The Annals of Statistics
2014, Vol. 42, No. 4, 1511–1545
DOI: 10.1214/14-AOS1218

NONPARAMETRIC RIDGE ESTIMATION

BY CHRISTOPHER R. GENOVESE[1], MARCO PERONE-PACIFICO[2], ISABELLA VERDINELLI[2] AND LARRY WASSERMAN[3]

Carnegie Mellon University, Sapienza University of Rome, Carnegie Mellon University and Sapienza University of Rome, and Carnegie Mellon University

We study the problem of estimating the ridges of a density function. Ridge estimation is an extension of mode finding and is useful for understanding the structure of a density. It can also be used to find hidden structure in point cloud data. We show that, under mild regularity conditions, the ridges of the kernel density estimator consistently estimate the ridges of the true density. When the data are noisy measurements of a manifold, we show that the ridges are close and topologically similar to the hidden manifold. To find the estimated ridges in practice, we adapt the modified mean-shift algorithm proposed by Ozertem and Erdogmus [*J. Mach. Learn. Res.* **12** (2011) 1249–1286]. Some numerical experiments verify that the algorithm is accurate.

1. Introduction. Multivariate data in many problems exhibit intrinsic lower dimensional structure. The existence of such structure is of great interest for dimension reduction, clustering and improved statistical inference, and the question of how to identify and characterize this structure is the focus of active research. A commonly used representation for low-dimensional structure is a smooth manifold. Unfortunately, estimating manifolds can be difficult even under mild assumptions. For instance, the rate of convergence for estimating a manifold with bounded curvature blurred by homogeneous Gaussian noise, is *logarithmic* [Genovese et al. (2012a)], meaning that an exponential amount of data are needed to attain a specified level of accuracy. In this paper, we offer a way to circumvent this problem. We define an object, which we call a *hyper-ridge* set that can be used to approximate the low-dimensional structure in a data set. We show that the hyper-ridge set captures the essential features of the underlying low-dimensional structure while being estimable from data at a polynomial rate.

Let $X_1, \ldots, X_n$ be a sample from a probability density p defined on an open subset of D-dimensional Euclidean space and let $\hat{p}$ be an estimate of the density. We will define hyper-ridge sets (called ridges for short) for both p and $\hat{p}$, which

Received June 2013; revised March 2014.
[1]Supported by NSF Grant DMS-08-06009.
[2]Supported by Italian National Research Grant PRIN 2008.
[3]Supported by NSF Grant DMS-08-06009, Air Force Grant FA95500910373.
MSC2010 subject classifications. Primary 62G05, 62G20; secondary 62H12.
Key words and phrases. Ridges, density estimation, manifold learning.

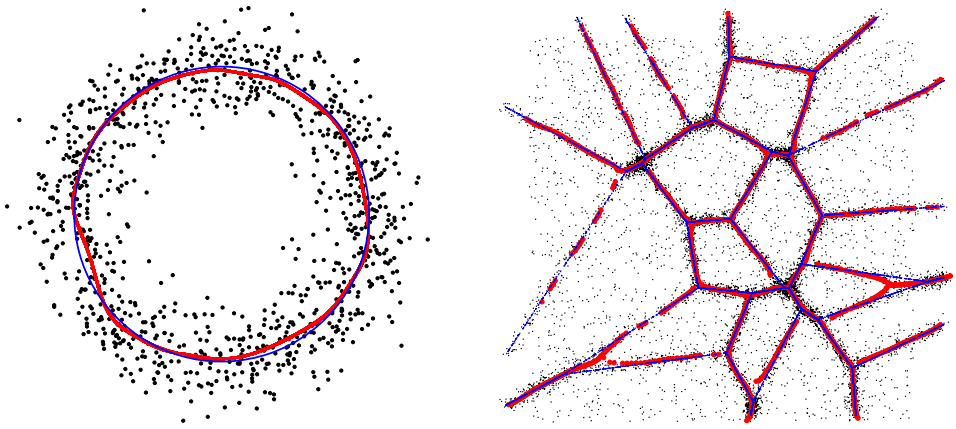

FIG. 1. *Synthetic data showing lower dimensional structure. The left plot is an example of the hidden manifold case. The right plot is an example of a hidden set consisting of intersecting manifolds.*

we denote by R and $\hat{R}$. We consider two cases that make different assumptions about p. In the *hidden manifold* case (see Figure 1), we assume that the density p is derived by sampling from a $d < D$ dimensional manifold M and adding D-dimensional noise. In the *density ridge* case, we look for ridges of a density without assuming any hidden manifold, simply as a way of finding structure in a point cloud, much like clustering. The goal in both cases is to estimate the hyper-ridge set. Although in the former case, we would ideally like to estimate M, this is not always feasible for reasonable sample sizes, so we use the ridge R as a *surrogate* for M. We focus on estimating ridges from point cloud data; we do not consider image data in this paper.

A formal definition of a ridge is given in Section 2. Let $1 \leq d < D$ be fixed. Loosely speaking, we define a d-dimensional hyper-ridge set of a density p to be the points where the Hessian of p has $D - d$ strongly negative eigenvalues and where the projection of the gradient on that subspace is zero. Put another way, the ridge is a local maximizer of the density when moving in the normal direction defined by the Hessian.

Yet another way to think about ridges is by analogy with modes. We can define a mode to be a point where the gradient is 0 and the second derivative is negative, that is, the eigenvalues of the Hessian are negative. The Hessian defines a $(D-d)$-dimensional normal space (corresponding to the $D - d$ smallest eigenvalues) and a d dimensional tangent space. A ridge point has a projected gradient (the gradient in the direction of the normal) that is 0 and eigenvalues in the normal space that are negative. Modes are simply 0 dimensional ridges.

EXAMPLE. A stylized example is shown in Figure 2. In this example, the density is $p(x) = \int_M \phi(x - z)w(z)\,dz$ where $x \in \mathbb{R}^2$, M is a circle in $\mathbb{R}^2$, w is a smooth (but nonuniform) density supported on M and ϕ is a two-dimensional

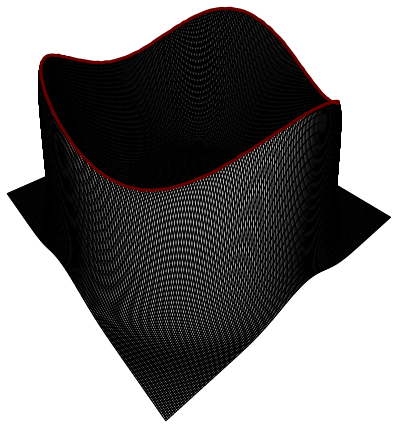

FIG. 2. *An example of a one dimensional ridge defined by a two-dimensional density p. The ridge R is a circle on the plane. The solid curve is the ridge, lifted onto p, that is, $\{(x, p(x)) : x \in R\}$.*

Gaussian with a variance σ^2 that is much smaller than the radius of the circle. The ridge R is a one-dimensional subset of $\mathbb{R}^2$. The figure has a solid curve to show the ridge lifted onto p, that is, the curve shows the set $\{(x, p(x)) : x \in R\}$. The ridge R does not coincide exactly with M due to the blurring by convolution with the Gaussian. In fact, R is a circle with slightly smaller radius than M. That is, R is a biased version of M. Figure 3 shows M and R.

Note that the density is not uniform over the ridge. Indeed, there can be modes (0-dimensional ridges) within a ridge. What matters is that the function rises sharply as we approach the ridge (strongly negative eigenvalue).

One of the main points of this paper is that R captures the essential features of M. If we can live with the slight bias in R, then it is better to estimate R since R can be estimated at a polynomial rate while M can only be estimated at a logarithmic rate. Throughout this paper, we take the dimension of interest d as fixed and given.

Many different and useful definitions of a "ridge" have been proposed; see the discussion of related work at the end of this section. We make no claim as to the uniqueness and optimality of ours. Our definition is motivated by four useful properties that we demonstrate in this paper:

1. If $\hat{p}$ is close to p, then $\hat{R}$ is close to R where $\hat{R}$ is the ridge of $\hat{p}$ and R is the ridge of p.
2. If the data-generating distribution is concentrated near a manifold M, then the ridge R approximates M both geometrically and topologically.
3. R can be estimated at a polynomial rate, even in cases where M can be estimated at only a logarithmic rate.

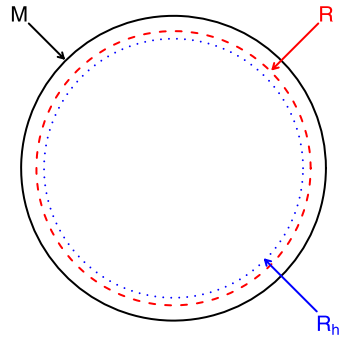

FIG. 3. *The outer circle denotes the manifold M. The dashed circle is the ridge R of the density p. The ridge is a biased version of M and acts as a surrogate for M. The inner circle R_h shows the ridge from a density estimator with bandwidth h. R can be estimated at a much faster rate than M.*

4. The definition corresponds essentially with the algorithm derived by Ozertem and Erdogmus (2011). That is, our definition provides a mathematical formalization of their algorithm.

Our broad goal is to provide a theoretical framework for understanding the problem of estimating hyper-ridge sets. In particular, we show that the ridges of a kernel density estimator consistently estimate the ridges of the density, and we find and upper bound on the rate of convergence. The main results of this paper are (stated here informally):

- *Stability* (Theorem 4). If two densities are sufficiently close together, their hyper-ridge sets are also close together.
- *Estimation* (Theorem 5). There is an estimator $\hat{R}$ such that

$$\mathsf{Haus}(R, \hat{R}) = O_P\left(\left(\frac{\log n}{n}\right)^{2/(D+8)}\right), \tag{1}$$

  where Haus is the Hausdorff distance, defined in equation (9). Moreover, $\hat{R}$ is topologically similar to R in the sense that small dilations of these sets are topologically similar.
- *Surrogate* (Theorem 7). In the Hidden Manifold case with small noise variance σ^2 and assuming M has no boundary, the hyper-ridge set of the density p satisfies

$$\mathsf{Haus}(M, R) = O\big(\sigma^2 \log(1/\sigma)\big) \tag{2}$$

  and R is topologically similar to M. Hence, when the noise σ is small, the ridge is close to M. Note that we treat M as fixed while $\sigma \to 0$. It then follows that

$$\mathsf{Haus}(M, \hat{R}) = O_P\left(\left(\frac{\log n}{n}\right)^{2/(D+8)}\right) + O\big(\sigma^2 \log(1/\sigma)\big). \tag{3}$$

This leaves open the question of how to locate the ridges of the density estimator. Fortunately, this latter problem has recently been solved by Ozertem and Erdogmus (2011) who derived a practical algorithm called the *subspace constrained mean shift (SCMS) algorithm* for locating the ridges. Ozertem and Erdogmus (2011) derived their method assuming that the underlying density function is known (i.e., they did not discuss the effect of estimation error). We, instead, assume the density is estimated from a finite sample and adapt their algorithm accordingly by including a denoising step in which we discard points with low density. This paper provides a statistical justification for, and extension to, their algorithm. We introduce a modification of their algorithm called SuRF (Subspace Ridge Finder) that applies density estimation, followed by denoising, followed by SCMS.

Related work. Zero dimensional ridges are modes and in this case ridge finding reduces to mode estimation and SCMS reduces to the mean shift clustering algorithm [Chacón (2012), Cheng (1995), Fukunaga and Hostetler (1975), Li, Ray and Lindsay (2007)].

If the hidden structure is a manifold, then the process of finding the structure is known as *manifold estimation* or *manifold learning*. There is a large literature on manifold estimation and related techniques. Some useful references are Niyogi, Smale and Weinberger (2008) Caillerie et al. (2011), Genovese et al. (2009, 2012a, 2012b, 2012c), Roweis and Saul (2000), Tenenbaum, de Silva and Langford (2000) and references therein.

The notion of ridge finding spans many fields. Previous work on ridge finding in the statistics literature includes Cheng, Hall and Hartigan (2004), Hall, Peng and Rau (2001), Wegman, Carr and Luo (1993), Wegman and Luo (2002) and Hall, Qian and Titterington (1992). These papers focus on visualization and exploratory analysis. An issue that has been discussed extensively in the applied math and computer science literature is how to define a ridge. A detailed history and taxonomy is given in the text by Eberly (1996). Two important classes of ridges are watershed ridges, which are global in nature, and height ridges, which are locally defined. There is some debate about the virtues of various definitions. See, for example, Norgard and Bremer (2012), Peikert, Günther and Weinkauf (2012). Related definitions also appear in the fluid dynamics literature [Schindler et al. (2012)] and astronomy [Aragón-Calvo et al. (2010), Sousbie et al. (2008)]. There is also a literature on Reeb graphs [Ge et al. (2011)] and metric graphs [Aanjaneya et al. (2012), Lecci, Rinaldo and Wasserman (2013)]. Metric graph methods are ideal for representing intersecting filamentary structure but are much more sensitive to noise than the methods in this paper. It is not our intent in this paper to argue that one particular definition of ridge is optimal for all purposes. Rather, we use a particular definition which is well suited for studying the statistical estimation of ridges.

More generally, there is a vast literature on hunting for structure in point clouds and analyzing the shapes of densities. Without attempting to be exhaustive, some representative work includes Davenport et al. (2010), Adams, Atanasov and Carlsson (2011), Bendich, Wang and Mukherjee (2012), Chazal et al. (2011), Klemelä (2009).

Throughout the paper, we use symbols like $C, C_0, C_1, c, c_0, c_1, \ldots$ to denote generic positive constants whose value may be different in different expressions.

