Submitted to the Annals of Statistics

MANIFOLD FITTING BY RIDGE ESTIMATION: A SUBSPACE-CONSTRAINED APPROACH

BY ZHIGANG YAO*, AND ZHENG ZHAI

National University of Singapore

Modern statistics encounters with the high dimensional data in ambient space in almost everywhere. Although a data point usually represents itself as a long vector or a big matrix, in principle they all can be viewed as points on or near an intrinsic manifold. Estimation of the underlying manifold has been emerging as a great interest to the statistics community. In this paper, we propose the Subspace-Constrained Ridge Estimator (SCRE) to estimate the underlying manifold. The essence of the local SCRE is to extend a local version of the kernel-density estimation (KDE) approach to the manifold fitting field. First, we demonstrate the connection between the ridge, essentially an approximation of the unknown manifold coming from the probability density function of the data, and the manifold itself. Second, we show that, under mild conditions, this new approach has more promising theoretical results than the classical KDE-based approaches with respect to Hausdorff margin to the manifold. Third, we further show that the necessity of nonlinear transformation composed with the SCRE estimator, in order to improve the estimated ridge. An algorithm is also provided which outputs an discrete ridge sets. Numerical simulated examples as well as real data sets demonstrate that the ridge obtained from our new approach does indeed have an improved average margin to the underlying manifold relative to other primary methods in the field of manifold fitting.

1. Introduction. Today, the process of dealing with high-dimensional data requires the exponential consumption of computational resources, and is therefore costly but effective. Yet, even so, challenges remain, as the high-dimensional data tend to lie near a low-dimensional sub-manifold of the ambient space, which is a phenomenon termed "the curse of dimensionality" by Fefferman et al. (2018). Although, the geometric foundation of statistical inference in nonlinear regression was brought to attention of the statistical community as early as the 1980s by Kass (1989), the role of geometry was not primarily explored back then. At present, there are numerous works concerning manifold learning, with different areas of focus, such as dimension reduction focusing on approximating the embedding (Roweis and Saul, 2000; Zhang and Zha, 2004; Donoho and Grimes, 2003; Zha and Zhang, 2007) and manifold approximation via fitting the unknown manifold (Genovese et al., 2012; Chen et al., 2015; Genovese et al., 2014). In another seemingly-related direction, there have been works on statistical methods on manifolds over the past decades, centering on finding the mode of distribution (Huckemann, Hotz and Munk, 2010; Chen et al., 2016a,b) and searching for the principal components of manifold (Hauberg, 2015; Huckemann and Ziezold, 2006; Hastie and Stuetzle, 1989).

There has also been some theoretical research into manifold learning using hypothesis testing (Fefferman, Mitter and Narayanan, 2016; Fefferman et al., 2018, 2019). For instance,

*Supported in part by Singapore MOE Tier 1 funding (R-155-000-210-114) and Tier 2 funding (R-155-xxx-xxx-xxx).

MSC 2010 subject classifications: Primary 00X00, 00X00; secondary 00X00.

Keywords and phrases: Manifold Fitting, Ridge estimation, Kernel density estimation, Derivatives estimation.

Fefferman, Mitter and Narayanan (2016) developed an algorithm that tests the manifold hypothesis and Fefferman et al. (2018) provided a condition and proved that the produced putative manifold $\mathcal{M}_o$'s Hausdorff distance to $\mathcal{M}$ is small, and its reach is not much smaller than that of $\mathcal{M}$. Mohammed and Narayanan proposed that the output manifold $\mathcal{M}$ can be defined by the set of points where the gradient of the approximate squared distance function (asdf) is orthogonal to the subspace spanned by the largest $D-d$ eigenvectors of the Hessian of the asdf (Mohammed and Narayanan, 2017).

Manifold-fitting algorithms intend to recover the low-dimensional manifold in the ambient space by supposing that the observation data is generated from the manifold with some noise added. There are many classical algorithms to find the lower-dimensional manifold, such as the principal flow (Panaretos, Pham and Yao, 2014) and the local principal curves and surfaces based on kernel-density estimation (Ozertem and Erdogmus, 2011). The principal flow attempts to fit a curve moving along the maximal variation of the data, subject to a smoothness constraint. The solution to the principal flow is reduced to an ODE problem via the Euler-Lagrange method. Ozertem and Erdogmus (2011) redefined the principal curves and surfaces in terms of the gradient and Hessian of the probability density estimate. They also provided an algorithm based on the classical KDE. Jung, Dryden and Marron (2012) propose a general framework to fit the high-dimensional data with principal nested spheres. In addition, there is a large amount of literature about searching for hidden structure with an application in the point clouds, nebular hypothesis, earthquake locations, and object-oriented objects (Adams, Atanasov and Carlsson, 2011; Davenport et al., 2010; Klemelä, 2009; Patrangenaru and Ellingson, 2015).

Overall, the problem of manifold estimation is treated as a statistical manifold-fitting problem under the observed data $\{x_i, i=1:n\}$. The observed data, it may be assumed, are constructed as

$$x_i = \tilde{x}_i + \epsilon_i, i=1:n, \tag{1}$$

where noiseless data $\{\tilde{x}_i, i=1:n\}$ are assumed to be sampled from some unknown manifold $\mathcal{M}$ and the noise vectors $\{\epsilon_i, i=1:n\}$ are i.i.d and each ϵ_i obeys the rules of some distributions, such as the multivariate Gaussian distribution supported on $\mathbb{R}^D$, whose density function $f(\epsilon_i)$ yields the form as

$$f(\epsilon_i) = \frac{1}{(2\pi\sigma^2)^{D/2}} \exp(-\|\epsilon_i\|_2^2/2\sigma^2).$$

Thus, the manifold-estimation problem is aimed at recovering a set of new samples $\{\hat{x}_i, i=1:n'\}$, which we label $\hat{\mathcal{G}}$ from the data $\{x_i\}$ such that the recovered samples in $\hat{\mathcal{G}}$ stay in the real manifold $\mathcal{M}$ as much as possible. Normally, we can obtain $\hat{\mathcal{G}}$ by constructing an estimator using the data:

$$\hat{\mathcal{G}} = \hat{\mathcal{G}}(x_1, ..., x_n|\Theta), \tag{2}$$

where Θ represents the parameters used in the estimator. It is worth noting that the number of recovered samples in $\hat{\mathcal{G}}$ may be different from that of the data because we can always resample to find more points. Here, $\hat{\mathcal{G}}$ is named as a ridge estimator (discrete set). Regardless of the methods/algorithms used to obtain $\hat{\mathcal{G}}$, this estimator is, in principle, a subset of the ridge (continuous function) defined for the manifold $\mathcal{M}$; the connections between them will be elaborated in the later sections.

1.1. *The connection between $\mathcal{M}$ and the probability density function.* We use the probability density function as a useful tool to describe the manifold $\mathcal{M}$ for the following three reasons.

- The domain of the density function can be exactly on the manifold.
- The geometric property can be described by the analytic differential forms derived from the density function, such as the gradient and Hessian matrix.
- The probability that a point x belongs to the manifold can be determined from the density function. As a result, we can recover the manifold from the density function through sampling-rejection methods.

REMARK. The discussion on the relation among the principal curve/surface, the ridge with the Hessian or the gradient of the probability density function can be found in Ozertem and Erdogmus (2011); Genovese et al. (2014); Mohammed and Narayanan (2017). In the low-dimensional structure estimation approach, KDE, as a plug-in method, is commonly used in place of the unknown probability density function because of its concise form.

In this way, the manifold-fitting problem is closely related to the analysis of the density function. We will now explain the reasons in details.

The manifold $\mathcal{M}$ can also be viewed as a particular lower-dimensional data distribution in the ambient space. At the population level, we can also use a density function $p(x)$ to describe the data distribution for $\mathcal{M}$ in the ambient space:

$$p(x) \geq 0, \quad \int_{x \in \mathcal{M}} p(x)dx = 1, \quad p(x) = 0, \forall x \notin \mathcal{M}.$$

It should be noted that, even though $p(x)$ is not continuous for x in every direction (such as the directions in the normal space of $\mathcal{M}$), we can think of $p(x)$ as continuous and differentiable in the directions parallel with the manifold. Knowing $p(x)$ is equivalent to obtaining the information about $\mathcal{M}$, because we can recover the manifold by re-sampling and rejecting the points with small values of $p(x)$.

Clearly, it is possible to judge whether a point x is on $\mathcal{M}$ by directly evaluating $p(x)$. The simplest way is to check the condition $p(x) \neq 0$: if true, $x \in \mathcal{M}$. Another useful characteristic is that, when $x \in \mathcal{M}$, we know the gradient at x lies exactly in the tangent space of $\mathcal{M}$ at x, i.e. $\nabla p(x) \in \mathcal{T}_x$. For the density function $p(x)$, we can get the tangent space from the eigenvalue decomposition of the subspace-constrained Hessian matrix of $p(x)$ at x. The discontinuous property of $p(x)$ makes it impossible to derive the derivative in the normal directions of $\mathcal{M}$. To overcome this problem, we use the smooth density function $p_h(x)$ through convolution:

$$p_h(x) = \int p(y)K_h(x-y)dy. \tag{3}$$

Note that the resultant density function $p_h(x)$ is continuous and smooth in every direction. We would also need to choose a proper h such that the following conditions are satisfied:

- For $x \notin \mathcal{M}$, the gradient $\nabla p_h(x)$ points to $P_{\mathcal{M}}(x)$ (the projected nearest point of x in $\mathcal{M}$).
- For $x \in \mathcal{M}$, the gradient $\nabla p_h(x)$ lies in the space spanned by the top d eigenvectors of $H_{p_h}(x)$, i.e., the tangent space of $\mathcal{M}$ at x.
- There is τ, such that when $x \in \mathcal{M} \oplus \tau$, the projection $P(x) = U(x)U(x)^T$ corresponding to the top d eigenvectors of the Hessian $p_h(x)$ changes continuously when $x \to P_{\mathcal{M}}(x)$, where $\mathcal{M} \oplus \tau = \cup_{y \in \mathcal{M}} B_\tau^D(y) = \cup_{y \in \mathcal{M}} \{x | \|x-y\|_2 \leq \tau\}$.

At the sample level, we will need to use the empirical version, instead of (3), as follows:

$$\hat{p}_{n,h}(x) = \frac{1}{n}\sum_{k=1}^{n} K_h(x - x_k).$$

where h is the bandwidth and n is the number of samples. In order to ensure $\int \hat{p}_{n,h}(x)dx = 1$, we need to normalize it within $K_h(x - x_k)$. Usually, when we do not distinguish the bandwidth in each dimension, $K_h(x-x_k)$ can yield a simple KDE form, $1/h^D K(x_i - x/h)$. Assume $K(u)$ is a kernel that satisfies

$$\int K(u)du = 1, \int uK(u)du = 0, \int uu^T K(u)du = I\{1/D \int \|u\|_2^2 K(u)du\}.$$

For ease of notation, from this point we will omit the subscript n in $\hat{p}_{n,h}(x)$ (i.e. render it as $\hat{p}_h(x)$), when we do not need to emphasize n.

REMARK. $\hat{p}_h(x)$ is a random function depending on the observations $\{x_k\}$. $\hat{p}_h(x)$ can be used to estimate $p_h(x)$ is unbiased, i.e. the expectation of $\hat{p}_h(x)$ is $p_h(x)$. Even though we can use $\hat{p}(x)$ to approximate $p_h(x)$, it is not guaranteed that the gradient and Hessian of $\hat{p}_h(x)$ will approximate those of $p(x)$ well enough, as shown in the work of Sasaki et al. (2017). It is well known that $\mathrm{E}(\hat{p}_h(x)) - p(x) = O(h^2)$, and this is also true for the derivatives.

1.2. *Subspace-Constrained Derivatives.* Note that $p(x)$ is only smooth and differentiable in the direction parallel with $\mathcal{M}$. For $v \in \mathcal{T}_x$, recall that the directional derivative of $p(x)$ is

$$\partial_v p(x) = \lim_{t\to 0} \frac{p(x + vt) - p(x)}{t}.$$

Similarly, we have the second-order directional derivative $\partial_{v_1,v_2} p(x)$ by recurrently differentiating from the first-order derivative. For the tangent space $\mathcal{T}_x$ of $\mathcal{M}$ at x, we can find any orthogonal basis $V(x) = \{v_1(x), v_2(x), ..., v_d(x)\}$. Then the Hessian constrained in the tangent space is

$$H_V(x) = V(x)XV(x)^T.$$

where X is a $d \times d$ matrix with the i, j-th element as $X_{ij} = \{\partial_{v_i,v_j} p(x)\}, 0 \le i, j \le d$. Since X is a real symmetric matrix, we can find the eigenvalue decomposition

$$X = U(x)\Lambda U(x)^T.$$

where $U(x)$ is an $d \times d$ orthonormal matrix and Λ is a $d \times d$ diagonal matrix. Next, when we use the columns of $M(x) = V(x)U(x)$ as the new basis in the tangent space, we will get

$$H_V(x) = M(x)\Lambda M(x)^T.$$

After defining the subspace-constrained gradient and Hessian, we can extent the non-zero domain of $p(x)$ to $\mathcal{M} \oplus \tau$:

$$\bar{p}(x) = p(\tilde{x}) + (x - \tilde{x})^T M(\tilde{x})M^T(\tilde{x})\nabla p_M(\tilde{x}) + \frac{1}{2}(x - \tilde{x})^T M(\tilde{x})\Lambda M(\tilde{x})^T(x - \tilde{x}).$$

The constructed $\bar{p}(x)$ has several good properties:

- First, for any $x_1, x_2 \in \mathcal{M} \oplus \tau$, if $P_{\mathcal{M}}(x_1) = P_{\mathcal{M}}(x_2)$, we have $\bar{p}(x_1) = \bar{p}(x_2)$.
- Second, the gradient and Hessian remain unchanged with x moving in the normal space, which is a constant vector or matrix equal to the subspace-constrained gradient and Hessian of $p(x)$ at the projected point $P_{\mathcal{M}}(x)$.

To be more concrete, the set corresponding to a large $\bar{p}(x)$ will produce a geometric structure $\bar{\mathcal{M}}$ much thicker than $\mathcal{M}$. Even though we can extend the non-zero domain of $p(x)$ to get $\bar{p}(x)$, $\bar{p}(x)$ is still not continuous and not smooth on the boundary of $\mathcal{M} \oplus \tau$.

1.3. *Hessian Matrix Decomposition for $p_h(x)$.* For this reason, we can consider the convolution form $p_h(x)$, which is smooth and continuous everywhere.

A few calculation steps make it clear that $p_h(x) = p(x) + O(h^2)$. Because $p(x)$ only has derivatives along the manifold when $x \in \mathcal{M}$ and $p_h(x)$ are differentiable in every direction, we need to split the Hessian of $p_h(x)$ into two parts based on the tangent space of $p(x)$. In other words, we consider the derivatives in two independent spaces respectively. For the Hessian in the tangent space, we have

$$M(x)^T H_{p_h}(x) M(x) = \Lambda + O(h^2)\mathbb{1}_{d\times d},$$
$$M(x)_\perp^T H_{p_h}(x) M(x)_\perp = O(h^2)\mathbb{1}_{\{D-d\}\times\{D-d\}},$$
$$M(x)_\perp^T H_{p_h}(x) M(x) = O(h^2)\mathbb{1}_{\{D-d\}\times d}.$$

Merging the above three equalities, we conclude that $H_{p_h}(x)$ can be written as a main part and a small part concerning h:

$$H_{p_h}(x) = M(x)\Lambda M(x)^T + O(h^2)\mathbb{1}_{D\times D}.$$

The existence of the perturbation term $O(h^2)\mathbb{1}_{D\times D}$ in $H_{p_h}(x)$ makes it difficult to recover the tangent-space orthonormal matrix $M(x)$ directly when h gradually becomes large. To fix this problem, in this paper we propose a local approach to construct a better estimator and explain the effectiveness of our approach in reducing the effect of the perturbation. Our approach recovers the subspace $M(x)$ more accurately.

1.4. *Ridge and Manifold.* The most common examples of the manifold $\mathcal{M}$ are d-dimensional structures embedded in a D-dimensional space, such as $\mathcal{M}$ being the ball surface ($d = 2, D = 3$) or the torus ($d = 2, D = 3$). To deal with the manifold, we define a function

$$\phi(x) : \mathbb{R}^D \to \mathbb{R}^+,$$

and construct a set

$$R = \{x | \omega(\phi(x), \nabla\phi(x), H_\phi(x)) = 0\}$$

such that R will approximate $\mathcal{M}$ well. Here, we refer ω to some abstract relations or formulas. When $\phi(x)$ is selected as the density function, the points on $\mathcal{M}$ satisfy the subspace-constrained optimal condition. This is because, for $x \in \mathcal{M}$, $\phi(x)$ will decrease when x moves along any direction v as long as $v \perp \mathcal{T}_{\mathcal{M}}(x)$, i.e. x is a local optimal for the normal space of $\mathcal{M}$ at x. We call the set R consisting of points satisfying the subspace-constrained optimal condition the *Ridge*, which is similar to the ridge of a mountain in the real world.

The manifold-estimation problem is closely related to ridge estimation for the following reasons. First, the ridge is the set of points that satisfies the subspace-constrained optimal condition of a function $p(x)$, and, as a result, it can be smooth as long as the derivatives of $p(x)$ are continuous. Second, the ridge can easily yield a geometry structure of any dimension since the dimension size is controlled by the dimension of the constrained subspace. This property is also another key reason for using the ridge to estimate the manifold. For these two reasons, the ridge-estimation method bridges the gap between the discrete samples and the smooth manifold.

In this paper, we propose to estimate the ridge using the solution manifold of a local kernel density. The ridge obtained is shown to have a smaller distance (average margin or Hausdorff distance) from the underlying manifold where the observations are generated. We explain this phenomenon mainly from two perspectives: the bias of derivatives and the stability of the eigenspace corresponding to the Hessian matrix. In the first part of the paper, we provide the theoretical analysis of the bias of derivatives. In the second part, we show the relationship between a nonlinear transformation and the rank-one modification of the Hessian matrix. We also use a numerical experiment to verify the effectiveness of our method by measuring the ridge error between the estimated and the ideal hidden manifold.

