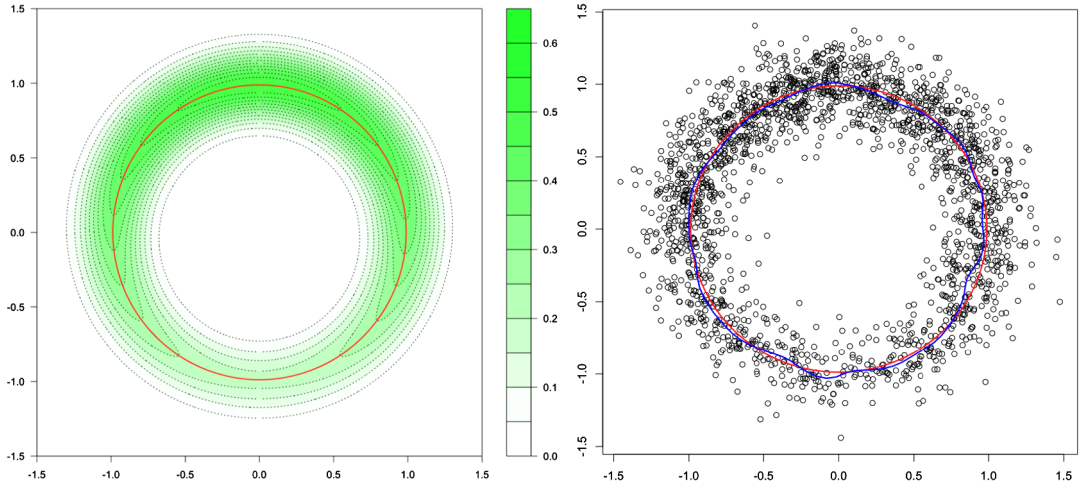


Figure 1. Left: contour plot of a density function, where the red solid curve is a ridge and the dotted lines are contour lines; Right: simulated data points from the density function and the estimated ridge (blue solid curve).

where $a_n > 0$, $b_n \in \mathbb{R}$ and $Q_n(x)$ is a normalizing matrix. Here determining Q_n, a_n and b_n is critical to guarantee that $\widehat{C}_{n,h}(a_n, b_n)$ has a desired asymptotic coverage probability for $\mathcal{M}$ or $\mathcal{M}_h$ as $n \to \infty$ and $h \to 0$. The basic idea for our VV approach is as follows. We consider density ridges as the intersection of the zero-level sets of the functions $V^T \nabla f$ and a sublevel set of λ_{r+1}. When we use plug-in estimators for these functions, we allow their values to vary in a range (specified by a_n and b_n), which implicitly defines a neighborhood near $\widehat{\mathcal{M}}$. The shape of this neighborhood is envisioned as a tube around $\widehat{\mathcal{M}}$ with varying local radii. This tube is geometrically different from the one with constant radius based on the asymptotic distribution of $d_H(\widehat{\mathcal{M}}, \mathcal{M}_h)$, which is the Hausdorff distance (belonging to HV) between $\widehat{\mathcal{M}}$ and $\mathcal{M}_h$. As seen from its definition given in (1.1) and (1.2), ridge estimation mainly involves the estimation of the density gradient and Hessian. Between these two major components, the rate $\sqrt{nh^{d+4}}$ in (1.3) follows from the rate of convergence of the Hessian, which is $1/\sqrt{nh^{d+4}}$ (ignoring the bias). Note that the rate of convergence of the gradient is $1/\sqrt{nh^{d+2}}$, which is much faster than that of the Hessian, and makes the Hessian estimation a dominant component in ridge estimation. We note in passing that this statement also applies to the asymptotic properties of $d_H(\widehat{\mathcal{M}}, \mathcal{M}_h)$ (see [8]).

The asymptotic validity of the confidence regions for $\mathcal{M}_h$ and $\mathcal{M}$ in the form of (1.3) will be shown through the following steps, which are also the main results in the paper. Note that if we write $B_n(x) = \|Q_n(x)\widehat{V}(x)^T \nabla \widehat{f}(x)\|$, then $\mathcal{M}_h \subset \widehat{C}_{n,h}(a_n, b_n)$ is equivalent to $\sqrt{nh^{d+4}} \sup_{x \in \mathcal{M}_h} B_n(x) \le a_n$ and $\sup_{x \in \mathcal{M}_h} \widehat{\lambda}_{r+1}(x) < b_n$. Under some regularity assumptions one can show that

(i) the distribution of $\sqrt{nh^{d+4}} \sup_{x \in \mathcal{M}_h} B_n(x)$ equals that of $\sup_{g \in \mathcal{F}_h} \mathbb{G}_n(g)$ asymptotically, where $\mathbb{G}_n$ is an empirical process and $\mathcal{F}_h$ is a class of functions, which is induced by some linear functionals of the second derivatives of kernel density estimators;
(ii) the distribution of $\sup_{g \in \mathcal{F}_h} \mathbb{G}_n(g)$ is asymptotically the same as that of $\sup_{g \in \mathcal{F}_h} \mathbb{B}(g)$, where $\mathbb{B}$ is a locally stationary Gaussian process indexed by $\mathcal{F}_h$;
(iii) the distribution of $\sup_{g \in \mathcal{F}_h} \mathbb{B}(g)$ is derived by applying the extreme value theory of χ-fields indexed by manifolds developed in our companion work [37].

Then a_n is determined by the above approximations and distributional results and b_n is chosen such that $\sup_{x \in \mathcal{M}_h} \widehat{\lambda}_{r+1}(x) < b_n$ holds with probability tending to one. In fact, one can show that $P(\mathcal{M}_h \subset \widehat{C}_{n,h}(a_n, b_n)) = e^{-e^{-z}} + o(1)$ with $b_n = 0$ and $a_n = \frac{z+c}{\sqrt{2r \log(h^{-1})}} + \sqrt{2r \log(h^{-1})}$, for some $c > 0$ depending on f, K, and $\mathcal{M}_h$. This type of result is similar to the confidence bands for univariate probability density functions developed in the classical work [4]. The derivation for $\mathcal{M}$ is similar except that we have to deal with the bias in the estimation.

The way that we study ridge estimation is naturally connected to the literature of level set estimation (see, e.g., [7,22,31,35,36,46]), which mainly focuses on density functions and regression functions. Confidence regions for level sets have been studied in [9,30,42]. It is clear that technically a ridge is a more sophisticated object to study than a density or regression level set, not only because the former involves the estimation of eigen-decomposition of Hessians and their interplay with gradients, but also a ridge is viewed as the intersection of level sets of multiple functions if $d - r \ge 2$. To our knowledge there are no nonparametric distributional results for the estimation of intersections of density or regression level sets in the literature. In addition to the papers mentioned above, previous work on ridge estimation also includes [2,10,16,21,28,32,48,49].

The paper is organized as follows. We first introduce our notation and assumptions in Section 2. In Section 3.1, we develop the asymptotic confidence regions for $\mathcal{M}_h$ following the procedure listed above. Specifically, steps (i)–(iii) are established in Proposition 3.3, and Theorems 3.4, and 3.7, respectively. In Section 3.2, we use bias correction methods to extend the results to asymptotic confidence regions for $\mathcal{M}$. The confidence regions involve unknown surface integrals on ridges. In Section 3.3,

we show the asymptotic validity of the confidence regions with these unknown quantities replaced by their plug-in estimators. In particular, Corollary 3.10, as our main result from a statistical perspective, gives a data-driven asymptotic confidence regions for ridges. For technical reasons, the consideration of critical points on ridges are deferred until Section 3.4, where we also discuss different choices of b_n. The proofs are given in Section 5 and the supplementary material [38].

2. Notation and assumptions

We first give the notation used in the paper. For a real matrix A and compatible vectors u and v, denote $\langle u, v\rangle_A = u^T Av$. Also we write $\langle u, u\rangle_A = \|u\|_A^2$ and $\|u\|$ is the Euclidian norm of u. Let $\|A\|_F$ be the Frobenius norm of A and $\|A\|_{\max} = \max_{i,j} |a_{ij}|$ where $A = (a_{ij})$. Let A^+ be the Moore-Penrose pseudoinverse of A (see page 36, [29]), which always exists and is unique. For a positive integer m, let $\boldsymbol{I}_m$ be the $m \times m$ identity matrix. For a vector field $W : \mathbb{R}^m \to \mathbb{R}^n$, let $\boldsymbol{R}(W) = \int_{\mathbb{R}^m} W(x)W(x)^T\,dx \in \mathbb{R}^{n\times n}$, assuming the integral is well defined. For a $d \times d$ matrix A, $\mathrm{vec}(A)$ vectorizes A by stacking the columns of A into a $d^2 \times 1$ column vector, while $\mathrm{vech}(A)$ only vectorizes the lower triangular part of A into a $d(d+1)/2 \times 1$ column vector. The duplication matrix D is such that $\mathrm{vec}(A) = D\,\mathrm{vech}(A)$ for a symmetric matrix A. The matrix D does not depend on A and is unique for dimension d (and we have suppressed d in the notation). For example, when $d = 2$ and $A = \left(\begin{smallmatrix} a_{11} & a_{12} \\ a_{12} & a_{22}\end{smallmatrix}\right)$, using the above notation we have

$$\mathrm{vech}(A) = (a_{11}, a_{12}, a_{22})^T, \qquad \mathrm{vec}(A) = (a_{11}, a_{12}, a_{12}, a_{22})^T, \quad \text{and} \quad D = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}^T.$$

For two matrices A and B, let $A \otimes B$ be the Kronecker product between A and B (see page 31 of [29]). For a real symmetric matrix A, let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ be the smallest and largest eigenvalues of A, respectively.

For a smooth function $K : \mathbb{R}^d \to \mathbb{R}$, let ∇K and $\nabla^2 K$ be its gradient and Hessian, respectively, and we denote $d^2K = \mathrm{vech}\,\nabla^2 K$. Let $\mathbb{Z}_+$ be the set of non-negative integers. For $m \in \mathbb{Z}_+$, we use $\mathcal{H}_m$ to denote the m-dimensional normalized Hausdorff measure. Let $\mathcal{B}(x, t) = \{y \in \mathbb{R}^d : \|y - x\| \le t\}$ be the ball centered at x with radius $t > 0$. For a set $M \subset \mathbb{R}^d$ and $\epsilon > 0$, let $M \oplus \epsilon = \bigcup_{x\in M} \mathcal{B}(x, \epsilon)$, which is the ϵ-enlarged set of M. For $m \in \mathbb{Z}_+$, let $\mathbb{S}^m = \{x \in \mathbb{R}^{m+1} : \|x\| = 1\}$ be the unit m-sphere. For any subset $\mathcal{A} \subset \mathbb{R}^d$, let $\mathbf{1}_{\mathcal{A}}$ be the indicator function of $\mathcal{A}$. Let $\mathrm{int}(\mathcal{A})$ and $\partial\mathcal{A}$ be the interior and boundary of $\mathcal{A}$, respectively.

Given an i.i.d. sample $X_1, \ldots X_n$ from the probability density function f on $\mathbb{R}^d$, denote the kernel density estimator

$$\widehat{f}(x) = \widehat{f}_{n,h}(x) = \frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \quad x \in \mathbb{R}^d, \tag{2.1}$$

where $h > 0$ is a bandwidth and K is a twice differentiable kernel density function on $\mathbb{R}^d$. The notation h is used as a default bandwidth unless otherwise indicated, and we suppress the subscripts n, h in the kernel density estimator and all the quantities induced by it (so that $\widehat{V} = \widehat{V}_{n,h}$ and $\widehat{\lambda}_{r+1} = \widehat{\lambda}_{r+1,n,h}$, for example). Let $f_h(x) = \mathbb{E}\widehat{f}(x)$ and let $v_{1,h}(x), \ldots, v_{d,h}(x)$ be orthonormal eigenvectors of $\nabla^2 f_h(x)$, with corresponding eigenvalues $\lambda_{1,h}(x) \ge \lambda_{2,h}(x) \ge \cdots \ge \lambda_{d,h}(x)$. Also write $V_h(x) = (v_{r+1,h}(x), \ldots, v_{d,h}(x))$. We focus on ridge estimation on a compact subset $\mathcal{H}$ of $\mathbb{R}^d$, which is assumed to be the hypercube $[0, 1]^d$ for simplicity, and all the ridge definitions $\mathcal{M}$, $\widehat{\mathcal{M}}$ and $\mathcal{M}_h$ are restricted on $\mathcal{H}$, such as $\mathcal{M}_h = \{x \in \mathcal{H} : V_h(x)^T \nabla f_h(x) = 0, \lambda_{r+1,h}(x) < 0\}$.

For $\gamma = (\gamma_1, \ldots, \gamma_d)^T \in \mathbb{Z}_+^d$, let $|\gamma| = \gamma_1 + \cdots + \gamma_d$. For a function $g : \mathbb{R}^d \to \mathbb{R}$ with $|\gamma|$th partial derivatives, define

$$g^{(\gamma)}(x) = \frac{\partial^{|\gamma|}}{\partial^{\gamma_1} x_1 \cdots \partial^{\gamma_d} x_d} g(x), \quad x \in \mathbb{R}^d. \tag{2.2}$$

Let $\boldsymbol{R} := \boldsymbol{R}(d^2 K)$. For $\delta > 0$, define

$$\mathcal{N}_\delta(\mathcal{M}) = \left\{ x \in \mathcal{H} : \left\| \nabla f(x)^T V(x) \right\| \le \delta, \lambda_{r+1}(x) < 0 \right\}. \tag{2.3}$$

For a bandwidth $h > 0$, let $\gamma_{n,h}^{(k)} = \sqrt{\frac{\log n}{n h^{d+2k}}}$, which is the rate of convergence of $\sup_{x \in \mathbb{R}^d} |\widehat{f^{(\gamma)}}(x) - f_h^{(\gamma)}(x)|$ for $|\gamma| = k \in \mathbb{Z}_+$ under standard assumptions. We use the following assumptions in the construction of confidence regions for ridges.

Assumptions:

(F1) f is four times continuously differentiable on $\mathcal{H} \oplus \eta_0$ for some $\eta_0 > 0$.

(F2) There exists $\delta_0 > 0$ such that $\mathcal{N}_{\delta_0}(\mathcal{M}) \subset \mathrm{int}(\mathcal{H})$ and the following is satisfied. For all $x \in \mathcal{N}_{\delta_0}(\mathcal{M})$, the smallest $d - r$ eigenvalues of $\nabla^2 f(x)$ are simple, i.e., $\lambda_r(x) > \lambda_{r+1}(x) > \cdots > \lambda_d(x)$. In particular, $\lambda_r(x) > \lambda_{r+1}(x)$ for all $x \in \mathcal{H}$.

(F3) $\{x \in \mathcal{H} : \lambda_{r+1}(x) = 0, V(x)^T \nabla f(x) = 0\} = \varnothing$.

(F4) When $d - r = 1$, we require that $\|\nabla(\nabla f(x)^T v_d(x))\| > 0$ for all $x \in \mathcal{N}_{\delta_0}(\mathcal{M})$; When $d - r \ge 2$, we require that $\nabla(\nabla f(x)^T v_i(x))$, $i = r+1, \ldots, d$ are linearly independent for all $x \in \mathcal{N}_{\delta_0}(\mathcal{M})$.

(K1) The kernel function K is a spherically symmetric probability density function on $\mathbb{R}^d$ with $\mathcal{B}(0, 1)$ as its support. It has continuous partial derivatives up to order 4.

(K2) For any open ball $\mathcal{S}$ with positive radius contained in $\mathcal{B}(0, 1)$, the coordinate functions of $s \mapsto \mathbf{1}_{\mathcal{S}}(s) d^2 K(s)$ are linearly independent as functions.

(K3) If $d = 2$, we require that $a_K := \frac{\int_{\mathbb{R}^d} [K^{(\rho_1)}(s)]^2 \, ds}{\int_{\mathbb{R}^d} [K^{(\rho_2)}(s)]^2 \, ds} > 1$, where $\rho_1 = (3, 0, \ldots, 0)^T \in \mathbb{Z}_+^d$ and $\rho_2 = (2, 1, 0, \ldots, 0)^T \in \mathbb{Z}_+^d$; If $d \ge 3$, we require $\frac{1}{a_K} \le b_K := \frac{\int_{\mathbb{R}^d} [K^{(\rho_3)}(s)]^2 \, ds}{\int_{\mathbb{R}^d} [K^{(\rho_2)}(s)]^2 \, ds} < 1$, where $\rho_3 = (1, 1, 1, 0, \ldots, 0)^T \in \mathbb{Z}_+^d$.

Remark 2.1.

(i) Note that ridges are defined using the second derivatives of densities. Assumption (F1) requires the existence of two additional orders of derivatives. This is similar to other work on the distributional results of ridge estimation (see [8,40]).

(ii) Assumptions (F2)–(F3) exclude some scenarios that are on the boundary of the class of density functions we consider (note that these assumptions only exclude some equalities). Here we give some brief discussion of the implications of these assumptions.
   (a) Assumption (F2) requires that the smallest $d - r$ eigenvalues of $\nabla^2 f(x)$ for $x \in \mathcal{M}$ all have multiplicity one, in order to have the differentiability of the functions $v_i(x)^T \nabla f(x)$ for $i = r+1, \ldots, d$, which generally does not hold when the eigenvalues are repeated.
   (b) Assumption (F3) avoids the existence of some degenerate ridge points. Such points have zero first and second directional derivatives along v_{r+1} and so they are almost like ridge points. This assumption has been used in [8,19,40].

