

KERNEL DENSITY ESTIMATION ON EMBEDDED MANIFOLDS WITH BOUNDARY

BY TYRUS BERRY AND TIMOTHY SAUER

We consider practical density estimation from large data sets sampled on manifolds embedded in Euclidean space. Existing density estimators on manifolds typically require prior knowledge of the geometry of the manifold, and all density estimation on embedded manifolds is restricted to compact manifolds without boundary. First, motivated by recent developments in kernel-based manifold learning, we show that it is possible to estimate the density on an embedded manifold using only the Euclidean distances between the data points in the ambient space. Second, we extend the theory of local kernels to a larger class of manifolds which includes many noncompact manifolds. This theory reveals that estimating the density without prior knowledge of the geometry introduces an additional bias term which depends on the extrinsic curvature of the embedding. Finally, we develop a boundary correction method that does not require any prior knowledge of the location of the boundary. In particular, we develop statistics which provably estimate the distance and direction of the boundary, which allows us to apply a cut-and-normalize boundary correction. By combining multiple cut-and-normalize estimators we introduce a consistent kernel density estimator that has uniform bias on manifold and boundary.

1. Introduction. Nonparametric density estimation has become an important tool in statistics with a wide range of applications to machine learning, especially for high-dimensional data. The increasing size and complexity of measured data creates the possibility of understanding increasingly complicated phenomena for which there may not be sufficient ‘first principles’ understanding to enable effective parametric modeling. The exponential relationship between model complexity (often quantified as dimensionality) and data requirements, colloquially known as the *curse of dimensionality*, demands that new and innovative priors be developed. A particularly effective assumption is the *geometric prior*, which assumes that the data lies on a manifold that is embedded in the ambient Euclidean space where the data is sampled. The geometric prior is nonparametric in that it does not assume a particular manifold or parametric form, merely that the data is restricted to lying on *some* manifold. This prior allows us to separate the *intrinsic* dimensionality of the manifold, which may be low, from the *extrinsic*

Keywords and phrases: kernel density estimation, manifold learning, boundary correction, geometric prior

dimensionality of the ambient space which is often high.

Recently the geometric prior has received some attention in the density estimation field [12, 23, 17, 21], although use of these methods remains restricted for several reasons. For example, the methods of [12, 23] require the structure of the manifold to be known *a priori*. However, in the applied harmonic analysis literature a method known as *diffusion maps* has been introduced which learns the structure of the manifold from the data [1, 7]. In this article, we use these results and generalizations to a broader class of kernel functions [4] to define practical algorithms for kernel density estimation on embedded manifolds which do not require any prior knowledge about the manifold.

A second restriction of the current literature (including the results in [7, 4, 21]) is the assumption that the manifold be compact. The new proofs presented here yield more natural assumptions on the geometry of the embedded manifold. These new assumptions include all compact manifolds, but also allow many non-compact manifolds, such as any linear manifold, which implies that standard kernel density estimation theory on Euclidean spaces is included a special case. For ease of exposition, we will assume the dimension of the manifold is known, although this is not necessary: In Appendix C we include a practical method of empirically tuning the bandwidth parameter that also estimates the dimension.

A third, and perhaps most significant limitation of applying existing manifold density estimators to real problems, is the restriction to manifolds without boundary. One exception is the special case of subsets of the real line where the location of the boundary is assumed to be known. This case has been thoroughly studied, and consistent estimators have been developed [13, 14, 27, 6, 15, 20].

In the following, we introduce a consistent kernel density estimator for manifolds with (unknown) boundary that has the same asymptotic bias in the interior as on the boundary. The first obstacle to such an estimator is that a conventional kernel does not integrate to one near the boundary. Therefore the normalization factor must be corrected in a way that is based on the distance to the boundary, which is not known *a priori*.

To locate the boundary, we couple the standard kernel density estimator (KDE) with a second calculation, a kernel weighted average of the vectors from every point in the data set to every other point, which we call the boundary direction estimator (BDE). We present asymptotic analysis of the BDE that shows that if the base point is near a boundary, the negative of the resulting average vector will point toward the nearest point on the boundary. We also use the asymptotic expansion of this vector to find a

lower bound on the distance to the boundary. Our new density estimate at this base point does not include the data which lie beyond the lower bound in the direction of the boundary. This creates a virtual boundary in the tangent space which is simply a hyperplane (dimension one less than the manifold) at a known distance from the base point. Creating a known virtual boundary allows us to bypass the above obstacle – we can now renormalize the kernel so that it integrates exactly to one at each base point, similar to the cut-and-normalize kernels that are used when the boundary is *a priori* known. For points in the interior (or for manifolds without boundary), the lower bound on the distance to the boundary goes to infinity in the limit of large data, and we recover the standard kernel density estimation formula. Moreover, using standard methods of constructing higher order kernels, we find a formula for a kernel density estimate with the same asymptotic bias for interior points and points on the boundary.