1.5. *l-SCRE vs SCRE.* In this paper, our objective is to obtain a ridge that is closer to the underlying manifold, 'closer' here referring to the average margin or the Hausdorff distance. Previous works have shown that the Hausdorff distance between the estimated ridge and the real manifold is highly related to the approximation error for the gradients and the Hessians. This relation can be seen in the ridge definition, which shows that the ridge is a level set satisfying a specific gradient-Hessian condition, as the Hausdorff distance between two sets is the largest margin that corresponds to the L_∞ norm in functional space. Besides the Hausdorff-distance measurement, we also consider the average margin, which stands for the overall approximation level between them. The average margin corresponds to the L_2 or L_1 norm. Unlike the Hausdorff distance, the average margin is less prone to be affected by the noise of some outliers. Therefore, when we are more interested in evaluating the overall approximating standard, it is better to use the average margin than the Hausdorff distance.

To obtain a ridge that yields a much smaller average margin, we concentrate mainly on the following two questions:

- How do we reduce the approximation error for each element in the gradient and Hessian, i.e. the first and second order of the partial derivatives, by defining a new density function?
- How do we make a modification with respect to the Hessian matrix such that we can derive an algorithm that can yield a more robust performance?

The first approach is an analysis from a microscopic perspective, and the second approach is an analysis from a macroscopic perspective. More specifically, we define a local version of the SCRE (*l*-SCRE) using only the partial of the nearest samples, and also give the theoretical result for the bias of derivatives in the first step. Following this, we adopt nonlinear transformation for the *l*-SCRE to obtain a stable Hessian matrix. 'Stable Hessian' refers to the Hessian matrix field in which the eigenvectors' direction does not shift frequently with different locations of interest.

The advantage of *l*-SCRE compared with SCRE is that the local strategy makes the smooth parameter h selection process much easier, as the local domain of $\mathcal{M}$ tends to be a linear subspace. In an extreme case, if the samples of interest are located in an affine space, any local combination weight w_k that sums to 1 is also a point in the affine space. Therefore, everything in the smooth parameter h range from 0 to ∞ becomes acceptable. This phenomenon can also be verified in Section 5.

1.6. *l-SCRE in a Nutshell.* Our intuition is from the observation of the convergence-speed comparison between two functions. Suppose there is a radial basis function $\phi(r)$ that satisfies $\phi(r) \to 0$ and $\phi(r)r^2 \to 0$ when $r \to +\infty$. First, consider a scalar quantity

$$Q(x) = \sum_i \phi(\|x_i - x\|),$$

which is often seen in constructing SCRE for density estimation, and the weighted covariance matrix $C(x) = \sum_k \phi(\|x_i - x\|)(x_i - x)(x_i - x)^T$, which is often used to estimate the local principal vector field. When extracting the squared norm, we see

$$C(x) = \sum_i \|x_i - x\|^2 \phi(\|x_i - x\|) \frac{(x_i - x)(x_i - x)^T}{\|x_i - x\|^2},$$

where $\frac{(x_i - x)(x_i - x)^T}{\|x_i - x\|^2}$ is a rank-one projection matrix constructed by sample x_i and x. Therefore, the weight for sample x_i is $\|x_i - x\|^2 \phi(\|x_i - x\|)$. We know the weight $\|x_i - x\|^2 \phi(\|x_i - x\|)$ is more sensitive to the samples that have a large scale of $\|x_i - x\|$ when computing $C(x)$ than the component $\phi(\|x_i - x\|)$ for constructing $Q(x)$. Therefore, it is

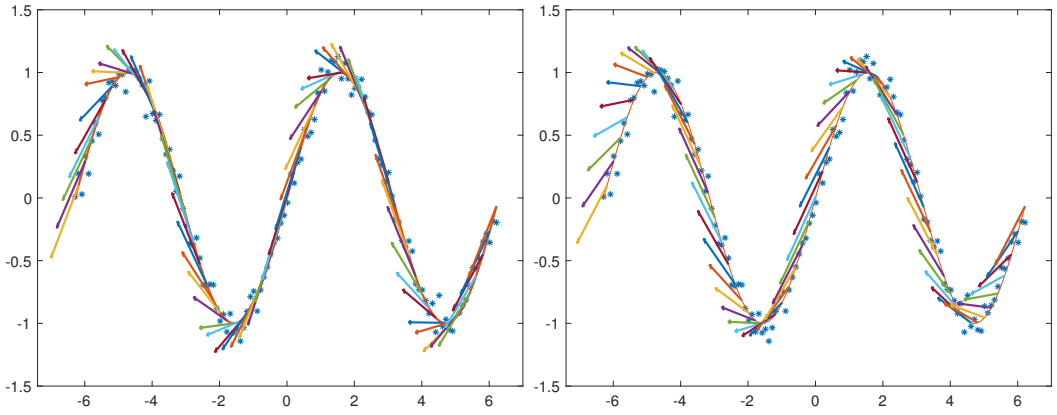

FIG 1. *Illustration of the vector field. Left: using partial nearest samples Right: using all samples*

worthwhile to remove some samples with a large scale to improve $C(x)$, because the degree of improvement for $C(x)$ is larger than the deterioration for $Q(x)$.

The effect of improving the quality of $C(x)$ by neglecting samples with a large scale can be seen in the vector field in Figure 1. The vector field is constructed from the dominant eigenvector of the weighted covariance matrix $C(x)$. In the left diagram, the set S is the k-nearest neighborhood of x, and in the right diagram the set S represents all the samples. From the figure, we know that by selecting the nearest samples we can get a weighted covariance matrix to approximate the vector field better.

Based on the above intuition, we show that constructing the *l*-SCRE by neglecting the far-away samples will help to reduce the point-wise bias of the derivatives. The reduced-bias phenomenon will further lead us to a ridge that is 'closer' than that obtained by using all the samples.

1.7. *Main Contributions.* Our contributions consist mainly of the following:

1. We propose to use the nearest samples to construct a local version of the SCRE method, and in turn use that to construct the ridge-estimator function. The proposed *l*-SCRE can obtain a ridge that is closer to the original manifold.
2. We provide the theoretical result of the component-wise bias and variance corresponding to the *l*-SCRE and prove that, under mild conditions, the local kernel function can indeed reduce the bias of derivatives more effectively than the KDE-based approaches.
3. By building the connection Hessian of the estimated density function with the empirical covariance matrix, we show the necessity of rank-one modification to obtain a stable tangent space estimator, which further improves the performance of the *l*-SCRE.

Finally, we provide numerical examples to show the effectiveness of our strategy, and also show that the vector field corresponding to our method is smoother and more appropriate for the manifold-estimation problem.

1.8. *Notations.* The most frequently used symbols and notations are listed in Table 1.

1.9. *Related Work.* Ozertem and Erdogmus (2011) proposed a KDE-based subspace mean-shift algorithm (SCMS) to find the principal curve projection where the kernel density estimator is defined as a data-related covariance sum of kernel functions

$$\hat{p}(x) = \frac{1}{n}\sum_i G_{\Sigma_i}(x - x_i).$$

TABLE 1
Symbols and Notations

Symbols	Meaning and Explanation
d,D	The dimension of the manifold and the dimension of the ambient space
E, Var	The expectation and the variance operator
$\mathcal{M}, \hat{\mathcal{G}}$	The true unknown manifold and the ridge estimator (discrete set)
$\mathcal{M}_{\hat{G}}$	The projection of the ridge estimator $\hat{\mathcal{G}}$ onto $\mathcal{M}$
$p_n(x), g_n(x), H(x)$	The smooth density function, its gradient, and the Hessian Matrix
$\hat{p}_{r,h}(x), g_{\hat{p}_{r,h}}(x), H_{\hat{p}_{r,h}}(x)$	The local KDE estimator, its gradient, and the Hessian Matrix
Π, Π^c	The projection onto the top d eigenvectors and its orthogonal complement
f	The nonlinear increasing, nonnegative, and concave function
$J(x)$	The semi-positive definite matrix, which is a key component of the Hessian
$C(x)$, $H(x)$	The weighted sample covariance matrix and the Hessian matrix at x
Π_H	The projection matrix constructed by the top d eigenvectors of H
$\mathcal{D}_r(x_0)$	The D-dimensional ball centered at x_0 with a radius of r
$\mathcal{N}(x, r)$	The samples in the neighborhood of x with a radius of r

Note that $\hat{p}(x)$ depends only on the n. By setting $\nabla\hat{p}(x) = 0$, we can obtain the mean-shift iteration:

$$x \leftarrow m(x) = (\sum_i c_i \Sigma_i^{-1})^{-1} \sum_i c_i \Sigma_i^{-1} x_i.$$

Directly applying the mean-shift algorithm will make x converge to the local maximum of $\hat{p}(x)$, which we call a mode. By restricting x to converge to the d dimension ridge defined by $\hat{p}(x)$, the subspace-constrained mean-shift iteration Ozertem and Erdogmus (2011) is given as

$$x \leftarrow m(x) = V(x)_\perp V(x)_\perp^T (\sum_i c_i \Sigma_i^{-1})^{-1} \sum_i c_i \Sigma_i^{-1} x_i,$$

where $V(x)_\perp$ is the eigenvectors corresponding to the $D-d$ largest eigenvectors of the Hessian matrix $\Sigma^{-1} = -\hat{p}^{-1}(x)\hat{H}(x) + \hat{p}^{-2}(x)\hat{g}(x)\hat{g}(x)^T$, where $\hat{H}(x), \hat{g}(x)$ are the Hessian and gradient of $\hat{p}(x)$.

Myhre et al. (2016) used the spectral decomposition on the Hessian matrix $H(x) = Q(x)\Lambda(x)Q(x)^T$, and, furthermore, decomposed $Q(x)$ into $Q(x) = [Q_\perp(x), Q_\parallel(x)]$, where $Q_\perp(x)$ is the eigenvectors corresponding to the d largest eigenvalues of $H(x)$. Then, projecting points onto a density ridge can be rendered as the following initial value problem:

$$\frac{dx(t)}{dt} = Q_\perp(x(t))Q_\perp^T(x(t))g(x(t)), \quad x(0) = x_0,$$

where $g(x(t))$ is the gradient of $P(x)$ at $x(t)$.

Mohammed and Narayanan (2017) also performed the gradient-ascent algorithm in the direction of $V(x)V(x)^T g(x)$, with V consisting of the eigenvectors corresponding to the $D-d$ largest eigenvalues, and $g(x)$ being the gradient of the approximate squared-distance function (asdf). The asdf is defined as a minus log transform of density function $p(x)$ or a weighted summation of squared distance for each cylinder.

Besides being defined with a kernel-density function, the manifold can also be defined with the zero points $\{x|F(x) = \mathbf{0}\}$ of a vector-valued function $F(x) : R^D \to R^D$. Such methods, as used by Fefferman et al. (2018); Mohammed and Narayanan (2017); Yao and Xia (2019), normally require estimating the local tangent space of the points residing in the neighborhood of the interested point x. The vector-valued function $F(x)$ is defined differently in works with different objectives. For example, Fefferman et al. (2018) define the function $F(x)$ as

$$F(x) = \Pi_x \sum_i \alpha_i(x)\Pi^i(x - x_i).$$

Furthermore, Yao and Xia (2019) simplify the two-step projection format into $F(x) = \Pi_x \sum_i \alpha_i(x)(x - x_i)$ by neglecting the inner projection Π^i. They also show the simplified version, which can also achieve good performance.

The work of Chen (2020), albeit tangentially related to the above, provides some theoretical results, such as the smoothness theorem, stability theorem, and convergence property of the gradient flow on the solution manifold. The solution manifold is defined as a set of points satisfying the abstract form $\{x|\Psi(x) = 0\}$.

2. Manifold Estimation. We adopt the classical kernel-density estimation approach to handle our manifold-fitting problem, as the majority of previous works do. From the kernel-density perspective, the manifold can be viewed as points that satisfy some specific ridge conditions of the estimated density function.

2.1. *Ridge Definition.* There are two main definitions of a ridge, both of which involve describing the eigenspace of the Hessian and the gradient.

DEFINITION 2.1. A point x is on the d-dimensional ridge R of its probability density function when the gradient $g(x)$ is orthogonal to at least $D - d$ eigenvectors of $H(x)$ and the corresponding $D - d$ eigenvalues are negative. See Ozertem and Erdogmus (2011)

This definition gives us only a condition to test whether a point x lies on the d-dimensional ridge. Given any x, it does not necessarily tell us how to pull it onto the d-dimensional ridge. There are $D - d$ eigenvectors of $H(x)$ that are orthogonal to the gradient $g(x)$, but it does not tell us the eigenvalues of $H(x)$ to which the eigenvectors correspond.

In some other works, such as Genovese et al. (2014), the ridge is defined more concretely by restriction of the specific eigenvectors. Given any density function $p(x)$, we can define the ridge as the subset of the domain that satisfies

$$R = \{x|\Pi^{\perp}(H(x))g(x) = 0, \lambda_{d+1}(H(x)) < 0\}, \tag{4}$$

where $H(x)$ is the Hessian matrix of $p(x)$, $g(x)$ is its gradient, and $\Pi^{\perp}$ is the projection onto the eigenspace spanned by the $D - d$ smallest eigenvalues.

At the sample level, given an empirical density function $\hat{p}(x)$ ($\hat{p}_h(x)$ or $\hat{p}_{r,h}(x)$), we will have the estimated ridge

$$\hat{R}_h = R(\hat{p}_h(x)) = \{x|\Pi^{\perp}(H_{\hat{p}_h}(x))g_{\hat{p}_h}(x) = 0, \lambda_{d+1}(H_{\hat{p}_h}(x)) < 0\} \quad \text{or}$$
$$\hat{R}_{r,h} = R(\hat{p}_{r,h}(x)) = \{x|\Pi^{\perp}(H_{\hat{p}_{r,h}}(x))g_{\hat{p}_{r,h}}(x) = 0, \lambda_{d+1}(H_{\hat{p}_{r,h}}(x)) < 0\},$$

respectively.

REMARK. It is worth noting that both the ridge $\hat{R}$ and the estimated ridges, such as $\hat{R}_h$ or $\hat{R}_{r,h}$ are continuous sets.

When $H(x)$ has exactly $D-d$ negative eigenvalues, these two definitions have no practical differences. However, when $H(x)$ has more than $D - d$ negative eigenvalues, then, from Definition 2.1, we have multiple choices to build the space to satisfy the ridge definition. Under this condition, the ridge in (4) is the only choice, and this ridge, corresponding to definition (4), is a subset of Definition 2.1.

From the definition, it can be observed that the ridge set R is closely related to the Hessian $H(x)$ and the gradient $g(x)$. The following section will show the connection between the distance between the ridges and the bias of the Hessian and gradient.

2.2. *Ridge Estimation and Error Measurement.* In practice, we use the ridge R as an approximation of the unknown manifold $\mathcal{M}$ numerically. However, the definition of a ridge in (4) is arguably too descriptive and unrealistic. We can only judge whether a point resides on the ridge, but the definition provides an explicit continuous function to describe the ridge. In reality, starting from a random sample given in advance, we will need an algorithm to push the data points onto the ridge so that the resultant points satisfy the ridge definition. To be specific, what we obtain is a set of modified discrete samples, $\hat{\mathcal{G}}$, satisfying the ridge definition. To this end, we may further specify the ridge estimator $\hat{\mathcal{G}}$ in (2) by

$$\hat{\mathcal{G}} = \text{Algorithm}(\{x_i, i = 1 : n\}, \hat{R})$$

where the estimated $\hat{R}$ can be $\hat{R}_h$ or $\hat{R}_{r,h}$. Note that the resultant $\hat{\mathcal{G}}$ relies on the initial data points $\{x_i\}$, the estimated ridge $\hat{R}$, and the actual implementation (algorithm) involved in determining how to process the data $\{x_i\}$ by moving them onto the defined ridge. Obviously, we know $\hat{\mathcal{G}} \subset \hat{R}$.

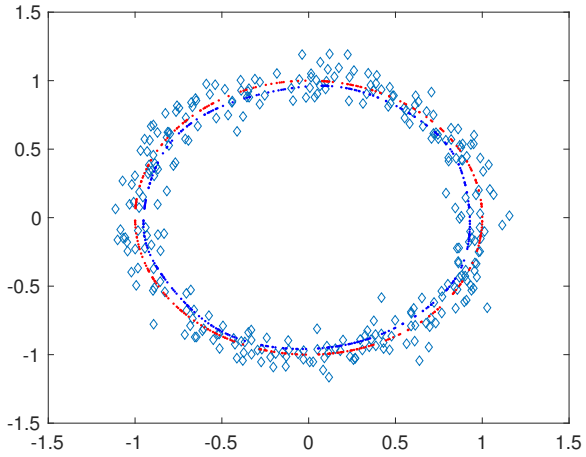

FIG 2. *Illustration of the $\hat{\mathcal{G}}, \mathcal{M}, \mathcal{M}_{\hat{\mathcal{G}}}$ and the Hausdorff and average margin using the example of a toy circle. The blue diamond markers represent the observation points. The blue dots represent the computed points that satisfy the ridge condition, i.e. $\hat{\mathcal{G}}$. The red dots represent the projection of $\hat{\mathcal{G}}$ onto $\mathcal{M}$, denoted by $\mathcal{M}_{\hat{\mathcal{G}}}$.*

Because $\hat{\mathcal{G}}$ is a discrete set and $\mathcal{M}$ is supposed to be continuous and smooth, the two sets have different cardinalities. Consider an extreme case where $\hat{\mathcal{G}}$ is a discrete subset of $\mathcal{M}$. Clearly, the largest distance from $\mathcal{M}$ to $\hat{\mathcal{G}}$ is $\sup_{x\in\mathcal{M}} \inf_{y\in\hat{\mathcal{G}}} \|x - y\|_2$, which is largely affected by the distribution of samples in $\hat{\mathcal{G}}$.