2. Model and ridges. In this section, we describe our assumptions about the data and give a formal definition of hyper-ridge sets, which we call *ridges* from now on. Further properties of ridges are stated and proved in Section 4.

We start with a point cloud $X_1, \ldots, X_n \in \mathbb{R}^D$. We assume that these data comprise a random sample from a distribution P with density p, where p has at least five bounded, continuous derivatives. This is all we assume for the density ridge case. In the hidden manifold case, we assume further that P and p are derived from a d-dimensional manifold M by convolution with a noise distribution, where $d < D$. Specifically, we assume that M is embedded within a compact subset $\mathcal{K} \subset \mathbb{R}^D$ and that

$$P = (1-\eta)\,\mathrm{Unif}(\mathcal{K}) + \eta(W \star \Phi_\sigma), \tag{4}$$

where $0 < \eta \leq 1$, $\mathrm{Unif}(\mathcal{K})$ is a uniform distribution on $\mathcal{K}$, $\star$ denotes convolution, W is a distribution supported on M, and Φ_σ is a Gaussian distribution on $\mathbb{R}^D$ with zero mean and covariance σI_D. While we could consider a more general noise distribution in (4), we focus on the common assumption of Gaussian noise. In that case, a hidden manifold M can only be estimated at a logarithmic rate [Genovese et al. (2012b)], so ridge estimators are particularly valuable. (Even when M can be estimated at a polynomial rate, ridge estimators are often easier in practice than estimating the manifold, which would involve deconvolution.)

The data generating process under model (4) is equivalent to the following steps:

1. Draw B from a Bernoulli(η).
2. If $B = 0$, draw X from a uniform distribution on $\mathcal{K}$.
3. If $B = 1$, let $X = Z + \sigma\varepsilon$ where $Z \sim W$ and ε is additional noise.

Points X_i drawn from $\mathrm{Unif}(\mathcal{K})$ represent background clutter. Points X_i drawn from $W \star \Phi_\sigma$ are noisy observations from M. When M consists of a finite set of points, this can be thought of as a clustering model.

2.1. *Definition of ridges.* As in Ozertem and Erdogmus (2011), our definition of ridges relies on the gradient and Hessian of the density function p. Recall that $0 < d < D$ is fixed throughout. Given a function $p : \mathbb{R}^D \to \mathbb{R}$, let $g(x) = \nabla p(x)$ denote its gradient and $H(x)$ its Hessian matrix, at x. Let

$$\lambda_1(x) \geq \lambda_2(x) \geq \cdots \geq \lambda_d(x) > \lambda_{d+1}(x) \geq \cdots \geq \lambda_D(x) \tag{5}$$

denote the eigenvalues of $H(x)$ and let $\Lambda(x)$ be the diagonal matrix whose diagonal elements are the eigenvalues. Write the spectral decomposition of $H(x)$ as $H(x) = U(x)\Lambda(x)U(x)^T$. Let $V(x)$ be the last $D-d$ columns of $U(x)$ (i.e., the

columns corresponding to the $D-d$ smallest eigenvalues). If we write $U(x) = [V_\diamond(x) : V(x)]$ then we can write $H(x) = [V_\diamond(x) : V(x)]\Lambda(x)[V_\diamond(x) : V(x)]^T$. Let $L(x) \equiv L(H(x)) = V(x)V(x)^T$ be the projector onto the linear space defined by the columns of $V(x)$. We call this the local normal space and the space spanned by $L^\perp(x) = I - L(x) = V_\diamond(x)V_\diamond(x)^T$ is the local tangent space. Define the *projected gradient*

$$G(x) = L(x)g(x). \tag{6}$$

If the vector field $G(x)$ is Lipschitz then by Theorem 3.39 of Irwin (1980), G defines a global flow as follows. The flow is a family of functions $\phi(x,t)$ such that $\phi(x,0) = x$ and $\phi'(x,0) = G(x)$ and $\phi(x, s+t) = \phi(\phi(x,t), s)$. The flow lines, or integral curves, partition the space (see Lemma 2) and at each x where $G(x)$ is nonnull, there is a unique integral curve passing through x. Thus, there is one and only one flow line through each nonridge point. The intuition is that the flow passing through x is a gradient ascent path moving toward higher values of p. Unlike the paths defined by the gradient g which move toward modes, the paths defined by the projected gradient G move toward ridges. The SCMS algorithm, which we describe later, can be thought of as approximating the flow with discrete, linear steps $x_{k+1} \leftarrow x_k + hG(x_k)$. [A proof that the linear interpolation of these points approximates the flow in the case $d = 0$ is given in Arias-Castro, Mason and Pelletier (2013).]

A map $\pi : \mathbb{R} \to \mathbb{R}^D$ is an integral curve with respect to the flow of G if

$$\pi'(t) = G\big(\pi(t)\big) = L\big(\pi(t)\big)g\big(\pi(t)\big). \tag{7}$$

Definition: The *ridge* R of dimension d is given by $R = \{x : \|G(x)\| = 0, \lambda_{d+1}(x) < 0\}$.

Note that the ridge consists of the destinations of the integral curves: $y \in R$ if $\lim_{t\to\infty} \pi(t) = y$ for some π satisfying (7).

Our definition is motivated by Ozertem and Erdogmus (2011) but is slightly different. They first define the d-critical points as those for which $\|G(x)\| = 0$. They call a critical point regular if it is d-critical but not $(d-1)$-critical. Thus, a mode within a one-dimensional ridge is not regular. A regular point with $\lambda_{d+1} < 0$ is called a principal point. According to our definition, the ridge lies between the critical set and the principal set. Thus, if a mode lies on a one-dimensional ridge, we include that point as part of the ridge.

2.2. *Assumptions*. We now record the main assumptions about the ridges that we will require for the results.

Assumption (A0) *differentiability.* For all x, $g(x)$, $H(x)$ and $H'(x)$ exist.

Assumption (A1) *eigengap.* Let $B_D(x,\delta)$ denote a D-dimensional ball of radius δ centered at x and let $R \oplus \delta = \bigcup_{x\in R} B_D(x,\delta)$. We assume that there exists $\beta > 0$ and $\delta > 0$ such that, for all $x \in R \oplus \delta$, $\lambda_{d+1}(x) < -\beta$ and $\lambda_d(x) - \lambda_{d+1}(x) > \beta$.

Assumption (A2) *path smoothness.* For each $x \in R \oplus \delta$,

$$\|L^{\perp}(x)g(x)\|\|H'(x)\|_{\max} < \frac{\beta^2}{2D^{3/2}}, \tag{A2}$$

where $H'(x) = \frac{d\,\mathrm{vec}(H(x))}{dx^T}$, $L^{\perp} = I - L$ and $\|A\|_{\max} = \max_{j,k} |A_{jk}|$.

Condition (A1) says that p is sharply curved around the ridge in the $D-d$ dimensional space normal to the ridge. To give more intuition about the condition, consider the problem of estimating a mode in one dimension. At a mode x, we have that $p'(x) = 0$ and $p''(x) < 0$. However, the mode cannot be uniformly consistently estimated by only requiring the second derivative to be negative since $p''(x)$ could be arbitrarily close to 0. Instead, one needs to assume that $p''(x) < -\beta$ for some positive constant β. Condition (A1) may be thought of as the analogous condition for a ridge. (A2) is a third derivative condition which implies that the paths cannot be too wiggly. (A2) also constrains the gradient from being too steep in the perpendicular direction. Note that these conditions are local: they hold in a size δ neighborhood around the ridge.

3. Technical background. Now we review some background. We recommend that the reader quickly skim this section and then refer back to it as needed.

3.1. *Distance function and Hausdorff distance.* We let $B(x, r) \equiv B_D(x, r)$ denote a D-dimensional open ball centered at $x \in \mathbb{R}^D$ with radius r. If A is a set and x is a point then we define the *distance function*

$$d_A(x) = d(x, A) = \inf_{y \in A} \|x - y\|, \tag{8}$$

where $\|\cdot\|$ is the Euclidean norm. Given two sets A and B, the *Hausdorff distance* between A and B is

$$\mathsf{Haus}(A, B) = \inf\{\varepsilon : A \subset B \oplus \varepsilon \text{ and } B \subset A \oplus \varepsilon\} = \sup_x \big|d_A(x) - d_B(x)\big|, \tag{9}$$

where

$$A \oplus \varepsilon = \bigcup_{x \in A} B_D(x, \varepsilon) = \{x : d_A(x) \le \varepsilon\} \tag{10}$$

is called the ε-*dilation* of A. The dilation can be thought of as a smoothed version of A. For example, if there are any small holes in A, these will be filled in by forming the dilation $A \oplus \varepsilon$.

We use Hausdorff distance to measure the distance between sets for several reasons: it is the most commonly used distance between sets, it is a very strict distance and is analogous to the familiar L_∞ distance between functions for sets.

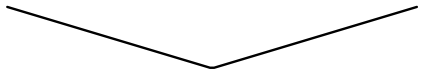

FIG. 4. *A straight line as infinite reach. A line with a corner, as in this figure, has* 0 *reach but has positive* μ*-reach.*

3.2. *Topological concepts.* This subsection follows Chazal, Cohen-Steiner and Lieutier (2009) and Chazal and Lieutier (2005). The *reach* of a set K, denoted by $\mathsf{reach}(K)$, is the largest $r > 0$ such that each point in $K \oplus r$ has a unique projection onto K. A set with positive reach is, in a sense, a smooth set without self-intersections.

Now we describe a generalization of reach called μ-reach. The key point is simply that the μ-reach is weaker than reach. The full details can be found in the aforementioned references. Let A be a compact set. Following Chazal and Lieutier (2005) define the gradient $\nabla_A(x)$ of $d_A(x)$ to be the usual gradient function whenever this is well defined. However, there may be points x at which d_A is not differentiable in the usual sense. In that case, define the gradient as follows. For $x \in A$ define $\nabla_A(x) = 0$ for all $x \in A$. For $x \notin A$, let $\Gamma(x) = \{y \in A : \|x - y\| = d_A(x)\}$. Let $\Theta(x)$ be the center of the unique smallest closed ball containing $\Gamma(x)$. Define $\nabla_A(x) = \frac{x-\Theta(x)}{d_A(x)}$.

The *critical points* are the points at which $\nabla_A(x) = 0$. The *weak feature size* $\mathsf{wfs}(A)$ is the distance from A to its closest critical point. For $0 < \mu < 1$, the μ*-reach* $\mathsf{reach}_\mu(A)$ is $\mathsf{reach}_\mu(A) = \inf\{d : \chi(d) < \mu\}$ where $\chi(d) = \inf\{\|\nabla_A(x)\| : d_A(x) = d\}$. It can be shown that reach_μ is nonincreasing in μ, that $\mathsf{wfs}(A) = \lim_{\mu\to 0} \mathsf{reach}_\mu(A)$ and that $\mathsf{reach}(A) = \lim_{\mu\to 1} \mathsf{reach}_\mu(A)$.

As a simple example, a circle C with radius r has $\mathsf{reach}(R) = r$. However, if we bend the circle slightly to create a corner, the reach is 0 but, provided the kink is not too extreme, the μ-reach is still positive. As another example, a straight line as infinite reach. Now suppose we add a corner as in Figure 4. This set has 0 reach but has positive μ-reach.

Two maps $f : A \to B$ and $g : A \to B$ are *homotopic* if there exists a continuous map $H : [0, 1] \times A \to B$ such that $H(0, x) = f(x)$ and $H(1, x) = g(x)$. Two sets A and B are homotopy equivalent if there are continuous maps $f : A \to B$ and $g : B \to A$ such that the following is true: (i) $g \circ f$ is homotopic to the identity map on A and (ii) $f \circ g$ is homotopic to the identity map on B. In this case we write $A \cong B$. Sometimes A fails to be homotopic to B but A is homotopic to $B \oplus \delta$ for every sufficiently small $\delta > 0$. This happens because $B \oplus \delta$ is slightly smoother than B. If $A \cong B \oplus \delta$ for all small $\delta > 0$, we will say that A and B are *nearly homotopic* and we will write $A \lessapprox B$.

The following result [Theorem 4.6 in Chazal, Cohen-Steiner and Lieutier (2009)] says that if a set K is smooth and $\widetilde{K}$ is close to K, then a smoothed version of $\widetilde{K}$ is nearly homotopy equivalent to K.

THEOREM 1 [Chazal, Cohen-Steiner and Lieutier (2009)]. *Let K and $\widetilde{K}$ be compact sets and let $\varepsilon = \mathsf{Haus}(\widetilde{K}, K)$. If*

$$\varepsilon < \frac{\mu^2\,\mathsf{reach}_\mu(K)}{5\mu^2+12} \quad \text{and} \quad \frac{4\varepsilon}{\mu^2} \le \alpha < \mathsf{reach}_\mu(K) - 3\varepsilon \tag{11}$$

then $(\widetilde{K} \oplus \alpha) \approx\!\!\approx K$.

3.3. *Matrix theory.* We make extensive use of matrix theory as can be found in Stewart and Sun (1990), Bhatia (1997), Horn and Johnson (2013) and Magnus and Neudecker (1988).

Let A be an $m \times n$ matrix. Let A_{jk} denote an element of the matrix. Then the Frobenius norm is $\|A\|_F = \sqrt{\sum_{j,k} A_{jk}^2}$ and the operator norm is $\|A\| = \sup_{\|x\|=1} \|Ax\|$. We define $\|A\|_{\max} = \max_{j,k} |A_{jk}|$. It is well known that $\|A\| \le \|A\|_F \le \sqrt{n}\|A\|$, that $\|A\|_{\max} \le \|A\| \le \sqrt{mn}\|A\|_{\max}$ and that $\|A\|_F \le \sqrt{mn}\|A\|_{\max}$.