In Section 2 we place our results in the context of the long history of nonparametric density estimation by reviewing the bridge between KDE theory and a growing literature on the geometric prior in the applied harmonic analysis community. Indeed, the deep connections between KDE and sampling theory were realized as early as 1958 in [33], and these connections continue to be explored today. In Section 3 we extend the results of local kernel theory introduced in [4] to a larger class of embedded manifolds which includes many non-compact manifolds such as Euclidean spaces. The fundamental strategy of [4] evolved from [7] and its generalization to anisotropic kernels in [30] and the new proofs in Section 3 are motivated by the elegant methods of [23, 17]. The boundary correction method using BDE is introduced in Section 4, and the results are demonstrated on several illuminating examples. We conclude with a brief discussion in Section 5.

2. Background. Assume one is given N samples $\{X_i\}_{i=1}^N$ (often assumed to be independent) of a probability distribution on R^n with a density function $f(x)$. The problem of nonparametric density estimation is to find an estimator $f_N(x)$ that approximates the true density function. A fundamental result of [26] shows that such an estimator $f_N(x)$ cannot be unbiased for finite N , so the focus has been on finding *consistent* estimators that satisfy $\lim_{N \rightarrow \infty} f_N(x) = f(x)$, called *asymptotically unbiased* estimators. (This result does have an exception in the case that the density is band limited, but this is not a common assumption.) The estimator f_N is typically constructed

as

$$(1) \quad f_N(x) = \frac{1}{N} \sum_{i=1}^N K_N(x, X_i)$$

where K_N is a smooth function of two variables known as the kernel function, motivating the name kernel density estimation (KDE). The formulation (1) was introduced for univariate densities in [22] and generalized to multivariate densities in [19]. It was shown in [33] that as $N \rightarrow \infty$ the kernel must converge to the delta function $\delta(x - X_i)$ in order for f_N to be a consistent estimator, and [33] also motivated the symmetric formulation,

$$(2) \quad f_N(x) = \frac{1}{Nh_N^n} \sum_{i=1}^N K\left(\frac{x - X_i}{h_N}\right)$$

where the kernel function is defined via a univariate shape function K_N and $h_N \rightarrow 0$ as $N \rightarrow \infty$. In both (1) and (2) the kernel function must be normalized to integrate to 1 for each N to have a consistent estimator. An alternative estimator can be formulated in terms of Fourier functions, which was shown to be fundamentally equivalent (via Mercer's theorem) in [33].

The first analysis of the bias and variance of the estimator (2) was conducted in [22] for univariate densities and generalized to multivariate densities in [5]. Many generalizations have been proposed, such as variable bandwidth kernels that allow h_N to also depend on i , and higher order kernels which have asymptotically smaller bias; an excellent overview can be found in [31]. The fundamental result presented in [31] is that the asymptotic bias of the estimator is $\mathcal{O}(h^p)$ where p is the order of the kernel and the asymptotic variance is $\mathcal{O}(N^{-1}h^{-n})$. Intuitively, by shrinking the bandwidth h , our kernel becomes closer to a delta function, which decreases the bias but increases the variance error. Balancing the variance and the squared bias requires $h = \mathcal{O}\left(N^{-\frac{1}{n+2p}}\right)$ and this results in an average mean squared error (AMSE) of $\mathcal{O}\left(N^{-\frac{p}{n+p}}\right)$. This clearly illustrates the curse of dimensionality, since controlling this error as the ambient dimension n increases requires exponential growth in the amount of data N .

One possibility for alleviating the curse of dimensionality is to increase the order p of the kernel [28]. Of course, increasing the order of the kernel places additional smoothness constraints on the density. In fact, it is possible to design kernels of infinite order for densities with sufficiently fast decay in the Fourier domain [24]. Common constructions of higher-order kernels often allow the possibility of negative density estimates, which is a significant

weakness, although this problem was solved by an alternative formulation in [32], which guarantees positive estimates.

The geometric prior is an alternative solution to the curse of dimensionality. Since the geometric prior assumes that the density is supported on a submanifold of Euclidean space, we may assume that the intrinsic dimensionality is small even when the extrinsic dimensionality is large. Nonparametric density estimation on manifolds essentially began with Hendriks [12], who modernized the Fourier approach of [33] using a generalized Fourier analysis on compact Riemannian manifolds without boundary, based on the eigenfunctions of the Laplace-Beltrami operator. The limitation of [12] in practice is that it requires the eigenfunctions of the Laplace-Beltrami operator on the manifold to be known, which is equivalent to knowing the entire geometry. A kernel-based method of density estimation was introduced in [23]. In this case the kernel was based on the geodesic distance between points on the manifold, which is again equivalent to knowing the entire geometry.