To diminish the random factor in the distance measurement between $\hat{\mathcal{G}}$ and $\mathcal{M}$, we extract a subset from $\mathcal{M}$, $\mathcal{M}_{\hat{\mathcal{G}}} = \pi_{\mathcal{M}}(\hat{\mathcal{G}})$, and consider the Hausdorff distance between $\hat{\mathcal{G}}$ and $\mathcal{M}_{\hat{\mathcal{G}}}$ (see Figure 2). Since the projection $\pi_{\mathcal{M}}(\hat{\mathcal{G}})$ from $\hat{\mathcal{G}}$ to $\mathcal{M}_{\hat{G}}$ is onto, the Hausdorff distance equals the quasi-Hausdorff distance

$$\text{Haus}(\hat{\mathcal{G}}, \mathcal{M}_{\hat{\mathcal{G}}}) = \max_{\hat{x}_k\in\hat{\mathcal{G}}} \min_{\tilde{y}_s\in\mathcal{M}_{\hat{\mathcal{G}}}} \|\hat{x}_k - \tilde{y}_s\|_2. \tag{5}$$

The Hausdorff distance is the largest margin between $\hat{\mathcal{G}}$ and $\mathcal{M}_{\hat{\mathcal{G}}}$. The largest margin is dominated by only one point in $\hat{\mathcal{G}}$, which cannot reflect the overall margin level. So we also

consider the average margin, defined as

$$\text{Marg}(\hat{\mathcal{G}}, \mathcal{M}_{\hat{\mathcal{G}}}) = \frac{1}{|\hat{\mathcal{G}}|} \sum_{\hat{x}_k \in \hat{\mathcal{G}}} \min_{\tilde{y}_s \in \mathcal{M}_{\hat{\mathcal{G}}}} \|\hat{x}_k - \tilde{y}_s\|_2, \tag{6}$$

where $|\hat{\mathcal{G}}|$ represents the number of samples in $\hat{\mathcal{G}}$. The distances, whether Hausdorff or average margin, are too complicated to bound or optimize directly. Because $\hat{\mathcal{G}} \subset \hat{R}$, we also need to connect $\mathcal{M}$ with some ridge under the following assumption:

CLAIM 1. *For any lower-dimensional manifold $\mathcal{M}$, there is a corresponding density function $p(x)$ such that*

1. *For the ridge $R(p(x))$ derived from $p(x)$: $R(p(x)) = \mathcal{M}$.*
2. *The projection of $\hat{G}$ onto $\mathcal{M}$ satisfies $\mathcal{M}_{\hat{G}} \subset R(p(x))$.*
3. *The distance between $R(p(x))$ and $\hat{R}(x)$ is comparable, where $\hat{R}(x)$ is a ridge derived from some empirical functions.*

REMARK. Since $\mathcal{M}$ can also be considered as a ridge of some density function $p(x)$, the problem of estimating the distance between $\mathcal{M}_{\hat{\mathcal{G}}}$ and $\hat{\mathcal{G}}$ can be converted to one of estimating the distance between two ridges. This assumption can be verified by selecting an appropriate form of $p(x)$, for instance by restricting $p(x)$ to nonzero only on $\mathcal{M}$.

Next, we connect the distance between any pair of ridges R_1, R_2 with the analytic quantity (such as gradient and Hessian) of their corresponding density functions $p_1(x), p_2(x)$. Specifically, the distance between R_1 and R_2 is of the same order as the error of the corresponding derivatives.

LEMMA 2.2. *For any R_1, R_2, and any point $x_1 \in R_1$, the pairwise distance from x_1 to R_2 yields the order of:*

$$\min_{x_2 \in R_2} \|x_1 - x_2\|_2 = O(\|H_1(x_1) - H_2(x_1)\|_F + \|g_1(x_1) - g_2(x_1)\|_2)$$

where $H_1(x_1), g_1(x_1)$ are the Hessian and gradient of some estimated density function $p_1(x_1)$ evaluated at x_1; $H_2(x_1)$ and $g_2(x_1)$ are the Hessian and gradient of the density function of $p_2(x)$ evaluated at x_1, respectively.

The proof of Lemma 2.2 can be found in the supplementary material.

REMARK. Lemma 2.2 builds the bridge between the geometric quantities and the analytic differential quantities. Through this lemma, we can easily convert the problem of searching for a 'closer' ridge to the problem of searching for a 'better' approximation of the Hessian and gradient. As illustrated in Figure 3, this conversion arguably give our work a stronger foundation and greater versatility.

Using Claim 1 and Lemma 2.2, we have, for any $\hat{x}_k \in \hat{\mathcal{G}}$, a $\tilde{y} \in \mathcal{M}$ that is the end point of the unit speed curve $\gamma(t)$ starting from $\hat{x}_k$. Meanwhile, both $\hat{x}_k$ and $\tilde{y}$ satisfy the ridge definition in (4). Thus, we know that

$$\min_{\tilde{y}_s \in \mathcal{M}_{\hat{\mathcal{G}}}} \|\hat{x}_k - \tilde{y}_s\|_2 \leq \|\hat{x}_k - \tilde{y}\|_2 = O(\|H_{\hat{p}}(\hat{x}_k) - H(\hat{x}_k)\|_F + \|g_{\hat{p}}(\hat{x}_k) - g(\hat{x}_k)\|_2). \tag{7}$$

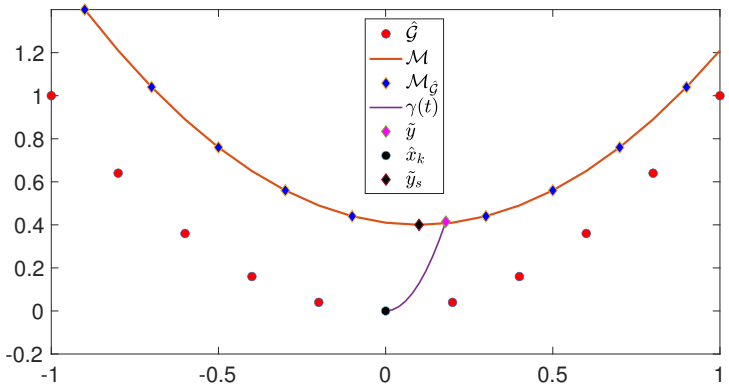

FIG 3. *Illustration of the relationship between* $\hat{\mathcal{G}}, \mathcal{M}, \mathcal{M}_{\hat{\mathcal{G}}}, \gamma(t), \hat{x}_k, \tilde{y}, \tilde{y}_s$

REMARK. The distance between two discrete sets $\hat{\mathcal{G}}, \mathcal{M}_{\hat{\mathcal{G}}}$ is converted to the distance between $\hat{\mathcal{G}}$ and $\mathcal{M}$ by the unit speed curve $\gamma(t)$. $\gamma(t)$ starts from a point in $\hat{\mathcal{G}}$ (such as $\hat{x}_k$) with a direction of $\Pi_{H_p}(\gamma(t))g_p(\gamma(t))$. The point at which $\gamma(t)$ goes through $\mathcal{M}$ is denoted by $\tilde{y}$. The details of the construction method of the curve $\gamma(t)$ can be found in the supplementary material.

From (5), by maximizing the distance with respect to $x_k \in \hat{\mathcal{G}}$ in (7), we know that $\mathrm{Haus}(\hat{\mathcal{G}}, \mathcal{M}_{\hat{\mathcal{G}}})$ yields a uniform upper bound by the sup norm at the right side:

$$\mathrm{Haus}(\hat{\mathcal{G}}, \mathcal{M}_{\hat{\mathcal{G}}}) = O(\sup_{\hat{x}\in\hat{\mathcal{G}}}(\|H_{\hat{p}}(\hat{x}) - H(\hat{x})\|_F + \|g_{\hat{p}}(\hat{x}) - g(\hat{x})\|_2)), \tag{8}$$

where $\hat{p}$ is the estimated density, which can be chosen as $\hat{p}_h(x)$, $\hat{p}_{r,h}(x)$, or any other empirical forms.

EXAMPLE. When we approximate $\hat{p}(x), g_{\hat{p}}(x), H_{\hat{p}}(x)$ using the kernel-density function as $\hat{p}_h(x), g_{\hat{p}_h}(x), H_{\hat{p}_h}(x)$, the pairwise error for the derivatives yields the following result:

$$(|g_{\hat{p}_h}(x) - g_p(x)|)_i = O(h^2) + O_p(\sqrt{\frac{1}{nh^{D+2}}}),$$

$$(|H_{\hat{p}_h}(x) - H_p(x)|)_{ij} = O(h^2) + O_p(\sqrt{\frac{1}{nh^{D+4}}}).$$

The empirical-process theory gives the uniform error as

$$\sup_x(|g_{\hat{p}_h}(x) - g_p(x)|)_i = O(h^2) + O_p(\sqrt{\frac{\log n}{nh^{D+2}}}),$$

$$\sup_x(|H_{\hat{p}_h}(x) - H_p(x)|)_{ij} = O(h^2) + O_p(\sqrt{\frac{\log n}{nh^{D+4}}}).$$

The error rates for the pairwise and uniform differ only in the stochastic term, with a scale of $\sqrt{\log n}$. See Chen (2017); Arias-Castro, Mason and Pelletier (2016); Genovese et al. (2014).

Unlike the upper bound of the Hausdorff distance shown in (8), the average margin $\mathrm{Margin}(\hat{\mathcal{G}}, M_{\hat{\mathcal{G}}})$ is closely related to the point-wise approximation error of $O(\|H_{\hat{p}_h}(x) -$

$H_p(x)\|_F + \|g_{\hat{p}_h}(x) - g_p(x)\|_2)$. To reduce the average margin, we should focus on finding a new density function to improve the point-wise error with respect to the gradient and Hessian, and hope that this improvement of the error will lead to good performance of the estimated ridge.

To our knowledge, the most widely used nonparametric approach to approximate $p(x)$ is the KDE method. Therefore, we will review the classical KDE approach to approximate a density function, and derive the details of the corresponding derivatives' approximation error.

2.3. *Derivatives for KDE.* The derivatives' bias generated from the kernel-density function is highly related to the distance between the ridges constructed by $p(x)$ and $\hat{p}_h(x)$, respectively. We introduce the bias of the first-order and second-order derivatives when using the kernel-density function to approximate the real density function.

2.3.1. *Derivative Bias.* To derive the bias of the first-order and second-order derivatives, we need to repeatedly implement the integration by the substitution method on the multiple-variable function. For the sake of completeness, we first recall some results for the bias and variance with respect to the derivatives. Subsequently, we show that using $\hat{p}_{r,h}(x)$ can improve the bias and variance results more than using $\hat{p}_h(x)$.

THEOREM 2.3. *The bias of the first order and second order of the $\hat{p}_h(x)$ is*

$$|\mathrm{E}(\partial_{x_s}\hat{p}_h(x)) - \partial_{x_s}p(x)| = \frac{h^2|\Delta(\partial_{x_s}p(x))|}{2D}\int \|u\|_2^2 K(u)du + o(h^2),$$

$$|\mathrm{E}(\partial_{x_s}\partial_{x_t}\hat{p}_h(x)) - \partial_{x_s}\partial_{x_t}p(x)| = \frac{h^2|\Delta(\partial_{x_s}\partial_{x_t}p(x))|}{2D}\int \|u\|_2^2 K(u)du + o(h^2).$$

where Δ is the Laplace-Beltrami operator.

The details of the procedure for proof can be found in the supplementary material.

REMARK. The Laplace-Beltrami operator is the divergence of the gradient. For any $f(x)$, $\Delta f(x)$ is the summation of the diagonal elements of $H_f(x)$ (the Hessian of $f(x)$), which is also the summation of the eigenvalues of $H_f(x)$.

REMARK. It should be noted that in (2.3) the square term h^2 originates from the fact that the positive kernel is of order 2, which is fixed and cannot be improved. However, we can make $\int \|u\|_2^2 K(u)du$ much smaller by making a small modification-*truncating the kernel*-as discussed below.

2.3.2. *Derivative Variance.* After we get the bias of the derivatives, in order to gain a better understanding of the stochastic property of the derivatives, we should also consider the variance. In this section, we give the variance result of the first and second derivatives of the KDE under the i.i.d. samples assumption.

Let $\phi_s(x)$ and $\phi_{s,t}(x)$ be

$$\phi_s(x) = p(x)\int (\partial_{u_s}K(u))^2 du, \quad \phi_{s,t}(x) = p(x)\int (\partial_{u_s}\partial_{u_t}K(u))^2 du$$

THEOREM 2.4. *The variance of the first- and second-order derivatives for $\hat{p}_h(x)$ has the following bound:*

$$\mathrm{E}|\partial_{x_s}\hat{p}_h(x) - \mathrm{E}(\partial_{x_s}\hat{p}_h(x))| = \sqrt{\frac{\phi_s(x)}{nh^{D+2}}} + O(\frac{1}{n^{1/2}h^{(D+1)/2}}),$$

$$\mathrm{E}|\partial_{x_s}\partial_{x_t}\hat{p}_h(x) - \mathrm{E}(\partial_{x_s}\partial_{x_t}\hat{p}_h(x))| = \sqrt{\frac{\phi_{s,t}(x)}{nh^{D+4}}} + O(\frac{1}{n^{1/2}h^{(D+3)/2}}).$$

The details of the procedure for proof can be found in the supplementary material.

2.3.3. *Optimal-Parameter Dilemma.* The bandwidth parameter h controls the smoothness of the density function and its derivatives. We can derive the best possible h to obtain the best approximation error for the kernel-density function. However, we cannot select a single optimum h to be the optimum solution for the multiple objectives, such as the first-order and second-order derivatives.

Using the triangle inequality of absolute function, we derive the upper bound for $|\partial_{x_s}\hat{p}_h(x) - \partial_{x_s}p(x)|$ by

$$|\partial_{x_s}\hat{p}_h(x) - \partial_{x_s}p(x)| \leq |\mathrm{E}(\partial_{x_s}\hat{p}_h(x)) - \partial_{x_s}p(x)| + |\mathrm{E}(\partial_{x_s}\hat{p}_h(x)) - \partial_{x_s}\hat{p}_h(x)|.$$

The first term $|\mathrm{E}(\partial_{x_s}\hat{p}_h(x)) - \partial_{x_s}p(x)|$ is deterministic and the second term $|\mathrm{E}(\partial_{x_s}\hat{p}_h(x)) - \partial_{x_s}\hat{p}_h(x)|$ is a random variable. From (2.3) and (2.4), we know the upper bound of $|\partial_{x_s}\hat{p}_h(x) - \partial_{x_s}p(x)|$ is

$$|\partial_{x_s}\hat{p}_h(x) - \partial_{x_s}p(x)| = \frac{h^2|\Delta(\partial_{x_s}p(x))|}{2D}\int \|u\|_2^2 K(u)du + O_p(\sqrt{\frac{\phi_s(x)}{nh^{D+2}}}).$$

Clearly, for the first-order partial derivative, the optimum h should minimize the objective $C_1(x)h^2 + C_2(x)\frac{1}{\sqrt{nh^{D+2}}}$. Thus, $h = O(n^{-\frac{1}{D+6}})$.

$$|\partial_{x_s}\partial_{x_t}\hat{p}_h(x) - \partial_{x_s}\partial_{x_t}p(x)| = \frac{h^2|\Delta(\partial_{x_s}\partial_{x_t}p(x))|}{2D}\int \|u\|_2^2 K(u)du + O_p(\sqrt{\frac{\phi_{s,t}(x)}{nh^{D+4}}}).$$

In the same way, for the second-order partial derivative, the optimum h should minimize the objective $C_3(x)h^2 + C_4(x)\frac{1}{\sqrt{nh^{D+4}}}$ as well. Thus, $h = O(n^{-\frac{1}{D+8}})$. Clearly, we cannot minimize the first and second derivative approximations at the same time.

Next, we bring in another parameter, r, to restrict our objective function such that we can construct the *l*-SCRE by using only the samples with better quality in a surrounding neighborhood of x.

3. *l*-SCRE. In this section, we propose the estimator derived from a local kernel-density function, which is called *l*-SCRE. We consider a type of generalized, locally defined kernel, $K_r(\frac{x-x_i}{h})$, controlled by two parameters r and h, h being the bandwidth that controls the distribution of each of the samples x_i in a soft manner, and r determining which partition of the samples should be used in our kernel-density function.

$$\hat{p}_{r,h}(x) = \frac{1}{nh^D}\sum_i K(\frac{x-x_i}{h})I(\|x-x_i\|_2 \leq r), \tag{9}$$

where $I(\cdot)$ is the indicator function, which equals 1 when the condition is met, and 0 otherwise. The necessity of bringing in the kernel with two parameters is a result of the fact that, in some cases, we will fall into the dilemma of having to use a large h to get a smoother

estimation, while, at the same time, needing a smaller h to keep a low bias for the estimated density function and even the derivatives.

It should be noted that in (9) we consider only the samples in the ball with a radius of r; this can also be regarded as a truncated kernel function using the following definition:

$$K_r(\frac{x-x_i}{h}) = K(\frac{x-x_i}{h})I(\|x-x_i\|_2 \leq r). \tag{10}$$

Note that $K_r(\frac{x-x_i}{h})$ is not differentiable where $\|x-x_i\|_2 = r$. From (10), we also know that for x such that $\|x-x_i\| < r$,

$$|\frac{\partial}{\partial x_s}K_r(\frac{x-x_i}{h})| \leq |\frac{\partial}{\partial x_s}K(\frac{x-x_i}{h})|. \tag{11}$$

Clearly, $K(u)$ satisfies $\int K(u)du = 1$. When necessary, we can also renormalize $K_r(u)$ by a constant value $c > 1$, such that

$$\int cK_r(u)du = c\int_{\|u\|\leq r/h} K(u)du = 1.$$

When selecting $K(u)$ as the exponential decreasing function $K(u) = \frac{1}{(2\pi)^{d/2}}\exp(-\frac{1}{2}\|u\|_2^2)$, the error for a non-normalized $K_r(u)$ will be small. Using the Gaussian integral formula, we know

$$\int_{\|u\|\leq r/h} K(u)du = 1 - O(\exp(-(r^2/h^2))). \tag{12}$$

When r is larger than h, the residue part $O(\exp(-(r^2/h^2)))$ will become ignorable.