(iii) When $d - r = 1$, assumption (F4) is related to the margin assumption in the literature of level set estimation [35]. In addition, as we consider ridges as intersections of level sets when $d - r \ge 2$,

this assumption guarantees the transversality of the intersecting manifolds. Assumption (F4) holds, for example, if f satisfies assumptions (A1) and (P1) in [8], which require $\lambda_2(x) \le -\beta_1$, $\lambda_1(x) \ge \beta_0 - \beta_1$ and $\|\nabla f_h(x)\| \max_{|\alpha|=3} |f_h^{(\alpha)}(x)| \le \beta_0(\beta_1 - \beta_2)$ for some constants $\beta_0, \beta_1, \beta_2 > 0$, for all x in a neighborhood of $\mathcal{M}_h$ (see their Lemma 2).

(iv) Assumptions (K1)–(K3) are for the kernel function K. In particular (K2) can guarantee that $\boldsymbol{R}$ is positive definite. In general one can show that $a_K \ge 1 \ge b_K$ (see Lemma B.1 in the supplementary material [38]). In fact, if we assume that $K^{(\rho_1)}$, $K^{(\rho_2)}$ and $K^{(\rho_3)}$ are linearly independent as functions, then the condition $a_K > 1 > b_K$ required in (K3) is satisfied. This can be easily seen from the proof of Lemma B.1. One can show that the following kernel density function is an example that satisfies (K1)–(K3):

$$K(x) = c_d\big(1 - \|x\|^2\big)^5 \mathbf{1}_{\mathcal{B}(0,1)}(x), \quad x \in \mathbb{R}^d,$$

where c_d is a normalizing constant.

3. Main results

In the literature, the following assumption or even stronger ones are used to get distributional results for ridge estimation, for example, assumption (F7) of [40].

(F5) $\|\nabla f(x)\| \ne 0$, for all $x \in \mathcal{M}$.

In other words, it is assumed that $\mathcal{M}$ does not contain any critical points of f. This assumption excludes many important scenarios in practice because (F5) implies that f does not have local modes on $\mathcal{H}$.

Our confidence regions for $\mathcal{M}_h$ and $\mathcal{M}$ eventually do not require assumption (F5). But the critical points and regular points on ridges need to be treated in different ways, because for critical points the estimation is mainly determined by the gradient of f, while the estimation of regular ridge points depends on both the gradient and Hessian. It is known that the estimation of Hessian has a slower rate of convergence than the critical points using kernel type estimators, which results in different behaviors of regular ridge and critical points. To deal with this issue, the strategy we use is to construct confidence regions for the set of critical points and regular ridge points individually and then combine them (see Section 3.4). For convenience we will first exclude critical points from our consideration and tentatively assume (F5).

3.1. Asymptotic confidence regions for $\mathcal{M}_h$

Given any $0 < \alpha < 1$, we first study how to determine a_n and b_n to make $\widehat{C}_{n,h}(a_n, b_n)$ an asymptotic $100(1-\alpha)\%$ confidence region for $\mathcal{M}_h$. The following lemma shows some basic properties of $\mathcal{M}$ as well as $\mathcal{M}_h$. For any subset $\mathcal{L} \subset \mathbb{R}^d$ and $x \in \mathbb{R}^d$, let $d(x, \mathcal{L}) = \inf_{y \in \mathcal{L}} \|x - y\|$. A point $u \in \mathcal{L}$ is called a normal projection of x onto $\mathcal{L}$ if $\|x - u\| = d(x, \mathcal{L})$. For $x \in \mathcal{L}$, let $\Delta(\mathcal{L}, x)$ denote the *reach* of $\mathcal{L}$ at x (see [15]), which is the largest $r \ge 0$ such that each point in $\mathcal{B}(x, r)$ has a unique normal projection onto $\mathcal{L}$. The reach of $\mathcal{L}$ is defined as $\Delta(\mathcal{L}) := \inf_{x \in \mathcal{L}} \Delta(\mathcal{L}, x)$. If $\mathcal{L}$ is a manifold, its reach reflects the curvature of $\mathcal{L}$ and the width of its nearly self-intersecting structure [1,3,15].

Lemma 3.1. *Under assumptions* (F1)–(F4) *and* (K1), *we have*

(i) $\mathcal{M}$ *is an* r*-dimensional compact manifold without boundary and with* $\Delta(\mathcal{M}) > 0$.

When h is small enough, we have

(ii) *for any fixed $0 < \delta \le \delta_0$, with δ_0 given in* (F2), *we have $\mathcal{M}_h \subset \mathcal{N}_\delta(\mathcal{M})$, where $\mathcal{N}_\delta(\mathcal{M})$ is defined in* (2.3);

(iii) $\inf_{x\in\mathcal{M}_h}[\lambda_{j-1,h}(x) - \lambda_{j,h}(x)] > \beta_0$, $j = r+1,\dots,d$, *and* $\sup_{x\in\mathcal{M}_h}\lambda_{r+1,h}(x) < -\beta_0$ *for some constant $\beta_0 > 0$ that does not depend on h*;

(iv) *$\mathcal{M}_h$ is an r-dimensional compact manifold without boundary and with $\Delta(\mathcal{M}_h) > \beta_1$ for some constant $\beta_1 > 0$ that does not depend on h.*

Remark 3.1. Property (iii) states that $\lambda_{r+1,h}$ is uniformly bounded away from zero on $\mathcal{M}_h$. As we show in Lemma A.1 in the supplementary material [38], $\widehat{\lambda}_{r+1}$ is a strongly uniform consistent estimator of $\lambda_{r+1,h}$ under our assumptions, that is, $\sup_{x\in\mathcal{H}}|\widehat{\lambda}_{r+1}(x) - \lambda_{r+1,h}(x)| = o(1)$ almost surely, which implies that with probability one $\widehat{\lambda}_{r+1}$ has the same sign as $\lambda_{r+1,h}$ on $\mathcal{M}_h$ for large n. This allows us to use $b_n = 0$ in $\widehat{C}_{n,h}(a_n, b_n)$, and focus on the behavior of $\widehat{V}(x)^T\nabla\widehat{f}(x)$ on $\mathcal{M}_h$ to choose a_n so that $\widehat{C}_{n,h}(a_n, b_n)$ in (1.3) is an asymptotic confidence region for $\mathcal{M}_h$. Also see Section 3.4 for different choices of b_n.

Note that $V_h(x)^T\nabla f_h(x) = 0$ for all $x \in \mathcal{M}_h$ by the definition of ridges. We need to study the behavior of $\widehat{V}(x)^T\nabla\widehat{f}(x) = \widehat{V}(x)^T\nabla\widehat{f}(x) - V_h(x)^T\nabla f_h(x)$ for $x \in \mathcal{M}_h$. The following proposition shows the asymptotic normality of this difference, which can be uniformly approximated by a linear form of $d^2\widehat{f}(x) - d^2 f_h(x)$. This is not surprising because the difference depends on the estimation of eigenvectors of the Hessian, which has a slower rate of convergence than the estimation of the gradient. Note that each unit eigenvector has two possible directions. Without loss of generality, for $i = r+1,\dots,d$, suppose that we fix the orientations of $\widehat{v}_i(x)$, $v_{i,h}(x)$ and $v_i(x)$ in such a way that they vary continuously for x in a neighborhood of $\mathcal{M}$ and have pairwise acute angles. By treating the ith unit eigenvector as a vector-valued function of $d \times d$ symmetric matrices, the application of matrix calculus (see [29]) gives the following first order approximation:

$$\widehat{v}_i(x) - v_{i,h}(x) \approx \Xi_i(x)\operatorname{vec}\bigl[\nabla^2\widehat{f}(x) - \nabla^2 f_h(x)\bigr] = \Xi_i(x)D\bigl[d^2\widehat{f}(x) - d^2 f_h(x)\bigr],$$

where $\Xi_i(x) = v_i(x)^T \otimes (\lambda_i \boldsymbol{I}_d - \nabla^2 f(x))^+$ is a $d \times d^2$ matrix representing the first derivatives in the linear approximation, and D is the duplication matrix defined in Section 2. By ignoring the error caused by the gradient estimation, which has a faster rate than the Hessian estimation, we approximately have that for $x \in \mathcal{M}_h$ and $i = r+1,\dots,d$,

$$\begin{aligned}\widehat{v}_i(x)^T\nabla\widehat{f}(x) &= \widehat{v}_i(x)^T\nabla\widehat{f}(x) - v_{i,h}(x)^T\nabla f_h(x)\\ &\approx \bigl[\widehat{v}_i(x) - v_{i,h}(x)\bigr]^T\nabla f(x)\\ &\approx m_i(x)^T\bigl[d^2\widehat{f}(x) - d^2 f_h(x)\bigr],\end{aligned} \tag{3.1}$$

where $m_i(x) = D^T\Xi_i(x)^T\nabla f(x)$. It turns out that for $i = r+1,\dots,d$, $m_i(x)$ has the following form given by

$$m_i(x) = D^T\left(v_i(x) \otimes \sum_{j=1}^{r}\left[\frac{v_j(x)^T\nabla f(x)}{\lambda_i(x) - \lambda_j(x)}v_j(x)\right]\right), \tag{3.2}$$

which are $d(d+1)/2$ dimensional column vectors. Let $M(x) = (m_{r+1}(x),\dots m_d(x))$, which is a $[d(d+1)/2] \times (d-r)$ matrix. The following result shows the asymptotic behavior of $\widehat{V}(x)^T\nabla\widehat{f}(x)$ on

$\mathcal{M}_h$ with a first-order approximation, where the matrix $M(x)^T$ can be viewed as the Jacobian matrix with respect to the perturbation of the Hessian of the density at x.

Proposition 3.2. *Under assumptions* (F1)–(F5), (K1), *and* (K2), *as* $\gamma_{n,h}^{(2)} \to 0$ *and* $h \to 0$, *we have*

$$\sup_{x\in\mathcal{M}_h} \big\| \widehat{V}(x)^T \nabla \widehat{f}(x) - M(x)^T \big[d^2 \widehat{f}(x) - d^2 f_h(x) \big] \big\| = O_p\big(\gamma_{n,h}^{(1)} + \big(\gamma_{n,h}^{(2)}\big)^2\big), \tag{3.3}$$

and there exists a constant $\delta_1 \in (0, \delta_0]$ *such that for all* $x \in \mathcal{N}_{\delta_1}(\mathcal{M})$,

$$\sqrt{nh^{d+4}} M(x)^T \big[d^2 \widehat{f}(x) - d^2 f_h(x) \big] \xrightarrow{D} \mathcal{N}_{d-r}\big(0,\, f(x)\Sigma(x)\big), \quad \text{as } n \to \infty, \tag{3.4}$$

where $\Sigma(x) = M(x)^T \boldsymbol{R} M(x)$ *is a positive definite matrix for all* $x \in \mathcal{N}_{\delta_1}(\mathcal{M})$, *and* $x \in \mathcal{M}_h$, *when* h *is small enough.*

Remark 3.2. The result in (3.4), especially the form of $\Sigma(x)$ in the variance, is a direct consequence of Theorem 3 of [13], which says

$$\sqrt{nh^{d+4}} \big[d^2 \widehat{f}(x) - d^2 f_h(x) \big] \xrightarrow{D} \mathcal{N}_{d(d+1)/2}\big(0,\, f(x)\boldsymbol{R}\big), \quad \text{as } n \to \infty. \tag{3.5}$$

For a positive definite matrix A, let $A^{1/2}$ be its square root such that $A^{1/2}$ is also positive definite and $A = A^{1/2}A^{1/2}$. It is known that $A^{1/2}$ is uniquely defined. The asymptotic normality result in (3.4) suggests that we can standardize $\widehat{V}(x)^T \nabla \widehat{f}(x)$ by left multiplying the matrix $Q(x) := [f(x)\Sigma(x)]^{-1/2}$, which is unknown and can be further estimated by a plug-in estimator $Q_n(x) := [\widehat{f}(x)\widehat{\Sigma}(x)]^{-1/2}$ as specified below. Let $\widehat{\Sigma}(x) = \widehat{M}(x)^T \boldsymbol{R} \widehat{M}(x)$ with $\widehat{M}(x) = (\widehat{m}_{r+1}(x), \ldots \widehat{m}_d(x))$, where

$$\widehat{m}_i(x) = D^T \left(\widehat{v}_i(x) \otimes \sum_{j=1}^{r} \left[\frac{\widehat{v}_j(x)^T \nabla \widehat{f}(x)}{\widehat{\lambda}_i(x) - \widehat{\lambda}_j(x)} \widehat{v}_j(x) \right] \right).$$

We can show that $Q_n(x)$ is a consistent estimator of $Q(x)$ for all $x \in \mathcal{N}_{\delta_1}(\mathcal{M})$ (see the proof of Proposition 3.3 in the supplementary material [38]), for δ_1 given in Proposition 3.2. Then in view of Proposition 3.2, for any $x \in \mathcal{M}_h$, $Q_n(x)\widehat{V}(x)^T \nabla \widehat{f}(x)$ asymptotically behaves like a $(d-r)$-dimensional standard normal random vector. In fact, the distribution of $\sup_{x\in\mathcal{M}_h} \| Q_n(x)\widehat{V}(x)^T \nabla \widehat{f}(x) \|$ can be approximated by the extreme value distributions of a sequence of Gaussian random fields. The standardization by using $Q_n(x)$ is related to the appearance of surface integrals over $\mathcal{M}_h$ in these extreme value distributions (see (3.22) and Theorem 3.7 below). Heuristically, the surface integral stands for the summation of the contribution of the standard normal random variables indexed by all the points on $\mathcal{M}_h$. An alternative approach, which is not pursued here, is to directly consider the distribution of $\sup_{x\in\mathcal{M}_h} \| \widehat{V}(x)^T \nabla \widehat{f}(x) \|$ (without standardization), which can be approximated by the extreme value distributions of Gaussian random fields with varying variances. Typically, for this type of Gaussian random fields, the asymptotic extreme value distributions are only related to the behaviors of the Gaussian random fields at the locations where the maximum variance is achieved, instead of the behaviors over the entire index set. See [26], for example. We only consider the approach with standardization, which requires the estimation of surface integrals (see Section 3.3) to construct confidence regions for ridges. It is expected that the other approach without standardization involves the estimation of the modes of the variance functions of the approximating Gaussian random fields.

Let

$$B_n(x) = \big\| Q_n(x)\widehat{V}(x)^T \nabla \widehat{f}(x) \big\| = \big\| \widehat{V}(x)^T \nabla \widehat{f}(x) \big\|_{[\widehat{f}(x)\widehat{\Sigma}(x)]^{-1}}. \tag{3.6}$$

We consider the following form of confidence regions for $\mathcal{M}_h$, which is slightly more formal than (1.3). For any $a_n \geq 0$ and $b_n \in \mathbb{R}$, let

$$\widehat{C}_{n,h}(a_n, b_n) = \big\{ x \in \mathcal{H} : \sqrt{nh^{d+4}} B_n(x) \leq a_n, \text{ and } \widehat{\lambda}_{r+1}(x) < b_n \big\}. \tag{3.7}$$

We first consider $b_n = 0$ for the reason given in Remark 3.1 and for simplicity write $\widehat{C}_{n,h}(a_n) = \widehat{C}_{n,h}(a_n, 0)$. For any $\alpha \in (0,1)$, we want to find a sequence $a_{n,h,\alpha}$ such that $\mathbb{P}(\mathcal{M}_h \subset \widehat{C}_{n,h}(a_{n,h,\alpha})) \to 1-\alpha$, that is, $\widehat{C}_{n,h}(a_{n,h,\alpha})$ is an asymptotic $100(1-\alpha)\%$ confidence region for $\mathcal{M}_h$. Let

$$D_n(x) = \big\| Q(x)M(x)^T \big(d^2 \widehat{f}(x) - d^2 f_h(x)\big) \big\|. \tag{3.8}$$

The following proposition indicates that the behaviors of the suprema of $B_n(x)$ and $D_n(x)$ on $\mathcal{M}_h$ are close, and hence $a_{n,h,\alpha}$ can be determined by the distribution of $\sqrt{nh^{d+4}} \sup_{x\in\mathcal{M}_h} D_n(x)$.

Proposition 3.3. *Under assumptions* (F1)–(F5), (K1), *and* (K2), *as* $\gamma_{n,h}^{(2)} \to 0$ *and* $h \to 0$, *we have*

$$\sup_{x\in\mathcal{M}_h} D_n(x) = O_p\big(\gamma_{n,h}^{(2)}\big), \tag{3.9}$$

$$\sup_{x\in\mathcal{M}_h} B_n(x) - \sup_{x\in\mathcal{M}_h} D_n(x) = O_p\Big(\big(\gamma_{n,h}^{(2)}\big)^2 + \gamma_{n,h}^{(1)}\Big). \tag{3.10}$$

Remark 3.3. When $r = 1$, for $i = 2, \ldots, d$ and $x \in \mathcal{M}$, $m_i(x)$ in (3.2) can be simplified to

$$m_i(x) = \frac{\|\nabla f(x)\|}{\lambda_i(x) - \lambda_1(x)} D^T \big(v_i(x) \otimes v_1(x)\big).$$

Correspondingly, we can replace $\widehat{m}_i(x)$ in $B_n(x)$ by $\widetilde{m}_i(x) = \frac{\|\nabla \widehat{f}(x)\|}{\widehat{\lambda}_i(x) - \widehat{\lambda}_1(x)} D^T (\widehat{v}_i(x) \otimes \widehat{v}_1(x))$, and the conclusion in this proposition is not changed, following the same proof of this proposition and the fact that the Hausdorff distance between $\mathcal{M}$ and $\mathcal{M}_h$ is of the order $O(h^2)$ (see Lemma 3.11 below).

We need to find the asymptotic distribution of $\sqrt{nh^{d+4}} \sup_{x\in\mathcal{M}_h} D_n(x)$. In particular, we will show that for any $z \in \mathbb{R}$ there exists β_h such that,

$$\mathbb{P}\Big\{ \sqrt{2\log\big(h^{-1}\big)} \Big(\sqrt{nh^{d+4}} \sup_{x\in\mathcal{M}_h} D_n(x) - \beta_h \Big) \leq z \Big\} \to e^{-e^{-z}}.$$

To this end, we will represent $\sqrt{nh^{d+4}} D_n(x)$ as an empirical process and approximate its supremum by the extreme value of a Gaussian process defined on a class of functions.