More recently, a method which uses kernels defined on the tangent space of the manifold was introduced [17]. However, evaluating the kernel of [17] requires lifting points on the manifold to the tangent space via the exponential map, which yet again is equivalent to knowing the geometry of the manifold. (See, for example, [25] which shows that the Riemannian metric can be recovered from either the Laplace-Beltrami operator, the geodesic distance function, or the exponential map.) The results of [12, 23, 17], in addition to being restricted to compact manifolds without boundary, are limited to manifolds which are known a priori, and cannot be applied to data lying on an unknown manifold embedded in Euclidean space.

The insight of [7] was that as the bandwidth h decreases and the kernel function approaches a delta function, the kernel is essentially zero outside a ball of radius h . Inside this ball, the geodesic distance on the embedded manifold and the Euclidean distance in the ambient space are equal up to an error which is higher order in h . This fact follows directly for compact manifolds. Although it is not true for general manifolds, a weaker condition than compactness is sufficient as we will show below. The equivalence of the ambient and geodesic distances on small scales suggests that for kernels with sufficiently fast decay at infinity, the approaches of [23, 17, 21] are equivalent. This fact first came to light in [7], although the focus was on estimating the Laplace-Beltrami operator on the unknown manifold, so the authors did not emphasize their density estimate result or analyze the variance of their estimate. The fact was later pointed out in [21], where the bias and variance of the kernel density estimate were computed.

The results of [7] were restricted to shape function kernels of the form

$K\left(\frac{\|x-X_i\|}{h_N}\right)$, and were later generalized to anisotropic kernels in [30] and then to local kernels in [4]. Local kernels can have an arbitrary functional form $K_h(x, y)$ and require only that the kernel is *local* in the sense that the kernel decays exponentially as y moves away from x – explicitly we require $K_h(x, x + hz) \leq a \exp(-b\|z\|)$ for some $a, b > 0$. The key result is that the moments of the kernel combine with the Riemannian metric inherited from the embedding, to define an intrinsic geometry of the kernel on the manifold. In [4], local kernels were used to estimate the Laplace-Beltrami operator with respect to this intrinsic geometry. For density estimation, we will see that the density estimate will be written relative to the volume form of the intrinsic geometry. We note that the class of local kernels includes all compactly supported kernels, such as the Epanechnikov kernel [9] and other similar kernels which are often used in density estimation. In Section 3, we generalize the results of [4] to a large class of manifolds and present significantly simplified proofs motivated by [23, 17].

The topic of kernel density estimation near a boundary has been thoroughly explored in case the distribution is supported on a subinterval $[b, \infty]$ of the real line, and with the assumption that b is known. Using a naive kernel results in an estimate that is not even consistent near b . An early method that achieved consistency on the boundary was the cut-and-normalize method [10], although the bias was only order h on the boundary, despite being order h^2 in the interior. Various alternatives were proposed to obtain bias uniform over the domain and boundary. These methods include reflecting the data [29, 16], generalized jackknifing [13, 14, 27, 18], translation-based methods [11], and the use of specially-designed asymmetric kernels from the beta and gamma distributions [6, 15, 20]. The cut-and-normalize method was extended to order h^2 on the boundary in [18]. The goal of Section 4 is to generalize the cut-and-normalize approach to an order h^2 method, including boundary and interior, on embedded manifolds where the location of the boundary is not known.

3. KDE on Embedded Manifolds. In this section we extend the results of [23, 17] to the case of an embedded manifold without boundary where the geometry is unknown. Given data $\{x_i\}_{i=1}^N \subset \mathbb{R}^n$ sampled from a probability distribution with density $\tilde{f} : \mathbb{R}^n \rightarrow [0, \infty)$, we say that the density satisfies the geometric prior if the set $\mathcal{M} = \{x \in \mathbb{R}^n : \tilde{f}(x) > 0\}$ is a differentiable manifold. Notice that the data points are constrained to the manifold by definition, and this is the key to manifold learning following [7, 4]. The dimension m of \mathcal{M} will be called m the intrinsic dimension and n the extrinsic dimension of the density (therefore $m \leq n$). For the purposes

of exposition, we will assume m is known (however, see Appendix C for an effective empirical method of obtaining m).