3.1. *Parameter Setup.* Suppose the true density function $p(x)$ is a constant on $\mathcal{M}$ and the observations $\{x_i\}$ are evenly sampled from $p(x)$. However, the shape of the manifold $\mathcal{M}$ can be very complicated, for instance a twisted two-dimensional surface embedded in the three-dimensional ambient space. In this case, the kernel-density estimated function $\hat{p}_h(x)$ will have the following characteristics:

1. If the manifold in the neighborhood of $\mathcal{D}_r(x)$ is complicated, $\hat{p}_h(x)$ will be large because the neighborhood of x will contain more observations.
2. If the manifold in the neighborhood of $\mathcal{D}_r(x)$ is relatively simple, for instance a flat surface, $\hat{p}_h(x)$ will be small because the neighborhood of x will contain far fewer observations.

Recall that n is the total number of observations. The expected number of observations in $\mathcal{N}(x_0, r)$ is

$$\mathrm{E}(\sum_i I(x_i \in \mathcal{N}(x_0, r))) = n\int_{u\in\mathcal{M}\cap\mathcal{D}_r(x_0)} p(u)du.$$

Based on the above analysis, we know that a larger value of $\hat{p}(x)$ implies a complicated structure of $\mathcal{M}$ in the neighborhood of x, and a smaller value of $\hat{p}(x)$ implies a simple structure of $\mathcal{M}$ in the neighborhood of x. For the complicated area of $\mathcal{D}_r(x) \cap \mathcal{M}$, we should focus on an area $\mathcal{D}_r(x)$ with a small radius r. For the simple area of $\mathcal{D}_r(x) \cap \mathcal{M}$, we can enlarge r, which could speed up our fitting process.

In this way, we can select r adaptively and think of it as a function of the true density function $p(x)$; this is denoted by $r(x)$. We determine $r(x)$ by requiring the accumulating density function

$$\phi(x, r) = \int_{y\in\mathcal{M}\cap\mathcal{D}_r(x)} p(y)dy,$$

to be a constant θ. The implicit function $\phi(x,r)=\theta$ can determine the function $r(x)$. Note that we can approximate the accumulating density function $\phi(x,r)$ with the samples:

$$\phi(x,r)\approx\frac{1}{n}\sum_{k}I(x_k\in\mathcal{N}(x,r)).$$

In real computational cases, $r(x)$ can be chosen adaptively as the smallest radius that still fulfils the condition that the set $\{x_i|\|x-x_i\|\leq r\}$ contain exactly m elements, where m is a predefined parameter.

As with the bias for the derivative of $\hat{p}_h(x)$, we also have the bias for the derivative $\hat{p}_{r,h}(x)$. It is easy to show that, under some mild conditions, the bias of the derivatives of $\hat{p}_{r,h}(x)$ is smaller than that of $\hat{p}_h(x)$.

3.2. *Upper Bound of the Bias.* In the previous section, we derived the upper bound of the bias for the classical $\hat{p}_h(x)$. Here, we present the bias result of $\hat{p}_{r,h}(x)$, and show that, under some mild conditions, the bias of the derivatives $\hat{p}_{r,h}(x)$ is smaller than that of $\hat{p}_h(x)$.

Given any parameter r,h and a density function $p(x)$, we can define two functions $B_s(x|r,h,p)$ and $B_{s,t}(x|r,h,p)$ as

$$B_s(x|r,h,p)=\frac{h^2|\Delta(\partial_{x_s}p(x))|}{2D}\int_{\|u\|\leq r/h}\|u\|_2^2K(u)du+|\partial_{x_s}p(x)|\int_{\|u\|\geq r/h}K(u)du$$

$$B_{s,t}(x|r,h,p)=\frac{h^2|\Delta(\partial_{x_s}\partial_{x_t}p(x))|}{2D}\int_{\|u\|\leq r/h}\|u\|_2^2K(u)du+|\partial_{x_s}\partial_{x_t}p(x)|\int_{\|u\|\geq r/h}K(u)du$$

REMARK. From the definition of $B_s(x|r,h,p)$ and $B_{s,t}(x|r,h,p)$, if $\|u\|_2>1$, the first term will become a major part because of the existence of the term $\int_{\|u\|\leq r/h}\|u\|_2^2K(u)du$. Taking the limit for $r\to+\infty$, we will of course have

$$\lim_{\|r\|\to+\infty}B_s(x|r,h,p)=\frac{h^2|\Delta(\partial_{x_s}p(x))|}{2D}\int\|u\|_2^2K(u)du, \tag{13}$$

where the right side is the bias of the first derivative for $\hat{p}_h(x)$.

LEMMA 3.1. *For the derivatives of $\hat{p}_{r,h}(x)$, we have the bias relationship for first- and second-order derivatives as*

$$|\mathrm{E}(\partial_{x_s}\hat{p}_{r,h}(x))-\partial_{x_s}p(x)|\leq B_s(x|r,h,p),$$
$$|\mathrm{E}(\partial_{x_s}\partial_{x_t}\hat{p}_{r,h}(x))-\partial_{x_s}\partial_{x_t}p(x)|\leq B_{s,t}(x|r,h,p).$$

Furthermore, if

$$r\geq\max\{h,\sqrt{\frac{2|\partial_{x_s}p(x)|}{|\Delta(\partial_{x_s}p(x))|}},\sqrt{\frac{2|\partial_{x_s}\partial_{x_t}p(x)|}{|\Delta(\partial_{x_s}\partial_{x_t}p(x))|}}\}, \tag{14}$$

the bound of the pairwise derivatives' bias for $\hat{p}_{r,h}(x)$ will be bounded by that of $\hat{p}_h(x)$

$$|\mathrm{E}(\partial_{x_s}\hat{p}_{r,h}(x))-\partial_{x_s}p(x)|\leq|\mathrm{E}(\partial_{x_s}\hat{p}_h(x))-\partial_{x_s}p(x)|, \tag{15}$$

$$|\mathrm{E}(\partial_{x_s}\partial_{x_t}\hat{p}_{r,h}(x))-\partial_{x_s}\partial_{x_t}p(x)|\leq|\mathrm{E}(\partial_{x_s}\partial_{x_t}\hat{p}_h(x))-\partial_{x_s}\partial_{x_t}p(x)|. \tag{16}$$

The proof can be found in the supplementary material.

REMARK. To obtain (15) and (16), we show that, under the conditions of (14), the upper bounds $B_s(x|r,h,p)$ and $B_{s,t}(x|r,h,p)$ are less than the pairwise biases $|\mathrm{E}(\partial_{x_s}\hat{p}_h(x)) - \partial_{x_s}p(x)|$ and $|\mathrm{E}(\partial_{x_s}\partial_{x_t}\hat{p}_h(x)) - \partial_{x_s}\partial_{x_t}p(x)|$, respectively. Thus, we have

$$|\mathrm{E}(\partial_{x_s}\hat{p}_{r,h}(x)) - \partial_{x_s}p(x)| \leq B_s(x|r,h,p) \leq |\mathrm{E}(\partial_{x_s}\hat{p}_h(x)) - \partial_{x_s}p(x)| \tag{17}$$

The procedure for obtaining the result for $|\mathrm{E}(\partial_{x_s}\partial_{x_t}\hat{p}_{r,h}(x)) - \partial_{x_s}\partial_{x_t}p(x)|$ is similar.

3.3. *Ridge-Improvement Justification.* We are interested in the ratio of the bias improvement. From (17), we have

$$\frac{|\mathrm{E}(\partial_{x_s}\hat{p}_{r,h}(x)) - \partial_{x_s}p(x)|}{|\mathrm{E}(\partial_{x_s}\hat{p}_h(x)) - \partial_{x_s}p(x)|} \leq \frac{B_s(x|r,h,p)}{|\mathrm{E}(\partial_{x_s}\hat{p}_h(x)) - \partial_{x_s}p(x)|}$$

For notational convenience, we bring in three notations, $\gamma, \mu_s(x,r,h), \mu_{s,t}(x,r,h)$, to simplify the result in Lemma 3.1.

$$\gamma(r,h) = \frac{\int_{\|u\|\geq r/h} \|u\|_2^2 K(u)du}{2\int_{\|u\|\geq r/h} K(u)du}, \quad \mu_s(x,r,h) = \frac{B_s(x|r,h,p)}{|\mathrm{E}(\partial_{x_s}\hat{p}_h(x)) - \partial_{x_s}p(x)|},$$

$$\mu_{s,t}(x,r,h) = \frac{B_{s,t}(x|r,h,p)}{|\mathrm{E}(\partial_{x_s}\partial_{x_t}\hat{p}_h(x)) - \partial_{x_s}\partial_{x_t}p(x)|}$$

REMARK. Clearly, there exists $r_0 = h$ such that, when $r > r_0(\|u\|_2 > 1)$, we have $\|u\|_2^2 K(u) \geq K(u)$, which implies $\gamma(r_0,h) \geq 1/2$.

REMARK. The observation that $K(u)$'s value approaches 0 as $\|u\|_2$ increases is important for our choice of r. Because of (13), we have the following:

$$\lim_{r\to\infty} \mu_s(x,r,h) = 1, \lim_{r\to\infty} \mu_{s,t}(x,r,h) = 1.$$

The scalars $\mu_s(x,r,h)$ and $\mu_{s,t}(x,r,h)$ decrease as r increases. Thus, we need to choose a relatively smaller r_0 to ensure that *l*-SCRE is more effective than SCRE.

REMARK. For any $x \in \mathbb{R}^D$, there is $r_0(x,h,s)$ (depending on x, h, and s), such that, when $r > r_0(x,h,s)$, the following inequality holds:

$$2D|\partial_{x_s}p(x)| \int_{\|u\|\geq r/h} K(u)du < |\Delta(\partial_{x_s}p(x))|h^2 \int_{\|u\|\geq r/h} \|u\|_2^2 K(u)du,$$

Similarly, for any $x \in \mathbb{R}^D$, there is $r_0(x,h,s,t)$ (depending on x, h, s, t), such that, when $r > r_0(x,h,s,t)$, the following inequality holds:

$$2D|\partial_{x_s}\partial_{x_t}p(x)| \int_{\|u\|\geq r/h} K(u)du < |\Delta(\partial_{x_s}\partial_{x_t}p(x))|h^2 \int_{\|u\|\geq r/h} \|u\|_2^2 K(u)du$$

Using the above two inequalities and the form of $\mu_s(x,r,h), \mu_{s,t}(x,r,h)$, we know that

$$r_0(x,h) = \max\{\max_s r_0(x,h,s), \max_{s,t} r_0(x,h,s,t)\},$$

such that, when $r \geq r_0(x,h)$, for any s,t, we have

$$0 < \mu_s(x,r,h) < 1, \quad 0 < \mu_{s,t}(x,r,h) < 1.$$

Substituting $\mu_s(x,r,h), \mu_{s,t}(x,r,h)$ into the results in Lemma 3.1 will naturally lead to Lemma 3.2.

LEMMA 3.2. *For any $x \in \mathbb{R}^D$, there is $r_0(x,h)$ (depending on x, h), such that, when $r > r_0(x,h)$, for any s,t, the following two inequalities hold:*

$$|\mathrm{E}(\partial_{x_s}\hat{p}_{r,h}(x)) - \partial_{x_s}p(x)| \leq \frac{\mu_s(x,r,h)h^2|\Delta(\partial_{x_s}p(x))|}{2D}\int \|u\|_2^2 K(u)du,$$

$$|\mathrm{E}(\partial_{x_s}\partial_{x_t}\hat{p}_{r,h}(x)) - \partial_{x_s}\partial_{x_t}p(x)| \leq \frac{\mu_{s,t}(x,r,h)h^2|\Delta(\partial_{x_s}\partial_{x_t}p(x))|}{2D}\int \|u\|_2^2 K(u)du,$$

where the parameters are $\mu_s(x,r,h) \in (0,1), \mu_{s,t}(x,r,h) \in (0,1)$.

REMARK. From comparing the above results with (2.3) and (2.3), we can conclude that the bias for the derivatives of $\hat{p}_{r,h}(x)$ is reduced with a scale of $\mu_s(x,r,h) \in (0,1)$ and $\mu_{s,t}(x,r,h) \in (0,1)$. In the following subsection, we determine that the derivative's variance of $\hat{p}_{r,h}(x)$ is also bounded by that of $\hat{p}_h(x)$.

3.3.1. *Variance of Upper Bound.* Recall that the variance of the partial derivative is

$$\mathrm{Var}(\partial_{x_s}\hat{p}_h(x)) = \frac{1}{nh^{D+2}}(p(x)\int(\partial_{u_s}K(u))^2du + O(h)).$$

Here, we can show that the variance of the derivative of $\hat{p}_{r,h}(x)$ also has a similar upper bound to that of $\hat{p}_h(x)$.

THEOREM 3.3. *The variance of derivative of $\hat{p}_{r,h}(x)$ is controlled by*

$$\mathrm{Var}(\partial_{x_s}\hat{p}_{r,h}(x)) \leq \frac{1}{nh^{D+2}}(p(x)\int(\partial_{u_s}K(u))^2du + O(h)).$$

The proof can be found in the supplementary material. By combining the bias and variance, we have the overall approximation for the derivative of $\hat{p}_{r,h}(x)$.

3.3.2. *Estimation for $\hat{p}_{r,h}(x)$'s Derivatives.* Using the analysis of bias and variance and Chebyshev's inequality, we obtain the overall approximation error for the derivatives of $\hat{p}_{r,h}(x)$ satisfying

$$\begin{aligned}
&|\partial_{x_s}\hat{p}_{r,h}(x) - \partial_{x_s}p(x)| \\
\leq&|\partial_{x_s}p(x) - \mathrm{E}(\partial_{x_s}\hat{p}_{r,h}(x))| + (|\partial_{x_s}\hat{p}_{r,h}(x) - \mathrm{E}(\partial_{x_s}\hat{p}_{r,h}(x))|) \\
\leq&\mu_s(x,r,h)(\frac{h^2|\Delta(\partial_{x_s}p(x))|}{2D}\int \|u\|_2^2 K(u)du) + O_p(\sqrt{\frac{\phi_s(x)}{nh^{D+2}}}).
\end{aligned} \tag{18}$$

Similarly, for the second-order differential form, we have

$$\begin{aligned}
&|\partial_{x_s}\partial_{x_t}\hat{p}_{r,h}(x) - \partial_{x_s}\partial_{x_t}p(x)| \\
\leq&|\partial_{x_s}\partial_{x_t}p(x) - \mathrm{E}(\partial_{x_s}\partial_{x_t}\hat{p}_{r,h}(x))| + |\partial_{x_s}\partial_{x_t}\hat{p}_{r,h}(x) - \mathrm{E}(\partial_{x_s}\partial_{x_t}\hat{p}_{r,h}(x))| \\
\leq&\mu_{s,t}(x,r,h)(\frac{h^2|\Delta(\partial_{x_s}\partial_{x_t}p(x))|}{2D}\int \|u\|_2^2 K(u)du) + O_p(\sqrt{\frac{\phi_{s,t}(x)}{nh^{D+4}}}).
\end{aligned} \tag{19}$$

REMARK. To derive an optimum with respect to h, we can abbreviate the parts in (18), (19) that are not related to h. For example, in Lemma 3.4, we abbreviate $\frac{\Delta(\partial_{x_s}p(x))}{2D}$ as a_1 and $\sqrt{\phi_s(x)}$ as a_2.

3.4. *Ridge-Estimator Analysis.* Based on the relatively smaller approximation error, we can expect the ridge obtained from *l*-SCRE to be at a smaller distance from the underlying manifold.

From Lemma 3.2, we know for any x there is $r_0(x,h)$ such that, when $r \geq r_0(x,h)$, we have $\mu_s(x,r,h) \in (0,1), \mu_{s,t}(x,r,h) \in (0,1)$ for any s,t. To derive a uniform bound for distance in pairwise point measurement, we let the partial uniform supremum be denoted by

$$\mu(r,h) = \sup_x \max\{\max_s \mu_s(x,r,h), \max_{s,t} \mu_{s,t}(x,r,h)\}.$$

Corresponding to $\mu(r,h)$, there is $r_0(h)$ such that, when we have $r > r_0(h)$, we will still get $\mu(r,h) \in (0,1)$. To get a further result, a stronger assumption may require $\mu(r,h)$ to be upper-bounded by some number in $(0,1)$ when we choose r properly.

Since the form of $\mu(r,h)$ is complicated, it is not easy to get an optimum of h from an objective with a term related to $\mu(r,h)$. To simplify our analysis, we make the following assumption:

LEMMA 3.4. *We have two functions* $\nu(h) = a_0 h^2 + a_1\sqrt{\frac{1}{nh^{D+m}}}$ *and* $\nu_\ell(h) = \ell a_0 h^2 + a_1\sqrt{\frac{1}{nh^{D+m}}}$ *with* $m = 2, 4, \ell \in (0,1]$. *Then, the optimal minimums of them have the following relationship:* $\min_h \nu_\ell(h) = \ell^{\frac{D+2}{D+6}} \min_h \nu(h)$.

REMARK. The point-wise and uniform derivatives' variance differs with a term $\sqrt{\log n}$. This term can also be taken into consideration by choosing a different form of a_1. In other words, when the Hausdorff distance is anything other than the average margin, we can assume $a_1 = \bar{a}_1 * \sqrt{\log n}$, where $\bar{a}_1$ is the parameter considered for the average-margin measurement.

REMARK. From Lemma 3.4, we know that by reducing the bias with a scalar of ℓ we will have the optimal minimum solution reduced with a scalar of $\ell^{\frac{D+m}{D+m+4}}$. It should be noted that, with the increase in the dimension D, we will obtain a smaller $\ell^{\frac{D+m}{D+m+4}}$. Lemma 3.4 proves that we can improve the distance between the ridges by reducing the derivatives' bias corresponding to the density function. As before, the details can be found in the supplementary material.

From Lemma 3.4, we know that, by replacing the SCRE with *l*-SCRE, the upper bound of the margin or Hausdorff of an optimal ridge will be reduced by at least a scalar $\ell^{\frac{D+2}{D+6}}$.