For any $z \in \mathbb{R}^{d-r} \backslash \{0\}$, let $A(x,z) = M(x)Q(x)z$. Notice that $\sqrt{f(x)} \|A(x,z)\|_{\boldsymbol{R}} = \|z\|$. Let $g_{x,z}(\cdot) = \frac{1}{\sqrt{h^d}} \langle A(x,z), d^2 K(\frac{x-\cdot}{h}) \rangle$, and define the class of functions

$$\mathcal{F}_h = \big\{ g_{x,z}(\cdot) : x \in \mathcal{M}_h, z \in \mathbb{S}^{d-r-1} \big\}. \tag{3.11}$$

Consider the local empirical process $\{\mathbb{G}_n(g_{x,z}) : g_{x,z} \in \mathcal{F}_h\}$, where

$$\mathbb{G}_n(g_{x,z}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \big[g_{x,z}(X_i) - \mathbb{E} g_{x,z}(X_1) \big].$$

Due to the elementary result $\|v\| = \sup_{z \in \mathbb{S}^{d-r-1}} v^T z$ for any $v \in \mathbb{R}^{d-r}$, we can write $\sqrt{nh^{d+4}} D_n(x) = \sup_{z \in \mathbb{S}^{d-r-1}} \mathbb{G}_n(g_{x,z})$. Hence

$$\sqrt{nh^{d+4}} \sup_{x \in \mathcal{M}_h} D_n(x) = \sup_{g_{x,z} \in \mathcal{F}_h} \mathbb{G}_n(g_{x,z}). \tag{3.12}$$

Using similar arguments as given in [11], the supremum of the empirical process in (3.12) can be approximated by the supremum of a Gaussian process, as shown in the following theorem. Let $\mathbb{B}$ be a centered Gaussian process on $\mathcal{F}_h$ such that for all $g_{x,z}, g_{\tilde{x},\tilde{z}} \in \mathcal{F}_h$,

$$\mathbb{E}\big(\mathbb{B}(g_{x,z}) \mathbb{B}(g_{\tilde{x},\tilde{z}})\big) = \operatorname{Cov}\big(g_{x,z}(X_1), g_{\tilde{x},\tilde{z}}(X_1)\big).$$

Theorem 3.4. *Under assumptions* (F1)–(F5), (K1), *and* (K2), *as* $\gamma_{n,h}^{(0)} \log^4 n \to 0$ *and* $h \to 0$ *we have*

$$\sup_{t>0} \Big| \mathbb{P}\Big(\sqrt{nh^{d+4}} \sup_{x \in \mathcal{M}_h} D_n(x) < t \Big) - \mathbb{P}\Big(\sup_{g \in \mathcal{F}_h} \mathbb{B}(g) < t \Big) \Big| = o(1). \tag{3.13}$$

Remark 3.4.

(i) In the derivation of the asymptotic distribution of the maximal deviation of density function estimation, [4] uses a sequence of Gaussian approximations. When extending the idea to multivariate density function estimation, [44] imposes an assumption that requires f to be d times continuously differentiable in order to use the Rosenblatt transformation (see [43]). This type of assumption is further used in related work for Gaussian approximation to maximal deviation in multivariate regression function estimation (see [25]). In fact, if one is willing to impose a similar assumption in our context (that is, f is $d+2$ times continuously differentiable, because ridges are defined using up to the second derivatives of f), then it can be verified that the Gaussian process $\mathbb{B}(g_{x,z})$ has the following representation:

$$\mathbb{B}(g_{x,z}) \overset{D}{=} \int_{\mathbb{R}^d} g_{x,z}(s) \, d\boldsymbol{B}\big(M(s)\big),$$

where $\boldsymbol{B}$ is the d-dimensional Brownian bridge, and M is the Rosenblatt transformation. In fact, by using a sequence of Gaussian approximations similar to those given in [44], we can show the following approximation holds.

$$\mathbb{B}(g_{x,z}) \overset{D}{\approx} \sqrt{f(x)} \int_{\mathbb{R}^d} g_{x,z}(s) \, d\boldsymbol{W}(s) =: U(x,z), \tag{3.14}$$

where $\boldsymbol{W}$ is the d-dimensional Wiener process. Instead of following the approach in [44], we directly find out the limiting extreme value distribution of $\mathbb{B}(g_{x,z})$, which is shown to be locally stationary (see Definition 3.1). This allows us to use a less stringent smoothness condition on f.

(ii) Let $w_x(\cdot) = \frac{1}{\sqrt{h^d}} Q(x)^T M(x)^T d^2 K(\frac{x-\cdot}{h})$, so that $g_{x,z}(\cdot) = z^T w_x(\cdot)$. Note that here the scaling factor $\frac{1}{\sqrt{h^d}}$ can guarantee that $\operatorname{Var}(w_x(X_1)) = \boldsymbol{I}_{d-r} + o(1)$ as $h \to 0$. Also let $S_h(x) =$

$(S_{1,h}(x), \ldots, S_{d-r,h}(x))^T$ be a vector of centered Gaussian random fields indexed by $\mathcal{M}_h$ such that $\mathbb{E}(S_h(x)S_h(\widetilde{x})^T) = \mathrm{Cov}(w_x(X_1), w_{\widetilde{x}}(X_1))$, for $x, \widetilde{x} \in \mathcal{M}_h$. Then it is clear that $\sup_{g\in\mathcal{F}_h} \mathbb{B}(g) = \sup_{x\in\mathcal{M}_h} \|S_h(x)\|$, where approximately $\|S_h(x)\|^2 \sim \chi^2_{d-r}$ for any $x \in \mathcal{M}_h$, because $\mathrm{Var}(S_h(x)) = \boldsymbol{I}_{d-r} + o(1)$, i.e., $S_{1,h}(x), \ldots, S_{d-r,h}(x)$ are asymptotically independent when $d - r \geq 2$. Note the standardization in $S_h(x)$ is only pointwise, and if $x - \widetilde{x} = o(h)$ then $S_{i,h}(x)$ and $S_{j,h}(\widetilde{x})$ are asymptotically dependent in general for $i \neq j$ when $d - r \geq 2$. Overall $\|S_h(x)\|^2$ is approximately a χ^2 field indexed by $\mathcal{M}_h$, as a sum of squares of Gaussian fields with *cross dependence*, whereas independence of the Gaussian fields is usually assumed in the literature of extreme value theory for χ^2 fields (see, e.g., [33]). This dependence structure has an effect on the form of the final extreme value distribution result (see Remark 3.6 below).

The confidence region for $\mathcal{M}_h$ that we seek relies on the asymptotic distribution of $\sup_{g\in\mathcal{F}_h} \mathbb{B}(g)$, for which we need the following definition and probability result. Suppose that n_1 and n_2 are positive integers and $0 < \alpha_1, \alpha_2 \leq 2$.

Definition 3.1 (Local equi-$(\alpha_1, D^{(1)}_{t,v}, \alpha_2, D^{(2)}_{t,v})$-stationarity). Let $\{Z_h(t, v), (t, v) \in \mathcal{S}_{h,1} \times \mathcal{S}_{h,2}\}_{h\in\mathbb{H}}$ be a class of random fields, where $\mathbb{H}$ is an index set, and $\mathcal{S}_{h,i}$ is a compact subset of $\mathbb{R}^{n_i}$ for $i = 1, 2$. We say that this class is locally equi-$(\alpha_1, D^{(1)}_{t,v}, \alpha_2, D^{(2)}_{t,v})$-stationary, if the following conditions hold. For any $t \in \mathcal{S}_{h,1}$, $v \in \mathcal{S}_{h,2}$ and $h \in \mathbb{H}$, there exist non-degenerate matrices $D^{(1)}_{t,v}$ and $D^{(2)}_{t,v}$ such that for $t_1, t_2 \in \mathcal{S}_{h,1}$ and $v_1, v_2 \in \mathcal{S}_{h,2}$, as $\max\{\|t_1 - t\|, \|t_2 - t\|\}/h \to 0$ and $\max\{\|v_1 - v\|, \|v_2 - v\|\} \to 0$,

$$\text{(i)}\quad \mathrm{Cov}\big(Z_h(t_1, v_1), Z_h(t_2, v_2)\big) = 1 - \left[\left\|\frac{1}{h} D^{(1)}_{t,v}(t_1 - t_2)\right\|^{\alpha_1} + \left\|D^{(2)}_{t,v}(v_1 - v_2)\right\|^{\alpha_2}\right]\big(1 + o(1)\big),$$

uniformly in $t \in \mathcal{S}_{h,1}$, $v \in \mathcal{S}_{h,2}$, $h \in \mathbb{H}$, and

$$\text{(ii)}\quad 0 < \inf \lambda_{\min}\big(\big[D^{(i)}_{t,v}\big]^T D^{(i)}_{t,v}\big) \leq \sup \lambda_{\max}\big(\big[D^{(i)}_{t,v}\big]^T D^{(i)}_{t,v}\big) < \infty, \quad i = 1, 2,$$

where the infimum and supremum are taken over $(t, v) \in \mathcal{S}_{h,1} \times \mathcal{S}_{h,2}$, and $h \in \mathbb{H}$.

We consider $1 \leq r_1 < n_2$ and $1 \leq r_2 < n_2$ below. Let $H^{(r_i)}_{\alpha_i}$, $i = 1, 2$ be the generalized Pickands' constant of Gaussian fields (see the appendix of [37]). The following result is given as Theorem 3.1 in our companion work [37]. For the convenience of the reader, we also give a very brief sketch of proof in Appendix B in the supplementary material [38]. For a differentiable submanifold $\mathcal{S}$ of $\mathbb{R}^d$, at each $u \in \mathcal{S}$, let $T_u\mathcal{S}$ denote the tangent space of $\mathcal{S}$ at u. Let $\Lambda(T_u\mathcal{S})$ be a matrix with orthonormal columns that span $T_u\mathcal{S}$, that is, the orthogonal projection matrix onto $T_u\mathcal{S}$. For an $n \times r$ matrix M with $r \leq n$, we denote by $\|M\|^2_r$ the sum of squares of all minor determinants of order r.

Theorem 3.5. *With some fixed $h_0 \in (0, 1)$, for $0 < h \leq h_0$ and $i = 1, 2$, let $\mathcal{M}^{(i)}_h$ be an r_i-dimensional compact submanifold of $\mathbb{R}^{n_i}$ with $\inf_{0<h\leq h_0} \Delta(\mathcal{M}^{(i)}_h) > 0$, and $0 < \inf_{0<h\leq h_0} \mathcal{H}_{r_i}(\mathcal{M}^{(i)}_h) \leq \sup_{0<h\leq h_0} \mathcal{H}_{r_i}(\mathcal{M}^{(i)}_h) < \infty$. Let $\{Z_h(t, v) : (t, v) \in \mathcal{M}^{(1)}_h \times \mathcal{M}^{(2)}_h\}_{h\in(0,h_0]}$ be a class of centered locally equi-$(\alpha_1, D^{(1)}_{t,v}, \alpha_2, D^{(2)}_{t,v})$-stationary Gaussian random fields with $0 < \alpha_1, \alpha_2 \leq 2$, and all the components of $D^{(i)}_{t,v}$ continuous in t and v. For $x > 0$, let*

$$Q(x) = \sup_{0<h\leq h_0} \big\{\big|r_h(t_1, t_2, v_1, v_2)\big| : (t_1, v_1), (t_2, v_2) \in \mathcal{M}^{(1)}_h \times \mathcal{M}^{(2)}_h, \|t_1 - t_2\| > hx\big\},$$

where $r_h(t_1, t_2, v_1, v_2)$ denotes the covariance between $Z_h(t_1, v_1)$ and $Z_h(t_2, v_2)$. Suppose that, for any $x > 0$, there exists $\eta > 0$ such that

$$Q(x) < \eta < 1. \tag{3.15}$$

Furthermore, assume that there exist $x_0 > 0$ and a function $v(\cdot)$ such that for all $x > x_0$,

$$Q(x)\left|(\log x)^{2(r_1/\alpha_1 + r_2/\alpha_2)}\right| \le v(x), \tag{3.16}$$

where v is a monotonically decreasing function, such that, for any $p > 0$, $v(x^q) = O(v(x)) = o(1)$ and $v(x)x^q \to \infty$ as $x \to \infty$. Let

$$\begin{aligned}\beta_h = {} & \left(2r_1 \log\frac{1}{h}\right)^{\frac{1}{2}} + \left(2r_1 \log\frac{1}{h}\right)^{-\frac{1}{2}} \\ & \times \left[\left(\frac{r_1}{\alpha_1} + \frac{r_2}{\alpha_2} - \frac{1}{2}\right)\log\log\frac{1}{h} \right. \\ & \left. + \log\left\{\frac{(2r_1)^{\frac{r_1}{\alpha_1}+\frac{r_2}{\alpha_2}-\frac{1}{2}}}{\sqrt{2\pi}} H_{\alpha_1}^{(r_1)} H_{\alpha_2}^{(r_2)} I_h\big(\mathcal{M}_h^{(1)} \times \mathcal{M}_h^{(2)}\big)\right\}\right],\end{aligned} \tag{3.17}$$

where

$$I_h\big(\mathcal{M}_h^{(1)} \times \mathcal{M}_h^{(2)}\big) = \int_{\mathcal{M}_h^{(2)}} \int_{\mathcal{M}_h^{(1)}} \big\| D_{s,u}^{(1)} \Lambda\big(T_s \mathcal{M}_h^{(1)}\big)\big\|_{r_1} \big\| D_{s,u}^{(2)} \Lambda\big(T_u \mathcal{M}_h^{(2)}\big)\big\|_{r_2} \, d\mathcal{H}_{r_1}(s)\, d\mathcal{H}_{r_2}(u).$$

Then for any $z \in \mathbb{R}$,

$$\lim_{h\to 0} \mathbb{P}\left\{\sqrt{2r_1 \log\frac{1}{h}}\Big(\sup_{v\in\mathcal{M}_h^{(2)}} \sup_{t\in\mathcal{M}_h^{(1)}} Z_h(t, v) - \beta_h\Big) \le z\right\} = e^{-e^{-z}}. \tag{3.18}$$

For $g \in \mathcal{F}_h$, let $\sigma_g = \sqrt{\mathrm{Var}(\mathbb{B}(g))}$. The standardization of the functions in $\mathcal{F}_h$ gives $\sigma_g = 1 + o(1)$ as $h \to 0$, and we can show that

$$\sup_{g\in\mathcal{F}_h} \mathbb{B}(g) \approx \sup_{g\in\mathcal{F}_h} \sigma_g^{-1}\mathbb{B}(g) = \sup_{(x,z)\in\mathcal{M}_h\times\mathbb{S}^{d-r-1}} \sigma_{g_{x,z}}^{-1}\mathbb{B}(g_{x,z}).$$

To find the asymptotic distribution of $\sup_{g\in\mathcal{F}_h}\mathbb{B}(g)$, we will apply Theorem 3.5 to the Gaussian field $\sigma_{g_{x,z}}^{-1}\mathbb{B}(g_{x,z})$, which is indexed by the manifold $\mathcal{M}_h \times \mathbb{S}^{d-r-1}$. It is critical to calculate the covariance structure of $\sigma_g^{-1}\mathbb{B}(g)$, $g \in \mathcal{F}_h$, and verify it has the desired properties (especially the local stationarity condition) to apply Theorem 3.5. For any $g_{x,z}, g_{\widetilde{x},\widetilde{z}} \in \mathcal{F}_h$ (which means $x, \widetilde{x} \in \mathcal{M}_h$ and $z, \widetilde{z} \in \mathbb{S}^{d-r-1}$), let $r_h(x, \widetilde{x}, z, \widetilde{z})$ be the correlation coefficient between $\mathbb{B}(g_{x,z})$ and $\mathbb{B}(g_{\widetilde{x},\widetilde{z}})$.

Proposition 3.6. *Let $\Delta x = \widetilde{x} - x$ and $\Delta z = \widetilde{z} - z$. Under assumptions* (F1)–(F5), (K1), *and* (K2), *as $h \to 0$, $\Delta z \to 0$ and $\Delta x/h \to 0$, we have*

$$r_h(x, \widetilde{x}, z, \widetilde{z}) = 1 - \frac{1}{2}\|\Delta z\|^2 - \frac{1}{2h^2}\Delta x^T \Omega(x, z)\Delta x + o\left(\left\|\frac{\Delta x}{h}\right\|^2 + \|\Delta z\|^2\right), \tag{3.19}$$

where

$$\Omega(x,z)=\int_{\mathbb{R}^d}\left[\nabla d^2K(u)\right]^T A(x,z)A(x,z)^T\nabla d^2K(u)\,du, \tag{3.20}$$

and the o-term in (3.19) *is uniform in* $x\in\mathcal{M}_h$, $z\in\mathbb{S}^{d-r-1}$, *and* $h\in(0,h_0]$ *for some* $h_0>0$.

Remark 3.5.

(i) When $d-r=1$, we have $z,\widetilde{z}\in\{1,-1\}$ and $\Delta z\equiv 0$, and then (3.19) should be understood as

$$r_h(x,\widetilde{x},z,\widetilde{z})=1-\frac{1}{2h^2}\Delta x^T\Omega(x,1)\Delta x+o\left(\left\|\frac{\Delta x}{h}\right\|^2\right). \tag{3.21}$$

(ii) The geometric interpretation of $\Omega(x,z)$ is as follows. Recall that if higher-order smoothness of f is assumed, the Gaussian process $\mathbb{B}(g)$, $g\in\mathcal{F}_h$ can be approximated by a Gaussian field U given in (3.14), which is differentiable. It can be shown that the matrix $\operatorname{diag}(\frac{1}{h^2}\Omega(x,z),\boldsymbol{I}_{d-r})$ is the leading term of $\operatorname{Var}(\nabla U(x,z))$ by using Itô's lemma, where $\frac{1}{h^2}\Omega(x,z)$ corresponds to the variance of the partial gradient of $U(x,z)$ with respect to x.