Since \tilde{f} is supported on a set of Lebesgue measure zero when $m < n$, it makes more sense to consider the density $f : \mathcal{M} \rightarrow (0, \infty)$ which defines a probability distribution on the manifold \mathcal{M} . We assume that f is written with respect to the volume form dV of the Riemannian metric that \mathcal{M} inherits from the ambient space (this metric is simply the Euclidean metric on the tangent spaces considered as subspaces of the ambient Euclidean space). The assumption that the data points x_i are realizations of random variables X_i sampled from f yields

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N K(x, x_i) = \mathbb{E}[K(x, X_i)] = \int_{\mathcal{M}} K(x, y) f(y) dV(y)$$

for any kernel function $K(x, \cdot) \in L^2(\mathcal{M}, f dV)$.

In order to find the correct density $f(x)$ using a kernel density estimate we will require assumptions on both the manifold \mathcal{M} and the kernel K . We will assume the kernel function $K : [0, \infty) \times \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ has the form $K(h, x, y)$ where the first parameter $h \in [0, \infty)$ is the bandwidth of the kernel. We define the moments of the kernel by

$$(3) \quad m_{\vec{\alpha}}(x) = \lim_{h \rightarrow 0} \int_{\mathbb{R}^m} \left(\prod_{j=1}^m z_j^{\alpha_j} \right) K(h, x, x + hz) dz$$

where $|\alpha| \equiv \sum_{j=1}^m \alpha_j$ is the order of the moment. For convenience we also define the covariance matrix $C(x) \in \mathbb{R}^{m \times m}$ of a kernel by

$$C_{ij}(x) = \lim_{h \rightarrow 0} \int_{\mathbb{R}^m} z_i z_j K(h, x, x + hz) dz,$$

since this will be a particularly important moment.

DEFINITION 3.1. *We say that K is a local kernel if all the moments of K are finite.*

This definition of a local kernel is more general than that of [4]. Intuitively, a local kernel must decay as $z = (y - x)/h$ becomes large, and this decay must be faster than any polynomial in the components of z . In particular, any kernel which is dominated by an exponential function $K(h, x, y) < a \exp(-b||x - y||/h)$ for some $a, b > 0$ will be a local kernel. This includes any kernel which is compactly supported in y for all x .

In [4], it was shown that a local kernel defines an *intrinsic geometry* on the manifold \mathcal{M} that depends on the geometry inherited from the ambient space and the normalized covariance matrix $\hat{A}(x) \equiv C(x)/m_0(x)$ of the local kernel. The theory of [4] was restricted to compact manifolds, so in Appendix A we extend those results to the wider class of manifolds described below with a significantly simplified proof. It was also shown in [4] that we can access any geometry on \mathcal{M} with a *prototypical kernel* of form

$$(4) \quad \begin{aligned} K_A(h, x, y) &= \pi^{-m/2} \exp\left(-\frac{(y-x)^T A(x)^{-1}(y-x)}{h^2}\right) \\ &= \pi^{-m/2} \exp\left(-\left\|A(x)^{-1/2}\left(\frac{y-x}{h}\right)\right\|^2\right) \end{aligned}$$

where $A(x)$ is a symmetric positive definite $n \times n$ matrix. The zeroth moment $m_0(x)$ is not typically known for anisotropic kernels. For the prototypical kernel it requires computing the determinant of $\hat{A}(x)$ which is the restriction of $A(x)$ to the tangent space (see Appendix A for details). While $A(x)$ is typically known, we do not want to assume that the tangent spaces are known a priori. So in practice, most density estimation will use an isotropic kernel, where the covariance matrix is a known multiple of the identity matrix. Local kernels can be used to estimate the density relative to different volume forms on the manifold, which is discussed in Appendix A. However, most applications are interested in estimating the density relative to the volume form that \mathcal{M} inherits from the ambient space \mathbb{R}^n , so we restrict our attention to this case.

Having established the necessary assumptions on the kernel function, we turn to the requirements for the manifold. Recall that the exponential map $\exp_x : T_x \mathcal{M} \rightarrow \mathcal{M}$ maps a tangent vector \vec{s} to $\gamma(\|\vec{s}\|)$ where γ is the arclength-parametrized geodesic starting at x with initial velocity $\gamma'(0) = \vec{s}/\|\vec{s}\|$. The injectivity radius $\text{inj}(x)$ of a point x is the maximum radius for which a ball in $T_x \mathcal{M}$ is mapped diffeomorphically into \mathcal{M} by \exp_x . In order to convert integrals over the entire manifold into integrals over the tangent space, we will use the decay of the kernel to localize the integral and then change variables using the exponential map. This requires that for a sufficiently small localization region (meaning h sufficiently small) the exponential map is a diffeomorphism. Therefore, the first requirement for kernel density estimation will be that the injectivity radius is non-zero.