THEOREM 3.5. *If the optimal* h *is chosen with respect to the sample size* n*, and then the upper bound with respect to the average margin or Hausdorff distance of the ridge obtained from l-SCRE will be reduced by at least a scalar* $\ell^{\frac{D+2}{D+6}}$*, i.e.*

$$\mathrm{Marg}(\hat{\mathcal{G}}_{l-\mathrm{SCRE}}, \mathcal{M}_{\hat{\mathcal{G}}_{l-\mathrm{SCRE}}}) \leq \ell^{\frac{D+2}{D+6}} \mathrm{Marg}(\hat{\mathcal{G}}_{\mathrm{SCRE}}, \mathcal{M}_{\hat{\mathcal{G}}_{\mathrm{SCRE}}}).$$

$$\mathrm{Haus}(\hat{\mathcal{G}}_{l-\mathrm{SCRE}}, \mathcal{M}_{\hat{\mathcal{G}}_{l-\mathrm{SCRE}}}) \leq \ell^{\frac{D+2}{D+6}} \mathrm{Haus}(\hat{\mathcal{G}}_{\mathrm{SCRE}}, \mathcal{M}_{\hat{\mathcal{G}}_{\mathrm{SCRE}}}),$$

where $\ell \in (0,1]$. *Furthermore, if* r *is selected as* $r = r_1$ *such that* $\ell = \sup_h \mu(r_1,h) < 1$*, the factor will satisfy* $\ell^{\frac{D+2}{D+6}} < 1$ *strictly.*

REMARK. From the proof of Lemma 3.4, we can see the optimal $h = O(n^{-\frac{1}{D+8}})$. When increasing the samples n to infinity, the optimal h will tend to be close to or equal to 0. With the optimal h chosen, the number of effective samples around x equals $O(n^{\frac{D+7}{D+8}})$.

REMARK. As the dimension D increases, the scalar $\ell^{\frac{D+2}{D+6}}$ will become smaller and smaller. This asymptotical behavior shows that, as the dimension of the ambient space D increases, the effect of l-SCRE will become more and more apparent compared with SCRE.

REMARK. As the radius r_1 increases to $+\infty$, the distinction between the performances of l-SCRE and SCRE will gradually disappear. In other words, l-SCRE will gradually become SCRE when $r_1 \to +\infty$. Thus, SCRE is a special case of l-SCRE.

3.5. *Confidence Regions for $\mathcal{M}$.* Here, we give the confidence region for $\mathcal{M}$. $x \in \mathcal{M}$ satisfies $V^T(H(x))g(x) = 0$, where V is a orthonormal matrix of shape $D \times (D-d)$, and each of its columns is the eigenvector corresponding to the $d+1$ to D largest eigenvalues of $H(x)$. Thus, we write $V(H(x))$ to indicate that V is a function depending on $H(x)$.

Meanwhile, if we can approximate $V^T(\hat{H}(x)) - V^T(H(x))$ by a first-order approximation with $\mathrm{vech}(\hat{H}(x) - H(x))$, i.e.

$$V^T(\hat{H}(x)) - V^T(H(x)) \approx M(x)\mathrm{vech}(\hat{H}(x) - H(x)), \tag{20}$$

multiplying by $\hat{g}(x)$ on both sides in (20) leads to

$$V^T(\hat{H}(x))\hat{g}(x) - V^T(H(x))\hat{g}(x) \approx M(x)\mathrm{vech}(\hat{H}(x) - H(x))\hat{g}(x).$$

Note that $V^T(H(x))\hat{g}(x) = V^T(H(x))(\hat{g}(x) - g(x))$, which is of the same order as $\hat{g}(x) - g(x)$.

Based on the above analysis, when $h \to 0$, $M(x)\mathrm{vech}(\hat{H}(x) - H(x))\hat{g}(x)$ can be regarded as a major part of $V^T(\hat{H}(x))\hat{g}(x)$. Let the variance matrix of $M(x)\mathrm{vech}(\hat{H}(x) - H(x))\hat{g}(x)$ be denoted by $\hat{\Gamma}$. Using the eigenvalue decomposition, we have $\hat{Q}(x)$ such that $\hat{Q}(x)\hat{Q}(x)^T = \hat{\Gamma}^{-1}$. Using $\hat{Q}(x)$, we can define a region in a form similar to that in Qiao (2020):

$$\hat{C}_{r,h}(a_n, b_n) = \{x : \sqrt{nd^{D+4}}\|\hat{Q}(x)V^T(\hat{H}(x))\hat{g}(x)\| \le a_n, \lambda_{d+1}(\hat{H}(x)) \le b_n\},$$

REMARK. Note that the condition $\mathcal{M} \subset \hat{C}_{r,h}(a_n, b_n)$ is equivalent to

$$\sup_{x\in\mathcal{M}} \sqrt{nd^{D+4}}\|\hat{Q}(x)V^T(\hat{H}(x))\hat{g}(x)\|_2 \le a_n, \quad \sup_{x\in\mathcal{M}} \lambda_{d+1}(\hat{H}(x)) \le b_n.$$

To cover $\mathcal{M}$ with the confidence region $\hat{C}_{r,h}(a_n, b_n)$, we just need to make the above relation satisfy probability.

THEOREM 3.6. *For any $\alpha \in (0,1)$, there exist $a_n(\alpha), b_n(\alpha)$ such that, when $n \to \infty$, we have*

$$P(\mathcal{M} \subset \hat{C}_{r,h}(a_n(\alpha), b_n(\alpha))) \ge 1 - \alpha.$$

To derive $a_n(\alpha)$, please refer to the proof in the supplementary material. The procedure to derive $b_n(\alpha)$ is similar to that for $a_n(\alpha)$.

REMARK. Note that, for any point $x \in \mathcal{M}$, both of these conditions are satisfied:

$$\hat{Q}(x)V^T(H(x))g(x) = 0, \quad \lambda_{d+1}(H(x)) \le 0.$$

Using the above relation, we can convert the condition of $\hat{C}_{r,h}(a_n, b_n)$ into the supremum of the estimated error on the manifold. Because $\hat{Q}(x)V^T(H(x))g(x) = 0, \forall x \in \mathcal{M}$,

$$\sup_{x\in\mathcal{M}} \sqrt{nd^{D+4}}\|\hat{Q}(x)V^T(\hat{H}(x))\hat{g}(x)\|_2 \le a_n,$$

which is equivalent to

$$\sup_{x\in\mathcal{M}} \sqrt{nd^{D+4}}\|\hat{Q}(x)V^T(\hat{H}(x))\hat{g}(x) - \hat{Q}(x)V^T(H(x))g(x)\|_2 \le a_n. \tag{21}$$

Similarly, because of $-\lambda_{d+1}(H(x)) \ge 0, \forall x \in \mathcal{M}$, we have the condition

$$\sup_{x\in\mathcal{M}} (\lambda_{d+1}(\hat{H}(x)) - \lambda_{d+1}(H(x))) \le b_n, \tag{22}$$

which is sufficient for $\sup_{x\in\mathcal{M}} \lambda_{d+1}(\hat{H}(x)) \le b_n$. To obtain the confidence ridge of $\mathcal{M}$, we just need to perform a confidence analysis on (21) and (22).

In this section, we consider the advantage of the bias and variance for the kernel-density function $\hat{p}_{r,h}(x)$ corresponding to *l*-SCRE over those of the classical KDE corresponding to SCRE. The results show that we can indeed obtain a ridge closer than that obtained from using the *l*-SCRE.

Next, instead of the analyzing the element-wise partial derivatives, from a higher perspective, we give the spectral property of $H(x) = \nabla\nabla\hat{p}_{r,h}(x)$ with the smooth movement of x, and show the necessity of transformation with a concave function.

4. Transformation of $\hat{p}_{r,h}(x)$. In this section, we are concerned mainly with the relationship between two ridges, $R(\hat{p}_{r,h}(x))$ and $R(f(\hat{p}_{r,h}(x)))$, which are the ridges derived from two functions, $\hat{p}_{r,h}(x)$ and $f(\hat{p}_{r,h}(x))$, respectively. Here, we perform a deep analysis of the effectiveness of nonlinear function f acting on the density function $\hat{p}_{r,h}(x)$. As Theorem 4.9 demonstrates, the Hausdorff distance between the ridge and the manifold projection would be reduced via transformation with a nonlinear concave function.

The transformation of the density function is a composite function f acting on $\hat{p}_{r,h}(x)$ as $f(\hat{p}_{r,h}(x))$. Usually, f is chosen to be a function that has a first- and second-order derivative satisfying $f'(x) > 0$ and $f''(x) < 0$. The transformation has two main benefits:

1. Since the transformation of $\hat{p}_{r,h}(x)$ corresponds to a rank-one modification to the Hessian matrix, the ridge generated after the transformation will leave out some singular points, compared with the ridge obtained before the transformation.
2. The eigenspace of the Hessian of $f(\hat{p}_{r,h}(x))$ becomes more stable and has an intuitive explanation as a tangent space. However, the eigenspace of the Hessian of $\hat{p}_{r,h}(x)$ is more complicated because of the perturbation effect of the scale of a rank-one matrix.

4.1. *The Weighted-Covariance Form.* When we restrict the form of $K_h(u)$ as a radial basis class $\phi(-\|u\|_2^2/h^2)$, the Hessian matrix $\nabla\nabla\hat{p}_{r,h}(x)$ yields a simple form of a covariance matrix added via a scaling matrix:

$$H_{\hat{p}_{r,h}}(x) = \frac{4}{nh^{D+4}}\{\sum_{i\in\mathcal{I}_r} \phi''(-\|\frac{x-x_i}{h}\|^2)(x-x_i)(x-x_i)^T + \gamma_r(x)I\}, \tag{23}$$

where $\gamma_r(x) = -\frac{h^2}{2}\sum_{i\in\mathcal{I}_r} \phi'(-\|\frac{x-x_i}{h}\|_2^2)$. Since the shifting caused by an identity matrix multiplied by a scalar does not affect the eigenspace, the former component

$$\psi_r(x) = \sum_{i\in\mathcal{I}_r} \phi''(-\|\frac{x-x_i}{h}\|^2)(x-x_i)(x-x_i)^T \tag{24}$$

determines the spectral property of $H_{\hat{p}_{r,h}}(x)$. For simplicity, we introduce $w_h(x, x_i)$ by normalizing the weights of $\phi''(-\|\frac{x-x_i}{h}\|^2)$ in (24), which will also keep the eigenspace unchanged.

$$w_h(x, x_i) = \phi''(-\|\frac{x-x_i}{h}\|^2)/\sum_{i\in I_r} \phi''(-\|\frac{x-x_i}{h}\|^2).$$

After the normalization, we have $\sum_{i\in\mathcal{I}_r} w_h(x,x_i)=1$. With this notation, the semi-definite covariance matrix $\psi(x)$ can be simplified as a weighted summation of a rank-one matrix:

$$J_r(x)=\sum_{i\in\mathcal{I}_r} w_h(x,x_i)(x-x_i)(x-x_i)^T. \tag{25}$$

The matrix fields $J_r(x)$ and $\psi_r(x)$ share the same eigenspace because the two matrices are connected by

$$J_r(x)\sum_{i\in I_r}\phi''(-\|\frac{x-x_i}{h}\|^2)=\psi_r(x).$$

The weights for different samples in $J_r(x)$ sum to 1, which makes the analysis and notation much easier in the following context.

Since the matrix field $J_r(x)$ is parameterized by x, the eigenspace corresponding to the largest d eigenvalues is largely affected by x. Next, we express the locally weighted mean $c_r(x)$ as

$$c_r(x)=\sum_{i\in\mathcal{I}_r} w_h(x,x_i)x_i. \tag{26}$$

Note that, in an r-ball of x, $\mathcal{M}\cap B_x(r)$ is approximately a local, lower-dimensional affine space. As a result, the weighted mean $c_r(x)$ also resides approximately on the affine space. Since $x_i-x=c_r(x)-x+(x_i-c_r(x))$, by substituting this into (25) we have:

$$\begin{aligned} J_r(x)=&\sum_{i\in\mathcal{I}_r} w_h(x,x_i)(c_r(x)-x)(c_r(x)-x)^T+...\\ &+\sum_{i\in\mathcal{I}_r} w_h(x,x_i)(x_i-c_r(x))(x_i-c_r(x))^T+...\\ &+\sum_{i\in\mathcal{I}_r} w_h(x,x_i)\{(c_r(x)-x)(x_i-c_r(x))^T+(x_i-c_r(x))(c_r(x)-x)^T\}. \end{aligned} \tag{27}$$

Using the definition of $c_r(x)$, we can see that

$$\sum_{i\in\mathcal{I}_r} w_h(x,x_i)(x_i-c_r(x))=\sum_{i\in\mathcal{I}_r} w_h(x,x_i)x_i-\sum_{i\in\mathcal{I}_r} w_h(x,x_i)c_r(x)=0.$$

Thus, we need to consider only the remaining two parts of (27):

$$J_r(x)=(c_r(x)-x)(c_r(x)-x)^T+\sum_{i\in I_r} w_h(x,x_i)(x_i-c_r(x))(x_i-c_r(x))^T. \tag{28}$$

Let the modified weighted covariance matrix $C_r(x)$ be denoted by

$$C_r(x)=\sum_{i\in\mathcal{I}_r} w_h(x,x_i)(x_i-c_r(x))(x_i-c_r(x))^T,$$

From (28), we can see that $J_r(x)$ consists of the following two parts:

1. *The subspace spanned by* $\{x_i-c_r(x), i\in\mathcal{I}_r\}$. When r is small enough, the samples in $\{x_i, i\in\mathcal{I}_r\}$ reside approximately in a d-dimensional affine space. Thus, $c_r(x)$ lies in the convex hull of the affine space, which also belongs to the affine space. As a result, the principal eigenvectors of $C_r(x)$ approximately span the tangent space of $\mathcal{M}$.
2. *The one-dimensional affine space spanned by* $c_r(x)-x$. This space can be seen as a distorted component added to $C_r(x)$, which will blur or hide some important information in the space spanned by the top d component eigenvectors of $J_r(x)$.

Clearly, the eigenspace $\mathcal{V}_d$, spanned by eigenvectors corresponding to the top d largest eigenvalues of $J_r(x)$, is largely affected by the scale of $c_r(x) - x$. When $c_r(x) - x$ is large, it will have a much larger influence on the space of $\mathcal{V}_d$, and vice versa.

We give an example to illustrate the influence of the scale of $c_r(x) - x$ on the eigenspace of $J_r(x)$ in Figure 4. The blue dash curve represents the hidden unknown manifold $\mathcal{M}$. The red star points stand for the observation points x_i, which are assumed to be sampled from the manifold $\mathcal{M}$ and disturbed by some noise. The middle red dot represents the location of x, which we want to move towards the manifold from far away. The two orthogonal arrows represent the vectors $\{\lambda_k u_k, k = 1, 2\}$, where λ_k and u_k are the k-th eigenvalue and eigenvector of $J_r(x)$, respectively.

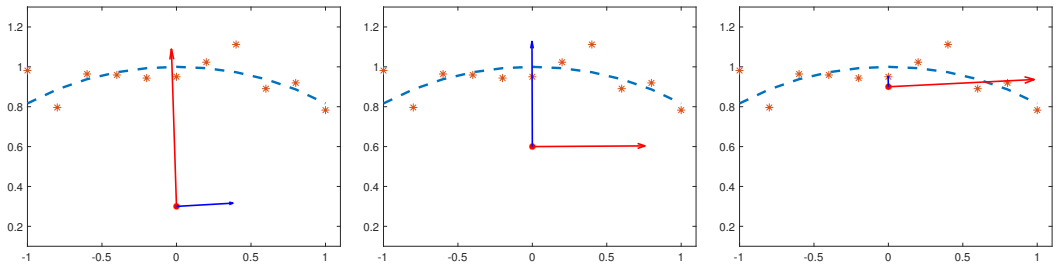

FIG 4. *The process of the variation of $J_r(x)$'s eigenspace, with x approaching the manifold*

The left diagram in the figure shows that, when x is far away from the data points, the space spanned by the principal eigenvectors of $J_r(x)$ resides approximately in the normal space of the manifold. As x comes near the ridge, we can see that the two eigenvalues of $J_r(x)$ will be equal, as shown in the middle diagram. If x continues to approach the ridge, the principal eigenvector will become parallel with the ridge.

The subspace-constrained shifting method requires x to move in the normal space of the manifold to get the projection x^* of x onto the manifold. However, as x is far away from the manifold and x^* is unknown, to overcome this difficulty we can use the space spanned by the eigenvectors of $J_r(x)$, corresponding to some particular eigenvectors, as an approximation of the normal space at x^*. We show that this may cause some error because, as x approaches the manifold, the magnitude of λ_d/λ_{d+1} will change dramatically. Thus, picking the eigenvector corresponding to the largest eigenvalue may cause a sudden turn in the orthogonal direction. The unpredictable directional behavior of the principal eigenvector of $J_r(x)$ causes the searching lower-dimensional ridge algorithm to converge badly. Fortunately, this difficulty can be overcome by using a nonlinear transformation, as introduced in the next section. Through modification with a rank-one matrix, we can obtain a new $J_r(x)$ with a very stable eigenspace, i.e. the order of the first and second eigenvectors will not be reversed with the moving of x.

4.2. *Eigenspace Analysis for $C_r(x)$ and $J_r(x)$.* Since the convex combination of samples in $\mathcal{I}_r$ spans a convex hull next to $\mathcal{M}$, $c_r(x)$ is in the convex hull, which is supposed to be next to $\mathcal{M}$ as well when r is relatively small. We prefer to use the eigenvectors corresponding to the d principal eigenvalues of $C_r(x)$, instead of $J_r(x)$, to approximate the tangent space. The reasons for this will be explained in detail in the following section.