To construct a confidence region for $\mathcal{M}_h$, we will use the distribution of $\sup_{g\in\mathcal{F}_h}\mathbb{B}(g)$. The distribution depends on the geometry of the manifold $\mathcal{M}_h\times\mathbb{S}^{d-r-1}$, through a surface integral on the manifold specifically defined below, which is originated from Theorem 3.5. For any nice (meaning the following is well-defined) set $\mathcal{A}\subset\mathcal{H}$, define

$$c_h^{(d,r)}(\mathcal{A})=\log\left\{\frac{r^{(d-2)/2}}{2\pi^{d/2}}\int_{\mathbb{S}^{d-r-1}}\int_{\mathcal{M}_h\cap\mathcal{A}}\left\|\Omega(x,z)^{1/2}\Lambda(T_x\mathcal{M}_h)\right\|_r d\mathcal{H}_r(x)\,d\mathcal{H}_{d-r-1}(z)\right\}. \tag{3.22}$$

For simplicity we write $c_h^{(d,r)}=c_h^{(d,r)}(\mathcal{H})$. The quantity $c_h^{(d,r)}$ reflects the integrated local variability of the approximating Gaussian fields over the index set, as explained below. By the Cauchy-Binet formula (see page 214 in [5]), the integrand in the above double integral can also be written as $[\det(\operatorname{diag}(J_{x,z}^{(1)},J_{x,z}^{(2)}))]^{1/2}=[\det(J_{x,z}^{(1)})]^{1/2}\times[\det(J_{x,z}^{(2)})]^{1/2}$, where

$$\begin{aligned}J_{x,z}^{(1)}&=\left[\Lambda(T_x\mathcal{M}_h)\right]^T\Omega(x,z)\Lambda(T_x\mathcal{M}_h),\\ J_{x,z}^{(2)}&=\left[\Lambda\left(T_z\mathbb{S}^{d-r-1}\right)\right]^T\boldsymbol{I}_{d-r}\Lambda\left(T_z\mathbb{S}^{d-r-1}\right).\end{aligned}$$

Note that $J_{x,z}^{(2)}=\boldsymbol{I}_{d-r-1}$ and $\det(J_{x,z}^{(2)})=1$. In view of Remark 3.5(ii), $\operatorname{diag}(J_{x,z}^{(1)},J_{x,z}^{(2)})$ can be interpreted as the covariance matrix of the orthogonal projection of the gradient of the approximating Gaussian fields onto the tangent space $T_x\mathcal{M}_h\times T_z\mathbb{S}^{d-r-1}$, up to a scaling factor. So the integrand in (3.22) quantifies the local variability of the approximating Gaussian fields as the square root of the determinant of this projected covariance matrix.

In the context of confidence bands for density functions on the unit interval or hypercube [4,44], the counterpart of the quantity $c_h^{(d,r)}$, denoted by c_K, is a constant only depending on the kernel function K (see, e.g., Theorem 2 in [44]). The connection between c_K and $c_h^{(d,r)}$ is as follows. The error in the kernel density estimation can be approximated by *stationary* Gaussian fields, so that the variance of the gradient of the Gaussian fields, denoted by ι_K, is a constant. In fact c_K can also be understood as an integral, but since its integrand ι_K is a constant, the integral is simplified to a constant that is proportional to the volume of index set, which equals one in [4,44] since the unit interval and hypercube

are considered. In the setting of ridge estimation, the form of the surface integral in $c_h^{(d,r)}$ arises for the following reasons: (1) the approximating Gaussian fields is *locally stationary* (see Definition 3.1), which means that their covariances depend on the locations, so that the integrand in $c_h^{(d,r)}$ is not a constant in general; (2) ridges are low-dimensional manifolds, and so the projections to the tangent spaces are involved in the integrand. Also see [40] for a similar line integral that appears in the asymptotic distribution of ridge estimation in the case of $d = 2$ and $r = 1$.

For $z, c \in \mathbb{R}$, let

$$b_h(z,c) = \frac{z}{\sqrt{2r\log(h^{-1})}} + \sqrt{2r\log\bigl(h^{-1}\bigr)} + \frac{1}{\sqrt{2r\log(h^{-1})}}\biggl[\frac{d-2}{2}\log\log\bigl(h^{-1}\bigr) + c\biggr]. \quad (3.23)$$

Note that the quantity $b_h(z, c_h^{(d,r)})$ in the following theorem corresponds to $\frac{z}{\sqrt{2r\log(h^{-1})}} + \beta_h$, where β_h is given in (3.17), with $r_1 = r$, $r_2 = d - r - 1$, and $\alpha_1 = \alpha_2 = 2$. For any $\alpha \in (0,1)$, let $z_\alpha = -\log[-\log(1-\alpha)]$ so that $e^{-e^{-z_\alpha}} = 1 - \alpha$. By applying Theorem 3.5 to the class of Gaussian fields $\{\mathbb{B}(g) : g \in \mathcal{F}_h\}$, the following theorem gives an asymptotic confidence region for $\mathcal{M}_h$.

Theorem 3.7. *Under assumptions* (F1)–(F5), *and* (K1)–(K3), *as* $\gamma_{n,h}^{(2)}\log n \to 0$ *and* $h \to 0$, *we have*

$$\mathbb{P}\Bigl(\sup_{g\in\mathcal{F}_h}\mathbb{B}(g) \le b_h\bigl(z, c_h^{(d,r)}\bigr)\Bigr) \to e^{-e^{-z}}. \quad (3.24)$$

This implies that for any $\alpha \in (0,1)$, *as* $n \to \infty$,

$$\mathbb{P}\bigl(\mathcal{M}_h \subset \widehat{C}_{n,h}\bigl(b_h\bigl(z_\alpha, c_h^{(d,r)}\bigr)\bigr)\bigr) \to 1 - \alpha, \quad (3.25)$$

where $\widehat{C}_{n,h}$ *is defined in* (3.7).

Remark 3.6. We give more discussion on the quantity $c_h^{(d,r)}$. When $d - r = 1$, we have $\mathbb{S}^0 = \{-1, 1\}$ and $\mathcal{H}_0$ is the counting measure, and so

$$c_h^{(d,r)} = \log\biggl\{\frac{r^{(d-2)/2}}{\pi^{d/2}}\int_{\mathcal{M}_h}\bigl\|\Omega(x,1)^{1/2}\Lambda(T_x\mathcal{M}_h)\bigr\|_r\,d\mathcal{H}_r(x)\biggr\}.$$

When $d - r \ge 2$, for any $x \in \mathcal{M}_h$, $\int_{\mathbb{S}^{d-r-1}}\|\Omega(x,z)^{1/2}\Lambda(T_x\mathcal{M}_h)\|_r\,d\mathcal{H}_{d-r-1}(z)$ is a hyperelliptic integral. Note the cross dependence in the Gaussian fields discussed in Remark 3.4(ii) is also reflected in $c_h^{(d,r)}$, where the integrals on $\mathcal{M}_h$ and $\mathbb{S}^{d-r-1}$ are not independent.

The confidence regions for $\mathcal{M}_h$ given in (3.25) is a theoretical result depending on the unknown quantity $c_h^{(d,r)}$. In what follows we address a few important questions: (i) confidence regions for $\mathcal{M}$ by correcting the bias (Section 3.2); (ii) data-driven confidence regions for $\mathcal{M}_h$ and $\mathcal{M}$ by consistently estimating $c_h^{(d,r)}$ (Section 3.3); (iii) different choices of b_n and modified confidence regions for $\mathcal{M}_h$ and $\mathcal{M}$ when assumption (F5) is relaxed (Section 3.4).

3.2. Asymptotic confidence regions for $\mathcal{M}$

We consider asymptotic confidence regions for $\mathcal{M}$ in this section. The difference between $\mathcal{M}$ and $\mathcal{M}_h$ is attributed to the bias in kernel type estimation. In Section 3.1, we focused on $\mathcal{M}_h$ by only

considering the stochastic variation B_n, which is of order $O_p(\gamma_{n,h}^{(2)})$. As we show in Lemma 3.8 below, the bias part in ridge estimation is of order $O(h^2)$. Usually there are two approaches to dealing with the bias in kernel type estimation: implicit bias correction using an undersmoothing bandwidth and explicit bias correction (see, e.g., [20]). The former makes the bias asymptotically negligible compared with the stochastic variation in the estimation, while the latter directly debiases the estimator by estimating the higher order derivatives in the leading terms of the bias using additional kernel estimation, which also means that the latter usually requires stronger assumptions on the smoothness of the underlying functions (see, e.g., [50]). We use both methods to construct asymptotic confidence regions for $\mathcal{M}$.

The next lemma gives the asymptotic form of the bias in ridge estimation. Let $\mu_K = \int_{\mathbb{R}^d} s_1^2 K(s)\,ds$, where $s = (s_1, \ldots, s_d)^T$. Let Δ_L be the Laplacian operator, that is, $\Delta_L \xi(x) = \sum_{i=1}^d \frac{\partial^2 \xi(x)}{\partial x_i^2}$, for a twice differentiable function ξ on $\mathbb{R}^d$. If ξ is a vector-valued function, then Δ_L applies to each element of ξ.

Lemma 3.8. *Under assumptions* (F1)–(F4) *and* (K1), *as* $h \to 0$, *we have*

$$V_h(x)^T \nabla f_h(x) - V(x)^T \nabla f(x) = \frac{1}{2} h^2 \mu_K \beta(x) + R_h,$$

where $\beta(x) = \{M(x)^T [\Delta_L d^2 f(x)] \nabla f(x) + V(x)^T [\Delta_L \nabla f(x)]\}$ *and* $R_h = o(h^2)$, *uniformly in* $x \in \mathcal{N}_{\delta_0}(\mathcal{M})$. *When both* f *and* K *are six times continuously differentiable, we have* $R_h = O(h^4)$, *uniformly in* $x \in \mathcal{N}_{\delta_0}(\mathcal{M})$.

Undersmoothing requires the use of a small bandwidth h such that $\gamma_{n,h}^{(4)} \to \infty$. One can also explicitly correct the bias by using a debiased estimator. For a bandwidth $l > 0$, let

$$\widehat{\beta}_{n,l}(x) = \big\{ \widehat{M}_{n,l}(x)^T \big[\Delta_L d^2 \widehat{f}_{n,l}(x)\big] \nabla \widehat{f}_{n,l}(x) + \widehat{V}_{n,l}(x)^T \big[\Delta_L \nabla \widehat{f}_{n,l}(x)\big] \big\},$$

where we have brought the subscripts n, l back to the kernel estimators to show their dependence on a different bandwidth l. For $a_n \geq 0$ and $b_n \in \mathbb{R}$, let

$$\begin{aligned} &\widehat{C}^{\mathrm{bc}}_{n,h,l}(a_n, b_n) \\ &\quad = \left\{ x \in \mathcal{H} : \sqrt{nh^{d+4}} \left\| Q_n(x) \left[\widehat{V}(x)^T \nabla \widehat{f}(x) - \frac{1}{2} h^2 \mu_K \widehat{\beta}_{n,l}(x) \right] \right\| \leq a_n, \right. \\ &\qquad \left. \text{and } \widehat{\lambda}_{r+1}(x) < b_n \right\}, \end{aligned} \tag{3.26}$$

and denote $\widehat{C}^{\mathrm{bc}}_{n,h,l}(a_n) = \widehat{C}^{\mathrm{bc}}_{n,h,l}(a_n, 0)$ for simplicity. Define

$$c^{(d,r)} = \log \left\{ \frac{r^{(d-2)/2}}{2\pi^{d/2}} \int_{\mathbb{S}^{d-r-1}} \int_{\mathcal{M}} \big\| \Omega(x,z)^{1/2} \Lambda(T_x \mathcal{M}) \big\|_r \, d\mathcal{H}_r(x)\, d\mathcal{H}_{d-r-1}(z) \right\},$$

where we simply replace the domain of integration $\mathcal{M}_h$ by $\mathcal{M}$ in $c_h^{(d,r)}$.

Theorem 3.9. *Suppose assumptions* (F1)–(F5), *and* (K1)–(K3) *hold. Also assume that* $\gamma_{n,h}^{(2)} \log n \to 0$ *and* $h \to 0$. *For any* $\alpha \in (0,1)$ *we have the following.*

(i) Undersmoothing: *As* $\gamma_{n,h}^{(4)} / \log n \to \infty$,

$$\mathbb{P}\big(\mathcal{M} \subset \widehat{C}_{n,h}\big(b_h\big(z_\alpha, c^{(d,r)}\big)\big)\big) \to 1 - \alpha. \tag{3.27}$$

(ii) Explicit bias correction: *Assume both f and K are six times continuously differentiable. As $(h/l)\log n \to 0$ and $\gamma_{n,h}^{(4)}/(l^2 \log n) \to \infty$,*

$$\mathbb{P}\big(\mathcal{M} \subset \widehat{C}_{n,h,l}^{\mathrm{bc}}\big(b_h\big(z_\alpha, c^{(d,r)}\big)\big)\big) \to 1-\alpha. \tag{3.28}$$

Remark 3.7. We emphasize that the method in (i) is feasible because we only require $\gamma_{n,h}^{(2)} \log n \to 0$ and $h \to 0$ for n and h in the results in Section 3.1. As a comparison, the Hausdorff distance based approach for $\mathcal{M}_h$ developed in [8] requires an oversmoothing bandwidth such that $\gamma_{n,h}^{(4)} \to 0$, which implies that the bias dominates the stochastic variation in ridge estimation using the Hausdorff distance if a second order kernel is used, and hence the approach using an undersmoothing bandwidth is not applicable in their method.

3.3. Estimating the unknowns

The surface integrals $c_h^{(d,r)}$ and $c^{(d,r)}$ are unknown quantities that need to be estimated in order to make the confidence regions in Theorems 3.7 and 3.9 computable with data. For a bandwidth $l > 0$, we use the following plug-in estimators. Let

$$\begin{aligned}
\widehat{A}_{n,l}(x,z) &= \widehat{M}_{n,l}(x)\big[\widehat{f}_{n,l}(x)\widehat{\Sigma}_{n,l}(x)\big]^{-1/2} z, \\
\widehat{\Omega}_{n,l}(x,z) &= \int_{\mathbb{R}^d} \nabla d^2 K(u)^T \widehat{A}_{n,l}(x,z)\widehat{A}_{n,l}(x,z)^T \nabla d^2 K(u)\, du, \\
\widehat{\mathcal{M}}_{n,l} &= \big\{x \in \mathcal{H} : \widehat{V}_{n,l}(x)^T \nabla \widehat{f}_{n,l}(x) = 0, \widehat{\lambda}_{r+1,n,l}(x) < 0\big\}.
\end{aligned}$$

Note that the bandwidth l here is not necessarily the same one as used for explicit bias correction in Section 3.8. But we do need a similar condition for them so the same bandwidth l is used for simplicity. For any nice set $\mathcal{A} \subset \mathcal{H}$, let

$$\widehat{c}_{n,l}^{(d,r)}(\mathcal{A}) = \log\left\{\frac{r^{(d-2)/2}}{2\pi^{d/2}} \int_{\mathbb{S}^{d-r-1}} \int_{\widehat{\mathcal{M}}_{n,l}\cap\mathcal{A}} \big\|\widehat{\Omega}_{n,l}(x,z)^{1/2}\Lambda(T_x\widehat{\mathcal{M}}_{n,l})\big\|_r \, d\mathcal{H}_r(x)\, d\mathcal{H}_{d-r-1}(z)\right\}.$$

For simplicity we denote $\widehat{c}_{n,l}^{(d,r)} = \widehat{c}_{n,l}^{(d,r)}(\mathcal{H})$. To prove the confidence regions for $\mathcal{M}_h$ and $\mathcal{M}$ are still valid after replacing $b_h(z_\alpha, c_h^{(d,r)})$ and $b_h(z_\alpha, c^{(d,r)})$ by $b_h(z, \widehat{c}_{n,l}^{(d,r)})$, we need to show that $\widehat{c}_{n,l}^{(d,r)}$ is a consistent estimator of $c_h^{(d,r)}$ and $c^{(d,r)}$. The proof uses similar ideas as in [39], where the focus is on the estimation of surface integrals of density level sets, which are $(d-1)$-dimensional manifolds embedded in $\mathbb{R}^d$. Since we view density ridges as the intersections of $d-r$ level sets (in a broad sense to include $d-r=1$), the methods in [39] are extended in our proof. The data-driven confidence regions are given in the following corollary.