The vec operator converts a matrix into a vector by stacking the columns. Thus, if A is $m \times n$ then $\mathsf{vec}(A)$ is a vector of length mn. Conversely, given a vector a of length mn, let $[[a]]$ denote the $m \times n$ matrix obtained by stacking a columnwise into matrix form. We can think of $[[a]]$ as the "anti-vec" operator.

If A is $m \times n$ and B is $p \times q$ then the Kronecker $A \otimes B$ is the $mp \times nq$ matrix

$$\begin{bmatrix} A_{11}B & \cdots & A_{1n}B \\ \vdots & & \vdots \\ A_{m1}B & \cdots & A_{mn}B \end{bmatrix}. \tag{12}$$

If A and B have the same dimensions, then the Hadamard product $C = A \circ B$ is defined by $C_{jk} = A_{jk}B_{jk}$.

For matrix calculus, we follow the conventions in Magnus and Neudecker (1988). If $F : \mathbb{R}^D \to \mathbb{R}^k$ is a vector-valued map then the Jacobian matrix will be denoted by $F'(x)$ or dF/dx. This is the $D \times k$ matrix with $F'(x)_{jk} = \partial F_i(x)/\partial x_j$. If $F : \mathbb{R}^D \to \mathbb{R}^{m\times p}$ is a matrix-valued map then $F'(x)$ is a $mp \times D$ matrix defined by

$$F'(x) \equiv \frac{dF}{dx^T} = \frac{d\,\mathrm{vec}(F(x))}{dx^T}. \tag{13}$$

If $F : \mathbb{R}^{n\times q} \to \mathbb{R}^{m\times p}$ then the derivative is a $mp \times nq$ matrix given by

$$F'(X) \equiv \frac{dF}{dX} = \frac{d\,\mathrm{vec}(F(X))}{d\,\mathrm{vec}(X)^T}.$$

We then have the following product rule for matrix calculus: if $F : \mathbb{R}^D \to \mathbb{R}^{m\times p}$ and $G : \mathbb{R}^D \to \mathbb{R}^{p\times q}$ then

$$\frac{dF(x)G(x)}{dx} = \big(G^T(x) \otimes I_m\big)F'(x) + \big(I_q \otimes F(x)\big)G'(x).$$

Also, if $A(x) = f(x)I$ then $A'(x) = \mathsf{vec}(I) \otimes (\nabla f(x))^T$ where ∇f denotes the gradient of f.

The following version of the Davis–Kahan theorem is from von Luxburg (2007). Let H and $\widetilde{H}$ be two symmetric, square $D \times D$ matrices. Let Λ be the diagonal matrix of eigenvalues of H. Let $S \subset \mathbb{R}$ and let V be the matrix whose columns are the eigenvectors corresponding to the eigenvalues of H in S and similarly for $\widetilde{V}$ and $\widetilde{H}$. Let

$$\beta = \min\{|\lambda - s| : \lambda \in \Lambda \cap S^c, s \in S\}. \tag{14}$$

According to the Davis–Kahan theorem,

$$\|VV^T - \widetilde{V}\widetilde{V}^T\| \leq \frac{\|H - \widetilde{H}\|_F}{\beta}. \tag{15}$$

Let H be a $D \times D$ square, symmetric matrix with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_D$. Let $\widetilde{H}$ be another square, symmetric matrix with eigenvalues $\widetilde{\lambda}_1 \geq \cdots \geq \widetilde{\lambda}_D$. By Weyl's theorem [Theorem 4.3.1 of Horn and Johnson (2013)], we have that

$$\lambda_n(\widetilde{H} - H) + \lambda_i(H) \leq \lambda_i(\widetilde{H}) \leq \lambda_i(H) + \lambda_1(\widetilde{H} - H). \tag{16}$$

It follows easily that

$$\big|\lambda_i(H) - \lambda_i(\widetilde{H})\big| \leq \|H - \widetilde{H}\| \leq D\|H - \widetilde{H}\|_{\max}. \tag{17}$$

4. Properties of ridges. In this section, we examine some of the properties of ridges as they were defined in Section 2 and show that, under appropriate conditions, if two functions are close together then their ridges are close and are topologically similar.

4.1. *Arclength parameterization.* It will be convenient to parameterize the gradient ascent paths by arclength. Thus, let $s \equiv s(t)$ be the arclength from $\pi(t)$ to $\pi(\infty)$:

$$s(t) = \int_t^\infty \|\pi'(u)\|\, du. \tag{18}$$

Let $t \equiv t(s)$ denote the inverse of $s(t)$. Note that

$$t'(s) = -\frac{1}{\|\pi'(t(s))\|} = -\frac{1}{\|L(\pi(t(s)))g(\pi(t(s)))\|} = -\frac{1}{\|G(\pi(t(s)))\|}. \tag{19}$$

Let $\gamma(s) = \pi(t(s))$. Then

$$\gamma'(s) = -\frac{G(\gamma(s))}{\|G(\gamma(s))\|}, \tag{20}$$

which is a restatement of (7) in the arclength parameterization.

In what follows, we will often abbreviate notation by using the subscript s in the following way: $G_s = G(\gamma(s))$, $H_s = H(\gamma(s)), \ldots$, and so forth.

4.2. *Differentials.* We will need derivatives of g, H, and L. The derivative of g is the Hessian H. Recall from (13) that $H'(x) = \frac{d\,\mathsf{vec}(H(x))}{dx^T}$. We also need derivatives along the curve γ. The derivative of a functions f along γ is

$$\dot{f}_{\gamma(s)} \equiv \dot{f}_s = \lim_{\varepsilon \to 0} \frac{f(\gamma(s+\varepsilon)) - f(\gamma(s))}{\varepsilon}. \tag{21}$$

Thus, the derivative of the gradient g along γ is

$$\dot{g}_{\gamma(s)} \equiv \dot{g}_s = \lim_{\varepsilon \to 0} \frac{g(\gamma(s+\varepsilon)) - g(\gamma(s))}{\varepsilon} = H_s \gamma_s' = -\frac{H_s G_s}{\|G_s\|}. \tag{22}$$

We will also need the derivative of H in the direction of a vector z which we will denote by

$$H'(x; z) \equiv \lim_{\varepsilon \to 0} \frac{H(x+\varepsilon z) - H(x)}{\varepsilon}.$$

We can write an explicitly formula for $H'(x; z)$ as follows. Note that the elements of H' are the partial derivatives $\partial H_{jk}(x)/\partial x_\ell$ arranged in a $D^2 \times D$ matrix. Hence, $H'(x; z) = [[H'(x)z]]$. (Recall that $[[a]]$ stacks a vector into a matrix.) Note that $[[H'(x)z]]$ is a $D \times D$ matrix.

Recall that $L(x) \equiv L(H(x)) = V(x)V(x)^T$. The collection $\{L(x) : x \in \mathbb{R}^D\}$ defines a matrix field: there is a matrix $L(x)$ attached to each point x. We will need the derivative of this field along the integral curves γ. For any $x \notin R$, there is a unique path γ and unique $s > 0$ such that $x = \gamma(s)$. Define

$$\begin{aligned} \dot{L}_s \equiv \dot{L}(x) &\equiv \lim_{\varepsilon \to 0} \frac{L(H(\gamma(s+\varepsilon))) - L(H(\gamma(s)))}{\varepsilon} \\ &= \lim_{t \to 0} \frac{L(H + tE) - L(H)}{t}, \end{aligned} \tag{23}$$

where $H = H(\gamma(s))$ and $E = (d/ds)H(\gamma(s)) = H'(x; z)$ with $z = \gamma'(s)$.

4.3. *Uniqueness of the γ paths.*

LEMMA 2. *Conditions* (A0)–(A2) *imply that, for each $x \in (R \oplus \delta) - R$, there is a unique path γ passing through x.*

PROOF. We will show that the vector field $G(x)$ is Lipschitz over $R \oplus \delta$. The result then follows from Theorem 3.39 of Irwin (1980). Recall that $G = Lg$ and g is differentiable. It suffices to show that L is differentiable over $R \oplus \delta$. Now $L(x) = L(H(x))$. It may be shown that, as a function of H, L is Frechet differentiable. And H is differentiable by assumption. By the chain rule, L is differentiable as a function of x. Indeed, dL/dx is the $D^2 \times D$ matrix whose jth column is $\mathsf{vec}(L^\dagger E_j)$ where $E_j = [[H' e_j]]$, $L^\dagger$ denotes the Frechet derivative, and e_j is the vector which is 1 in the jth coordinate and zero otherwise. □

4.4. *Quadratic behavior.* Conditions (A1) and (A2) imply that the function p has quadratic-like behavior near the ridges. This property is needed for establishing the convergence of ridge estimators. In this section, we formalize this notion of quadratic behavior. Give a path γ, define the function

$$\xi(s) = p\big(\pi(\infty)\big) - p\big(\pi\big(t(s)\big)\big) = p\big(\gamma(0)\big) - p\big(\gamma(s)\big). \tag{24}$$

Thus, ξ is simply the drop in the function p along the curve γ as we move away from the ridge. We write $\xi_x(s)$ if we want to emphasize that ξ corresponds to the path γ_x passing through the point x. Since $\xi : [0, \infty) \to [0, \infty)$, we define its derivatives in the usual way, that is, $\xi'(s) = d\xi(s)/ds$.

LEMMA 3. *Suppose that* (A0)–(A2) *hold. For all* $x \in R \oplus \delta$, *the following are true*:

1. $\xi(0) = 0$.
2. $\xi'(s) = \|G(\gamma(s))\|$ *and* $\xi'(0) = 0$.
3. *The second derivative of* ξ *is*:

$$\xi''(s) = -\frac{G_s^T H_s G_s}{\|G_s\|^2} + \frac{g_s^T \dot{L}_s G_s}{\|G_s\|}. \tag{25}$$

4. $\xi''(s) \geq \beta/2$.
5. $\xi(s)$ *is nonincreasing in* s.
6. $\xi(s) \geq \frac{\beta}{4}\|\gamma(0) - \gamma(s)\|^2$.

PROOF. 1. The first condition $\xi(0) = 0$ is immediate from the definition.

2. Next,

$$\xi'(s) = -\frac{dp(\gamma(s))}{ds} = -g_s\gamma_s' = \frac{g_s^T G_s}{\|G_s\|} = \frac{g_s^T L_s g_s}{\|G_s\|}$$
$$= \frac{g_s^T L_s L_s g_s}{\|G_s\|} = \frac{G_s^T G_s}{\|G_s\|} = \|G_s\|.$$

Since the projected gradient is 0 at the ridge, we have that $\xi'(0) = 0$.

3. Note that $(\xi'(s))^2 = \|G_s\|^2 = G_s^T G_s = g_s^T L_s g_s \equiv a(s)$. Differentiating both sides of this equation, we have that $2\xi'(s)\xi''(s) = a'(s)$, and hence

$$\xi''(s) = \frac{a'(s)}{2\xi'(s)} = \frac{a'(s)}{2\|G_s\|}.$$

Now

$$a'(s) = (\dot{g}_s)^T L_s g_s + g_s^T \dot{L}_s g_s + g_s^T L_s \dot{g}_s = 2(\dot{g}_s)^T L_s g_s + g_s^T \dot{L}_s g_s. \tag{26}$$

Since $L_s L_s = L_s$ we have that $\dot{L}_s = L_s \dot{L}_s + \dot{L}_s L_s$, and hence

$$g_s^T \dot{L}_s g_s = g_s^T L_s \dot{L}_s g_s + g_s^T \dot{L}_s L_s g_s = G_s^T \dot{L}_s g_s + g_s^T \dot{L}_s G_s = 2 g_s^T \dot{L}_s G_s.$$

Therefore,

$$a'(s) = 2(\dot{g}_s)^T L_s g_s + 2 g_s^T \dot{L}_s G_s. \tag{27}$$

Recall that $\dot{g}_s = -\frac{H_s G_s}{\|G_s\|}$. Thus,

$$\xi''(s) = \frac{a'(s)}{2\|G_s\|} = -\frac{G_s^T H_s G_s}{\|G_s\|^2} + \frac{g_s^T \dot{L}_s G_s}{\|G_s\|}. \tag{28}$$

4. The first term in $\xi''(s)$ is $-\frac{G_s^T H_s G_s}{\|G_s\|^2}$. Since G is in the column space of V, $G_s^T H_s G_s = G_s^T (V_s \Lambda_s V_s^T) G_s$ where $\Lambda_s = \operatorname{diag}(\lambda_{d+1}(\gamma(s)), \ldots, \lambda_D(\gamma(s)))$. Hence, from (A1),

$$\frac{G_s^T H_s G_s}{\|G_s\|^2} = \frac{G_s^T (V_s \Lambda_s V_s^T) G_s}{\|G_s\|^2} \le \lambda_{\max}\big(V_s \Lambda_s V_s^T\big) < -\beta$$

and thus

$$-\frac{G_s^T H_s G_s}{\|G_s\|^2} \ge \beta.$$

Now we bound the second term $\frac{g_s^T \dot{L}_s G_s}{\|G_s\|}$. Since $L_s + L_s^\perp = I$ and $L_s G_s = G_s$, we have $g_s^T \dot{L}_s G_s = g_s^T L_s \dot{L}_s G_s + g_s^T L_s^\perp \dot{L}_s G_s = g_s^T L_s \dot{L}_s L_s G_s + g_s^T L_s^\perp \dot{L}_s L_s G_s$. Now $|g_s^T L_s \dot{L}_s L_s G_s| = 0$. To see this, note that $L_s L_s = L_s$ implies $L_s \dot{L}_s + \dot{L}_s L_s = \dot{L}_s$ implies $L_s \dot{L}_s L_s + \dot{L}_s L_s = \dot{L}_s L_s$ implies $L_s \dot{L}_s L_s = 0$. To bound $g_s^T L_s^\perp \dot{L}_s L_s G_s$ we proceed as follows. Let $E = (d/ds) H(\pi(\gamma(s))) = H'(x; z)$ with $z = \gamma'(s)$. Then, from Davis–Kahan,

$$\begin{aligned}
\big|g_s^T L_s^\perp \dot{L}_s L_s G_s\big| &= \lim_{t \to 0} \frac{|g_s^T L_s^\perp (L(H + tE) - L(H)) L_s G_s}{t} \\
&\le \big\|L^\perp g_s\big\| \lim_{t \to 0} \frac{\|(L(H + tE) - L(H))\|}{t} \|G_s\| \\
&\le \frac{\|L^\perp g_s\| \|E\| \|G_s\|}{\beta}.
\end{aligned}$$

Note that $\|H'(x; z)\| \le D \|H'(x; z)\|_{\max} \le D^{3/2} \|H'(x)\|_{\max} \|z\| = D^{3/2} \times \|H'(x)\|_{\max}$. So $\frac{|g_s^T \dot{L}_s G_s|}{\|G_s\|} \le D^{3/2} \|L^\perp g\| \|H'\|_{\max} / \beta$ which is less than $\beta/2$ by (A2). Therefore, $\xi''(s) \ge \beta - (\beta/2) = \beta/2$.