The second requirement is that the ratio

$$R(x, y) = \frac{\|x - y\|}{d_I(x, y)}$$

is bounded away from zero for y sufficiently close to x , where $\|x - y\|$ is the Euclidean distance and $d_I(x, y)$ is the intrinsic distance, which is defined as the infimum of the lengths of all differentiable paths connecting x and y . When some path attains this infimum it is called a geodesic path and the distance is called the geodesic distance $d_g(x, y)$. We use the intrinsic distance since it is defined for all pairs of points, whereas the geodesic distance may technically be undefined when there is no path that attains the infimum. The reason we will require $R(x, y)$ to be bounded away from zero is that the local kernel is defined in the ambient space, which makes it practical to implement. But the theory requires that the kernel decays exponentially in the geodesic distance, meaning that the kernel is localized on the manifold, not just the ambient space. (The kernels of [23, 17] explicitly depend on the geodesic distance in order to obtain this decay.)

In order to estimate the density f at a point $x \in \mathcal{M}$ we require the injectivity radius $\text{inj}(x)$ to be non-zero and the ratio $R(x, y)$ to be bounded away from zero near x , which motivates the following definition.

DEFINITION 3.2. *We say that a point $x \in \mathcal{M} \subset \mathbb{R}^n$ is tangible if $\text{inj}(x) > 0$ and within a sufficiently small neighborhood N of x , $\inf_{y \in N} R(x, y) > 0$.*

We are mainly interested in manifolds for which every point is tangible.

DEFINITION 3.3. *An embedded manifold $\mathcal{M} \subset \mathbb{R}^n$ is tangible if every $x \in \mathcal{M}$ is tangible. If there exist lower bounds for $\text{inj}(x)$ and $\inf_{y \in \mathcal{M}} R(x, y)$ that are independent of x , then \mathcal{M} is called uniformly tangible.*

For example, every compact manifold as well as linear manifolds such as \mathbb{R}^n are uniformly tangible. This implies that standard KDE theory on Euclidean spaces as well as existing density estimation on manifolds are included in our theory. In addition to unifying these previous KDE theories, our theory applies to the large class of noncompact uniformly tangible manifolds.

An example where uniform tangibility fails is the 1-dimensional manifold in \mathbb{R}^2 given by $(r(\theta) \cos \theta, r(\theta) \sin \theta)$ where $r(\theta) = 1 - 1/\theta$ and $\theta \in [1, \infty)$. Then for any $\theta \in [1, \infty)$, set $\theta_n = \theta + 2\pi n$. The distances $d_g(\theta_n, \theta_{n+1})$ approach 2π as $n \rightarrow \infty$, whereas $\|\theta_n - \theta_{n+1}\|$ goes to zero. Thus the ratio $R(\theta_n, \theta_{n+1})$ is not uniformly bounded below on the manifold. However, every point on this manifold is tangible.

We can now state our main result, which proposes a practical algorithm for KDE that does not require any knowledge of the embedded manifold. The following corollary to Theorem A.1 in Appendix A allows kernel density

estimation in the spirit of [23, 17] but using a kernel defined on the ambient space without assuming that we know the geometry of \mathcal{M} . As mentioned above, we intuitively expect these kernels to be equivalent since a local kernel depends only on nearby points for which asymptotically the Euclidean and geodesic distances are equal up to an error of higher order in h .

COROLLARY 3.4 (Isotropic Kernels). *Let \tilde{f} be a density supported on an m -dimensional tangible manifold $\mathcal{M} \subset \mathbb{R}^n$ without boundary. Let $\tilde{f} = f dV$ where dV is the volume form on \mathcal{M} inherited from the embedding and $f \in \mathcal{C}^2(\mathcal{M})$ is bounded above by a polynomial. Let $K : [0, \infty) \times \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ be a local kernel with zeroth moment $m_0(x)$, second moment $C_{ij}(x) = m_0(x)\delta_{ij}$, and first and third moments equal to zero. If X_i are independent samples of f , then*

$$f_{h,N}(x) \equiv \frac{1}{Nm_0(x)h^m} \sum_{i=1}^N K(h, x, X_i)$$

is a consistent estimator of $f(x)$ with bias

$$\mathbb{E}[f_{h,N}(x) - f(x)] = \frac{h^2}{2}(\Delta f(x) + \frac{1}{3}f(x)(R(x) + \tilde{R}(x))) + \mathcal{O}(h^4)$$

where Δ is the Laplace-Beltrami operator with respect to the metric that \mathcal{M} inherits from the ambient space, $R(x)$ is the scalar curvature and $\tilde{R}(x)$ is the trace of the extrinsic curvature in Theorem A.1. The variance of the estimator is

$$\text{var}(f_{h,N}(x) - f(x)) = \frac{h^{-m}}{N} \frac{m_0^2(x)}{m_0(x)} f(x) + \mathcal{O}(1/N)$$

where $m_0^2(x)$ is the zeroth moment of K^2 .