Corollary 3.10. *Suppose assumptions* (F1)–(F5), *and* (K1)–(K3) *hold, and assume that $\gamma_{n,h}^{(2)} \log n \to 0$, $\gamma_{n,l}^{(4)} \to 0$, $h \to 0$ and $l \to 0$. For any $\alpha \in (0,1)$ we have the following.*

(i) For $\mathcal{M}_h$:

$$\mathbb{P}\big(\mathcal{M}_h \subset \widehat{C}_{n,h}\big(b_h\big(z_\alpha, \widehat{c}_{n,l}^{(d,r)}\big)\big)\big) \to 1-\alpha. \tag{3.29}$$

(ii) For $\mathcal{M}$ using an undersmoothing bandwidth: *as* $\gamma_{n,h}^{(4)}/\log n \to \infty$,

$$\mathbb{P}\big(\mathcal{M} \subset \widehat{C}_{n,h}\big(b_h\big(z_\alpha, \widehat{c}_{n,l}^{(d,r)}\big)\big)\big) \to 1-\alpha. \tag{3.30}$$

(iii) For $\mathcal{M}$ using explicit bias correction: *Assume that both* f *and* K *are six time continuously differentiable. As* $(h/l)\log n \to 0$ *and* $\gamma_{n,h}^{(4)}/(l^2 \log n) \to \infty$,

$$\mathbb{P}\big(\mathcal{M} \subset \widehat{C}_{n,h,l}^{\mathrm{bc}}\big(b_h\big(z_\alpha, \widehat{c}_{n,l}^{(d,r)}\big)\big)\big) \to 1-\alpha. \tag{3.31}$$

3.4. Further improvements related to eigenvalues and critical points

We have considered the confidence regions in the form of $\widehat{C}_{n,h}(a_n, b_n)$ defined in (3.7) and $\widehat{C}_{n,h,l}^{\mathrm{bc}}(a_n, b_n)$ defined in (3.26) for some $a_n > 0$ and $b_n = 0$. So far our main focus has been on the determination of a_n, after the justification for the choice $b_n = 0$ given in Remark 3.1. In fact, one can use some nonpositive b_n as the upper bound of $\widehat{\lambda}_{r+1}$, to potentially make the confidence regions more efficient. This is because $\sup_{x\in\mathcal{M}} \lambda_{r+1}(x)$ is strictly bounded away from 0 under assumption (F3), which allows us to choose a nonpositive b_n such that $\sup_{x\in\mathcal{M}} \widehat{\lambda}_{r+1}(x) < b_n$ holds with probability tending to one under our assumptions. Here b_n is determined by using $\sup_{x\in\widehat{\mathcal{M}}} \widehat{\lambda}_{r+1}(x)$, and so we first need to give the rate of convergence of the Hausdorff distance between $\widehat{\mathcal{M}}$ and $\mathcal{M}$. For any two nonempty subsets A and B of $\mathbb{R}^d$, their Hausdorff distance is defined as

$$d_H(A, B) = \inf\big\{\epsilon > 0 : A \subset (B \oplus \epsilon) \text{ and } B \subset (A \oplus \epsilon)\big\}. \tag{3.32}$$

Lemma 3.11. *Suppose assumptions* (F1)–(F4) *and* (K1) *hold and* $h \to 0$. *Also assume that* $\gamma_{n,h}^{(2)} \to 0$ *for* $d - r = 1$ *and* $\gamma_{n,h}^{(3)} \to 0$ *for* $d - r \geq 2$. *Then there exists a constant* $C_0 > 0$ *such that*

$$d_H(\mathcal{M}, \mathcal{M}_h) \leq C_0 h^2, \tag{3.33}$$

$$\mathbb{P}\big(d_H(\widehat{\mathcal{M}}, \mathcal{M}_h) \leq C_0 \gamma_{n,h}^{(2)}\big) \to 1, \tag{3.34}$$

which implies that $\mathbb{P}(d_H(\widehat{\mathcal{M}}, \mathcal{M}) \leq C_0(\gamma_{n,h}^{(2)} + h^2)) \to 1$.

For $a, b \in \mathbb{R}$, denote $a \wedge b = \min(a, b)$. Let ν_n be a sequence such that $\nu_n \to \infty$ and define

$$\zeta_n^0 = \Big[\sup_{x\in\widehat{\mathcal{M}}} \widehat{\lambda}_{r+1}(x) + \nu_n \gamma_{n,h}^{(2)}\Big] \wedge 0, \tag{3.35}$$

$$\zeta_n = \Big[\sup_{x\in\widehat{\mathcal{M}}} \widehat{\lambda}_{r+1}(x) + \nu_n\big(\gamma_{n,h}^{(2)} + h^2\big)\Big] \wedge 0. \tag{3.36}$$

Proposition 3.12. *Suppose assumptions* (F1)–(F4) *and* (K1) *hold and* $h \to 0$. *Also assume that* $\gamma_{n,h}^{(2)} \to 0$ *for* $d - r = 1$ *and* $\gamma_{n,h}^{(3)} \to 0$ *for* $d - r \geq 2$. *Then*

$$\mathbb{P}\Big(\sup_{x\in\mathcal{M}_h} \widehat{\lambda}_{r+1}(x) \geq \zeta_n^0\Big) \to 0, \tag{3.37}$$

$$\mathbb{P}\Big(\sup_{x\in\mathcal{M}} \widehat{\lambda}_{r+1}(x) \geq \zeta_n\Big) \to 0. \tag{3.38}$$

Remark 3.8. The result in Proposition 3.12 immediately implies that we can use ζ_n^0 to replace 0 as b_n in the confidence regions that we construct in Corollary 3.10 for $\mathcal{M}_h$ (and use ζ_n for $\mathcal{M}$), if we additionally assume $\gamma_{n,h}^{(3)} \to 0$ for $d - r \geq 2$.

So far we have imposed assumption (F5) to exclude critical points on ridges from our consideration. The reason is that the behaviors of the estimators of critical points and regular ridge points are different in our approach. Below we remove assumption (F5), that is, we allow the existence of points x such that $\|\nabla f(x)\| = 0$ on $\mathcal{M}$. For any $0 < \eta < 1$, let $\mathcal{K}_{h,\eta} = \{x \in \mathcal{H} : \|\nabla f_h(x)\| \leq h^\eta\}$. Note that $\mathcal{M}_h = (\mathcal{M}_h \cap \mathcal{K}_{h,\eta}) \cup (\mathcal{M}_h \cap \mathcal{K}_{h,\eta}^{\complement})$. When h is small, the set $\mathcal{M}_h \cap \mathcal{K}_{h,\eta}$ is a small neighborhood near all the critical points on the ridge $\mathcal{M}_h$, and $\mathcal{M}_h \cap \mathcal{K}_{h,\eta}^{\complement}$ is the set of the remaining points on the ridge. Our strategy is to construct two regions to cover $\mathcal{M}_h \cap \mathcal{K}_{h,\eta}$ and $\mathcal{M}_h \cap \mathcal{K}_{h,\eta}^{\complement}$ separately and then combine them. For a sequence $\mu_n \to \infty$ such that $h\mu_n \to 0$, let $\mathcal{E}_{n,\eta} = \{x \in \mathcal{H} : \|\nabla \widehat{f}(x)\| \leq \mu_n \gamma_{n,h}^{(1)} + h^\eta\}$ and

$$\mathcal{G}_{n,\eta}^0 = \mathcal{E}_{n,\eta} \cap \{x \in \mathcal{H} : \widehat{\lambda}_{r+1}(x) < \zeta_n^0\},$$
$$\mathcal{G}_{n,\eta} = \mathcal{E}_{n,\eta} \cap \{x \in \mathcal{H} : \widehat{\lambda}_{r+1}(x) < \zeta_n\}.$$

Then $\mathcal{G}_{n,\eta}^0$ and $\mathcal{G}_{n,\eta}$ cover $\mathcal{M}_h \cap \mathcal{K}_{h,\eta}$ and $\mathcal{M} \cap \mathcal{K}_\eta$, respectively, with a large probability, where $\mathcal{K}_\eta = \{x \in \mathcal{H} : \|\nabla f(x)\| \leq h^\eta\}$. The following theorem gives the confidence regions for $\mathcal{M}_h$ and $\mathcal{M}$ without the assumption (F5), where we also incorporate a new choice for b_n as discussed in Remark 3.8.

Theorem 3.13. *Suppose assumptions* (F1)–(F4), *and* (K1)–(K3) *hold and there exists at least one point* $x_0 \in \mathcal{M}$ *such that* $\|\nabla f(x_0)\| > 0$. *Also we assume that* $\gamma_{n,h}^{(2)} \log n \to 0$ *for* $d - r = 1$ *and* $\gamma_{n,h}^{(3)} \to 0$ *for* $d - r \geq 2$; $\gamma_{n,l}^{(4)} \to 0$ *and* $l \to 0$. *Suppose* $0 < \eta < 1$, $\nu_n \to \infty$, $\mu_n \to \infty$ *and* $h\mu_n \to 0$. *For any* $\alpha \in (0, 1)$ *we have the following.*

(i) For $\mathcal{M}_h$:

$$\mathbb{P}\big(\mathcal{M}_h \subset \big[\widehat{C}_{n,h}\big(b_h\big(z_\alpha, \widehat{c}_{h,l}^{(d,r)}\big(\mathcal{E}_{n,\eta}^{\complement}\big)\big), \zeta_n^0\big) \cup \mathcal{G}_{n,\eta}^0\big]\big) \to 1 - \alpha. \tag{3.39}$$

(ii) For $\mathcal{M}$ using an undersmoothing bandwidth: *as* $\gamma_{n,h}^{(4)} / \log n \to \infty$,

$$\mathbb{P}\big(\mathcal{M} \subset \big[\widehat{C}_{n,h}\big(b_h\big(z_\alpha, \widehat{c}_{n,l}^{(d,r)}\big(\mathcal{E}_{n,\eta}^{\complement}\big)\big), \zeta_n\big) \cup \mathcal{G}_{n,\eta}\big]\big) \to 1 - \alpha. \tag{3.40}$$

(iii) For $\mathcal{M}$ using explicit bias correction: *Assume that both* f *and* K *are six time continuously differentiable. As* $(h/l) \log n \to 0$ *and* $\gamma_{n,h}^{(4)} / (l^2 \log n) \to \infty$,

$$\mathbb{P}\big(\mathcal{M} \subset \big[\widehat{C}_{n,h,l}^{\mathrm{bc}}\big(b_h\big(z_\alpha, \widehat{c}_{n,l}^{(d,r)}\big(\mathcal{E}_{n,\eta}^{\complement}\big)\big), \zeta_n\big) \cup \mathcal{G}_{n,\eta}\big]\big) \to 1 - \alpha. \tag{3.41}$$

Remark 3.9.

(i) The results in this theorem still hold if ζ_n^0 and ζ_n are replaced by 0 as discussed in Remark 3.1.

(ii) We use two sequences $\mu_n \to \infty$ and $\nu_n \to \infty$ in the construction of the confidence regions. One may choose $\mu_n = h^{-\mu}$ and $\nu_n = h^{-\nu}$ for some $0 < \mu < 1$ and $\nu > 0$ to satisfy the assumptions in the theorem. The need for using these tuning parameters μ, ν as well as η reflects the fact that the definition of ridges involves multiple components, that is, the eigenvectors and eigenvalues of the Hessian and the gradient (see (1.1) and (1.2)). These components have different asymptotic behaviors and roles in the estimation. The choice of the tuning parameters allows us to focus on the asymptotic behaviors related to the eigenvectors.

4. Discussion

In this paper, we develop asymptotic confidence regions for density ridges. We treat ridges as the intersections of some level sets and use the VV based approach. The construction of our confidence regions is based on Gaussian approximation of suprema of empirical processes and the extreme value distribution of suprema of χ-fields indexed by manifolds. It is known that the rate of convergence of this type of extreme value distribution is slow. As an alternative approach, we are working on developing a bootstrap procedure using the VV idea for the confidence regions.

Apparently our approach can also be used for the construction of confidence regions for the intersections of multiple functions in general (such as density function and regression functions). It's known that estimating such intersections has applications in econometrics. See, for example, [6].

5. Proofs

We give the proofs of Lemma 3.1 and Theorem 3.7 in this section. The proofs of Proposition 3.2, Proposition 3.3, Theorem 3.4, Proposition 3.6, Lemma 3.8, Theorem 3.9, Corollary 3.10, Lemma 3.11, Proposition 3.12, and Theorem 3.13 can be found in the supplementary material [38].

Proof of Lemma 3.1. Under assumption (F4), the claim that $\mathcal{M}$ is an r-dimensional manifold is a consequence of the constant-rank level set theorem (see Theorem 5.12 in [27]). Under assumption (F3), we can write $\mathcal{M} = \{x \in \mathcal{H} : V(x)^T \nabla f(x) = 0, \lambda_{r+1}(x) \le 0\}$, which is a compact set, whose boundary is $\partial\mathcal{M} = \{x \in \mathcal{H} : V(x)^T \nabla f(x) = 0, \lambda_{r+1}(x) = 0\} = \varnothing$. Next we show that $\mathcal{M}$ has positive reach. For any twice differentiable function η on an open subset $\mathcal{A} \subset \mathbb{R}^d$, let $\mathcal{L}_\eta = \{x \in \mathcal{A} : \eta(x) = 0\}$. Suppose that $\mathcal{L}_\eta$ is nonempty. The proof of Lemma 4.11 in [15] shows that for any $x \in \mathcal{L}_\eta$, the inequality

$$\Delta(\mathcal{L}_\eta, x) \ge \min\left\{\frac{\epsilon}{2}, \frac{\inf_{x\in\mathcal{L}_\eta\oplus\epsilon} \|\nabla\eta(x)\|}{\sup_{\mathcal{L}_\eta\oplus(2\epsilon)} \|\nabla^2\eta(x)\|_F}\right\}. \tag{5.1}$$

holds for all $\epsilon > 0$ such that $\mathcal{L}_\eta \oplus (2\epsilon) \subset \mathcal{A}$ and the right-hand side of (5.1) is well defined and positive. For $i = 1, \ldots, d-r$, let

$$p_i(x) = \nabla f(x)^T v_{r+i}(x) \quad \text{and} \quad l_i(x) = \nabla p_i(x), \tag{5.2}$$

and define sets $\mathcal{M}_i = \{x \in \mathcal{N}_{\delta_0}(\mathcal{M}) : p_i(x) = 0, \lambda_{r+1}(x) < 0\}$. Note that $\mathcal{M} = \cap_{i=1}^{d-r} \mathcal{M}_i$. Under assumption (F4), there exists $\delta_1 > 0$ such that $\mathcal{M} \oplus \delta_1 \subset \mathcal{N}_{\delta_0}(\mathcal{M})$, and there exists $\epsilon_1 > 0$ such that $\inf_{x\in\mathcal{N}_{\delta_0}(\mathcal{M})} \|l_i(x)\| > \epsilon_1$, for $i = 1, \ldots, d-r$. It is known that there exist second derivatives of the unit eigenvectors corresponding to simple eigenvalues as functions of symmetric matrices (see, e.g., [12]). Therefore with assumptions (F1) and (F2), the functions p_i, $i = 1, \ldots, d-r$ are twice differentiable and there exists a constant $0 < C < \infty$ such that $\sup_{x\in\mathcal{N}_{\delta_0}(\mathcal{M})} \|\nabla^2 p_i(x)\|_F < C$, for $i = 1, \ldots, d-r$. Then applying (5.1), we get

$$\inf_{u\in\mathcal{M}} \Delta(\mathcal{M}_i, u) \ge \min(\delta_1/4, \epsilon_1/C) =: C_0. \tag{5.3}$$

If $d-r = 1$, then (5.3) has given a positive lower bound of $\Delta(\mathcal{M})$. Next, we consider the case $d-r \ge 2$. The proof follows similar arguments as given in the proof of Theorem 4.12 in [15]. Specifically, let $b_1 = C_0$, and for $k = 2, \ldots, d-r$, let

$$b_k = \frac{1}{2}\min\left\{1, \inf_{x\in\mathcal{N}_{\delta_0}(\mathcal{M})} \inf_{(a_1,\ldots,a_k)^T\neq 0} \frac{\|\sum_{i=1}^k a_i l_i(x)\|}{\|\sum_{i=1}^{k-1} a_i l_i(x)\| + \|a_k l_k(x)\|}\right\}.$$

Note that $b_k > 0$, for $k = 2, \ldots, d - r$ under assumption (F4). Then using (5.3) and Theorem 4.10 of [15] inductively, we get

$$\inf_{u \in \mathcal{M}} \Delta\left(\bigcap_{i=1}^{k} \mathcal{M}_i, u\right) \geq b_1 \cdots b_k, \tag{5.4}$$

and hence $\Delta(\mathcal{M}) = \inf_{u \in \mathcal{M}} \Delta(\bigcap_{i=1}^{d-r} \mathcal{M}_i, u) \geq b_1 \cdots b_{d-r} > 0$. This is assertion (i).

Next, we show assertion (ii). Let $\delta_{\text{gap}} := \inf_{x \in \mathcal{H}}[\lambda_r(x) - \lambda_{r+1}(x)]$. Since $\mathcal{H}$ is compact and $\lambda_r - \lambda_{r+1}$ is continuous on $\mathcal{H}$, we have $\delta_{\text{gap}} > 0$ due to assumption (F2). Lemma A.1 in the supplementary material [38] implies that

$$\inf_{x \in \mathcal{H}}\big[\lambda_{r,h}(x) - \lambda_{r+1,h}(x)\big] = \delta_{\text{gap}} + O\big(h^2\big) \geq \frac{1}{2}\delta_{\text{gap}}, \tag{5.5}$$

when h is small enough. Then using the Davis–Kahan theorem (see, e.g., [47]) and Lemma A.1 in the supplementary material [38] leads to

$$\sup_{x \in \mathcal{H}} \big\| V(x)V(x)^T - V_h(x)V_h(x)^T \big\|_F \leq \frac{2\sqrt{2}\sup_{x \in \mathcal{H}} \|\nabla^2 f(x) - \nabla^2 f_h(x)\|_F}{\delta_{\text{gap}}} = O\big(h^2\big). \tag{5.6}$$

Noticing that $V(x)^T V(x) = \boldsymbol{I}_{d-r}$, we can write

$$\begin{aligned}
\sup_{x \in \mathcal{M}_h} \big\| V(x)^T \nabla f(x) \big\| &= \sup_{x \in \mathcal{M}_h} \big\| V(x)V(x)^T \nabla f(x) - V_h(x)V_h(x)^T \nabla f_h(x) \big\| \\
&\leq \sup_{x \in \mathcal{H}} \big\| V(x)V(x)^T \nabla f(x) - V_h(x)V_h(x)^T \nabla f_h(x) \big\| \\
&= O\big(h^2\big),
\end{aligned} \tag{5.7}$$

where we use (5.6) and Lemma A.1 in the supplementary material [38].