5. Follows from 2.

6. For some $0 \le \widetilde{s} \le s$,

$$\xi(s) = \xi(0) + s\xi'(0) + \frac{s^2}{2} \xi''(\widetilde{s}) = \frac{s^2}{2} \xi''(\widetilde{s}) \ge \frac{\beta s^2}{4}$$

from part (4). So

$$\xi(s) - \xi(0) \ge \frac{\beta}{4} s^2 \ge \frac{\beta}{4} \big\|\gamma(0) - \gamma(s)\big\|^2. \qquad \square$$

4.5. *Stability of ridges.* We now show that if two functions p and $\widetilde{p}$ are close, then their corresponding ridges R and $\widetilde{R}$ are close. We use $\widetilde{g}, \widetilde{H}, \ldots$ etc. to refer to the gradient, Hessian and so on, defined by $\widetilde{p}$. For any function $f : \mathbb{R}^D \to \mathbb{R}$, let $\|f\|_\infty = \sup_{x \in R \oplus \delta} |f(x)|$. Let

$$\varepsilon = \|p - \widetilde{p}\|_\infty, \qquad \varepsilon' = \max_j \|g_j - \widetilde{g}_j\|_\infty, \tag{29}$$

$$\varepsilon'' = \max_{jk} \|H_{jk} - \widetilde{H}_{jk}\|_\infty, \qquad \varepsilon''' = \max_{jk} \|H'_{jk} - \widetilde{H}'_{jk}\|_\infty. \tag{30}$$

THEOREM 4. *Suppose that* (A0)–(A2) *hold for* p *and that* (A0) *holds for* $\widetilde{p}$. *Let* $\psi = \max\{\varepsilon, \varepsilon', \varepsilon''\}$ *and let* $\Psi = \max\{\varepsilon, \varepsilon', \varepsilon'', \varepsilon'''\}$. *When* Ψ *is sufficiently small*:

(1) *Conditions* (A1) *and* (A2) *hold for* $\widetilde{p}$.

(2) *We have*: $\mathsf{Haus}(R, \widetilde{R}) \le \frac{2C\psi}{\beta}$.

(3) *If* $\mathsf{reach}_\mu(R) > 0$ *for some* $\mu > 0$, *then* $\widetilde{R} \oplus \frac{4\psi}{\mu^2} \approx R$.

PROOF. (1) Write the spectral decompositions $H = U\Lambda U^T$ and $\widetilde{H} = \widetilde{U}\widetilde{\Lambda}\widetilde{U}^T$. By (17), $|\lambda_j - \widetilde{\lambda}_j| \le D\|H - \widetilde{H}\|_{\max} \le D\varepsilon''$. Thus, $\widetilde{p}$ satisfies (A1) when ε'' is small enough. Clearly, (A2) also holds as long as Ψ is small enough.

(2) By the Davis–Kahan theorem (15),

$$\|L - \widetilde{L}\| \le \frac{\|H - \widetilde{H}\|_F}{\beta} \le \frac{D\|H - \widetilde{H}\|_{\max}}{\beta} \le \frac{D\varepsilon''}{\beta}.$$

For each x,

$$\begin{aligned}\|G(x) - \widetilde{G}(x)\| &= \|L(x)g(x) - \widetilde{L}(x)\widetilde{g}(x)\| \\ &\le \|(L(x) - \widetilde{L}(x))g(x)\| + \|\widetilde{L}(x)(\widetilde{g}(x) - g(x))\| \\ &\le \frac{D\|g(x)\|\varepsilon''}{\beta} + \varepsilon'.\end{aligned}$$

It follows that, $\|L - \widetilde{L}\| \le C\varepsilon''$ and $\sup_x \|G(x) - \widetilde{G}(x)\| \le C\psi$.

Now let $\widetilde{x} \in \widetilde{R}$. Thus, $\|\widetilde{G}(\widetilde{x})\| = 0$, and hence $\|G(\widetilde{x})\| \le C\psi$. Let γ be the path through $\widetilde{x}$ so that $\gamma(s) = \widetilde{x}$ for some s. Let $r = \gamma(0) \in R$. From part 2 of Lemma 3, note that $\xi'(s) = \|G(\widetilde{x})\|$. We have

$$C\psi \ge \|G(\widetilde{x})\| = \xi'(s) = \xi'(0) + s\xi''(u)$$

for some u between 0 and s. Since $\xi'(0) = 0$, from part 4 of Lemma 3, $C\psi \ge s\xi''(u) \ge \frac{s\beta}{2}$ and so $\|r - \widetilde{x}\| \le s \le \frac{2C\psi}{\beta}$. Thus, $d(\widetilde{x}, R) \le \|r - \widetilde{x}\| \le 2C\psi/\beta$.

Now let $x \in R$. The same argument shows that $d(x, \widetilde{R}) \le 2C\psi/\beta$ since (A1) and (A2) hold for $\widetilde{p}$.

(3) Choose any fixed $\kappa > 0$ such that $\kappa < \frac{\mu^2}{5\mu^2+12}$. When Ψ is sufficiently small, $\Psi \le \kappa\, \mathsf{reach}_\mu(K)$. Then $\widetilde{R} \oplus \frac{4\psi}{\mu^2} \approx R$ from Theorem 1. □

5. Ridges of density estimators. Now we consider estimating the ridges in the density ridge case (no hidden manifold). Let $X_1, \dots, X_n \sim P$ where P has density p and let

$$\hat{p}_h(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h^D} K\left(\frac{\|x - X_i\|}{h}\right) \tag{31}$$

be a kernel density estimator with kernel K and bandwidth h. Let $\hat{R}$ be the ridge defined by $\hat{p}$. In this section, we bound $\mathsf{Haus}(R, \hat{R})$. We assume that P is supported on a compact set $\mathcal{K} \subset \mathbb{R}^D$ and that p and its first, second and third derivatives vanish on the boundary of $\mathcal{K}$. (This ensures there is no boundary bias in the kernel density estimator.)

We assume that all derivatives of p up to and including fifth degree are bounded and continuous. We also assume the conditions on the kernel in Gine and Guillou (2002) which are satisfied by all the usual kernels. Results on $\|p(x) - \hat{p}_h(x)\|_\infty$ are given, for example, in Prakasa Rao (1983), Giné and Guillou (2002) and Yukich (1985). The results in those references imply that

$$\varepsilon \equiv \sup_{x\in\mathcal{K}} \|p(x) - \hat{p}(x)\|_\infty = O(h^2) + O_P\left(\sqrt{\frac{\log n}{nh^D}}\right).$$

For the derivatives, rates are proved in the sense of mean squared error by Chacón, Duong and Wand (2011). They can be proved in the L_∞ norm using the same techniques as in Prakasa Rao (1983), Giné and Guillou (2002) and Yukich (1985). The rates are:

$$\varepsilon' \equiv \max_j \sup_{x\in\mathcal{K}} |g_j(x) - \hat{g}_j(x)| = O(h^2) + O_P\left(\sqrt{\frac{\log n}{nh^{D+2}}}\right),$$

$$\varepsilon'' \equiv \max_{j,k} \sup_{x\in\mathcal{K}} |H_{j,k}(x) - \hat{H}_{j,k}(x)| = O(h^2) + O_P\left(\sqrt{\frac{\log n}{nh^{D+4}}}\right),$$

$$\varepsilon''' \equiv \sup_{x\in\mathcal{K}} \|H'(x) - \hat{H}'(x)\|_{\max} = O(h^2) + O_P\left(\sqrt{\frac{\log n}{nh^{D+6}}}\right).$$

[See Arias-Castro, Mason and Pelletier (2013), e.g.] Let $\psi_n = (\frac{\log n}{n})^{2/(D+8)}$. Choosing $h \asymp \sqrt{\psi_n}$ we get that $\varepsilon \asymp \varepsilon' \asymp \varepsilon'' \asymp O_P(\psi_n)$ and $\varepsilon''' = o_P(1)$. From Theorem 4 and the rates above we have the following.

THEOREM 5. *Let $\hat{R}^* = \hat{R} \cap (R \oplus \delta)$. Under the assumptions above and assuming that* (A1) *and* (A2) *hold, we have, with $h \asymp \sqrt{\psi_n}$ that*

$$\mathsf{Haus}(R, \hat{R}^*) = O_P(\psi_n). \tag{32}$$

If $\mathsf{reach}_\mu(R) > 0$ *then* $\hat{R}^* \oplus O(\psi_n) \approx\approx R$.

Let $\overline{p}_h(x) = \mathbb{E}(\hat{p}_h(x))$ and let R_h be the ridge set of $\overline{p}_h$. It may suffice for practical purposes to estimate R_h for some small $h > 0$. Indeed, as a corollary to Theorem 9 in the next section (letting R take the role of M and R_h take the role of R_σ) it follows that Hausdorff$(R, R_h) = O(h^2)$ and R_h is topologically similar to R. In this case, we can take h fixed rather than letting it tend to 0. For fixed h, we then have dimension-independent rates.

THEOREM 6. *Let $h > 0$ be fixed and let $\widetilde{\psi}_n = \sqrt{\log n/n}$. Let $\hat{R}^* = \hat{R} \cap (R \oplus \delta)$. Under the assumptions above and assuming that* (A1) *and* (A2) *hold for R_h we have, that*

$$\mathsf{Haus}\big(R_h, \hat{R}^*\big) = O_P(\widetilde{\psi}_n). \tag{33}$$

If $\mathsf{reach}_\mu(R_h) > 0$ *then* $\hat{R}^* \oplus O(\widetilde{\psi}_n) \approx\approx R$.

6. Ridges as surrogates for hidden manifolds. Consider now the case where $P_\sigma = (1-\eta)\,\mathrm{Unif}(\mathcal{K}) + \eta(W \star \Phi_\sigma)$ where W is supported on M. We assume in this section that M is a compact manifold without boundary. We also assume that W has a twice-differentiable density w respect to the uniform measure on M. (Here, w is a function on a smooth manifold and the derivatives are defined in the usual way, that is, with respect to any coordinate chart.) We also assume that w is bounded away from zero and ∞. In this section, we add the subscript σ to the density, the gradient, etc. to emphasize the dependence on σ. For example, the density of P_σ is denoted by p_σ, the gradient by g_σ and the Hessian by H_σ.

We want to show that the ridge of p_σ is a surrogate for M. Specifically, we show that, as σ gets small, there is a subset $R_* \subset R$ in a neighborhood of M such that $\mathsf{Haus}(M, R_*) = O(\sigma^2 \log(1/\sigma))$ and such that $R_* \approx\approx M$. We assume that $\eta = 1$ in what follows; the extension to $0 < \eta < 1$ is straightforward. We also assume that M is a compact d-manifold with positive reach κ. We need to assume that M has positive reach rather than just positive μ-reach. The reason is that, when M has positive reach, the measure W induces a smooth distribution on the tangent space $T_x M$ for each $x \in M$. We need this property in our proofs but this property is lost if M only has positive μ-reach for some $\mu < 1$ due to the presence of unsmooth features such as corners.

The density of X is

$$p_\sigma(x) = \int_M \phi_\sigma(x - z)\, dW(z), \tag{34}$$

where $\phi_\sigma(u) = (2\pi)^{-D/2}\sigma^{-D}\,\exp(-\frac{\|u\|^2}{2\sigma^2})$. Thus, p_σ is a mixture of Gaussians. However, it is a rather unusual mixture; it is a *singular mixture of Gaussians* since the mixing distribution W is supported on a lower dimensional manifold.