We note that in [17] the term $\tilde{R}(x)$ does not appear, since their kernel was defined exactly on the tangent space, whereas here the bias depends on the extrinsic curvature due to the fact that our kernel is defined in the ambient space, which only approximates the tangent space in the limit $h \rightarrow 0$. For isotropic kernels the zeroth moment $m_0(x)$ is typically known, so one can easily use a local kernel to find the sampling measure with respect to the intrinsic geometry as in Corollary 3.4, or one can divide by $m_0(x)$ as in Corollary A.2 to obtain the sampling measure with respect to the geometry inherited from the embedding.

Corollary 3.4 shows that KDE is straightforward when X_i are random variables sampled according to a density $f(x)$ on an embedded manifold

with no boundary. In particular, when the goal is to find the density $f(x)$ written with respect to the volume form inherited from the embedding, using the result of Corollary 3.4 we have

$$(5) \quad f_h(x) \equiv \lim_{N \rightarrow \infty} \frac{1}{Nm_0(x)h^m} \sum_{i=1}^N K(h, x, X_i) = f(x) + \mathcal{O}(h^2).$$

Moreover, this holds for any local kernel K with zeroth moment $m_0(x)$ and covariance $C_{ij}(x) = m_0(x)I_{m \times m}$. For example, one may use the prototypical kernel

$$K(h, x, y) = \pi^{-m/2} \exp\left(-\frac{\|y - x\|^2}{h^2}\right),$$

which has $m_0(x) = 1$ and $C_{ij}(x) = I_{m \times m}$. Notice that (5) is simply a standard KDE formula in \mathbb{R}^n , except that h^m appearing in the denominator would normally be h^n . Intuitively, this is because the data lies on an m -dimensional subspace of the n -dimensional ambient space; so the true dimension is m . Similarly, the zeroth moment $m_0(x)$ depends on the intrinsic dimension m rather than the ambient space dimension n .

The fundamental idea of a practical KDE on embedded manifolds is that in the limit of small h , the kernel is localized in a very small region $\|y - x\| < h$, and in this region the geodesic distance (as used in [23]) is equal to the ambient Euclidean distance up to an error of higher order in h . Equivalently, the KDE summation approximates an integral over the local region $y \in N_h(x)$, which for h small is very close to an integral over the entire tangent space $T_x\mathcal{M}$ (as used in [17]). Of course, there is a price to pay for using the readily available Euclidean distance instead of the much more difficult geodesic distance. That price is the additional bias term $\tilde{R}(x)$, which is an extrinsic curvature measure that arises from the change of coordinates from the ambient Euclidean coordinates to the geodesic normal coordinates inside the kernel. By assuming the geometry of the manifold is known *a priori*, this change of coordinates is avoided in [23, 17] and their estimators do not contain this additional bias term. Of course, when the structure of the manifold is known, and the geodesic distance or geodesic coordinates can be explicitly computed, one should use these more accurate kernels. However, in applications where the geometry is not known beforehand, the slight additional bias of our method is likely to be unavoidable.

Corollary 3.4 allows us to estimate the density $f(x)$ for any tangible point $x \in \mathcal{M}$ for h sufficiently small and the number of points N sufficiently large. However, we are often interested in estimating the density at each of the sample points x_i using all the other sample points. For this to be practical,

we require the manifold to be uniformly tangible, otherwise as the number of data points increases, the lower bounds may decrease sufficiently quickly that we cannot take h to zero as $N \rightarrow \infty$.

EXAMPLE 3.5 (Density Estimation on Embedded Circle). In this example we demonstrate the difference between the methods of [23, 17] which assume the geometry is known, and our KDE for embedded manifolds. We use the simplest compact manifold without boundary, which is a circle parameterized by $\theta \in [0, 2\pi)$. We first sampled the circle by generating 20000 uniformly random points in $[0, 2\pi)$, so the initial sampling density is $q(\theta) = \frac{1}{2\pi}$. We then used the rejection method to generate samples of the density $f(\theta) = \frac{1}{4\pi}(2 + \sin(\theta))$. We define $M = \max_{\theta} \{f(\theta)/q(\theta)\} = 3/2$ and for each data point θ_i generate a uniform random number $\xi \in [0, 1]$. If $f(\theta_i)/q(\theta_i) \geq M\xi$ we accept θ_i as a sample of f and otherwise we reject the data point. After rejection sampling we were left with 13217 independent samples of f . The methods of [23, 17] use a kernel based on the geodesic distance between points, which is independent of the embedding. Since we know that the manifolds under consideration will be isometric embeddings of the unit circle, we can compute the geodesic distance