Let $\mathcal{M}^{(1)} = \{x \in \mathcal{H} : V(x)^T \nabla f(x) = 0\}$ and $\mathcal{M}^{(2)} = \{x \in \mathcal{H} : \lambda_{r+1}(x) < 0\}$. Then $\mathcal{M} = \mathcal{M}^{(1)} \cap \mathcal{M}^{(2)}$. For any $\delta > 0$, let $\mathcal{N}_\delta(\mathcal{M}^{(1)}) = \{x \in \mathcal{H} : \|V(x)^T \nabla f(x)\| \leq \delta\}$. Note that (5.7) implies that for any fixed $\delta \in (0, \delta_0]$, $\mathcal{M}_h \subset \mathcal{N}_\delta(\mathcal{M}^{(1)})$ when h is small enough. It suffices to show $\mathcal{M}_h \subset \mathcal{M}^{(2)}$, when h is small enough. Since $\mathcal{M}^{(1)}$ is a compact set and λ_{r+1} is continuous on $\mathcal{H}$, under assumption (F3) there exists $\beta_0 > 0$ such that $\inf_{x \in \mathcal{M}^{(1)}} |\lambda_{r+1}(x)| \geq 4\beta_0$, and hence there exists δ_2 with $0 < \delta_2 \leq \delta_0$ such that

$$\inf_{x \in \mathcal{N}_{\delta_2}(\mathcal{M}^{(1)})} \big|\lambda_{r+1}(x)\big| \geq 2\beta_0, \tag{5.8}$$

which further implies that $\inf_{x \in \mathcal{M}_h} |\lambda_{r+1}(x)| \geq 2\beta_0$, when h is small enough. Then we must have

$$\sup_{x \in \mathcal{M}_h} \lambda_{r+1}(x) \leq -2\beta_0, \tag{5.9}$$

since if there exists $x_0 \in \mathcal{M}_h$ such that $\lambda_{r+1}(x_0) \geq 2\beta_0$, then Lemma A.1 in the supplementary material [38] would lead to

$$\lambda_{r+1,h}(x_0) \geq \lambda_{r+1}(x_0) - \big|\lambda_{r+1}(x_0) - \lambda_{r+1,h}(x_0)\big| \geq 2\beta_0 + O\big(h^2\big) \geq \beta_0,$$

when h is small, which contradicts the definition of $\mathcal{M}_h$. Hence, $\mathcal{M}_h \subset [\mathcal{N}_{\delta_0}(\mathcal{M}^{(1)}) \cap \mathcal{M}^{(2)}] = \mathcal{N}_{\delta_0}(\mathcal{M})$, when h is small enough. This is assertion (ii).

For assertion (iii), it can be seen from (5.9) that

$$\sup_{x\in\mathcal{M}_h} \lambda_{r+1,h}(x) \le -\beta_0 \tag{5.10}$$

when h is small enough. Using a similar argument, we get that when h is small,

$$\inf_{x\in\mathcal{M}_h} \big[\lambda_{j-1,h}(x) - \lambda_{j,h}(x)\big] > \beta_0, \quad j = r+1,\dots,d, \tag{5.11}$$

by possibly decreasing β_0 to a smaller positive constant.

To show that $\mathcal{M}_h$ is an r-dimensional manifold without boundary and has positive reach when h is small in assertion (iv), we use a similar argument as given in the proof of assertion (i). We first show that there exists a constant $\delta_3 \in (0,\delta_0]$ such that $\mathcal{N}_{\delta_3}(\mathcal{M})$ is a compact set. Let $\mathcal{A} = \{x \in \mathcal{H} : \lambda_{r+1}(x) = 0\}$. If $\mathcal{A} = \varnothing$, then we simply take $\delta_3 = \delta_0$ and can write $\mathcal{N}_{\delta_0}(\mathcal{M}) = \{x \in \mathcal{H} : \|V(x)^T \nabla f(x)\| \le \delta_0, \lambda_{r+1}(x) \le 0\}$, which is a compact set. Otherwise, $\mathcal{A}$ is a compact nonempty set and we let $\delta_3^* = \inf_{x\in\mathcal{A}} \|V(x)^T \nabla f(x)\|$. Since $\|V^T \nabla f\|$ is a continuous function of x, we must have $\delta_3^* > 0$ under assumption (F3). Taking $\delta_3 = \min(\frac{1}{2}\delta_3^*, \delta_0)$, we can write $\mathcal{N}_{\delta_3}(\mathcal{M}) = \{x \in \mathcal{H} : \|V(x)^T \nabla f(x)\| \le \delta_3, \lambda_{r+1}(x) \le 0\}$, which is a compact set.

Next we show that f_h satisfies the similar properties as in the assumptions (F1)–(F4) for f, when h is small. First, f_h is four times continuous differentiable on $\mathcal{H}$ due to assumption (K1). Also it is easy to see from (5.8) that the set $\{x \in \mathcal{H} : V_h(x)^T \nabla f_h(x) = 0, \lambda_{r+1,h}(x) = 0\} = \varnothing$, when h is small enough. Hence we can write $\mathcal{M}_h = \{x \in \mathcal{H} : V_h(x)^T \nabla f_h(x) = 0, \lambda_{r+1,h}(x) \le 0\}$, which is a compact set. Similar to (5.11), we can show that there exists a constant $\beta_2 > 0$ such that $\inf_{x\in\mathcal{N}_{\delta_4}(\mathcal{M})}[\lambda_{j-1,h}(x) - \lambda_{j,h}(x)] > \beta_2$, $j = r+1,\dots,d$, for some $0 < \delta_4 \le \delta_0$ and $\inf_{x\in\mathcal{H}}[\lambda_{r,h}(x) - \lambda_{r+1,h}(x)] > \beta_2$, when h is small enough. It suffices to show that f_h satisfies a similar condition as given in assumption (F4) for f. With the notation in (5.2), let $L(x) = (l_1(x),\dots,l_{d-r}(x))$. Then assumption (F4) is equivalent to $\inf_{x\in\mathcal{M}} \det(L(x)^T L(x)) > 0$. Since $\mathcal{N}_{\delta_3}(\mathcal{M})$ is a compact set and $\det(L(x)^T L(x))$ is a continuous function on $\mathcal{H}$ under our assumptions, we can find $\epsilon_0 > 0$ such that

$$\inf_{x\in\mathcal{N}_{\delta_3}(\mathcal{M})} \det\big(L(x)^T L(x)\big) \ge \epsilon_0. \tag{5.12}$$

Let $L_h(x) = (l_{1,h}(x),\dots,l_{d-r,h}(x))$, where $l_{i,h}(x) = \nabla(\nabla f_h(x)^T v_{r+i,h}(x))$, $i = 1,\dots,d-r$. With (5.12) we have

$$\begin{aligned}
&\inf_{x\in\mathcal{N}_{\delta_3}(\mathcal{M})} \det\big(L_h(x)^T L_h(x)\big)\\
&\quad\ge \inf_{x\in\mathcal{N}_{\delta_3}(\mathcal{M})} \det\big(L(x)^T L(x)\big) - \sup_{x\in\mathcal{N}_{\delta_3}(\mathcal{M})} \big|\det\big(L(x)^T L(x)\big) - \det\big(L_h(x)^T L_h(x)\big)\big|\\
&\quad\ge \epsilon_0 - O\big(h^2\big),
\end{aligned} \tag{5.13}$$

where we use Lemma A.1 in the supplementary material [38] and Theorem 3.3 in [23], the latter giving a perturbation bound for matrix determinants. This then implies that there exists $\epsilon_1 > 0$ such that for h small enough, $\inf_{x\in\mathcal{N}_{\delta_3}(\mathcal{M})} \|l_{i,h}(x)\| > \epsilon_1$, and $l_{i,h}(x)$, $i = 1,\dots,d-r$ are linearly independent for all $x \in \mathcal{N}_{\delta_3}(\mathcal{M})$. The rest of the proof is omitted because it is similar to the proof of assertion (i). □

To prove Theorem 3.7, we need the following lemma.

Lemma 5.1. *Suppose assumptions* (F1)–(F5), *and* (K1)–(K3) *hold. There exists a constant* $\delta_1 > 0$ *such that for all* $x \in \mathcal{N}_{\delta_1}(\mathcal{M})$ *and* $z \in \mathbb{R}^{d-r}\backslash\{0\}$, $\Omega(x, z)$ *in* (3.20) *is positive definite.*

Proof of Lemma 5.1. We need to introduce some notation first. Recall that for any $d \times d$ symmetric matrix A, $\text{vech}(A)$ is the half-vectorization of A, that is, it vectorizes only the lower triangular part of A (including the main diagonal of A). Let $\text{diag}(A)$ be the vector of the diagonal entries of A and $\text{vech}_s(A)$ be the vectorization of the *strictly* lower triangular portion of A, which can be obtained from $\text{vech}(A)$ by eliminating all the diagonal elements of A. Let $\text{dvech}(A)$ be a vectorization of the lower triangular portion of A, such that $\text{dvech}(A) = (\text{diag}(A)^T, \text{vech}_s(A)^T)^T$. Let Q be a $[d(d+1)/2] \times [d(d+1)/2]$ matrix such that $\text{dvech}(A) = Q\,\text{vech}(A)$. Note that Q is nonsingular.

Let $\mathcal{I} = \mathcal{I}^d \cup \mathcal{I}^o$, where $\mathcal{I}^d = \{1, 2, \ldots, d\}$ and $\mathcal{I}^o = \{d+1, d+2, \ldots, d(d+1)/2\}$, that is, $\mathcal{I}^d$ and $\mathcal{I}^o$ are the index sets for $\text{diag}(A)$ and $\text{vech}_s(A)$ in $\text{dvech}(A)$, respectively. Suppose that we can write $A = (a_{l,m})_{1\le l,m\le d}$. Define a map $\pi = (\pi_1, \pi_2) : \mathcal{I} \to \mathcal{I}^d \times \mathcal{I}^d$ such that the kth element of $\text{dvech}(A)$ is $a_{\pi_1(k),\pi_2(k)}$, $k \in \mathcal{I}$. For $k_1, k_2 \in \mathcal{I}$, let $\pi_\Delta(k_1, k_2) = \{\pi_1(k_1), \pi_2(k_1)\}\Delta\{\pi_1(k_2), \pi_2(k_2)\}$, where Δ denotes the symmetric difference between two sets, i.e., $A\Delta B = (A\backslash B) \cup (B\backslash A)$ for any two sets A and B. For $i, j \in \mathcal{I}^d$, let $\pi_q^{-1}(i) = \{k \in \mathcal{I} : \pi_q(k) = i\}$, $q = 1, 2$, and

$$\pi^{-1}(i, j) = \begin{cases} \pi_1^{-1}(i) \cap \pi_2^{-1}(j) & \text{if } i \ge j, \\ \pi_1^{-1}(j) \cap \pi_2^{-1}(i) & \text{if } i < j. \end{cases}$$

Note that $\pi^{-1}(i, j) = \pi^{-1}(j, i)$. Let $\pi_\cup^{-1}(i) = \pi_1^{-1}(i) \cup \pi_2^{-1}(i)$, $i \in \mathcal{I}_d$. Let $\delta(i, j)$ be the Kronecker delta. For any set $\mathcal{J}$, let $\delta(i, \mathcal{J}) = \mathbf{1}_{\mathcal{J}}(i)$, which is an indicator function regarding whether $i \in \mathcal{J}$.

With $\delta_1 > 0$ given in Proposition 3.2, for $x \in \mathcal{N}_{\delta_1}(\mathcal{M})$, and $z \in \mathbb{R}^{d-r}\backslash\{0\}$, let $\widetilde{A}(x, z) = A(x, z)^T Q^{-1} =: (t_1(x, z), \ldots, t_{d(d+1)/2}(x, z))$. Recall that $\sqrt{f(x)}\|A(x, z)\|_{\boldsymbol{R}} = \|z\|$ for all $x \in \mathcal{N}_{\delta_1}$ and $z \in \mathbb{R}^{d-r}\backslash\{0\}$. Hence for any $x \in \mathcal{N}_{\delta_1}$ and $z \in \mathbb{R}^{d-r}\backslash\{0\}$,

$$t_k(x, z) \ne 0, \quad \text{for at least one } k \in \mathcal{I}. \tag{5.14}$$

Then we can write $\Omega(x, z) = \int [\nabla d^2 K(u)]^T Q^T \widetilde{A}(x, z)^T \widetilde{A}(x, z) Q \nabla d^2 K(u)\, du$, for which $\Omega_{i,j}(x, z)$ denotes the element at the ith row and jth column. Below we consider any $x \in \mathcal{N}_{\delta_1}$ and $z \in \mathbb{R}^{d-r}\backslash\{0\}$ and will suppress x and z in the notation. Let $\eta : \mathcal{I}^d \times \mathcal{I}^d \to \mathbb{Z}_+^d$ be a map such that for $(l, m) \in \mathcal{I}^d \times \mathcal{I}^d$, $\frac{\partial^2 K(u)}{\partial u_l\, \partial u_m} = K^{(\eta(l,m))}(u)$, $u \in \mathbb{R}^d$ (see (2.2)). Then

$$\begin{aligned} \Omega_{i,j} &= \sum_{(k_1,k_2)\in\mathcal{I}\times\mathcal{I}} w_{k_1k_2}^{(i,j)} t_{k_1} t_{k_2}, \\ &\text{where } w_{k_1k_2}^{(i,j)} = \int_{\mathbb{R}^d} \left[\frac{\partial}{\partial u_i} K^{(\eta(\pi(k_1)))}(u)\right]\left[\frac{\partial}{\partial u_j} K^{(\eta(\pi(k_2)))}(u)\right] du. \end{aligned} \tag{5.15}$$

Next, we will show that we can write

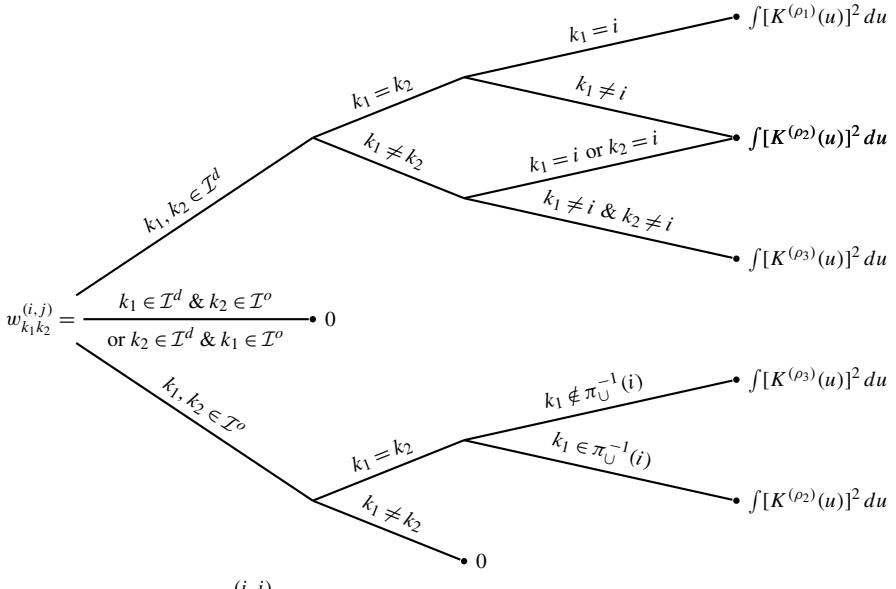
$$\Omega = \int_{\mathbb{R}^d} \left[K^{(\rho_2)}(s)\right]^2 ds\, P, \tag{5.16}$$

where ρ_2 is given in assumption (K3), $P = (p_{ij})$ is a $d \times d$ matrix depending on x and z, and P is positive definite under the given assumptions in this lemma. When $d = 2$, it follows from direct

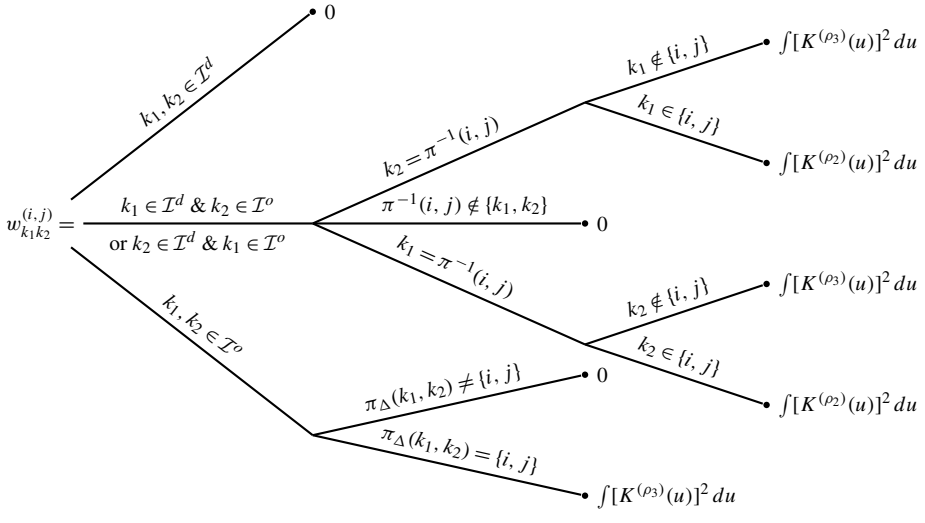
calculation using Lemma B.1 in the supplementary material [38] that the elements of P are given by

$$p_{11} = a_K t_1^2 + t_2^2 + t_3^2 + 2t_1 t_3,$$
$$p_{12} = p_{21} = 2t_1 t_2 + 2t_2 t_3,$$
$$p_{22} = a_K t_3^2 + t_2^2 + t_1^2 + 2t_1 t_3.$$

It is clear from Proposition 3.2 that P is positive definite, when we assume $a_K > 1$. We consider $d \geq 3$ below. Note that $w_{k_1k_2}^{(i,j)} \in \{\int [K^{(\rho_q)}(u)]^2\,du : q = 1, 2, 3\} \cup \{0\}$ and we can determine the values of $w_{k_1k_2}^{(i,j)}$ using Lemma B.1 in the supplementary material [38]. We split our discussion into two cases: $i = j$ and $i \neq j$. When $i = j$, the values of $w_{k_1k_2}^{(i,j)}$ can be determined by the following tree diagram.