Let $T_x M$ be the tangent space[4] to M at x and let $T_x^{\perp} M$ be the normal space to M at x. Define the *fiber* at $x \in M$ by $F_x = T_x^{\perp} M \cap B_D(x, r)$. A consequence of the fact that the reach κ is positive and M has no boundary is that, for any $0 < r < \kappa$, $M \oplus r$ can be written as a disjoint union

$$M \oplus r = \bigcup_{x \in M} F_x. \tag{35}$$

Let $r_\sigma > 0$ satisfy the following conditions:

$$r_\sigma < \sigma, \frac{r_\sigma}{\sigma^2} \to \infty, \qquad r_\sigma \log\left(\frac{1}{\sigma^{2+D}}\right) = o(1) \qquad \text{as } \sigma \to 0. \tag{36}$$

Specifically, take $r_\sigma = \alpha\sigma$ for some $0 < \alpha < 1$. Fix any $A \geq 2$ and define

$$K_\sigma = \sqrt{2\sigma^2 \log\left(\frac{1}{\sigma^{A+D}}\right)}. \tag{37}$$

THEOREM 7 (Surrogate theorem). *Suppose that* $\kappa = \mathsf{reach}(M) > 0$. *Let* R_σ *be the ridge set of* p_σ. *Let* $M_\sigma = M \oplus r_\sigma$ *and* $R_\sigma^* = R_\sigma \cap M_\sigma$. *For all small* $\sigma > 0$:

1. R_σ^* *satisfies* (A1) *and* (A2) *with* $\beta = c\sigma^{-(D-d+2)}$ *form some* $c > 0$.
2. $\mathsf{Haus}(M, R_\sigma^*) = O(K_\sigma^2)$.
3. $R_\sigma^* \oplus C K_\sigma^2 \approx\!\!\approx M$.

If R_σ *is instead taken to be the ridge set of* $\log p_\sigma$ *then the same results are true with* $\beta = c\sigma^{-2}$ *and* $M_\sigma = M \oplus \kappa$.

REMARK. Without the assumption that M has no boundary, there would be boundary effects of order K_σ. That is, the Hausdorff distance behaves like $O(K_\sigma)$ for points near the boundary and like $O(K_\sigma^2)$ for points not near the boundary.

The theorem shows that in a neighborhood of the manifold, there is a well-defined ridge, that the ridge is close to the manifold and is nearly homotopic to the manifold. It is interesting to compare the above result to recent work on finite mixtures of Gaussians [Carreira-Perpinan and Williams (2003), Edelsbrunner, Fasy and Rote (2012)]. In those papers, it is shown that there can be fewer or more modes than the number of Gaussian components in a finite mixture. However, for small σ, it is easy to see that for each component of the mixture, there is a nearby mode. Moreover, the density will be highly curved at those modes. Theorem 7 can be thought of as a version of the latter two facts for the case of manifold mixtures.

The theorem refers to the ridges defined by p_σ and the ridges defined by $\log p_\sigma$. Although the location of the ridge sets is the same for both cases, the behavior of

[4]Recall that the tangent space at a point x is the linear space spanned by the derivative vectors of smooth curves on the manifold through that point.

the function around the ridges is different. There are several reasons we might want to use $\log p$ rather than p. First, when p is Gaussian, the ridges of $\log p$ correspond to the usual principal components. Second, the surrogate theorem holds in an $O(1)$ neighborhood of M for the log-density whereas it only holds in an $O(\sigma)$ neighborhood of M for the density.

To prove the theorem, we need a preliminary result. Let

$$\widetilde{\sigma} = \sigma \log^3\left(\frac{1}{\sigma^{D+A}}\right). \tag{38}$$

Given a point x let $\hat{x}$ be its projection onto M. In what follows, if T is a matrix, then an expression of the form $T + O(r_n)$ is to be interpreted to mean $T + B_n$ where B_n is a matrix whose entries are of order $O(r_n)$. Let

$$\phi_\perp(u) = \frac{e^{-\|u\|^2/(2\sigma^2)}}{(2\pi)^{(D-d)/2}\sigma^{D-d}}, \qquad u \in \mathbb{R}^{D-d}. \tag{39}$$

LEMMA 8. *For all $x \in M_\sigma$,*

1. $p_\sigma(x) = \phi_\perp(x - \hat{x})(1 + O(\widetilde{\sigma}))$.
2. *Let* $p_{\sigma,B}(x) = \int_{M\cap B} \phi_\sigma(x - z)\, dW(Z)$. *Then* $p_{\sigma,B}(x) = \phi_\perp(x - \hat{x})(1 + O(\widetilde{\sigma}))$.
3. $g_\sigma(x) = -\frac{1}{\sigma^2} p_\sigma(x)((x - \hat{x}) + O(K_\sigma^2))$ *and* $\|g_\sigma(x)\| = O(\sigma^{-(D-d-1)})$.
4. *The eigenvalues of $H_\sigma(x)$ are*

$$\lambda_j(x) = \begin{cases} O(\widetilde{\sigma}), & j \le d, \\ -\dfrac{p_\sigma(x)}{\sigma^2}\left[1 - \dfrac{d_M^2(x)}{\sigma^2} + O(\widetilde{\sigma})\right], & j = d+1, \\ -\dfrac{p_\sigma(x)}{\sigma^2}\left[1 + O(\widetilde{\sigma})\right], & j > d+1. \end{cases} \tag{40}$$

5. *The projection matrix L_σ satisfies*

$$L_\sigma(x) = \left[\begin{pmatrix} 0_d & 0_{d,D-d} \\ 0_{D-d,d} & I_{D-d} \end{pmatrix}\right] + O(\widetilde{\sigma}).$$

6. *Projected gradient*:

$$G_\sigma(x) = -\frac{1}{\sigma^2}\big((x - \hat{x})\phi_\perp(x - \hat{x})\big(1 + O(\widetilde{\sigma})\big) + O_\perp\big(K_\sigma^2\big)\big),$$

where $O_\perp(K_\sigma^2)$ is a term of size $O(K_\sigma^2)$ in $T_x^\perp$.

7. *Gap*:

$$\lambda_d(x) - \lambda_{d+1}(x) \ge \frac{p_\sigma(x)}{\sigma^2}\big[1 - \alpha^2 + O(\widetilde{\sigma})\big]$$

and

$$\beta \equiv \inf_{x\in R\oplus\delta}\big[\lambda_d(x) - \lambda_{d+1}(x)\big] \ge c\sigma^{-(D=d+2)}$$

and $\lambda_{d+1}(x) \le -\beta$.

8. $\|H'_\sigma\|_{\max} = O(\sigma^{-(D+3-d)})$.

PROOF. The proof is quite long and technical and so we relegate it to the Appendix. □

PROOF OF THEOREM 7. Let us begin with the ridge based on p_σ.

(1) Condition (A1) follows from parts 8 and 1 of Lemma 8 together with equation (49).

To verify (A2), we use parts 3 and 8 of Lemma 8: we get, for all small σ, that

$$\|L^\perp g\|\|H'\|_{\max} \le \|g\|\|H'\|_{\max} \le \frac{c}{\sigma^{D-d-1}}\frac{1}{\sigma^{D+3-d}} < \frac{c^2\sigma^2}{\sigma^{2(D-d+2)}} \le \sigma^2\beta^2$$
$$\le \frac{\beta^2}{2D^{3/2}}$$

as required.

(2) Suppose that $x \in R^*_\sigma$. Then $\|G_\sigma(x)\| = 0$. Let $\hat{x}$ be the unique projection of x onto M. From part 6 of Lemma 8,

$$\|(x-\hat{x})\phi_\perp(x-\hat{x})(1+O(\widetilde{\sigma})) + O_\perp(K^2)\| = 0,$$

and hence $x = \hat{x} + O(K_\sigma^2)$.

Now let $\hat{x} \in M$. From the expression above, we see that $\|G_\sigma(\hat{x})\| = O(K_\sigma^2)$. Let γ be the path through x and let r be the destination of the path. Hence $\gamma(s) = x$ for some s and $\gamma(0) = r$. Now we use Lemma 3. Then $\|G\| = \xi'$ and

$$O(K_\sigma^2) = \xi'(s) = \xi'(s) - \xi'(0) = s\xi''(\widetilde{s}) \ge \|x-\hat{r}\|\xi''(\widetilde{s}) \ge \|x-\hat{r}\|\beta/2$$

and so $\|x-\hat{r}\| = O(K_\sigma^2)$. Hence, $\mathsf{Haus}(R_\sigma, M) = O(K_\sigma^2)$.

(3) Homotopy. This follows from part (2) and Theorem 1.

Now consider the ridges of $\log p_\sigma(x)$. The proof is essentially the same as the proof above. The main difference is the Hessian as we now explain. Note that the Hessian H^*_σ for $\log p_\sigma(x)$ is

$$H^*_\sigma(x) = \frac{1}{p_\sigma(x)}\left(H_\sigma(x) - \frac{1}{p_\sigma(x)}g_\sigma(x)g_\sigma^T(x)\right).$$

From Lemma 8, parts 3 and 4, it follows that (after an appropriate rotation),

$$H^*_\sigma(x) = -\frac{1}{\sigma^2}\left(\begin{bmatrix} O_d & 0_{d\times D-d} \\ 0_{D-d\times d} & I_{D-d}\end{bmatrix} + O(\widetilde{\sigma})\right).$$

Hence,

$$\lambda_{d+1}(x) = -\frac{1}{\sigma^2} + O(\widetilde{\sigma})$$

and

$$\lambda_{d+1}(x) - \lambda_d(x) = \frac{1}{\sigma^2} + O(\widetilde{\sigma}).$$

Notice in particular, that the dominant term of the smallest eigenvalue of $-\beta H_\sigma^*(x)$ is 1 whereas that the dominant term of the smallest eigenvalue of $-\beta H_\sigma(x)$ is 1 $d_M^2(x)/\sigma^2$ which is why we required $\|x-\hat{x}\|$ to be less than σ in Theorem 7. Here, we only require that $\|x-\hat{x}\| \le \kappa$. $\square$

We may now combine Theorems 4, 5, 6 and 7 to get the following.

COROLLARY 9. *Let $\hat{R}^*$ be defined as in Theorem 5. Then*

$$\mathsf{Haus}(\hat{R}^*, M) = O_P\left(\left(\frac{\log n}{n}\right)^{2/(D+8)}\right) + O(K_\sigma^2). \tag{41}$$

Similarly, if $\hat{R}^*$ be defined as in Theorem 6 then

$$\mathsf{Haus}(\hat{R}^*, M) = O_P\left(\sqrt{\frac{\log n}{n}}\right) + O(K_\sigma^2 + h^2). \tag{42}$$

7. SuRFing the ridge. Here, we discuss Subspace Ridge Finding (SuRF) by using density estimation, followed by denoising and then followed by the subspace constrained mean shift (SCMS) algorithm due to Ozertem and Erdogmus (2011). We will not go into great details about the algorthm; we refer the reader to Ozertem and Erdogmus (2011).

Let us begin by reviewing the mean shift algorithm. The *mean shift algorithm* [Cheng (1995), Comaniciu and Meer (2002), Fukunaga and Hostetler (1975)] is a method for finding the modes of a density by approximating the steepest ascent paths. The algorithm starts with a mesh of points and then moves the points along the gradient ascent trajectories toward local maxima.

Given a sample $X_1, \dots, X_n$ from p, consider the kernel density estimator

$$\hat{p}_h(x) = \frac{1}{n}\sum_{i=1}^n \frac{1}{h^D} K\left(\frac{\|x-X_i\|}{h}\right), \tag{43}$$

where K is a kernel and $h > 0$ is a bandwidth. Let $\mathcal{M} = \{v_1, \dots, v_m\}$ be a collection of mesh points. These are often taken to be the same as the data but in general they need not be. Let $v_j(1) = v_j$ and for $t = 1, 2, 3, \dots$ we define the trajectory $v_j(1), v_j(2), \dots,$ by

$$v_j(t+1) = \frac{\sum_{i=1}^n X_i K(\|v_j(t) - X_i\|/h)}{\sum_{i=1}^n K(\|v_j(t) - X_i\|/h)}. \tag{44}$$

It can be shown that each trajectory $\{v_j(t) : t = 1, 2, 3, \dots, \}$ follows the gradient ascent path and converges to a mode of $\hat{p}_h$. Conversely, if the mesh $\mathcal{M}$ is rich enough, then for each mode of $\hat{p}_h$, some trajectory will converge to that mode.

The SCMS algorithm mimics the mean shift algorithm but it replaces the gradient with the projected gradient at each step. The algorithm can be applied to $\hat{p}$ or

Subspace Constrained Mean Shift (SCMS)

1. For each x in the mesh, repeat the following steps until convergence:
2. For $i = 1, \ldots, n$ define

$$u_i = \frac{x - X_i}{h^2}, \qquad c_i = K\left(\frac{x - X_i}{h}\right),$$

$$g(x) = -\frac{1}{n}\sum_i c_i u_i, \qquad H(x) = \frac{1}{n}\sum_i c_i \left(u_i u_i^T - \frac{1}{h^2} I\right).$$

3. Decompose $H(x) = U(x)\Lambda(x)U^T(x)$, $U(x) = [V(x)V_\diamond(x)]$ and let $L(x) = V(x)V^T(x)$.
4. Update x: let $x \leftarrow L(x)[m(x) - x]$ where

$$m(x) = \frac{\sum_i c_i x_i}{\sum_i c_i}.$$

FIG. 5. *SCMS algorithm from Ozertem and Erdogmus (2011).*

any monotone function of $\hat{p}$. As we explained earlier, there are some advantages to using $\log \hat{p}$. Figure 5 gives the algorithm for the log-density. This is the version we will use in our examples. Figure 6 gives the full SuRF algorithm.

The SCMS algorithm provides a numerical approximation to the paths γ defined by the projected gradient. We illustrate the numerical algorithm in Section 8.

8. Implementation and examples. Here, we demonstrate ridge estimation in some two-dimensional examples. In each case, we will find the one-dimensional ridge set. Our purpose is to show proof of concept; there are many interesting implementation details that we will not address here. In each case, we use SuRF.