$$d_g(\theta_1, \theta_2) = \min\{(\theta_1 - \theta_2) \bmod 2\pi, (\theta_2 - \theta_1) \bmod 2\pi\}$$

and the density estimate of [23, 17] is

$$\hat{f}_{h,N}(\theta_j) = \frac{1}{Nh\sqrt{\pi}} \sum_{i=1}^N e^{-d_g(\theta_i, \theta_j)^2/h^2}.$$

On the other hand, our method does not assume that the geometry is known, so the geodesic distance is unavailable. Instead, we assume only an embedding of the manifold into Euclidean space and rely on the Euclidean distances to estimate the density. In this example we consider the family of isometric embeddings of the unit circle into \mathbb{R}^4 given by

$$x(\theta) = \frac{1}{\sqrt{1+k^2}} (\sin(\theta), \cos(\theta), \sin(k\theta), \cos(k\theta))^{\top}$$

which is isometric for any k since $Dx^{\top}Dx = 1$. We then compute the standard KDE for embedded manifolds,

$$f_{h,N}(\theta_j) = \frac{1}{Nh\sqrt{\pi}} \sum_{i=1}^N e^{-||x(\theta_i) - x(\theta_j)||^2/h^2}.$$

By varying k we can now demonstrate the influence of the extrinsic curvature on the KDE for embedded manifolds, which does not affect the density estimates of [23, 17]. Of course, in practice we will not have natural coordinates such as θ on our data set, so normally we consider $f_{h,N}$ to be a function of the embedded data points $x_j = x(\theta_j)$. We write everything in terms of θ in order to clearly and easily compare the two density estimates.

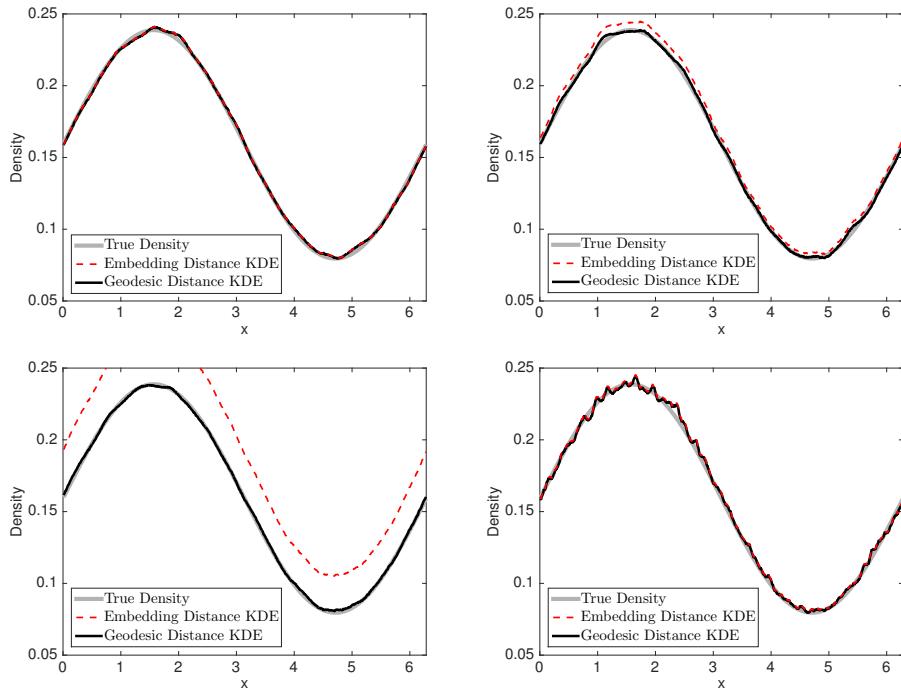


FIG 1. Comparison of the true density f (gray, solid) with the intrinsic density estimate $\hat{f}_{h,N}$ (black, solid) and the embedding KDE $f_{h,N}$ (black, dashed). Results are averaged over 20 experiments to illustrate the bias error (variance error is averaged out). Top, Left: $k = 1, h = \sqrt{0.02}$ Top, Right: $k = 4, h = \sqrt{0.02}$, Bottom, Left: $k = 5, h = \sqrt{0.02}$, Bottom, Right: $k = 5, h = \sqrt{0.002}$.