When $i \neq j$, the values of $w_{k_1k_2}^{(i,j)}$ can be determined by the following tree diagram.



Plugging these values of $w_{k_1k_2}^{(i,j)}$ into (5.15) we can show that the elements of the matrix P in (5.16) are given by

$$p_{ij} = \begin{cases} \Bigg[\displaystyle\sum_{(k_1,k_2)\in\mathcal{I}^d\times\mathcal{I}^d} a_K^{\delta(i,k_1)\delta(i,k_2)} b_K^{(1-\delta(i,k_1))(1-\delta(i,k_2))(1-\delta(k_1,k_2))} t_{k_1}t_{k_2} \\ \quad + \displaystyle\sum_{k\in\mathcal{I}^o} b_K^{1-\delta(k,\pi_\cup^{-1}(i))} t_k^2 \Bigg] & \text{if } i=j, \\ 2\displaystyle\sum_{k\in\mathcal{I}^d} b_K^{1-\delta(k,\{i,j\})} t_k t_{\pi^{-1}(i,j)} + b_K \displaystyle\sum_{k_1,k_2\in\mathcal{I}^o:\pi_\Delta(k_1,k_2)=\{i,j\}} t_{k_1}t_{k_2} & \text{if } i\neq j. \end{cases} \tag{5.17}$$

We will find a matrix L such that

$$P = LL^T + S, \tag{5.18}$$

where $S = (a_K - 1/b_K)\,\mathrm{diag}(t_1^2, t_2^2, \ldots, t_d^2)$. The matrix L is in the form of $L = (L_1, L_2, L_3)$ and the construction of L_1, L_2, and L_3 is as follows. First, $L_1 = (l_{ij}^{(1)})$ is a $d \times d$ matrix where

$$l_{ij}^{(1)} = \begin{cases} \dfrac{1}{\sqrt{b_K}} t_i + \sqrt{b_K} \displaystyle\sum_{k\in\mathcal{I}^d\setminus\{i\}} t_k & \text{if } i=j, \\ \sqrt{b_K}\, t_{\pi^{-1}(i,j)} & \text{if } i\neq j. \end{cases}$$

$L_2 = (l_{ij}^{(2)})$ is a $d \times \binom{d}{3}$ matrix, whose columns are constructed in the following way. For any $1 \le j_1 < j_2 < j_3 \le d$, a generic column $v = (v_1, \ldots, v_d)^T$ of L_2 is defined by

$$v_i = \begin{cases} \sqrt{b_K}\, t_{\pi^{-1}(j_2,j_3)} & \text{if } i = j_1, \\ \sqrt{b_K}\, t_{\pi^{-1}(j_1,j_3)} & \text{if } i = j_2, \\ \sqrt{b_K}\, t_{\pi^{-1}(j_1,j_2)} & \text{if } i = j_3, \\ 0 & \text{otherwise.} \end{cases} \tag{5.19}$$

$L_3 = (l_{ij}^{(3)})$ is a $d \times [d(d-1)]$ matrix consisting of $\binom{d}{2}$ paired columns. For any $1 \le j_1 < j_2 \le d$, each pair of the generic columns of L_3, denoted by $v^{(1)} = (v_1^{(1)}, \ldots, v_d^{(1)})^T$ and $v^{(2)} = (v_1^{(2)}, \ldots, v_d^{(2)})^T$, are defined by

$$v_i^{(1)} = \begin{cases} \sqrt{1-b_K}\, t_{j_2} & \text{if } i = j_1, \\ \sqrt{1-b_K}\, t_{\pi^{-1}(j_1,j_2)} & \text{if } i = j_2, \\ 0 & \text{otherwise,} \end{cases} \qquad v_i^{(2)} = \begin{cases} \sqrt{1-b_K}\, t_{\pi^{-1}(j_1,j_2)} & \text{if } i = j_1, \\ \sqrt{1-b_K}\, t_{j_1} & \text{if } i = j_2, \\ 0 & \text{otherwise.} \end{cases} \tag{5.20}$$

It is straightforward to verify that (5.18) holds with the above construction. The explicit expressions of P, L and S when $d = 3$ are given as an example in Appendix B of the supplementary material [38].

To show that P is positive definite, using (5.18) and assumption (K3) we only need to show that L is of full rank. This can be seen from the following procedure. Let e_i be the ith standard basis vector of $\mathbb{R}^d$, that is, its ith element is 1 and the rest are zeros. Denote $\widetilde{L}_1 = \frac{1}{\sqrt{b_K}}(t_1e_1, \ldots, t_de_d)$ and $\widetilde{L} = (\widetilde{L}_1, L_2, L_3)$. Below we show that there exists a non-singular $d \times d$ matrix M such that $\widetilde{L} = LM$, which implies that L and $\widetilde{L}$ have the same rank. Here M can be constructed by finding a sequence of

elementary column operations on L, which transform L_1 into $\widetilde{L}_1$. Let $l_i^{(1)}$ and $l_i^{(3)}$ be the ith columns of L_1 and L_3, respectively. The transformation is achieved by simply noticing that

$$l_i^{(1)} - \sum_{k: l_{ik}^{(3)} \in \mathcal{I}_d \setminus \{i\}} \sqrt{\frac{b_K}{1-b_K}} l_k^{(3)} = \frac{1}{\sqrt{b_K}} t_i e_i.$$

Below we will show that there exists at least one column of $\widetilde{L}_1$, L_2 or L_3 in the form of $\frac{1}{\sqrt{b_K}} t_k e_i$, $\sqrt{b_K} t_k e_i$ or $\sqrt{1-b_K} t_k e_i$ for some $t_k \neq 0$, for all $i = 1, 2, \ldots, d$, which implies that L is full rank. This is trivially true if none of $t_1, \ldots, t_d$ is zero. Now assume there is at least one of $t_1, \ldots, t_d$ is zero. Without loss of generality, assume $t_1 = 0$ and we would like to show that there exists at least one column of L_2 or L_3 in the form of

$$\sqrt{b_K} t_k e_1 \quad \text{or} \quad \sqrt{1-b_K} t_k e_1, \tag{5.21}$$

for some $t_k \neq 0$. In the construction of the paired columns $v^{(1)}$ and $v^{(2)}$ of L_3 given in (5.20), take $j_1 = 1$ and let j_2 be any integer such that $1 < j_2 \leq d$. If $t_{\pi^{-1}(1,j_2)} \neq 0$ then $v^{(2)}$ satisfies (5.21); otherwise if $t_{j_2} \neq 0$ then $v^{(1)}$ satisfies (5.21). If neither $v^{(1)}$ nor $v^{(2)}$ satisfies (5.21), then we must have $t_{\pi^{-1}(1,k)} = t_k = 0$ for all $k \in \mathcal{I}^d$ (note that $t_{\pi^{-1}(1,1)} = t_1$), which is what we assume for the rest of the proof. Now we consider the columns in L_2. For v given in (5.19) we take $j_1 = 1$ and let j_2 and j_3 be any two integers satisfying $1 < j_2 < j_3 \leq d$. Then there must exist $t_{\pi^{-1}(j_2,j_3)} \neq 0$ according to (5.14) so that v satisfies (5.21), because $\mathcal{I} = \mathcal{I}^d \cup \mathcal{I}^o$ and $\mathcal{I}^o = \{\pi^{-1}(i, j) : 1 \leq i < j \leq d\}$. □

Proof of Theorem 3.7. We first consider the case $d - r \geq 2$ and then briefly discuss the case $d - r = 1$ at the end of the proof. Recall that $\sigma_g = \sqrt{\mathrm{Var}(\mathbb{B}(g))}$ for $g \in \mathcal{F}_h$. First, we want to show

$$\lim_{h \to 0} \mathbb{P}\Big(\sup_{g \in \mathcal{F}_h} \sigma_g^{-1} \mathbb{B}(g) < b_h\big(z, c_h^{(d,r)}\big) \Big) = e^{-e^{-z}}. \tag{5.22}$$

We need to show that $B(x, z) := \sigma_{g_{x,z}}^{-1} \mathbb{B}(g_{x,z})$ for $g_{x,z} \in \mathcal{F}_h$ satisfies the conditions of the Gaussian fields in Theorem 3.5. Note that $r_h(x, \tilde{x}, z, \tilde{z})$ in (3.19) is the covariance between $B(x, z)$ and $B(\tilde{x}, \tilde{z})$. Proposition 3.6 and Lemma 5.1 can be used to verify that $B(x, z)$, $(x, z) \in \mathcal{M}_h \times \mathbb{S}^{d-r-1}$ is locally equi-$(\alpha_1, D_{x,z}^{(1)}, \alpha_2, D_{x,z}^{(2)})$-stationary (see Definition 3.1), where

$$\alpha_1 = \alpha_2 = 2, \qquad D_{x,z}^{(1)} = \frac{1}{\sqrt{2}} \Omega(x, z)^{1/2}, \quad \text{and} \quad D_{x,z}^{(2)} = \frac{1}{\sqrt{2}} \boldsymbol{I}_{d-r}. \tag{5.23}$$

Also recall that $\beta_1 > 0$ given in Lemma 3.1 is a lower bound of $\Delta(\mathcal{M}_h)$ for all $h \in (0, h_1]$ for some $h_1 > 0$. Without loss of generality we assume that $\beta_1 \leq \delta_0$. Then applying Lemma 3 in [24] we get $\sup_{h \in (0,h_1]} \mathcal{H}_{d-r}(\mathcal{M}_h) \leq \frac{d!}{(d-r)!} \beta_1^{r-d} \mathcal{H}_d(\mathcal{H}) < \infty$. Using Proposition 3.2, we can suppose h_1 is small enough such that $\mathcal{M}_h \subset \mathcal{N}_{\delta_1}(\mathcal{M})$ for all $h \in (0, h_1]$.

Note that (3.16) in Theorem 3.5 is clearly satisfied, simply because the kernel function K is assumed to have bounded support in assumption (K1). We only need to verify that r_h satisfies (3.15). For any $\lambda \in \mathbb{R}$, $x, \widetilde{x} \in \mathcal{M}_h$ and $z, \widetilde{z} \in \mathbb{S}^{d-r-1}$, let $\kappa(\lambda; X_1, x, \widetilde{x}, z, \widetilde{z}, h) = g_{x,z}(X_1) - \lambda g_{\widetilde{x},\widetilde{z}}(X_1)$ and

$$\zeta(\lambda; X_1, x, \widetilde{x}, z, \widetilde{z}, h) = \big[\kappa(\lambda; X_1, x, \widetilde{x}, z, \widetilde{z}, h) - \mathbb{E}\kappa(\lambda; X_1, x, \widetilde{x}, z, \widetilde{z}, h)\big]^2.$$

Denote $B(x,\widetilde{x},h)=\mathcal{B}(x,h)\cup\mathcal{B}(\widetilde{x},h)$. Using the boundedness of the support of K and the Cauchy-Schwarz inequality we have

$$\begin{aligned}
&\big[\mathbb{E}\kappa(\lambda;X_1,x,\widetilde{x},z,\widetilde{z},h)\big]^2\\
&\quad=\frac{1}{h^d}\left\{\int_{\mathbb{R}^d}\left[\left\langle A(x,z),d^2K\left(\frac{x-s}{h}\right)\right\rangle-\lambda\left\langle A(\widetilde{x},\widetilde{z}),d^2K\left(\frac{\widetilde{x}-s}{h}\right)\right\rangle\right]f(s)\,ds\right\}^2\\
&\quad=\frac{1}{h^d}\left\{\int_{B(x,\tilde{x},h)}\left[\left\langle A(x,z),d^2K\left(\frac{x-s}{h}\right)\right\rangle-\lambda\left\langle A(\widetilde{x},\widetilde{z}),d^2K\left(\frac{\widetilde{x}-s}{h}\right)\right\rangle\right]\sqrt{f(s)}\sqrt{f(s)}\,ds\right\}^2\\
&\quad\le\mathbb{E}\big[\kappa(\lambda;X_1,x,\widetilde{x},z,\widetilde{z},h)^2\big]F(x,\widetilde{x},h),
\end{aligned}$$

where $F(x,\widetilde{x},h)=\int_{B(x,\tilde{x},h)}f(s)\,ds=O(h^d)$, uniformly in $x,\tilde{x}\in\mathcal{M}_h$ for all $0<h\le h_1$. This implies that there exists $h_2\in(0,h_1]$ such that for all $0<h\le h_2$,

$$\begin{aligned}
\mathbb{E}\zeta(\lambda;X_1,x,\widetilde{x},z,\widetilde{z},h)&=\mathbb{E}\big[\kappa(\lambda;X_1,x,\widetilde{x},z,\widetilde{z},h)^2\big]-\big[\mathbb{E}\kappa(\lambda;X_1,x,\widetilde{x},z,\widetilde{z},h)\big]^2\\
&\ge\frac{1}{2}\mathbb{E}\big[\kappa(\lambda;X_1,x,\widetilde{x},z,\widetilde{z},h)^2\big].
\end{aligned}\tag{5.24}$$

Denote $\Delta x=\widetilde{x}-x$ and $\Delta z=\widetilde{z}-z$. Due to the bounded support of K we have

$$\begin{aligned}
\mathbb{E}\big[\kappa(\lambda;X_1,x,\widetilde{x},z,\widetilde{z},h)^2\big]&\ge\mathbb{E}\big\{\mathbf{1}_{\mathcal{B}(x,h)\setminus\mathcal{B}(\tilde{x},h)}(X_1)\kappa(\lambda;X_1,x,\widetilde{x},z,\widetilde{z},h)^2\big\}\\
&=\mathbb{E}\big\{\mathbf{1}_{\mathcal{B}(x,h)\setminus\mathcal{B}(\tilde{x},h)}(X_1)\kappa(0;X_1,x,\widetilde{x},z,\widetilde{z},h)^2\big\}\\
&=\mathbb{E}\big\{\mathbf{1}_{\mathcal{B}(x,h)\setminus\mathcal{B}(\tilde{x},h)}(X_1)\big[g_{x,z}(X_1)\big]^2\big\}\\
&=\int_{\mathcal{B}(0,1)\setminus\mathcal{B}(\Delta x/h,1)}\big\langle A(x,z),d^2K(s)\big\rangle^2f(x-hs)\,ds\\
&=f(x)\int_{\mathcal{B}(0,1)\setminus\mathcal{B}(\Delta x/h,1)}\big\langle A(x,z),d^2K(s)\big\rangle^2\,ds+O(h),
\end{aligned}\tag{5.25}$$

where in the last step we use a Taylor expansion for $f(x-hs)$ and the $O(h)$-term is uniform in $x,\widetilde{x}\in\mathcal{M}_h$ for all $0<h\le h_2$ and $z,\widetilde{z}\in\mathbb{S}^{d-r-1}$.

Note that for any $\delta>0$, if $\|\Delta x\|>h\delta$, then the set $\mathcal{B}(0,1)\setminus\mathcal{B}(\Delta x/h,1)$ contains a ball $\mathcal{B}^*$ with radius $\min(1,\delta/2)$. It follows that for any $x\in\mathcal{M}_h$, $0<h\le h_2$ and $z\in\mathbb{S}^{d-r-1}$,

$$\inf_{\|\Delta x\|>h\delta}\int_{\mathcal{B}(0,1)\setminus\mathcal{B}(\Delta x/h,1)}\big\langle A(x,z),d^2K(s)\big\rangle^2\,ds\ge\int_{\mathcal{B}^*}\big\langle A(x,z),d^2K(s)\big\rangle^2\,ds.$$

Recall that $\sqrt{f(x)}\|A(x,z)\|_{\boldsymbol{R}}=1$ and hence $A(x,z)\ne0$ for all $x\in\mathcal{N}_{\delta_1}(\mathcal{M})$ and $z\in\mathbb{S}^{d-r-1}$. Then $\int_{\mathcal{B}^*}\langle A(x,z),d^2K(s)\rangle^2\,ds>0$ under assumption (K2). Without loss of generality suppose that δ_1 is small enough such that $\mathcal{N}_{\delta_1}(\mathcal{M})$ is compact (see the proof of Lemma 3.1). Since $\mathbb{S}^{d-r-1}$ is also compact and $A(x,z)$ in continuous in $x\in\mathcal{N}_{\delta_1}(\mathcal{M})$ and $z\in\mathbb{S}^{d-r-1}$, we have

$$\inf_{0<h\le h_2}\inf_{x\in\mathcal{M}_h,z\in\mathbb{S}^{d-r-1}}\inf_{\|\Delta x\|>h\delta}\int_{\mathcal{B}(0,1)\setminus\mathcal{B}(\Delta x/h,1)}\big\langle A(x,z),d^2K(s)\big\rangle^2\,ds>0,$$

which by (5.24) and (5.25) further implies that for some $h_0 \in (0, h_2]$,

$$\inf_{\substack{x,\tilde{x}\in\mathcal{M}_h,z,\tilde{z}\in\mathbb{S}^{d-r-1}\\ \|\Delta x\|>h\delta,\|\Delta z\|>\delta,0<h\le h_0}} \mathbb{E}\zeta(\lambda; X_1, x, \widetilde{x}, z, \widetilde{z}, h) > 0. \tag{5.26}$$

Note that $\mathbb{E}\zeta(\lambda; X_1, x, \widetilde{x}, z, \widetilde{z}, h) = \lambda^2\sigma^2_{g_{\tilde{x},\tilde{z}}} - 2\lambda\operatorname{Cov}(g_{\tilde{x},\tilde{z}}(X_1), g_{x,z}(X_1)) + \sigma^2_{g_{x,z}}$, which is a quadratic polynomial in λ and its discriminant is given by

$$\sigma(x, \widetilde{x}, z, \widetilde{z}, h) = 4\operatorname{Cov}\big(g_{\tilde{x},\tilde{z}}(X_1), g_{x,z}(X_1)\big) - 4\sigma^2_{g_{x,z}}\sigma^2_{g_{\tilde{x},\tilde{z}}}.$$

Then (5.26) implies that

$$\sup_{\substack{x,\tilde{x}\in\mathcal{M}_h,z,\tilde{z}\in\mathbb{S}^{d-r-1}\\ \|\Delta x\|>h\delta,\|\Delta z\|>\delta,0<h\le h_0}} \sigma(x, \widetilde{x}, z, \widetilde{z}, h) < 0,$$

or equivalently,

$$\sup_{\substack{x,\tilde{x}\in\mathcal{M}_h,z,\tilde{z}\in\mathbb{S}^{d-r-1}\\ \|\Delta x\|>h\delta,\|\Delta z\|>\delta,0<h\le h_0}} \big|r_h(x, \widetilde{x}, z, \widetilde{z})\big| < 1.$$

Thus the condition in (3.15) has been verified.