To implement the method requires that we choose a bandwidth h for the kernel density estimator. There has been recent work on bandwidth selection for multivariate density estimators such as Chacón and Duong (2010, 2012) and Panaretos

Subspace Ridge Finder (SuRF)

1. Compute the kernel density estimator $\hat{p}(x)$ and choose a threshold t. By default, the bandwidth h is selected by Silverman's rule.
2. Select a mesh $\mathcal{M}$ of points. By default, we take $\mathcal{M} = \{X_1, \ldots, X_n\}$.
3. Denoise: remove point m from mesh if $\hat{p}(m) < t$. Let $\mathcal{M}'$ denote the remaining mesh points.
4. Apply SCMS to each point in $\mathcal{M}'$.

FIG. 6. *SuRF algorithm.*

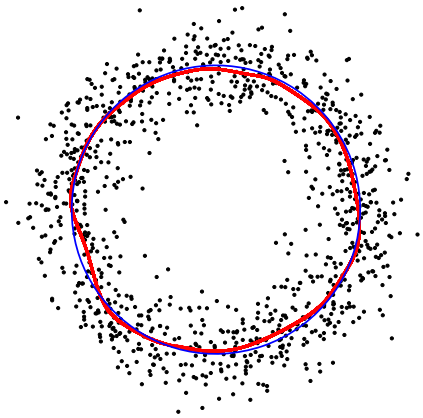

FIG. 7. *Estimated hyper-ridge set* (*red curve*) *from data generated from a circular manifold M* (*blue curve*) *of radius* 3. *The sample size is* 1000, *using Normal noise with* $\sigma = 0.5$. *The estimate is computed from a kernel density estimator using the Silverman Normal reference rule for the bandwidth. The starting points for the modified SCMS algorithm are taken the evaluation points of the density estimator excluding the points below* 25% *of the maximum estimated density.*

and Konis (2012). For the purposes of this paper, we simply use the Silverman rule [Scott (1992)].

Figures 7 through 10 show two examples of SuRF. In the first example, the manifold is a circle. Although the circle example may seem easy, we remind the

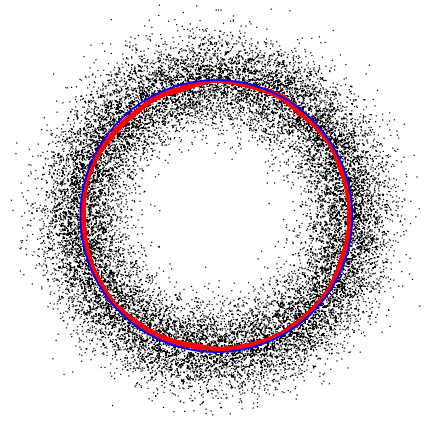

FIG. 8. *Estimated hyper-ridge set* (*red curve*) *from data generated from a circular manifold M* (*blue curve*) *of radius* 3. *The sample size is* 20,000, *using Normal noise with* $\sigma = 0.5$. *The estimate is computed from a kernel density estimator using the Silverman Normal reference rule for the bandwidth. The starting points for the modified SCMS algorithm are taken the evaluation points of the density estimator excluding the points below* 25% *of the maximum estimated density.*

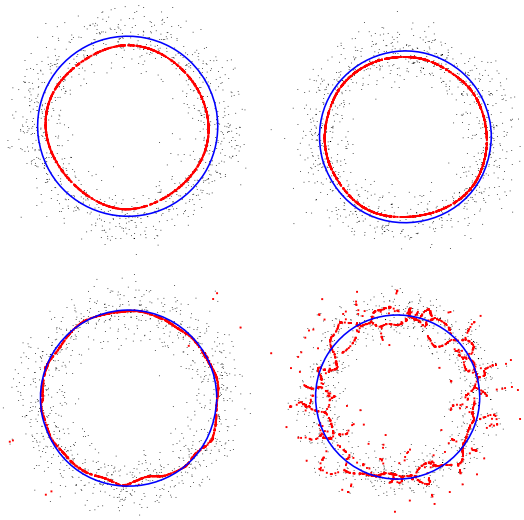

FIG. 9. *Effect of decreasing bandwidth. The data are i.i.d. samples from the same manifold as in the previous figure. Eventually we reach a phase transition where the structure of the estimator falls apart.*

reader that no existing statistical algorithms that we are aware of can, without prior assumptions, take a point cloud as input and find a circle, automatically.

The second example is a stylized "cosmic web" of intersecting line segments and with random background clutter. This is a difficult case that violates the as-

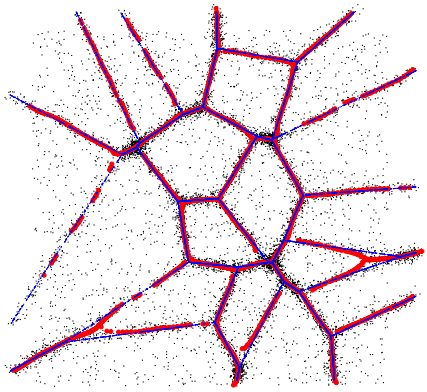

FIG. 10. *Data generated from a stylized "cosmic web" consisting of intersecting line segments and a uniform background clutter. Total sample size is* 10,000. *The starting points for the modified SCMS algorithm are taken the evaluation points of the density estimator excluding the points below* 5% *of the maximum estimated density.*

sumptions; specifically the underlying object does not have positive reach. The starting points for the SCMS algorithm are a subset of the grid points at which a kernel density estimator is evaluated. We select those points for which the estimated density is above a threshold relative to the maximum value.

Figure 9 shows the estimator for four bandwidths. This shows an interesting phenomenon. When the bandwidth h is large, the estimator is biased (as expected) but it is still homotopy equivalent to the true M. However, when h gets too small, we see a phase transition where the estimator falls apart and degenerates into small pieces. This suggests it is safer to oversmooth and have a small amount of bias. The dangers of undersmoothing are greater than the dangers of oversmoothing.

The theory in Section 6 required the underlying structure to have positive reach which rules out intersections and corners. To see how the method fares when these assumptions are violated, see Figure 10. While the estimator is far from perfect, given the complexity of the example, the procedure does surprisingly well.

9. Conclusion. We presented an analysis of nonparametric ridge estimation. Our analysis had two main components: conditions that guarantee that the estimated ridge converges to the true ridge, and conditions to relate the ridge to an underlying hidden manifold.

We are currently investigating several questions. First, we are finding the minimax rate for this problem to establish whether or not our proposed method is optimal. Also, Klemelä (2005) has derived mode estimation procedures that adapt to the local regularity of the mode. It would be interesting to derive similar adaptive theory for ridges. Second, the hidden manifold case required that the manifold had positive reach. We are working on relaxing this condition to allow for corners and intersections (often known as stratified spaces). Third, we are developing an extension where ridges of each dimension $d = 0, 1, \ldots$ are found sequentially and removed one at a time. This leads to a decomposition of the point cloud into structures of increasing dimension. Finally, there are a number of methods for speeding up the mean shift algorithm. We are investigating how to adapt these speedups for SuRF.

As we mentioned in the Introduction, there is recent work on metric graph reconstruction which is a way of modeling intersecting filaments [Aanjaneya et al. (2012), Lecci, Rinaldo and Wasserman (2013)]. These algorithms have the advantage of being designed to handle intersecting ridges. However, it appears that they are very sensitive to noise. Currently, we are investigating the idea of first running SuRF and then applying metric graph reconstruction. Preliminary results suggest that this approach may get the best of both approaches.

APPENDIX

The purpose of this appendix is to prove Lemma 8. Recall that the gradient is $g_\sigma(x) = -\frac{1}{\sigma^2}\int_M (x-z)\phi_\sigma(x-z)\,dW(z)$ and the Hessian is

$$
\begin{aligned}
H_\sigma(x) &= -\frac{1}{\sigma^2}\int_M \left(I - \frac{(x-z)(x-z)^T}{\sigma^2}\right)\phi_\sigma(x-z)\,dW(z) \\
&= -\frac{1}{\sigma^2}\left[p_\sigma(x)I - \frac{1}{\sigma^2}\int_M (x-z)(x-z)^T\phi_\sigma(x-z)\,dW(z)\right].
\end{aligned} \tag{45}
$$

We can partition M_σ into disjoint fibers. Choose an $x \in M_\sigma$ and let $\hat{x}$ be the unique projection of x onto M. Let $B = B(\hat{x}, K_\sigma)$. For any bounded function $f(x,z)$,

$$
\int_{M\cap B^c} f(x,z)\phi_\sigma(x-z)\,dW(z) \le \frac{C}{(2\pi)^{D/2}}\frac{e^{-K_\sigma^2/(2\sigma^2)}}{\sigma^D}W(B^c) \le C\sigma^A. \tag{46}
$$

Let $T = T_{\hat{x}}M$ denote the d-dimensional tangent space at $\hat{x}$ and let $T^\perp$ denote the $(D-d)$-dimensional normal space. For $z \in B \cap M$, let $\overline{z}$ be the projection of z onto T. Then

$$
x - z = (x - \hat{x}) + (\hat{x} - \overline{z}) + (\overline{z} - z) = d_M(x)u + (\hat{x} - \overline{z}) + R, \tag{47}
$$

where $u = (x-\hat{x})/d_M(x) \in T^\perp$ and $R = (\overline{z} - z)$. [Recall that d_M is the distance function; see (8).] For small enough σ, there is a smooth map h taking z to $\overline{z}$ that is a bijection $B \cap M$ and so the distribution W induces a distribution $\overline{W}$, that is, $\overline{W}(A) = W(h^{-1}(A))$. Let $\overline{w}$ denote the density of $\overline{W}$ with respect to Lebesgue measure μ_d on T. The density is bounded above and below and has two continuous derivatives.

LEMMA 10. *For every $x \in R \oplus \sigma$, $\sup_{z\in B}\|z - \overline{z}\| \le cK_\sigma^2$.*

PROOF. Choose any $z \in B$ and let $\overline{z}$ be its projection onto T. Because the reach is $\kappa > 0$, there exists a ball $S(a,\kappa) \subset \mathbb{R}^D$ such that a is in the plane defined $\hat{x}, z$ and $\overline{z}$, $S(a,\kappa)$ is tangent to the manifold at $\hat{x}$ and $S(a,\kappa)$ does not intersect M except at $\hat{x}$. Consider the line through z and $\overline{z}$ and let z_a be the point where the line intersects $S(a,\kappa)$. Now $\|z_a - \overline{z}\| \ge \|z - \overline{z}\|$ and by elementary geometry, $\|z_a - \overline{z}\| \le CK_\sigma^2$. □

Recall that $r_\sigma = \alpha\sigma$ with $0 < \alpha < 1$. Define the following quantities:

$$
\beta = \frac{e^{-\alpha^2/2}(1-\alpha^2)}{2\sigma^{D-d+2}}, \qquad \widetilde{\sigma} = \sigma\log^3\left(\frac{1}{\sigma^{D+A}}\right),
$$

$$
\phi_\perp(u) = \frac{e^{-\|u\|^2/(2\sigma^2)}}{(2\pi)^{(D-d)/2}\sigma^{D-d}}, \qquad \phi_\parallel(w) = \frac{e^{-\|w\|^2/(2\sigma^2)}}{(2\pi)^{d/2}\sigma^d},
$$

$$
p_{\sigma,B}(x) = \int_{M\cap B}\phi_\sigma(x-z)\,dW(z), \qquad B = B(\hat{x}, K_\sigma),
$$

where $u \in \mathbb{R}^{D-d}$ and $w \in \mathbb{R}^d$.

LEMMA 11. *We have that*

$$\phi_\sigma(x-z) = \phi_\perp(x-\hat{x})\phi_\parallel(\hat{x}-\overline{z})\big(1+O(\widetilde{\sigma})\big). \tag{48}$$

PROOF. First note that, for all $x \in R_\sigma$,

$$\frac{1}{(2\pi)^{(D-d)/2}}\frac{e^{-\alpha^2/2}}{\sigma^{D-d}} \le \phi_\perp(x-\hat{x}) \le \frac{1}{(2\pi)^{(D-d)/2}}\frac{1}{\sigma^{D-d}} \tag{49}$$

and so, $\phi_\perp(x-\hat{x}) \asymp \sigma^{-(D-d)}$ as $\sigma \to 0$. Now,

$$\|x-z\|^2 = \|x-\hat{x}\|^2 + \|\hat{x}-z\|^2 + \|\overline{z}-z\|^2 + 2\langle x-\hat{x}, \overline{z}-z\rangle,$$

we have that

$$\phi_\sigma(x-z) = \phi_\perp(x-\hat{x})\phi_\parallel(\hat{x}-\overline{z})e^{-\|z-\overline{z}\|^2/(2\sigma^2)}e^{-\langle x-\hat{x},\overline{z}-z\rangle/\sigma^2}.$$

Now $\|z-\overline{z}\|^2 = O(K_\sigma^4)$ and $|\langle x-\hat{x}, \overline{z}-z\rangle| \le \|x-\hat{x}\|\,\|\overline{z}-z\| = O(\sigma K_\sigma^2)$ and so

$$e^{-\|z-\overline{z}\|^2/(2\sigma^2)}e^{-\langle x-\hat{x},\overline{z}-z\rangle/\sigma^2} = \big(1+O(\widetilde{\sigma})\big). \qquad \square$$