Recall $J_r(x)$ in (28), which shares the same eigenvectors with the Hessian of $\hat{p}_{r,h}(x)$. We will now determine the difference between the eigenspaces of $J_r(x)$ and $C_r(x)$. To simplify our analysis, we make the following assumption:

ASSUMPTION 4.1. *Assume the vector $c_r(x) - x$ is orthogonal to the subspace spanned by the eigenvectors corresponding to the top d eigenvalues of the covariance matrix $C_r(x)$.*

DEFINITION 4.2. The distance between two subspaces $\mathcal{V}$ and $\mathcal{U}$ is defined as the operator norm of the error of two projection matrices, i.e. $D(\mathcal{V},\mathcal{U}) = \|P_{\mathcal{V}} - P_{\mathcal{U}}\|_2$.

With Assumption 4.1, and having $\mathcal{V}_d(C_r(x))$ denote the eigenspace spanned by the eigenvectors corresponding to the largest d eigenvalues of $C_r(x)$, we have the following theorem:

THEOREM 4.3. *If $\|c_r(x) - x\|_2^2 < \lambda_d(C_r(x)))$, the eigenspaces corresponding to the top d eigenvalues of $C_r(x)$ and $J_r(x)$ coincide, i.e. the distance*

$$D(\mathcal{V}_d(C_r(x)), \mathcal{V}_d(J_r(x))) = 0.$$

Otherwise, if $\|c_r(x) - x\|_2^2 \geq \lambda_d(C_r(x)))$, then $D(\mathcal{V}_d(C_r(x)), \mathcal{V}_d(J_r(x))) = 1$.

The details of the proof can be found in the supplementary material.

4.3. *The Singular-Point Phenomenon.* Here, we use the term 'singular point' to refer to the points that satisfy the ridge condition but are obviously not on the ridge. Suppose there exists x_0 such that $\|c_r(x) - x\|_2^2 > \lambda_d(J_r(x_0))$, $c_r(x_0) - x_0$ is in the space $\mathcal{V}_d(J_r(x))$, which is spanned by the eigenvectors corresponding to the top d eigenvalues of $J(x)$, and thus $c_r(x_0) - x_0 \perp \mathcal{V}_{D-d}(J_r(x))$. In this case, applying the subspace-constrained mean-shift algorithm

$$x_0' = x_0 + (I - P(\mathcal{V}_d(J_r(x))))(c_r(x_0) - x_0)$$

cannot move x_0 because of $(I - P(\mathcal{V}_d(J_r(x))))(c_r(x_0) - x_0) = 0$. This phenomenon will be demonstrated in Figure 4. To avoid this phenomenon, we introduce the density-function transformation idea and show the inclusion relation for the ridges before and after the transformation.

The nonlinear transformation to the density function corresponds to the rank-one modification to $J_r(x)$. With a proper chosen nonnegative, increasing, and concave function, the Hessian matrix for the transformed density function will diminish the effect of the uncontrollable term $(c_r(x) - x)(c_r(x) - x)^T$ in $J_r(x)$.

4.4. *Ridge Variation via Transformation.* In this section, we show that the transformation of the estimated density function by a monotonously increasing function, f, will result in a new ridge $R(f(p(x)))$, which is a subset of $R(p(x))$. In the following calculations, we can abbreviate $R(f(p(x)))$ and $R(p(x))$ as $R(f(p))$ and $R(p)$, respectively.

For two ridges defined by the density $p(x)$ and $f(p(x))$, where $f(y)$ is a monotonously increasing and concave function satisfying $f'(y) > 0$ and $f''(y) < 0$, we have

$$\begin{aligned} &\nabla f(p(x)) = f'(p(x))\nabla p(x), \\ &H_{f(p)}(x) = f''(p(x))\nabla p(x)\nabla^T p(x) + f'(p(x))H_p(x). \end{aligned} \tag{29}$$

Let the d-dimensional ridges $R(p)$ and $R(f(p))$, corresponding to $p(x)$ and $f(p(x))$, be defined as

$$\begin{aligned} R(p) &= \{x | \Pi^{\perp}_{H_p}(x)\nabla p(x) = 0, \lambda_{d+1}(H_p(x)) < 0\}, \\ R(f(p)) &= \{x | \Pi^{\perp}_{H_{f(p)}}(x)\nabla f(p(x)) = 0, \lambda_{d+1}(H_{f(p)}(x)) < 0\}. \end{aligned}$$

Next, we will provide two lemmas to show the relationship between the ridges before and after transformation:

1. If $\lambda > 0$, a rank-one modification expressed as λuu^T will enlarge the projection of u onto the eigenspace corresponding to the largest d eigenvalues.
2. For a rank-one modification expressed as λuu^T, if u is in some subspace spanned by the largest d eigenvectors and $\lambda > 0$, the rank-one modification will keep the eigenvalues corresponding to the orthogonal complement space unchanged.

These two lemmas are aligned with our intuition to some degree, which is a key step in proving the inclusion property of ridges.

LEMMA 4.4. *For any symmetric matrix B, let $A = B + \lambda uu^T, \forall \lambda \geq 0$. We have $\|\Pi_A u\|_2 \geq \|\Pi_B u\|_2$, where Π_A, Π_B are the projections onto the space spanned by the eigenvectors corresponding to the d largest eigenvalues of A and B, respectively.*

LEMMA 4.5. *For any symmetric matrix B, let $A = B + \lambda uu^T$, $\forall \lambda \geq 0$ and any nonzero vector $u \in \text{span}\{u_1(B(x)), u_2(B(x)), ..., u_d(B(x))\}$; the $(d+1)$-th to D-th largest eigenvalues of A and B yield*

$$\lambda_{d+k}(A) = \lambda_{d+k}(B), \quad k = 1, ..., D-d,$$

where $u_k(B(x))$ and $\lambda_k(B(x))$ are the eigenvector and eigenvalue corresponding to the k-th largest eigenvalues of $B(x)$.

The details of the proof of Lemma 4.4 and Lemma 4.5 are in the supplementary material.

REMARK. Lemma 4.4 is a key step in this section, as it shows that the rank-one modification λuu^T will enlarge the projection of u onto the space spanned by the component eigenvectors.

REMARK. Lemma 4.5 shows that the semi-positive rank-one modification in the subspace corresponding to the largest d eigenvalues will keep the remaining eigenvalues corresponding to the orthogonal complement subspace unchanged.

Using Lemma 4.4 and Lemma 4.5, we obtain Lemma 4.6:

LEMMA 4.6. *For any monotonously increasing and concave function $f(y)$, i.e. $f'(x) > 0, f''(x) \leq 0$, for $x \in R(f(p))$, the following are satisfied simultaneously:*

$$\lambda_{d+1}(H_p(x)) < 0,$$

$$\|\Pi^{\perp}_{H_p}(x)\nabla p(x)\|_2 \leq \|\Pi^{\perp}_{H_{f(p)}}(x)\nabla p(x)\|_2 = 0,$$

The condition $\|\Pi^{\perp}_{H_p}(x)\nabla p(x)\|_2 = 0$ implies $\Pi^{\perp}_{H_p}(x)\nabla p(x) = \mathbf{0}$, which in turn indicates that $x \in R(p)$. Thus, $R(f(p)) \subset R(p)$.

The details of the proof of lemma 4.6 can be found in the supplementary material.

REMARK. The inclusion property $R(f(p)) \subset R(p)$ is applicable to any density function $p(x)$. When $p(x)$ is selected to be the special formate at $\hat{p}_{r,h}(x)$, we will obtain the following results by applying Lemma 4.6 to the local density function $\hat{p}_{r,h}(x)$:

THEOREM 4.7. *For any monotonously increasing and concave function $f(y)$, we have $R(f(\hat{p}_{r,h})) \subset R(\hat{p}_{r,h})$. Furthermore,*

$$\|\Pi^{\perp}_{H_{\hat{p}_{r,h}}}(x)\nabla \hat{p}_{r,h}(x)\|_2 \leq \|\Pi^{\perp}_{H_{f(\hat{p}_{r,h})}}(x)\nabla \hat{p}_{r,h}(x)\|_2.$$

Numerically, we can set some small number ϵ such that at least 10^{-6} will act as a threshold for searching the ridge condition, because of the rounding error occurring with the computation. It can be observed that the ridge searching with a stop condition $\|\Pi^{\perp}_{H_{f(\hat{p}_{r,h})}}(x)\nabla\hat{p}_{r,h}(x)\|_2 \leq \epsilon$ is stronger than that with $\|\Pi^{\perp}_{H_{\hat{p}_{r,h}}}(x)\nabla\hat{p}_{r,h}(x)\|_2 \leq \epsilon$ because of the existing inequality in Theorem 4.7.

Clearly, the gradients of $p(x)$ and $f(p(x))$ only differ with a scale $f'(p(x))$. The scale difference will not affect the ridge from the ridge definition. As a consequence, we will analyze the Hessian matrices corresponding to $p(x)$ and $f(p(x))$, which will generate a difference between the ridges.

For any density function $p(x)$ and an increasing concave function $f(x)$, and taking the second derivative with respect to x, we have

$$H_{f(p)}(x) = f'(p(x))H_p(x) + f''(p(x))\nabla p(x)\nabla^T p(x) \tag{30}$$

Rearranging (30), we have the relationship between Hessian matrices:

$$H_p(x) = \frac{1}{f'(p(x))}(H_{f(p)}(x) - f''(p(x))\nabla p(x)\nabla^T p(x)). \tag{31}$$

From (31), we know the rank-one modification becomes even stronger when the term $\| - f''(p(x))\nabla p(x)\nabla p^T(x)\|_F^2$ is at a larger scale. Specifically, if we select $f(y) = \log(y)$, we have $f''(p(x)) = -1/p^2(x)$. Then, the rank-one modification term becomes

$$-f''(y)\nabla p(x)\nabla p^T(x) = \frac{1}{p^2(x)}\nabla p(x)\nabla p^T(x).$$

Note that

$$\| - f''(y)\nabla p(x)\nabla p^T(x)\|_F^{1/2} = \frac{\|\nabla p(x)\|_2}{p(x)}.$$

The difference between the principal eigenspaces (corresponding to the top d eigenvalues) of $H_p(x)$ and $H_{f(p)}$ depends on the balance of values for $\|\nabla p(x)\|_2$ and $p(x)$.

1. For x that has large $\|\nabla p(x)\|_2$ and small $p(x)$, the rank-one modification will make a big difference.
2. For x that has small $\|\nabla p(x)\|_2$ and large $p(x)$, the effect of the rank-one modification will be small.

EXAMPLE. We give an example that has a large scale $\|\nabla p(x)\|_2/p(x)$, estimating $\hat{p}(x)$ using the KDE function with a form of

$$\hat{p}(x) := \frac{1}{n}\sum_i K(-\|x - x_i\|_2^2/h^2).$$

Then, $\|\nabla\hat{p}(x)\|_2/\hat{p}(x)$ has a geometry interpretation as the distance from x to a point in the convex hull

$$\frac{\|\nabla\hat{p}(x)\|_2}{\hat{p}(x)} = \frac{2}{h^2}\|c_r(x) - x\|_2,$$

where $c_r(x)$ is defined in (26) as a point in the convex hull of $\{x_i, i \in I_r\}$. For any $x \notin \text{Conv}\{x_i, i \in I_r\}$, the distance $\|c_r(x) - x\|_2$ has a lower bound:

$$\|c_r(x) - x\|_2 \geq \|P_{\text{Conv}}(x) - x\|_2 = \|P^{\perp}_{\text{Conv}}x\|_2,$$

where $P_{\text{Conv}}(x)$ is the projection of x into the convex hull. As the initial point $x := x_0$ could be any point away from the convex hull, the value $\|P^{\perp}_{\text{Conv}}x_0\|_2$ could be at any large scale.

THEOREM 4.8. *The subtract set $R(p)/R(f(p))$ consists of points in the set:*

$$\{x|c_r(x) - x \in \text{span}\{u_1, ..., u_d\}, c_r(x) - x \notin \text{span}\{v_1, ..., v_d\}\},$$

where $\{u_1, ..., u_d\}$ and $\{v_1, ..., v_d\}$ are the eigenvectors corresponding to the top d eigenvalues of $H_p(x)$ and $H_{f(p)}(x)$, respectively.

REMARK. The points in $R(p)/R(f(p))$ are often far from our intended ridge. Neglecting this group of points will yield a ridge much closer than before.

We demonstrate that the ridge obtained from the nonlinear transformation can indeed be closer than the original one by showing that $\text{Haus}(R(f(p)), \mathcal{M}_{R(f(p))})$ is smaller than $\text{Haus}(R(p), \mathcal{M}_{R(p)})$.

THEOREM 4.9. *For the ridge $R(f(p))$ defined by the transformed nonlinear increasing and concave function f, we have*

$$\text{Haus}(R(f(p)), \mathcal{M}_{R(f(p))}) \leq \text{Haus}(R(p), \mathcal{M}_{R(p)}),$$

where $R(p)$ and $R(f(p))$ are the d-dimensional ridges corresponding to p and $f(p)$, and $\mathcal{M}_{R(p)}$ and $\mathcal{M}_{R(f(p))}$ are the projections of $R(p)$ and $R(f(p))$ onto $\mathcal{M}$, respectively.

The proof of Theorem 4.9 can be found in the supplementary material.

4.5. *Geometric Interpretation.* The angle between a vector v and space $\mathcal{S}$ is defined as the smallest angle between v and $u \in \mathcal{S}$. Under this condition, u is the projection of v onto the space of $\mathcal{S}$, i.e. $u = P_{\mathcal{S}}(v)$.

$$\theta(v, \mathcal{S}) = \arg\cos \frac{\langle v, P_{\mathcal{S}} v \rangle}{\|v\|_2 \|P_{\mathcal{S}} v\|_2} = \arg\cos(\|P_{\mathcal{S}} v\|_2 / \|v\|_2).$$

Let $\mathcal{S}_{H_p}$ and $\mathcal{S}_{H_{f(p)}}$ denote the spaces spanned by the eigenvectors corresponding to the top d eigenvalues of $H_p(x)$ and $H_{f(p)}(x)$, respectively.

From Lemma 4.6, we know

$$\nabla p(x)^T \Pi_{H_p}(x) \nabla p(x) \geq \nabla p(x)^T \Pi_{H_{f(p)}}(x) \nabla p(x). \tag{32}$$

The inequality (32) indicates $\|\Pi_{H_p}(x)\nabla p(x)\|_2 \geq \|\Pi_{H_{f(p)}}(x)\nabla p(x)\|_2$, which is equivalent to

$$\cos(\theta(\nabla p(x), \mathcal{S}_{H_p})) = \frac{\|\Pi_{H_p}(x)\nabla p(x)\|_2}{\|\nabla p(x)\|_2} \geq \frac{\|\Pi_{H_{f(p)}}(x)\nabla p(x)\|_2}{\|\nabla p(x)\|_2} = \cos(\theta(\nabla p(x), \mathcal{S}_{H_{f(p)}})),$$

which implies that, for any x and any increasing concave function f, we always have the inequality

$$\theta(\nabla p(x), \mathcal{S}_{H_p}(x)) \leq \theta(\nabla p(x), \mathcal{S}_{H_{f(p)}}(x)).$$

REMARK. The absolute value of $\cos(\theta(\nabla p(x), \mathcal{S}_{H_p}))$ stands for the size of the angle between the vector $\nabla p(x)$ and the subspace $\mathcal{S}_{H_p}$. If $|\cos(\theta(\nabla p(x), \mathcal{S}_{H_p}))| = 1$, the vector $\nabla p(x)$ is parallel with the subspace. If $|\cos(\theta(\nabla p(x), \mathcal{S}_{H_p}))| = 0$, the vector $\nabla p(x)$ is vertical in relation to the subspace. If $|\cos(\theta(\nabla p(x), \mathcal{S}_{H_p}))|$ approaches 1, this will make x satisfy the ridge condition $\Pi^{\perp}(x)\nabla g(x) = 0$ gradually.

5. Numerical Experiments. In this section, we will show the effectiveness of our method compared with the classical subspace-constrained methods for manifold fitting. The comparison is based on a uniform frame for all methods. Overall, the manifold algorithm can be split into two parts. For the first part, for any point x lying outside the manifold, construct an attraction force starting from x and ending with some point (often represented by the observations) on the manifold. For the second part, estimate the normal space and use the projection to modify the attraction force such that x's moving trajectory resembles the projected trajectory as much as possible. We briefly summarize the main methods considered for comparison below and we refer the readers to the Appendix B for the implementation details (B1-B2) and the Ridge-searching Algorithm for SCRE (B3).

Implementation: SCRE, *l*-SCRE, MFIT-i (Fefferman et al., 2018), and MFIT-ii (Yao and Xia, 2019). The MATLAB codes together with all numerical examples used in this paper are available on `https://zhigang-yao.github.io/research.html` (for SCRE, *l*-SCRE and MFIT-ii). We have implemented MFIT-i, since the authors of Fefferman et al. (2018) have not provided implementation due to the nature of their work has been purely abstract.

5.1. *Ridge Criteria.* The criterion for a point on a ridge is that it must satisfy the ridge condition. The ridge condition indicates that, when x is on the ridge, the attraction vector $F(x)$ resides in the tangent space. The tangent space is usually obtained from the top d eigenvectors of the covariance matrix. Conversely, if the ridge condition is not met at x, we can also expect the angle between the attraction vector $F(x)$ and the estimated eigenspace to be large, i.e. a small cos value.

An example of a one-dimensional curve embedded in a two-dimensional space is provided below. In this case, the tangent space and the normal space are both one-dimensional. We first sample some points (blue circles) that represent the observations $\{x_k\}$. Then, we randomly sample the filled red diamonds that stand for x. The value on the figure is computed as

$$\mathrm{s}(x) = |\cos(v(x), u(x))| = |\frac{\langle v(x), u(x)\rangle}{\|v(x)\|\|u(x)\|}|,$$

where $u(x)$ is the eigenvector corresponding to the largest eigenvalue of the covariance matrix and $v(x)$ is the gradient of the density function. Clearly, $u(x)$ is parallel with $v(x)$ when x is on the manifold, i.e. $\mathrm{s}(x) = 1$. We want $\mathrm{s}(x)$ to indicate the relationship between x and the ridge. When $\mathrm{s}(x) = 1$ the point x should lie approximately on the ridge, and the far-away point x makes $\mathrm{s}(x)$ approach 0.

Different methods have different forms of attraction and projection. For brevity, we list the covariance matrices corresponding to different methods in the table below.