With $\beta_h = \sqrt{2r\log(h^{-1})} + \frac{1}{\sqrt{2r\log(h^{-1})}}[\frac{d-2}{2}\log\log(h^{-1}) + c_h^{(d,r)}]$, Theorem 3.5 yields

$$\lim_{h\to 0}\mathbb{P}\Big\{\sqrt{2r\log\big(h^{-1}\big)}\Big(\sup_{g\in\mathcal{F}_h}\sigma_g^{-1}\mathbb{B}(g) - \beta_h\Big) \le z\Big\} = e^{-e^{-z}}, \tag{5.27}$$

where in the calculation of $c_h^{(d,r)}$ we use (5.23) and $H_m^{(2)} = \pi^{-m/2}$ for any $m \in \mathbb{Z}^+$, which is a well-known fact for Pickands' constant (see page 31 of [34]). This is (5.22).

For $g_{x,z} \in \mathcal{F}_h$ we have

$$\begin{aligned}
\sigma^2_{g_{x,z}} &= \mathbb{E}\big[g_{x,z}(X_1)^2\big] - \big[\mathbb{E}g_{x,z}(X_1)\big]^2\\
&= \frac{1}{h^d}\int_{\mathbb{R}^d}\left\langle A(x,z), d^2K\left(\frac{x-u}{h}\right)\right\rangle^2 f(u)\,du - \frac{1}{h^d}\left[\int_{\mathbb{R}^d}\left\langle A(x,z), d^2K\left(\frac{x-u}{h}\right)\right\rangle f(u)\,du\right]^2\\
&= \int_{\mathbb{R}^d}\big\langle A(x,z), d^2K(u)\big\rangle^2 f(x-hu)\,du - h^d\left[\int_{\mathbb{R}^d}\big\langle A(x,z), d^2K(u)\big\rangle f(x-hu)\,du\right]^2\\
&= 1 + O\big(h^2\big),
\end{aligned}$$

where the $O(h^2)$-term is uniform in $x \in \mathcal{M}_h$ for $0 < h \le h_0$ and $z \in \mathbb{S}^{d-r-1}$. Note that (5.27) implies that $\sup_{g\in\mathcal{F}_h}|\sigma_g^{-1}\mathbb{B}(g)| = O_p(\sqrt{\log(h^{-1})})$ and hence

$$\Big|\sup_{g\in\mathcal{F}_h}\mathbb{B}(g) - \sup_{g\in\mathcal{F}_h}\sigma_g^{-1}\mathbb{B}(g)\Big| \le \sup_{g\in\mathcal{F}_h}\big|(\sigma_g - 1)\sigma_g^{-1}\mathbb{B}(g)\big| = O_p\big(h^2\sqrt{\log\big(h^{-1}\big)}\big).$$

We then get (3.24) by using (5.27). By Theorem 3.4, for D_n defined in (3.8) we have

$$\mathbb{P}\Big(\sqrt{nh^{d+4}}\sup_{x\in\mathcal{M}_h} D_n(x) \le b_h\big(z, c_h^{(d,r)}\big)\Big) \to e^{-e^{-z}}. \tag{5.28}$$

Next we show (3.25). It follows from Lemma A.1 in the supplementary material [38] and Lemma 3.1 that

$$\mathbb{P}\big(\mathcal{M}_h \subset \big\{x \in \mathcal{H} : \widehat{\lambda}_{r+1}(x) < 0\big\}\big) \to 1. \tag{5.29}$$

Let $\widehat{C}^*_{n,h}(a) = \{x \in \mathcal{H} : \sqrt{nh^{d+4}} B_n(x) \le a\}$, for $a \ge 0$. Then by (5.29) we get

$$\sup_{a \ge 0} \big|\mathbb{P}\big(\mathcal{M}_h \subset \widehat{C}_{n,h}(a)\big) - \mathbb{P}\big(\mathcal{M}_h \subset \widehat{C}^*_{n,h}(a)\big)\big| \to 0. \tag{5.30}$$

Furthermore it is clear that $\mathbb{P}(\mathcal{M}_h \subset \widehat{C}^*_{n,h}(a)) = \mathbb{P}(\sqrt{nh^{d+4}} \sup_{x \in \mathcal{M}_h} B_n(x) \le a)$ for all $a \ge 0$. By applying Proposition 3.3 and (5.28), we finish the proof of (3.25) for the case $d - r \ge 2$. When $d - r = 1$, the covariance structure of $\mathbb{B}$ is simplified (see Remark 3.5). Then instead of using Theorem 3.5, we apply the main theorem in [41]. The rest of the proof is similar to the above. □

Acknowledgements

We would like to thank Anand Vidyashankar and three anonymous referees for their insightful comments that have led to significant improvements of the paper. The author's work is partially supported by NSF Grant No. 1821154 and Grant No. 1900061.

Supplementary Material

Supplement to "Asymptotic confidence regions for density ridges" (DOI: 10.3150/20-BEJ1261SUPP; .pdf). This supplementary material presents additional proofs that are not shown in Section 5 due to page constraints, as well as some miscellaneous results.

References

[1] Aamari, E., Kim, J., Chazal, F., Michel, B., Rinaldo, A. and Wasserman, L. (2019). Estimating the reach of a manifold. *Electron. J. Stat.* **13** 1359–1399. MR3938326 https://doi.org/10.1214/19-ejs1551

[2] Arias-Castro, E., Donoho, D.L. and Huo, X. (2006). Adaptive multiscale detection of filamentary structures in a background of uniform random points. *Ann. Statist.* **34** 326–349. MR2275244 https://doi.org/10.1214/009053605000000787

[3] Berenfeld, C., Harvey, J., Hoffmann, M. and Shankar, K. (2020). Estimating the reach of a manifold via its convexity defect function. arXiv:2001.08006.

[4] Bickel, P.J. and Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates. *Ann. Statist.* **1** 1071–1095. MR0348906

[5] Broida, J.G. and Williamson, S.G. (1989). *A Comprehensive Introduction to Linear Algebra*. Redwood City, CA: Addison-Wesley Company. Advanced Book Program. MR1045200

[6] Bugni, F.A. (2010). Bootstrap inference in partially identified models defined by moment inequalities: Coverage of the identified set. *Econometrica* **78** 735–753. MR2656646 https://doi.org/10.3982/ECTA8056

[7] Cadre, B. (2006). Kernel estimation of density level sets. *J. Multivariate Anal.* **97** 999–1023. MR2256570 https://doi.org/10.1016/j.jmva.2005.05.004

[8] Chen, Y.-C., Genovese, C.R. and Wasserman, L. (2015). Asymptotic theory for density ridges. *Ann. Statist.* **43** 1896–1928. MR3375871 https://doi.org/10.1214/15-AOS1329

[9] Chen, Y.-C., Genovese, C.R. and Wasserman, L. (2017). Density level sets: Asymptotics, inference, and visualization. *J. Amer. Statist. Assoc.* **112** 1684–1696. MR3750891 https://doi.org/10.1080/01621459.2016.1228536

[10] Cheng, M.-Y., Hall, P. and Hartigan, J.A. (2004). Estimating gradient trees. In *A Festschrift for Herman Rubin*. *Institute of Mathematical Statistics Lecture Notes – Monograph Series* **45** 237–249. Beachwood, OH: IMS. MR2126901 https://doi.org/10.1214/lnms/1196285394

[11] Chernozhukov, V., Chetverikov, D. and Kato, K. (2014). Gaussian approximation of suprema of empirical processes. *Ann. Statist.* **42** 1564–1597. MR3262461 https://doi.org/10.1214/14-AOS1230

[12] Dunajeva, O. (2004). The second-order derivatives of matrices of eigenvalues and eigenvectors with an application to generalized F-statistic. *Linear Algebra Appl.* **388** 159–171. MR2077857 https://doi.org/10.1016/j.laa.2003.08.019

[13] Duong, T., Cowling, A., Koch, I. and Wand, M.P. (2008). Feature significance for multivariate kernel density estimation. *Comput. Statist. Data Anal.* **52** 4225–4242. MR2432459 https://doi.org/10.1016/j.csda.2008.02.035

[14] Eberly, D. (1996). *Ridges in Image and Data Analysis*. Boston, MA: Kluwer.

[15] Federer, H. (1959). Curvature measures. *Trans. Amer. Math. Soc.* **93** 418–491. MR0110078 https://doi.org/10.2307/1993504

[16] Genovese, C., Perone-Pacifico, M., Verdinelli, I. and Wasserman, L. (2017). Finding singular features. *J. Comput. Graph. Statist.* **26** 598–609. MR3698670 https://doi.org/10.1080/10618600.2016.1260472

[17] Genovese, C.R., Perone-Pacifico, M., Verdinelli, I. and Wasserman, L. (2009). On the path density of a gradient field. *Ann. Statist.* **37** 3236–3271. MR2549559 https://doi.org/10.1214/08-AOS671

[18] Genovese, C.R., Perone-Pacifico, M., Verdinelli, I. and Wasserman, L. (2012). The geometry of nonparametric filament estimation. *J. Amer. Statist. Assoc.* **107** 788–799. MR2980085 https://doi.org/10.1080/01621459.2012.682527

[19] Genovese, C.R., Perone-Pacifico, M., Verdinelli, I. and Wasserman, L. (2014). Nonparametric ridge estimation. *Ann. Statist.* **42** 1511–1545. MR3262459 https://doi.org/10.1214/14-AOS1218

[20] Hall, P. (1992). Effect of bias estimation on coverage accuracy of bootstrap confidence intervals for a probability density. *Ann. Statist.* **20** 675–694. MR1165587 https://doi.org/10.1214/aos/1176348651

[21] Hall, P., Qian, W. and Titterington, D.M. (1992). Ridge finding from noisy data. *J. Comput. Graph. Statist.* **1** 197–211. MR1270818 https://doi.org/10.2307/1390716

[22] Hartigan, J.A. (1987). Estimation of a convex density contour in two dimensions. *J. Amer. Statist. Assoc.* **82** 267–270. MR0883354

[23] Ipsen, I.C.F. and Rehman, R. (2008). Perturbation bounds for determinants and characteristic polynomials. *SIAM J. Matrix Anal. Appl.* **30** 762–776. MR2421470 https://doi.org/10.1137/070704770

[24] Kim, J., Rinaldo, A. and Wasserman, L. (2019). Minimax rates for estimating the dimension of a manifold. *J. Comput. Geom.* **10** 42–95. MR3918925 https://doi.org/10.1214/19-ejs1551

[25] Konakov, V.D. and Piterbarg, V.I. (1984). On the convergence rate of maximal deviation distribution for kernel regression estimates. *J. Multivariate Anal.* **15** 279–294. MR0768499 https://doi.org/10.1016/0047-259X(84)90053-8

[26] Konstantinides, D.G., Piterbarg, V. and Stamatovic, S. (2004). Gnedenko-type limit theorems for cyclo stationary χ^2-processes. *Liet. Mat. Rink.* **44** 196–208. MR2116482 https://doi.org/10.1023/B:LIMA.0000033781.86969.c9

[27] Lee, J.M. (2013). *Introduction to Smooth Manifolds*, 2nd ed. *Graduate Texts in Mathematics* **218**. New York: Springer. MR2954043

[28] Li, W. and Ghosal, S. (2020). Posterior contraction and credible sets for filaments of regression functions. *Electron. J. Stat.* **14** 1707–1743. MR4083733 https://doi.org/10.1214/20-EJS1705

[29] Magnus, J.R. and Neudecker, H. (2007). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. 3rd ed. Chichester: Wiley. MR1698873

[30] Mammen, E. and Polonik, W. (2013). Confidence regions for level sets. *J. Multivariate Anal.* **122** 202–214. MR3189318 https://doi.org/10.1016/j.jmva.2013.07.017

[31] Mason, D.M. and Polonik, W. (2009). Asymptotic normality of plug-in level set estimates. *Ann. Appl. Probab.* **19** 1108–1142. MR2537201 https://doi.org/10.1214/08-AAP569

[32] Ozertem, U. and Erdogmus, D. (2011). Locally defined principal curves and surfaces. *J. Mach. Learn. Res.* **12** 1249–1286. MR2804600

[33] Piterbarg, V.I. (1994). High excursions for nonstationary generalized chi-square processes. *Stochastic Process. Appl.* **53** 307–337. MR1302916 https://doi.org/10.1016/0304-4149(94)90068-X

[34] Piterbarg, V.I. (1996). *Asymptotic Methods in the Theory of Gaussian Processes and Fields. Translations of Mathematical Monographs* **148**. Providence, RI: Amer. Math. Soc. Translated from the Russian by V.V. Piterbarg, Revised by the author. MR1361884 https://doi.org/10.1090/mmono/148

[35] Polonik, W. (1995). Measuring mass concentrations and estimating density contour clusters – an excess mass approach. *Ann. Statist.* **23** 855–881. MR1345204 https://doi.org/10.1214/aos/1176324626

[36] Polonik, W. and Wang, Z. (2005). Estimation of regression contour clusters – an application of the excess mass approach to regression. *J. Multivariate Anal.* **94** 227–249. MR2167913 https://doi.org/10.1016/j.jmva.2004.05.001

[37] Qiao, W. (2020). Extremes of locally stationary Gaussian and chi fields on manifolds. *Stochastic Process. Appl.* https://doi.org/10.1016/j.spa.2020.11.006

[38] Qiao, W. (2021). Supplement to "Asymptotic confidence regions for density ridges." https://doi.org/10.3150/20-BEJ1261SUPP

[39] Qiao, W. (2021). Nonparametric estimation of surface integrals on level sets. *Bernoulli* **27** 155–191. MR4177365 https://doi.org/10.3150/20-BEJ1232

[40] Qiao, W. and Polonik, W. (2016). Theoretical analysis of nonparametric filament estimation. *Ann. Statist.* **44** 1269–1297. MR3485960 https://doi.org/10.1214/15-AOS1405

[41] Qiao, W. and Polonik, W. (2018). Extrema of rescaled locally stationary Gaussian fields on manifolds. *Bernoulli* **24** 1834–1859. MR3757516 https://doi.org/10.3150/16-BEJ913

[42] Qiao, W. and Polonik, W. (2019). Nonparametric confidence regions for level sets: Statistical properties and geometry. *Electron. J. Stat.* **13** 985–1030. MR3934621 https://doi.org/10.1214/19-EJS1543

[43] Rosenblatt, M. (1952). Remarks on a multivariate transformation. *Ann. Math. Stat.* **23** 470–472. MR0049525 https://doi.org/10.1214/aoms/1177729394

[44] Rosenblatt, M. (1976). On the maximal deviation of k-dimensional density estimates. *Ann. Probab.* **4** 1009–1015. MR0428580 https://doi.org/10.1214/aop/1176995945

[45] Sousbie, T., Pichon, C., Colombi, S., Novikov, D. and Pogosyan, D. (2008). The 3D skeleton: Tracing the filamentary structure of the Universe. *Mon. Not. R. Astron. Soc.* **383** 1655–1670.

[46] Tsybakov, A.B. (1997). On nonparametric estimation of density level sets. *Ann. Statist.* **25** 948–969. MR1447735 https://doi.org/10.1214/aos/1069362732

[47] von Luxburg, U. (2007). A tutorial on spectral clustering. *Stat. Comput.* **17** 395–416. MR2409803 https://doi.org/10.1007/s11222-007-9033-z

[48] Wegman, E.J., Carr, D.B. and Luo, Q. (1993). Visualizing multivariate data. In *Multivariate Analysis*: *Future Directions* (C.R. Rao, ed.). Amsterdam: North-Holland.

[49] Wegman, E.J. and Luo, Q. (2002). Smoothings, ridges, and bumps. In *Proceedings of the ASA* (*Published on CD*). *Development of the Relationship Between Geometric Aspects of Visualizing Densities and Density Approximators*, *and a Discussion of Rendering and Lighting Models*, *Contouring Algorithms*, *Stereoscopic Display Algorithms*, *and Visual Design Considerations* 3666–3672. American Statistical Association.

[50] Xia, Y. (1998). Bias-corrected confidence bands in nonparametric regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **60** 797–811. MR1649488 https://doi.org/10.1111/1467-9868.00155

Received October 2019 and revised July 2020