PROOF OF LEMMA 8. 1. From (46), $p_\sigma(x) = \int_{M\cap B}\phi_\sigma(x-z)\,dW(z) + O(\sigma^A)$. Now

$$\begin{aligned}
&\int_{M\cap B}\phi_\sigma(x-z)\,dW(z)\\
&\quad= \big(1+O(\widetilde{\sigma})\big)\phi_\perp(x-\hat{x})\int_{M\cap B}\phi_\parallel(\hat{x}-\overline{z})\,dW(z)\\
&\quad= \big(1+O(\widetilde{\sigma})\big)\phi_\perp(x-\hat{x})\int_{TM\cap B}\phi_\parallel(\hat{x}-\overline{z})\overline{w}(\overline{z})\,d\mu_d(\overline{z})\\
&\quad= \big(1+O(\widetilde{\sigma})\big)\phi_\perp(x-\hat{x})\int_T \frac{1}{(2\pi)^{d/2}}e^{-\|t\|^2/2}\overline{w}(\hat{x}+\sigma t)\,d\mu_d(t),
\end{aligned}$$

where $A = \{t = (\overline{z}-\hat{x})/\sigma : \overline{z} \in B\}$. The volume of T is $O(\sigma^{D+A})$ and $T \to \mathbb{R}^d$ as $\sigma \to 0$. Also, $\overline{w}(\hat{x}+\sigma t) = \overline{w}(\hat{x}) + O(\sigma)$. Hence,

$$\int_T \overline{w}(\hat{x}+\sigma t)\,d\mu_d(t) = \big(\overline{w}(\hat{x})+O(\sigma)\big)\big(1-O\big(\sigma^{D+A}\big)\big)$$

and so

$$\int_{M\cap B}\phi_\sigma(x-z)\,dW(z) = \phi_\perp(x-\hat{x})\big(1+O(\widetilde{\sigma})\big)\big(\overline{w}(\hat{x})+O(\sigma)\big)\big(1-O\big(\sigma^{D+A}\big)\big)$$

and

$$\begin{aligned}
p_\sigma(x) &= \phi_\perp(x-\hat{x})\big(1+O(\widetilde{\sigma})\big)\big(\overline{w}(\hat{x})+O(\sigma)\big)\big(1-O\big(\sigma^{D+A}\big)\big) + O\big(\sigma^A\big)\\
&= \phi_\perp(x-\hat{x})\big(1+O(\widetilde{\sigma})\big).
\end{aligned}$$

2. $p_{\sigma,B}(x)$. This follows since in part 1 we showed that $p_{\sigma,B}(x) = p_\sigma(x) + O(\sigma^A)$.

3. For the gradient, we have

$$\begin{aligned}
-\sigma^2 g_\sigma(x) &= \int (x-z)\phi_\sigma(x-z)\,dW(z) \\
&= (x-\hat{x})\int \phi_\sigma(x-z)\,dW(z) \\
&\quad + \int (\hat{x}-\overline{z})\phi_\sigma(x-z)\,dW(z) \\
&\quad + \int (\overline{z}-z)\phi_\sigma(x-z)\,dW(z) \\
&= \mathrm{I} + \mathrm{II} + \mathrm{III}.
\end{aligned}$$

Now, $\mathrm{I} = (x-\hat{x})\,p_\sigma(x) = (x-\hat{x})\phi_\perp(x-\hat{x})(1+O(\widetilde{\sigma}))$ and

$$\begin{aligned}
\mathrm{II} &= \int_{M\cap B} (\hat{x}-\overline{z})\phi_\sigma(x-z)\,dW(z) + O\big(\sigma^A\big) \\
&= \big(1+O(\widetilde{\sigma})\big)\phi_\perp(x-\hat{x})\int_{M\cap B} (\hat{x}-\overline{z})\phi_\parallel(\hat{x}-\overline{z})\,dW(z) + O\big(\sigma^A\big).
\end{aligned}$$

For some u between $\hat{x}$ and $\overline{z}$, we have

$$\begin{aligned}
&\int_{M\cap B} (\hat{x}-\overline{z})\phi_\parallel(\hat{x}-\overline{z})\,dW(z) \\
&\quad = \sigma\int_{M\cap B} \frac{\hat{x}-\overline{z}}{\sigma}\phi_\parallel(\hat{x}-\overline{z})\,dW(z) \\
&\quad = \sigma\int_{h^{-1}(B)} \frac{\hat{x}-\overline{z}}{\sigma}\phi_\parallel(\hat{x}-\overline{z})\overline{w}(\overline{z})\,d\mu_d(\overline{z}) \\
&\quad = \sigma\int_A t\frac{1}{(2\pi)^{d/2}}e^{-\|t\|^2/2}\overline{w}(\hat{x}+\sigma t)\,d\mu_d(t) \\
&\quad = \sigma\int_A t\frac{1}{(2\pi)^{d/2}}e^{-\|t\|^2/2}\big[\overline{w}(\hat{x}) + \overline{w}'(\hat{x})\sigma t + \overline{w}''(u)\sigma^2 t^2/2\big]\,d\mu_d(t) \\
&\quad = O\big(\sigma^2\big),
\end{aligned}$$

where $A = \{t = (\overline{z}-\hat{x})/\sigma \in h^{-1}(B)\}$. Finally,

$$\begin{aligned}
\mathrm{III} &= \int_{M\cap B} (\overline{z}-z)\phi_\sigma(x-z)\,dW(z) + O\big(\sigma^A\big) \\
&= \phi_\perp(x-\hat{x})\,O\big(K_\sigma^2\big) + O\big(\sigma^A\big) \\
&= O\big(K_\sigma^2\big)\phi_\perp(x-\hat{x}).
\end{aligned}$$

Hence,

$$-\sigma^2 g_\sigma(x) = (x-\hat{x})p_\sigma(x) + O(\sigma^2) + \phi_\perp(x-\hat{x})O(K_\sigma^2)$$
$$= p_\sigma(x)\big((x-\hat{x}) + O(K_\sigma^2)\big)$$

and hence

$$g_\sigma(x) = -\frac{1}{\sigma^2}p_\sigma(x)\big((x-\hat{x}) + O(K_\sigma^2)\big).$$

It follow from part 1 that $\|g_\sigma(x)\| = O(\sigma^{-(D-d-1)})$.

4. To find the eigenvalues, we first approximate the Hessian. Without loss of generality, we can rotate the coordinates so that T is spanned by $e_1, \ldots, e_d$, $T^\perp$ is spanned by $e_{d+1}, \ldots, e_D$ and $u = (0, \ldots, 0, 1)$. Now,

$$-\frac{\sigma^2 H_\sigma(x)}{p_\sigma(x)} = I - \frac{\int (x-z)(x-z)^T \phi_\sigma(x-z)\,dW(z)}{\sigma^2 \int \phi_\sigma(x-z)\,dW(z)}$$

and

$$\int_{M\cap B^c} (x-z)(x-z)^T \phi_\sigma(x-z)\,dW(z) = O(\sigma^A).$$

Let $Q = \int_{M\cap B}(x-z)(x-z)^T\phi_\sigma(x-z)\,dW(z)$. Then, from (47), we have $Q = Q_1 + Q_2 + Q_3 + Q_4 + Q_5 + Q_6$ where

$$Q_1 = d_M^2(x)uu^T \int_{M\cap B} \phi_\sigma(x-z)\,dW(z),$$

$$Q_2 = \int_{M\cap B} (\hat{x}-\overline{z})(\hat{x}-\overline{z})^T \phi_\sigma(x-z)\,dW(z),$$

$$Q_3 = \int_{M\cap B} (\overline{z}-z)(\overline{z}-z)^T \phi_\sigma(x-z)\,dW(z),$$

$$Q_4 = \int_{M\cap B} (x-\hat{x})(\hat{x}-\overline{z})^T \phi_\sigma(x-z)\,dW(z),$$

$$Q_5 = \int_{M\cap B} (x-\hat{x})(\overline{z}-z)^T \phi_\sigma(x-z)\,dW(z),$$

$$Q_6 = \int_{M\cap B} (\hat{x}-\overline{z})(\overline{z}-z)^T \phi_\sigma(x-z)\,dW(z).$$

First, we note that

$$Q_1 = d_M^2(x)uu^T\phi_\perp(x-\hat{x})\big(1+O(\widetilde{\sigma})\big).$$

Next,

$$Q_2 = \int_{M\cap B} (\hat{x}-\overline{z})(\hat{x}-\overline{z})^T \phi_\sigma(x-z)\,dW(z)$$
$$= \big(1+O(\widetilde{\sigma})\big)\phi_\perp(\hat{x}-x)\int_{M\cap B} (\hat{x}-\overline{z})(\hat{x}-\overline{z})^T \phi_\parallel(\hat{x}-\overline{z})\,dW(z)$$

and

$$\begin{aligned}\int_{M\cap B}&(\hat{x}-\overline{z})(\hat{x}-\overline{z})^T\phi_{\|}(\hat{x}-\overline{z})\,dW(z)\\ &=\int_{h^{-1}(B)}(\hat{x}-\overline{z})(\hat{x}-\overline{z})^T\phi_{\|}(\hat{x}-\overline{z})\overline{w}(\overline{z})\,d\mu_d(\overline{z})\\ &=\overline{w}(\hat{x})\int_{h^{-1}(B)}(\hat{x}-\overline{z})(\hat{x}-\overline{z})^T\phi_{\|}(\hat{x}-\overline{z})\,d\mu_d(\overline{z})+O\big(K_\sigma^5\big).\end{aligned}$$

Next, with $t=(t_1,\ldots,t_d,0,\ldots,0)$,

$$\begin{aligned}&\int_{h^{-1}(B)}(\hat{x}-\overline{z})(\hat{x}-\overline{z})^T\phi_{\|}(\hat{x}-\overline{z})\,d\mu_d(\overline{z})\\ &\quad=\sigma^2\int_{\overline{B}}tt^T(2\pi)^{-d/2}e^{-\|t\|^2/2}\,d\mu_d(t)\\ &\quad=\sigma^2\bigg(\int tt^T(2\pi)^{-d/2}e^{-\|t\|^2/2}\,d\mu_d(t)\\ &\qquad\quad-\int_{\overline{B}^c}tt^T(2\pi)^{-d/2}e^{-\|t\|^2/2}\,d\mu_d(t)\bigg)\\ &\quad=\sigma^2\bigg(\begin{bmatrix}I_d&0\\0&0\end{bmatrix}+O\big(\sigma^{A+D}\big)\bigg)\end{aligned}$$

and so

$$\begin{aligned}Q_2=\big(1+O(\widetilde{\sigma})\big)&\phi_\perp(x-\hat{x})\sigma^2\\ &\times\bigg(\begin{bmatrix}I_d&0\\0&0\end{bmatrix}+O\big(\sigma^{A+D}\big)\bigg).\end{aligned}$$

A similar analysis on the remaining terms yields:

$$\begin{aligned}Q_3&=\big(1+O(\widetilde{\sigma})\big)\phi_\perp(x-\hat{x})O\big(K_\sigma^4\big),\\ Q_4&=\big(1+O(\widetilde{\sigma})\big)\phi_\perp(x-\hat{x})O\big(\sigma K_\sigma^2\big),\\ Q_5&=\big(1+O(\widetilde{\sigma})\big)\phi_\perp(x-\hat{x})O\big(\sigma K_\sigma^2\big),\\ Q_6&=\big(1+O(\widetilde{\sigma})\big)\phi_\perp(x-\hat{x})O\big(K_\sigma^3\big).\end{aligned}$$

Combining all the terms, we have

$$\begin{aligned}Q=\big(1+O(\widetilde{\sigma})\big)&\phi_\perp(x-\hat{x})\\ &\times\bigg(d_M^2(x)uu^T+\sigma^2\bigg(\begin{bmatrix}I_d&0\\0&0\end{bmatrix}+O\big(\sigma^{A+D}\big)\bigg)\bigg)\\ &+O\big(\sigma K_\sigma^2\big).\end{aligned}$$

Hence,

$$H_\sigma(x) = -\bigl(1 + O(\widetilde{\sigma})\bigr)\frac{p_\sigma(x)}{\sigma^2} \times \left(\left[\begin{array}{cccccccc} 0 & \cdots & 0 & 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & \ddots & \vdots & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \cdots & 0 & 0 & \cdots & \cdots & \cdots & 0 \\ \hline 0 & \cdots & 0 & 1 & 0 & \cdots & \cdots & 0 \\ 0 & \cdots & 0 & 0 & 1 & \cdots & \cdots & 0 \\ 0 & \cdots & 0 & 0 & 0 & \ddots & \cdots & 0 \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 1 - \frac{d_M^2(x)}{\sigma^2} \end{array}\right] + O(\widetilde{\sigma}) \right).$$

The result follows.

5. This follows from part 4 and the Davis–Kahan theorem.

6. From part 5, $L_\sigma(x) = L^\dagger + E$ where $L^\dagger = \left[\begin{smallmatrix} 0_{d\times d} & 0_{d,D-d} \\ 0_{D-d,d} & I_{D-d} \end{smallmatrix}\right]$ and $E = O(\widetilde{\sigma})$. Hence, $G_\sigma(x) = L_\sigma(x)g_\sigma(x) = (L^\dagger + E)g_\sigma(x)$ and the result follows from parts 3 and 4.