From Figure 5, we can draw the following conclusions:

- As the left diagram shows, $\mathrm{s}(x) = 1$ is a necessary but insufficient condition for x to be on the ridge for SCRE. The outliers that are far from the ridge may also make $s(x) = 1$, which is the main disadvantage of obtaining the ridge by SCRE without the nonlinear transformation.
- Theorem 4.7 is verified, as the values in the middle figure are uniformly smaller than the corresponding values in the left figure.
- *l*-SCRE has a better property, especially for the values emphasized in the rectangle shown in the right diagram. The outlier points have much smaller values of $\mathrm{s}(x)$ in the right diagram than those in the former two diagrams.

In Figure 6, we plot the vector field generated by the eigenvector corresponding to the smallest eigenvalue. This vector field is the geometric interpretation of the projection $\Pi^{\perp}$ in

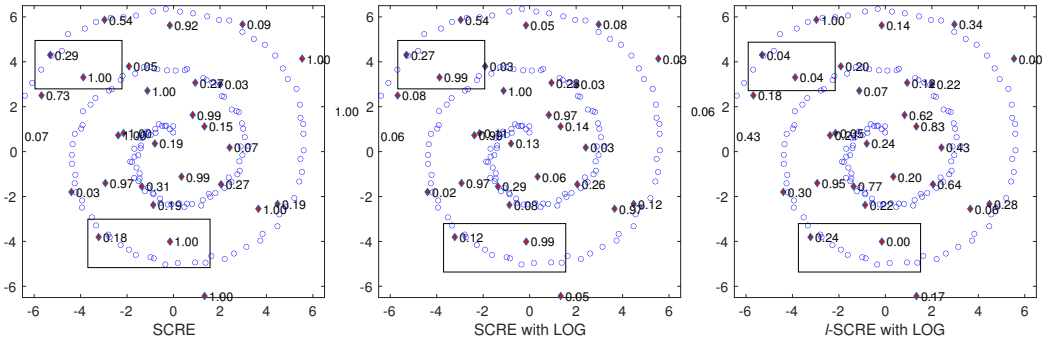

FIG 5. *Illustration of the Performances of Three Difference Covariance Matrices*

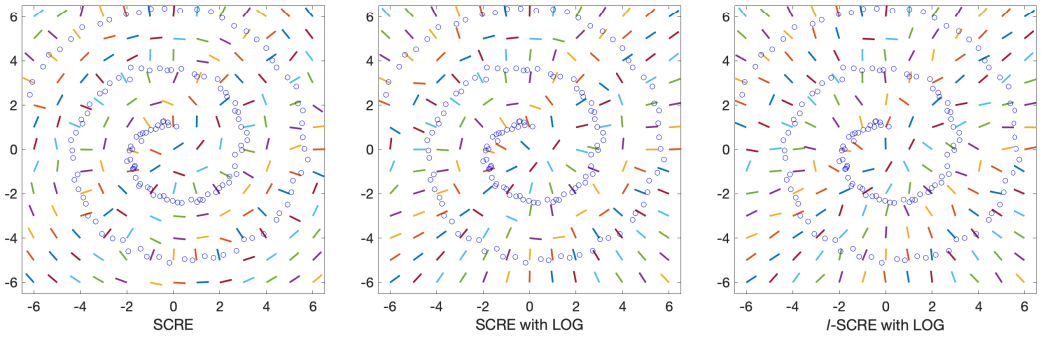

FIG 6. *Illustration of the vector field corresponding to the eigenvector of the second eigenvalue with respect to three difference covariance matrices*

the two-dimensional space. Because the constrained subspace in the two-dimensional space is one-dimensional, the vector field coincides with the trajectory of the outliers moving onto the ridge. We can conclude from the above figures that the vector field corresponding to *l*-SCRE is more smooth and points approximately to the projection onto the manifold, while the vector field of SCRE is not well behaved for samples not close to the manifold.

5.2. *Synthetic Data Set.* First, we provide a one-dimensional ring example to show the effectiveness of our method compared with SCRE. Because of the simplicity of the circle structure, it is easy to compute the real projection onto it. We assume that our hidden manifold $\mathcal{M}$ is the one-dimensional circle embedded in a two-dimensional space with a radius of 1 and center of $(0, 0)$.

The observations are generated in the following two steps. First, a uniform sample from $\mathcal{M}$ is used to obtain the ideal observations without noise $\tilde{y}_i, i = 1 : N$. Second, the independent noise from a two-dimensional normal distribution is added to get $y_i = \tilde{y}_i + \epsilon_i, i = 1 : N$. The observations are used to construct our SCRE and *l*-SCRE functions.

Generate a random mesh set $\mathcal{G}$ arbitrarily. For each point $x_k \in \mathcal{G}$, apply the subspace-constrained mean-shift algorithm to obtain a point $\hat{x}_k$ that satisfies the ridge definition. Thus, the computed ridge set is $\hat{\mathcal{G}} = \{\hat{x}_k\}$. Because we use different approaches to define our ridge sets, we have $\hat{\mathcal{G}}$ and $\hat{\mathcal{G}}_\ell$, which correspond to the SCRE and *l*-SCRE approaches.

For each point $\hat{x}$ belonging to $\hat{\mathcal{G}}$ or $\hat{\mathcal{G}}_\ell$, we can define the projection of $\hat{x}$ onto the real manifold $\mathcal{M}$ as $\pi_{\mathcal{M}}(\hat{x}) = \arg\min_{y \in \mathcal{M}} \|y - \hat{x}\|_2$. In the case of the special two-dimensional circle, the projection has an explicit form, $\pi_{\mathcal{M}}(\hat{x}) = \hat{x}/\|\hat{x}\|$. The average margin from $\hat{\mathcal{G}}$ or

$\hat{\mathcal{G}}_\ell$ to $\mathcal{M}$ is defined as

$$\text{Margin}(\hat{\mathcal{G}}, \mathcal{M}) = \frac{1}{|\hat{\mathcal{G}}|} \sum_{x_k \in \hat{\mathcal{G}}} \min_{y \in \mathcal{M}} \|x_k - y\|_2.$$

In Figure 7, the small blue diamonds $\diamond$ represent the observations that are used to construct our SCRE or l-SCRE estimator. The blue dots $\bullet$ represent the points that satisfy the ridge condition. The red dots $\bullet$ represent the projection of the blue dots onto the ideal manifold $\mathcal{M}$.

For the above two figures, the margin of the ridge corresponding to SCRE is shown in the left partition and the margin of the ridge corresponding to l-SCRE is shown in the right partition. In the two figures below, we show how the average margin and Hausdorff distance, between $\hat{\mathcal{G}}$ and $\mathcal{M}_{\hat{\mathcal{G}}}$, change with the parameter h when the neighborhood size parameter is fixed.

It is clear from Table 2, Table 3 and Figure 7 that, under the same parameter setting (such as h and noise level σ), the algorithm of l-SCRE yields a ridge with a margin much smaller than that of SCRE. The rate of increase of the distance measurement for l-SCRE is much lower than that for SCRE with the increasing h. This phenomenon is also explained in our theoretical analysis.

TABLE 2
The margin and Hausdorff between $\hat{\mathcal{G}}$ and $\mathcal{M}$ vary with h for SCRE, l-SCRE, MFIT-i and MFIT-ii on the 1-dimensional circle

	h	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Marg	SCRE	0.0117	0.0159	0.0287	0.0481	0.0756	0.1138	0.1704	0.2641	0.4427
	l-SCRE	0.0116	0.0124	0.0133	0.0137	0.0139	0.0140	0.0141	0.0141	0.0142
	MFIT-i	0.0203	0.0114	0.0123	0.0160	0.0222	0.0303	0.0401	0.0520	0.0662
	MFIT-ii	0.0219	0.0131	0.0102	0.0095	0.0146	0.0224	0.0313	0.0412	0.0519
Haus	SCRE	0.0344	0.0329	0.0436	0.0604	0.0913	0.1310	0.1935	0.2973	0.5027
	l-SCRE	0.0350	0.0289	0.0307	0.0322	0.0331	0.0336	0.0339	0.0341	0.0342
	MFIT-i	0.0952	0.0383	0.0303	0.0347	0.0393	0.0441	0.0531	0.0643	0.0828
	MFIT-ii	0.1058	0.0393	0.0306	0.0266	0.0336	0.0426	0.0560	0.0675	0.0740

TABLE 3
The margin and Hausdorff between $\hat{\mathcal{G}}$ and $\mathcal{M}$ vary with h for SCRE, l-SCRE, MFIT-i and MFIT-ii on the 2-dimensional sphere

	h	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Marg	SCRE	0.0320	0.0320	0.0558	0.0963	0.1560	0.2474	0.4062	0.6370	0.2451
	l-SCRE	0.0288	0.0292	0.0333	0.0357	0.0369	0.0376	0.0380	0.0383	0.0385
	MFIT-i	0.1429	0.0474	0.0297	0.0251	0.0325	0.0467	0.0645	0.0848	0.1097
	MFIT-ii	0.0292	0.0315	0.0330	0.0304	0.0310	0.0412	0.0558	0.0738	0.0924
Haus	SCRE	0.1478	0.0803	0.1070	0.1542	0.2114	0.3139	0.4923	0.7778	0.9165
	l-SCRE	0.0904	0.0828	0.0839	0.0907	0.0937	0.0952	0.0960	0.0966	0.0970
	MFIT-i	0.8873	0.7947	0.1251	0.1047	0.0806	0.0851	0.1024	0.1271	0.1594
	MFIT-ii	0.1127	0.1237	0.1723	0.1874	0.1440	0.1524	0.1636	0.1753	0.1791

As shown in Table 2 and Table 3, when the radius is too small, we cannot ensure that the domain which is centered at x and with a radius of h, contains enough samples to run both MFIT-i and MFIT-ii. By comparing the margin and Hausdorff distance, we can conclude that the result of l-SCRE is more stable and competitive than the results of related methods.

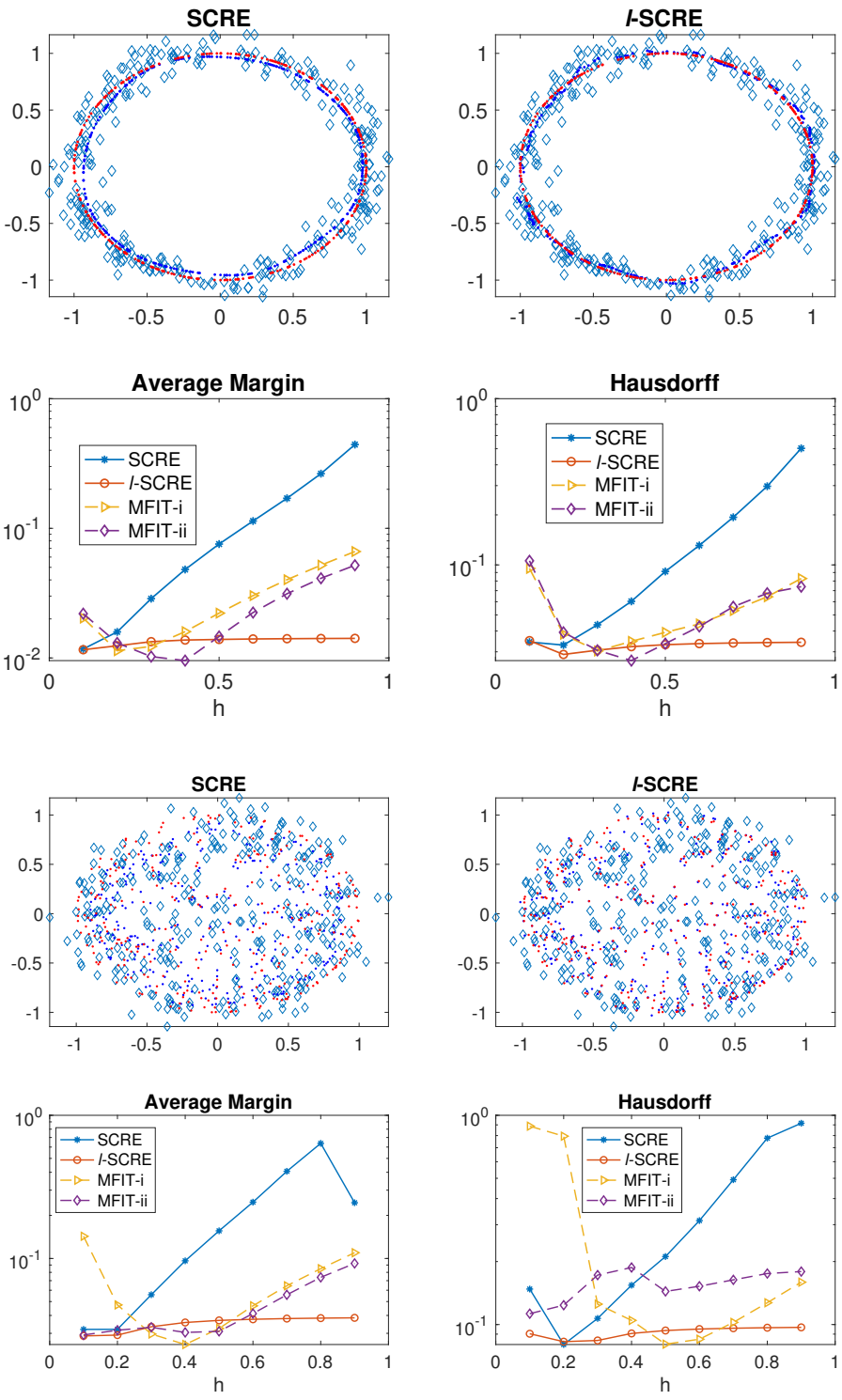

FIG 7. *Margin and Hausdorff illustration for ridges obtained from l-SCRE and SCRE on the 1D circle and the 2D sphere*

5.3. *Real Data Set.* We also test the performance of our algorithm using the Coil20 S. A. Nene and Murase (1996) dataset. The data is collected by rotating an object with 360

degrees. For each object, there are 72 images corresponding to a particular angle. Therefore, the images for each object can be supposed to reside on a one-dimensional manifold, since the underlying parameter is the angle.

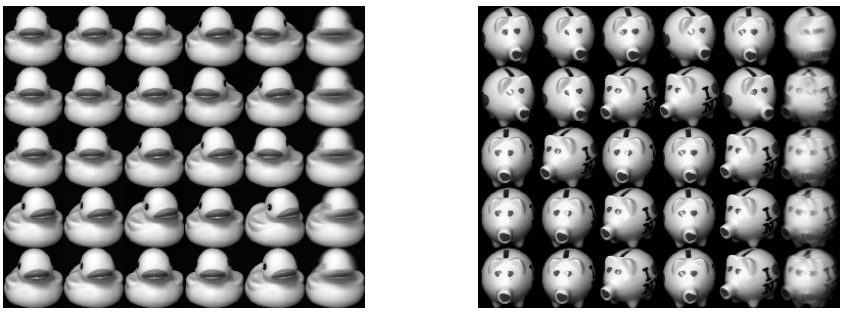

FIG 8. *Illustration of the interpolation for the one-dimensional Image Manifold*

Because we need to compute the eigenspace decomposition in the pixel space, the complexity is $O(n^3)$, where n is the number of pixels. To reduce the complexity, we subsample each image with a 2×2 window and obtain an image with the shape of 64×64.

We randomly generate initial figures and use our algorithm to obtain an approximated image that is supposed to lie on the manifold. As Figure 8 shows, the last column of each figure is the solution of an interpolation point corresponding to different initial figures. The five figures in the same line ahead of the interpolation figure are the five nearest neighborhoods.

6. Discussion. The manifold fitting is a challenging problem because of the unknown topology structure of the hidden manifold $\mathcal{M}$, which in turn causes three main difficulties. First, since there is no global functional structure for the manifold, we need to approximate within a local domain. Second, the number of observations that assume samples from the underlining manifold is very limited. Third, the noise in the observations often makes it difficult to find the true manifold.

In this work, we mainly consider how to improve the estimated ridge under some distance criteria. On many occasions, our sense is that the ridge that is closer to the true manifold is a better one. To improve the ridge, we propose an *l*-SCRE estimator and show that some particular form of nonlinear transformation is also useful in achieving our goals.

Most of the current research focuses on how to approximate the manifold linearly in a local area. However, to approximate the manifold in a large area, a nonlinear approximation may be better because the curvature of the manifold is also taken into consideration. Our subsequent work will concentrate on some nonlinear manifold approximation and projection techniques.

APPENDIX A: RIDGE-SEARCHING ALGORITHM

From the above analysis, we adopt the ridge-searching algorithm after transformation by a nonlinear function. Here, the kernel function is selected as $K_h(u) = \exp(-\|u\|_2^2/h^2)$ and the specific transformation function as $f(y) = \log(y)$. Thus, $f(\hat{p}_{r,h}(x))$ becomes

$$f(\hat{p}_{r,h}(x)) = \log(\frac{1}{nh^D} \sum_{i \in I_r} \exp(-\frac{\|x - x_i\|_2^2}{h^2})),$$

where the index set is $I_r = \{i | \|x - x_i\|_2 \leq r\}$, where r varies with x such that x's neighborhood $\mathcal{N}_{r,x}$ contains a constant number of samples. For notational convenience, we introduce $w_h(x, x_i)$ to represent the weight for x_i, and $c_{r,h}(x)$ to be the local shift mean.

$$w_h(x, x_i) = \exp(-\frac{\|x - x_i\|_2^2}{h^2}), \quad c_{r,h}(x) = \frac{1}{\sum_{i \in I_r} w_h(x, x_i)} \sum_{i \in I_r} w_h(x, x_i) x_i.$$

With the above notations, the gradient of $f(\hat{p}_{r,h}(x))$ is

$$\nabla f(\hat{p}_{r,h}(x)) = c_{r,h}(x) - x,$$

which is a vector starting from x and pointing towards the weighted mean $c_{r,h}(x)$. The Hessian of $f(\hat{p}_{r,h}(x))$ can be written as

$$H_{f(\hat{p}_{r,h})}(x) = \frac{4}{nh^4 \hat{p}_{r,h}(x)} (\sum_{i \in I_r} w_h(x, x_i)(x_i - c_{r,h}(x))(x_i - c_{r,h}(x))^T + \gamma(x) I),$$

where $\gamma(x) = -\frac{nh^2 \hat{p}_{r,h}(x)}{2}$. The algorithm can be summarized as follows:

1. *Determine the Radius:* For x, find s samples that are in the radius r ball of $B_x(r)$. Use the samples to construct the locally weighted center $c_{r,h}(x)$ and local Hessian $H_{f(\hat{p}_{r,h})}(x)$.
2. *Hessian Matrix Decomposition:* Apply spectral decomposition $H_{f(\hat{p}_{r,h})}(x)$ to obtain the projection $\Pi^{\perp} = \sum_{k=d+1}^{D} u_k u_k^T$ corresponding to the eigenspace of the smallest $D - d$ eigenvalues of $H_{f(\hat{p}_{r,h})}(x)$.
3. *Subspace-Constraint Iteration:* Apply the subspace-constrained mean-shift iteration as $x_{t+1} = x_t + \lambda \Pi^{\perp}(c_{r,h}(x_t) - x_t)$. The step size $\lambda \in (0, 1)$ is used to control the speed of convergence and the smoothness of the convergence path.

We claim that the convergence point x_c of $\{x_t, t = 1, 2, 3, ...\}$ will satisfy

$$\Pi^{\perp}(H(x_c)) g(x_c) = 0,$$

where the sequence $\{x_t\}$ is generated from the recursive form $x_{t+1} = x_t + \lambda \Pi^{\perp}(c_{r,h}(x_t) - x_t)$. If the sequence $\{x_t\}$ converges to x_c, letting $t \to +\infty$, we will have $x_c = x_c + \lambda \Pi^{\perp}(c_{r,h}(x_c) - x_c)$. We will then conclude that the convergence point x_c satisfies $\Pi^{\perp}(H(x_c)) g(x_c) = 0$.

Next, we generalize the ridge-searching algorithm into a uniform framework and highlight the fact that the iteration corresponding to kernel-density estimation also belongs to a special case of our uniform framework.

APPENDIX B: UNIFORM FRAMEWORK FOR MANIFOLD FITTING

In this section, we demonstrate some manifold-fitting algorithms under a uniform framework, and compare their strengths and weaknesses to those of related methods.

B.1. Attraction Force. In the setting of our problem, we have a set of samples $\{x_i^*\}$ drawn from a hidden manifold $\mathcal{M}$ and an outlier point x. Then, each $x_i^* - x$ is a vector starting from x and ending with x_i^* on the manifold. Different algorithms employ different strategies to construct the attraction force using the observations $\{x_i\}$. Of course, if x is located within the reach τ of $\mathcal{M}$, the ideal attraction point should be the unique projection of $x^* = \pi_{\mathcal{M}}(x)$. The form of $\pi_{\mathcal{M}}$ is unknown as the structure of the hidden manifold $\mathcal{M}$ is not specified. Different algorithms approximate x^* in different ways, from the observations.

- **SCRE**: For the ridges derived from the classical KDE function, the essence is to approximate x^* by the weighted summation as $\hat{x} = \sum_i w_h(x, x_i)x_i$ of all the samples $\{x_i\}$, where the weights decay at an exponential rate with the squared distance. The attraction force in this case is $F(x) = \hat{x} - x$. When fixing the observations, the only parameter h decides the error of $\hat{x} - x^*$. A larger h corresponds to a larger bias, and a smaller one corresponds to a smaller bias but with a less smooth ridge. When $h \to +\infty$, $\hat{x}$ tends to gradually become the equal-weighted center of the observations, and when $h \to 0$, $\hat{x}$ tends to become the nearest neighbor of x, i.e. $\hat{x} = \arg\min_{\{x_i\}} \|x - x_i\|_2$.
- ***l*-SCRE**: l-SCRE approximates x^* by a weighted local summation of samples by $\hat{x} = \sum_{i \in I_r} w_{r,h}(x, x_i)x_i$. The attraction force can also be adopted as the subtraction vector $F(x) = \hat{x} - x$. In the local neighborhood of x, the set $\mathcal{M} \cap B_x(r)$ can be regarded as an affine space $\mathcal{H}$, approximately. As a result, the summation $\hat{x}$ also belongs to the convex hull of $\{x_i\}$, which is also part of $\mathcal{H}$. Notably, when $h \to +\infty$, $\hat{x}$ tends to become the equal-weighted center of the partial observations in $B_x(r)$, and when $h \to 0$, $\hat{x}$ tends to become the nearest neighbor of samples in $B_x(r)$, i.e. $\hat{x} = \arg\min_{i \in \mathcal{I}_r} \|x - x_i\|_2$. Considering the weighted summation of samples in a bounded area will provide a fail-safe measure to ensure that $\hat{x}$ maintains a good quality to approximate x^*.
- **MFIT-i** : Unlike the above strategies, the manifold fitting (MFIT) (Fefferman et al., 2018) constructs the attraction force as a combination of the projection onto the tangent space of each point. The attraction force has an explicit form $F(x) = \sum_{i \in \mathcal{I}_r} \alpha_i(x)\Pi_i^{\perp}(x_i - x)$, where the projection $\Pi_i^{\perp}$ is the estimation of the normal space at point x_i and $\alpha_i(x)$ is the normalized weight corresponding to point x_i, as $\alpha_i(x) = \tilde{\alpha}_i(x)/\sum_i \tilde{\alpha}_i(x)$, with $\tilde{\alpha}_i(x)$ defined as $\tilde{\alpha}_i(x) = (1 - \|x - x_i\|_2^2/r^2)_+^{d+2}$.
- **MFIT-ii:** With similar settings to those of the above method, Yao and Xia propose in Yao and Xia (2019) that the attraction force can also be constructed with only a one-step projection as $F(x) = \sum_{i \in \mathcal{I}_r} \alpha_i(x)(x_i - x)$.

B.2. Regularization Space. Usually, the normal space (the orthogonal complement of the tangent space) is used to act as a constraint, such that the iteration of the mean-shift algorithm performs within a proper subspace. This subspace constraint tries to keep the component in the normal space and leave out or diminish the component of scale in the tangent space. All the normal space is obtained from the eigenvalue decomposition of a temporary matrix (Hessian or Combination of Normal Space).

- **SCRE**: For the ridges obtained from both SCRE and log-SCRE (SCRE transformed with a $\log$ function), we need to compute the second derivative for the function to obtain the Hessian matrix $H(x)$ or the corresponding covariance matrix $J(x)$. Then, the projection can be produced from the eigenvalue decomposition, and we can pick the eigenspace corresponding to the smallest $D - d$ eigenvalues. For SCRE, the covariance matrix is $J(x) = \sum_i w(x_i, x)(x_i - x)(x_i - x)^T$. For log-SCRE, the covariance matrix is $J(x) = \sum_i w(x_i, x)(x_i - c(x))(x_i - c(x))^T$, where $c(x)$ is the one-step shift mean of x.
- ***l*-SCRE**: In our l-SCRE approach, we build the semidefinite local covariance matrix as $C_r(x) = \sum_{i \in \mathcal{I}_r} w_{h,r}(x_i, x)(x_i - c_r(x))(x_i - c_r(x))^T$. The l-SCRE has two main advantages. First, the covariance matrix $C_r(x)$ in the local area of $B_x(r)$ is more similar to a low-rank matrix than one in a global area. As a result, we can easily recover the low-dimensional space to approximate $B_x(r) \cap \mathcal{M}$. Second, because we restrict our consideration to a small domain, the smooth parameter h can be chosen more easily than before.
- **MFIT-i**: Instead of computing the Hessian, the manifold-fitting strategy Fefferman et al. (2018) approximates the normal space at x with a weighted combination of projection matrix $\Pi_i^{\perp}$ as $A = \sum_i \alpha_i \Pi_i^{\perp}$ of the normal projection $\Pi_i^{\perp}$ at each point x_i in the neighborhood. Then, the $D - d$ principal components from the eigenvalue decomposition of A are

	Attraction	Projection Construction
SCRE	$c(x) - x$	$\sum_i w(x, x_i)(x - x_i)(x - x_i)^T$
l-SCRE	$c_r(x) - x$	$\sum_{i \in I_r} w(x, x_i)(c_r(x) - x_i)(c_r(x) - x_i)^T$
MFIT-i	$\sum_{i \in \mathcal{I}_r} \alpha_i(x) \Pi_i^{\perp}(x_i - x)$	$\sum_i \alpha_i \Pi_i^{\perp}$
MFIT-ii	$\sum_{i \in \mathcal{I}_r} \alpha_i(x)(x_i - x)$	$\sum_i \alpha_i \Pi_i^{\perp}$

Ridge-Searching Algorithm:

Input: $\{x_i, i = 1 : N\}$**, mesh of random sample initial points** $\{y_i, i = 1 : s\}$**, tolerance** ϵ
Output: $\{y_i^*, i = 1 : s\}$ **lies approximately on the manifold defined by** $\{x_i, i = 1 : N\}$

Iteration:

0. For each initial point x_i in the mesh:

   Iteration:

   1. Compute the attraction force $F(x)$ and the covariance matrix $J(x)$. Different methods could result in different constructions of $F(x)$ and $J(x)$.
   2. Estimate the tangent space from the covariance matrix $J(x)$ by performing an eigenvalue decomposition:
   $$J(x) = [V_d(x), V_{D-d}(x)]\Lambda(x)[V_d(x), V_{D-d}(x)]$$
   3. Compute the projection operator onto the normal space at $c(x)$ as:
   $$P_{\perp}(c(x)) = V_{D-d}(x)V_{D-d}(x)^T$$
   4. Update x as $\tilde{x} = x + \lambda P_{\perp}(c(x))F(x)$
   5. Check whether $\|\tilde{x} - x\|_2 > \epsilon$; if true, return to step 1 and let $x = \tilde{x}$. Otherwise, return to step 0.

regarded as the projection onto the normal space at x. The benefit of this approach is that the projection $\Pi_i^{\perp}$ at each point x_i is a constant matrix that does not depend on the location of x. However, because it is required to approximate a large number of tangent spaces, this approach needs a considerably large amount of computation resources.

- **MFIT-ii**: The approach for the regularization space in Yao and Xia (2019) is similar to that in **MFIT-i**.

B.3. Algorithm: Subspace-Constraint Iteration. It is difficult to project an outlier x onto a hidden manifold $\mathcal{M}$ that is represented by some observed samples, because we cannot provide an explicit form to exactly represent this projection procedure, $\pi_{\mathcal{M}}(x)$. Since $\pi_{\mathcal{M}}(x)$ is unknown, we can expect to estimate another point x_e, which is not far away from $\pi_{\mathcal{M}}(x)$.

For the projection $\pi_{\mathcal{M}}(x)$, we have $\pi_{\mathcal{M}}(x) - x \perp \mathcal{H}_{\pi_{\mathcal{M}}(x)}$. We would also like to expect $x_e - x \perp \mathcal{H}_{\pi_{\mathcal{M}}(x)}$, so that $x_e - x$ and $\pi_{\mathcal{M}}(x) - x$ reside in the same subspace, which is orthogonal to $\mathcal{H}_{\pi_{\mathcal{M}}(x)}$. As the tangent space $\mathcal{H}_{\pi_{\mathcal{M}}(x)}$ is also unknown, we need to give an estimation of it, denoted by $\mathcal{H}_e$. From our perspective, different manifold-fitting algorithms differ in their approaches to estimate the point x_e and the space of $\mathcal{H}_e$.

Usually, the estimation of x_e and $P_{\mathcal{H}_e}$ is a function of x. When x is far away from $\mathcal{M}$, the estimated x_e and $P_{\mathcal{H}_e}$ are less accurate. As a result, we use a parameter, λ, to adjust the step size. Our uniform manifold-fitting algorithm yields the following form:

$$x_{t+1} = x_t + \lambda P_{\mathcal{H}_e}^{\perp}(x_t)(x_e(x_t) - x_t).$$

The stopping condition is that $\|P_{\mathcal{H}_e}^{\perp}(x_e(x_t) - x_t)\| \leq \epsilon$. When the stopping condition is met, we assume that the current x is approximately on the ridge (or the hidden manifold).

B.4. Complexity Analysis. The computation cost comes mainly from the eigenvalue decomposition, which is $O((D-d)D^2)$ for the least $D-d$ eigenvectors of a $D \times D$ covariance matrix. SCRE and l-SCRE only need to compute the eigenvalue-decomposition problem once for each update. Because the MFIT needs the normal space projection at each sample in the neighborhood, it needs to compute $k+1$ eigenvalue problems for each step. Therefore, the complexity of MFIT is $k+1$ times that of the SCRE method.

SUPPLEMENTARY MATERIAL

Supplement to "Manifold fitting by ridge estimation: a subspace-constrained approach" (Yao and Zhai, 2021). We include all materials of proof omitted from the main text.

REFERENCES

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

ARIAS-CASTRO, E., MASON, D. and PELLETIER, B. (2016). On the Estimation of the Gradient Lines of a Density and the Consistency of the Mean-Shift Algorithm. *Journal of Machine Learning Research* **17** 1-28.

CHEN, Y.-C. (2017). A tutorial on kernel density estimation and recent advances. *Biostatistics & Epidemiology* **1** 161–187.

CHEN, Y.-C. (2020). Solution manifold and Its Statistical Applications. *arXiv preprint arXiv:2002.05297*.

CHEN, Y.-C., GENOVESE, C. R., WASSERMAN, L. et al. (2015). Asymptotic theory for density ridges. *The Annals of Statistics* **43** 1896–1928.

CHEN, Y.-C., GENOVESE, C. R., TIBSHIRANI, R. J., WASSERMAN, L. et al. (2016a). Nonparametric modal regression. *The Annals of Statistics* **44** 489–514.

CHEN, Y.-C., GENOVESE, C. R., WASSERMAN, L. et al. (2016b). A comprehensive approach to mode clustering. *Electronic Journal of Statistics* **10** 210–241.

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

DONOHO, D. L. and GRIMES, C. (2003). Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences* **100** 5591–5596.

FEFFERMAN, C., MITTER, S. and NARAYANAN, H. (2016). Testing the manifold hypothesis. *Journal of the American Mathematical Society* **29** 983–1049.

FEFFERMAN, C., IVANOV, S., KURYLEV, Y., LASSAS, M. and NARAYANAN, H. (2018). Fitting a putative manifold to noisy data. In *Conference On Learning Theory* 688–720.

FEFFERMAN, C., IVANOV, S., KURYLEV, Y., LASSAS, M. and NARAYANAN, H. (2019). Reconstruction and Interpolation of Manifolds. I: The Geometric Whitney Problem. *Foundations of Computational Mathematics* 1–99.

GENOVESE, C. R., PERONE-PACIFICO, M., VERDINELLI, I., WASSERMAN, L. et al. (2012). Manifold estimation and singular deconvolution under Hausdorff loss. *The Annals of Statistics* **40** 941–963.

GENOVESE, C. R., PERONE-PACIFICO, M., VERDINELLI, I., WASSERMAN, L. et al. (2014). Nonparametric ridge estimation. *The Annals of Statistics* **42** 1511–1545.

HASTIE, T. and STUETZLE, W. (1989). Principal curves. *Journal of the American Statistical Association* **84** 502–516.

HAUBERG, S. (2015). Principal curves on Riemannian manifolds. *IEEE transactions on pattern analysis and machine intelligence* **38** 1915–1921.

HUCKEMANN, S., HOTZ, T. and MUNK, A. (2010). Intrinsic shape analysis: Geodesic PCA for Riemannian manifolds modulo isometric Lie group actions. *Statistica Sinica* 1–58.

HUCKEMANN, S. and ZIEZOLD, H. (2006). Principal component analysis for Riemannian manifolds, with an application to triangular shape spaces. *Advances in Applied Probability* **38** 299–319.

JUNG, S., DRYDEN, I. L. and MARRON, J. (2012). Analysis of principal nested spheres. *Biometrika* **99** 551–568.

KASS, R. E. (1989). The Geometry of Asymptotic Inference. *Statistical Science* **4** 188–219.

KLEMELÄ, J. S. (2009). *Smoothing of multivariate data: density estimation and visualization* **737**. John Wiley & Sons.

MOHAMMED, K. and NARAYANAN, H. (2017). Manifold learning using kernel density estimation and local principal components analysis. *arXiv preprint arXiv:1709.03615*.

MYHRE, J. N., SHAKER, M., KABA, D., JENSSEN, R. and ERDOGMUS, D. (2016). Manifold unwrapping using density ridges. *arXiv preprint arXiv:1604.01602*.

OZERTEM, U. and ERDOGMUS, D. (2011). Locally defined principal curves and surfaces. *Journal of Machine learning research* **12** 1249–1286.

PANARETOS, V. M., PHAM, T. and YAO, Z. (2014). Principal flows. *Journal of the American Statistical Association* **109** 424–436.

PATRANGENARU, V. and ELLINGSON, L. (2015). *Nonparametric Statistics on Manifolds and Their Application to Object Data Analysis*.

QIAO, W. (2020). Asymptotic Confidence Regions for Density Ridges. *arXiv preprint arXiv:2004.11354*.

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

S. A. NENE, S. K. N. and MURASE, H. (1996). *Columbia Object Image Library (COIL-20)*.

SASAKI, H., KANAMORI, T., HYVÄRINEN, A., NIU, G. and SUGIYAMA, M. (2017). Mode-seeking clustering and density ridge estimation via direct estimation of density-derivative-ratios. *The Journal of Machine Learning Research* **18** 6626–6672.

YAO, Z. and XIA, Y. (2019). Manifold Fitting under Unbounded Noise. *arXiv preprint arXiv:1909.10228*.

YAO, Z. and ZHAI, Z. (2021). Supplementary material for "Manifold fitting by ridge estimation: a subspace-constrained approach".

ZHA, H. and ZHANG, Z. (2007). Continuum Isomap for manifold learnings. *Computational Statistics & Data Analysis* **52** 184–200.

ZHANG, Z. and ZHA, H. (2004). Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *SIAM journal on scientific computing* **26** 313–338.