7. These follow from part 4.

8. Now we turn to $\|H'_\sigma\|$. Let $\Delta = (x - z)$. We claim that

$$H' = \frac{1}{\sigma^4}\int \Bigl[(\Phi \otimes I)(I \otimes \Delta + \Delta \otimes I) - \frac{\phi_\sigma(\Delta)}{\sigma^2}\bigl(I \otimes \Delta\Delta^T\bigr)\bigl(\mathsf{vec}(I) \otimes \Delta^T\bigr)\Bigr] dW(z) + \frac{1}{\sigma^4}\int \phi_\sigma(\Delta)\bigl(\mathsf{vec}(I) \otimes \Delta^T\bigr)\, dW(z).$$

To see this, note first that $H = \frac{1}{\sigma^4}Q - \frac{1}{\sigma^2}A$ where

$$Q = \int (x - z)(x - z)^T \phi_\sigma(x - z)\, dW(z) \quad \text{and} \quad A = p_\sigma(x)I. \tag{50}$$

Note that $Q = \int (x - z)(x - z)^T \Phi\, dW(z)$ where $\Phi = \phi_\sigma(\Delta)I_D$. So

$$Q' = \int (d/dx)\bigl[(x - z)(x - z)^T \Phi\bigr] dW(z)$$

and

$$\frac{d}{dx}\bigl[(x - z)(x - z)^T \Phi\bigr] = \frac{d(x - z)(x - z)^T \Phi}{dx} = \frac{d\Delta\Delta^T \Phi}{d\Delta}.$$

Now $(d/dx)(\Delta\Delta^T\Phi) = (fg)'$ where $f = \Delta\Delta^T$ and $g = \Phi$ and so

$$\frac{d}{dx}(\Delta\Delta^T\Phi) = (\Phi\otimes I)\frac{d}{dx}(\Delta\Delta^T) + (I\otimes\Delta\Delta^T)\frac{d}{dx}\Phi$$
$$= (\Phi\otimes I)(I\otimes\Delta + \Delta\otimes I) - \frac{\phi_\sigma(\Delta)}{\sigma^2}(I\otimes\Delta\Delta^T)(\mathsf{vec}(I)\otimes\Delta^T).$$

Hence,

$$Q' = \int\left[(\Phi\otimes I)(I\otimes\Delta + \Delta\otimes I) - \frac{\phi_\sigma(\Delta)}{\sigma^2}(I\otimes\Delta\Delta^T)(\mathsf{vec}(I)\otimes\Delta^T)\right]dW(z).$$

By a similar calculation,

$$A' = -\frac{1}{\sigma^2}\int\phi_\sigma(\Delta)(\mathsf{vec}(I)\otimes\Delta^T)\,dW(z).$$

Thus,

$$H' = \frac{1}{\sigma^4}Q' - \frac{1}{\sigma^2}A'$$
$$= \frac{1}{\sigma^4}\int\Bigg[(\Phi\otimes I)(I\otimes\Delta + \Delta\otimes I) - \frac{\phi_\sigma(\Delta)}{\sigma^2}(I\otimes\Delta\Delta^T)(\mathsf{vec}(I)\otimes\Delta^T)\Bigg]dW(z)$$
$$+ \frac{1}{\sigma^4}\int\phi_\sigma(\Delta)(\mathsf{vec}(I)\otimes\Delta^T)\,dW(z).$$

Each of these terms is of order $O(\sup_{x\in M}\|w''(x)\|/\sigma^{D-d+1})$. Consider the first term

$$\frac{1}{\sigma^4}\int(\Phi\otimes I)(I\otimes\Delta)\,dW(z) = \frac{1}{\sigma^{4+D}}(2\pi)^{D/2}\int e^{-\|x-z\|^2/(2\sigma^2)}(I\otimes\Delta)\,dW(z)$$
$$= \frac{1}{\sigma^{3+D}(2\pi)^{D/2}}\int e^{-\|u\|^2/2}(I\otimes u)\,dW(z),$$

where $u = (x-z)/\sigma$. As in the proof of part 1, we can restrict to $B\cap M$, do a change of measure to $\overline{W}$ and the term is dominated by

$$\frac{1}{\sigma^{3+D-d}(2\pi)^{D/2}}\int_A e^{-\|u\|^2/2}(I\otimes u)\big[\overline{w}(\hat{x}) + \overline{w}'(\tilde{u})\sigma u\big]\,d\mu_d(t)$$
$$= \frac{C}{\sigma^{3+D-d}(2\pi)^{D/2}}.$$ □

The other terms may be bounded similarly.

Acknowledgements. The authors thank the reviewers for many suggestions that improved the paper. In particular, we thank the Associate Editor who suggested a simplified proof of Lemma 3.

REFERENCES

AANJANEYA, M., CHAZAL, F., CHEN, D., GLISSE, M., GUIBAS, L. and MOROZOV, D. (2012). Metric graph reconstruction from noisy data. *Internat. J. Comput. Geom. Appl.* **22** 305–325. MR2994583

ADAMS, H., ATANASOV, A. and CARLSSON, G. (2011). Morse theory in topological data analysis. Preprint. Available at arXiv:1112.1993.

ARAGÓN-CALVO, M. A., PLATEN, E., VAN DE WEYGAERT, R. and SZALAY, A. S. (2010). The spine of the cosmic web. *The Astrophysical Journal* **723** 364.

ARIAS-CASTRO, E., MASON, D. and PELLETIER, B. (2013). On the estimation of the gradient lines of a density and the consistency of the mean-shift algorithm. Unpublished manuscript.

BENDICH, P., WANG, B. and MUKHERJEE, S. (2012). Local homology transfer and stratification learning. In *Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms* 1355–1370. SIAM, New York.

BHATIA, R. (1997). *Matrix Analysis. Graduate Texts in Mathematics* **169**. Springer, New York. MR1477662

CAILLERIE, C., CHAZAL, F., DEDECKER, J. and MICHEL, B. (2011). Deconvolution for the Wasserstein metric and geometric inference. *Electron. J. Stat.* **5** 1394–1423. MR2851684

CARREIRA-PERPINAN, M. and WILLIAMS, C. (2003). On the number of modes of a Gaussian mixture. In *Scale Space Methods in Computer Vision* 625–640. Springer, New York.

CHACÓN, J. E. (2012). Clusters and water flows: A novel approach to modal clustering through morse theory. Preprint. Available at arXiv:1212.1384.

CHACÓN, J. and DUONG, T. (2012). Bandwidth selection for multivariate density derivative estimation, with applications to clustering and bump hunting. Preprint. Available at arXiv:1204.6160.

CHACÓN, J. E. and DUONG, T. (2010). Multivariate plug-in bandwidth selection with unconstrained pilot bandwidth matrices. *TEST* **19** 375–398. MR2677734

CHACÓN, J. E., DUONG, T. and WAND, M. P. (2011). Asymptotics for general multivariate kernel density derivative estimators. *Statist. Sinica* **21** 807–840. MR2829857

CHAZAL, F., COHEN-STEINER, D. and LIEUTIER, A. (2009). A sampling theorem for compact sets in Euclidean space. *Discrete Comput. Geom.* **41** 461–479. MR2486371

CHAZAL, F. and LIEUTIER, A. (2005). The λ-medial axis. *Graphical Models* **67** 304–331.

CHAZAL, F., OUDOT, S., SKRABA, P. and GUIBAS, L. J. (2011). Persistence-based clustering in Riemannian manifolds. In *Computational Geometry* (*SCG'*11) 97–106. ACM, New York. MR2919600

CHENG, Y. (1995). Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **17** 790–799.

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. IMS, Beachwood, OH. MR2126901

COMANICIU, D. and MEER, P. (2002). Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24** 603–619.

DAVENPORT, M. A., HEGDE, C., DUARTE, M. F. and BARANIUK, R. G. (2010). Joint manifolds for data fusion. *IEEE Trans. Image Process.* **19** 2580–2594. MR2798033

EBERLY, D. (1996). *Ridges in image and data analysis*. Kluwer Academic, Boston.

EDELSBRUNNER, H., FASY, B. T. and ROTE, G. (2012). Add isotropic Gaussian kernels at own risk: More and more resilient modes in higher dimensions. In *Computational Geometry* (*SCG'*12) 91–100. ACM, New York. MR3024704

FUKUNAGA, K. and HOSTETLER, L. D. (1975). The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Trans. Inform. Theory* **IT-21** 32–40. MR0388638

GE, X., SAFA, I., BELKIN, M. and WANG, Y. (2011). Data skeletonization via reeb graphs. In *Advances in Neural Information Processing Systems* 24 (J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira and K. Q. Weinberger, eds.) 837–845. Curran Associates, New York.

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

GENOVESE, C. R., PERONE-PACIFICO, M., VERDINELLI, I. and WASSERMAN, L. (2012a). The geometry of nonparametric filament estimation. *J. Amer. Statist. Assoc.* **107** 788–799. MR2980085

GENOVESE, C. R., PERONE-PACIFICO, M., VERDINELLI, I. and WASSERMAN, L. (2012b). Manifold estimation and singular deconvolution under Hausdorff loss. *Ann. Statist.* **40** 941–963. MR2985939

GENOVESE, C. R., PERONE-PACIFICO, M., VERDINELLI, I. and WASSERMAN, L. (2012c). Minimax manifold estimation. *J. Mach. Learn. Res.* **13** 1263–1291. MR2930639

GINÉ, E. and GUILLOU, A. (2002). Rates of strong uniform consistency for multivariate kernel density estimators. *Ann. Inst. Henri Poincaré Probab. Stat.* **38** 907–921. MR1955344

HALL, P., PENG, L. and RAU, C. (2001). Local likelihood tracking of fault lines and boundaries. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **63** 569–582. MR1858403

HALL, P., QIAN, W. and TITTERINGTON, D. M. (1992). Ridge finding from noisy data. *J. Comput. Graph. Statist.* **1** 197–211. MR1270818

HORN, R. A. and JOHNSON, C. R. (2013). *Matrix Analysis*, 2nd ed. Cambridge Univ. Press, Cambridge. MR2978290

IRWIN, M. C. (1980). *Smooth Dynamical Systems*. *Pure and Applied Mathematics* **94**. Academic Press, New York. MR0586942

KLEMELÄ, J. (2005). Adaptive estimation of the mode of a multivariate density. *J. Nonparametr. Stat.* **17** 83–105. MR2112688

KLEMELÄ, J. (2009). *Smoothing of Multivariate Data*: *Density Estimation and Visualization*. Wiley, Hoboken, NJ. MR2640738

LECCI, F., RINALDO, A. and WASSERMAN, L. (2013). Statistical analysis of metric graph reconstruction. Preprint. Available at arXiv:1305.1212.

LI, J., RAY, S. and LINDSAY, B. G. (2007). A nonparametric statistical approach to clustering via mode identification. *J. Mach. Learn. Res.* **8** 1687–1723. MR2332445

MAGNUS, X. and NEUDECKER, H. (1988). Matrix Differential Calculus. Wiley, New York.

NIYOGI, P., SMALE, S. and WEINBERGER, S. (2008). Finding the homology of submanifolds with high confidence from random samples. *Discrete Comput. Geom.* **39** 419–441. MR2383768

NORGARD, G. and BREMER, P.-T. (2012). Second derivative ridges are straight lines and the implications for computing lagrangian coherent structures. *Phys. D* **241** 1475–1476.

OZERTEM, U. and ERDOGMUS, D. (2011). Locally defined principal curves and surfaces. *J. Mach. Learn. Res.* **12** 1249–1286. MR2804600

PANARETOS, V. M. and KONIS, K. (2012). Nonparametric construction of multivariate kernels. *J. Amer. Statist. Assoc.* **107** 1085–1095. MR3010896

PEIKERT, R., GÜNTHER, D. and WEINKAUF, T. (2012). Comment on "Second derivative ridges are straight lines and the implications for computing lagrangian coherent structures". *Phys. D* **242** 65–66.

PRAKASA RAO, B. L. S. (1983). *Nonparametric Functional Estimation*. Academic Press, New York. MR0740865

ROWEIS, S. T. and SAUL, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science* **290** 2323–2326.

SCHINDLER, B., PEIKERT, R., FUCHS, R. and THEISEL, H. (2012). Ridge concepts for the visualization of Lagrangian coherent structures. In *Topological Methods in Data Analysis and Visualization II*. 221–235. Springer, Heidelberg. MR3025953

SCOTT, D. W. (1992). *Multivariate Density Estimation*: *Theory*, *Practice*, *and Visualization*. Wiley, New York. MR1191168

SOUSBIE, T., PICHON, C., COURTOIS, H., COLOMBI, S. and NOVIKOV, D. (2008). The three-dimensional skeleton of the SDSS. *The Astrophysical Journal Letters* **672** L1.

STEWART, G. W. and SUN, J. G. (1990). *Matrix Perturbation Theory*. Academic Press, Boston, MA. MR1061154

TENENBAUM, J. B., DE SILVA, V. and LANGFORD, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science* **290** 2319–2323.

VON LUXBURG, U. (2007). A tutorial on spectral clustering. *Stat. Comput.* **17** 395–416. MR2409803

WEGMAN, E., CARR, D. and LUO, Q. (1993). Visualizing multivariate data. In *Multivariate Analysis: Future Directions* 423–466. Elsevier, Washington, DC.

WEGMAN, E. 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.

YUKICH, J. E. (1985). Laws of large numbers for classes of functions. *J. Multivariate Anal.* **17** 245–260. MR0813235

C. R. GENOVESE
L. WASSERMAN
DEPARTMENT OF STATISTICS
CARNEGIE MELLON UNIVERSITY
PITTSBURGH, PENNSYLVANIA 15213
USA
E-MAIL: genovese@stat.cmu.edu
larry@stat.cmu.edu

M. PERONE-PACIFICO
DEPARTMENT OF STATISTICAL SCIENCES
SAPIENZA UNIVERSITY OF ROME
ROME
ITALY
E-MAIL: marco.peronepacifico@uniroma1.it

I. VERDINELLI
DEPARTMENT OF STATISTICS
CARNEGIE MELLON UNIVERSITY
PITTSBURGH, PENNSYLVANIA 15213
USA
AND
DEPARTMENT OF STATISTICAL SCIENCES
SAPIENZA UNIVERSITY OF ROME
ROME
ITALY
E-MAIL: isabella@stat.cmu.edu