In Figure 1 we clearly demonstrate the additional bias $\tilde{R}(x)$ which results from the extrinsic curvature of the embedding. For fixed $h = \sqrt{0.02}$ the bias is small for $k \leq 4$ but when $k = 5$ the bias becomes very large. We also show that this additional bias can be alleviated by taking a smaller $h = \sqrt{0.002}$. Of course, for practical density estimation, decreasing h will lead to larger variance error and hence will require more data. This shows the tradeoff between the kernels of [23, 17], which require the structure of the manifold

to be known, and the embedded KDE method, which does not require this information but has a larger bias.

4. Boundary Correction. The standard KDE formula (5) fails for manifolds with boundary because the domain of integration is no longer symmetric near the boundary. For a point x on the boundary, $\partial\mathcal{M}$, the integral over the local region $N_h(x)$ approximates the integral over the half space $T_x\mathcal{M} \cong \mathbb{R}^{m-1} \oplus \mathbb{R}^+$. The zeroth moment of local kernel $m_0(x)$ is defined to be the integral over \mathbb{R}^m , so dividing by this normalization constant will lead to a biased estimator even in the limit $h \rightarrow 0$. While technically the estimator is still asymptotically unbiased for all interior points, for fixed h the additional bias from using the incorrect normalization constant can be quite large for points within geodesic distance h of the boundary.

To fix the bias, we need to estimate the distance b_x and direction η_x to $\partial\mathcal{M}$ for every point x in \mathcal{M} . Our motivation is that if they are known, Theorem 4.1 below gives a consistent estimate of $f(x)$ both in the interior and the boundary. Next, we compute three more variants of the KDE computation (5) to estimate b_x and η_x , and to extend the second-order estimate of $f(x)$ everywhere. Figure 2 shows the proposed workflow.

First, in section 4.1 we compute the boundary direction estimator (BDE) denoted

$$(6) \quad \mu_{h,N}(x) \equiv \frac{1}{Nh^{m+1}} \sum_{i=1}^N K(h, x, X_i)(X_i - x).$$

The BDE is sensitive to the presence of the boundary, and we will combine the KDE (5) with the BDE (6) to derive estimates of b_x and η_x .

Second, with b_x and η_x known, in section 4.2 we approximate $\partial\mathcal{M}$ as a hyperplane in the tangent space to more accurately normalize a cut-and-normalize kernel denoted $f_{h,N}^c$. Third, section 4.3 repeats the cut-and-normalize kernel, with bandwidth $2h$, so that Richardson extrapolation can be used to decrease the order of the error of $f(x)$ to $O(h^2)$ at points x up to and including the boundary $\partial\mathcal{M}$.

4.1. Distance and Direction to the Boundary. Correcting the bias of the standard KDE (5) near the boundary requires computing the true zeroth moment of the kernel,

THEOREM 4.1 (KDE near the Boundary). *Under the same hypotheses as Corollary 3.4 except with \mathcal{M} a manifold with boundary $\partial\mathcal{M}$, let $x \in \mathcal{M}$ and let b_x be the geodesic distance to the boundary, and let $\eta_x \in T_x\mathcal{M}$ be a*

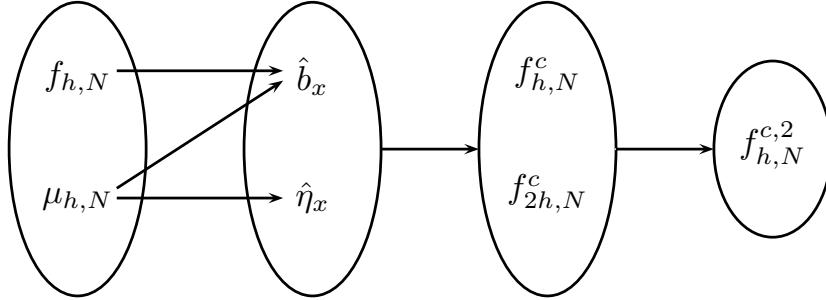


FIG 2. Workflow schematic. At each point x , the standard KDE (5) $f_{h,N}$ is combined with the boundary direction estimator BDE (6) $\mu_{h,N}$ to estimate the distance b_x to $\partial\mathcal{M}$. Cut-and-normalize estimators $f_{h,N}^c$ and $f_{2h,N}^c$ are calculated and combined to get the second-order estimate $f_{h,N}^{c,2}$.

unit vector in the direction of boundary. Then for h sufficiently small, η_x is well defined and

$$f_{h,N}^\partial(x) \equiv \frac{1}{Nm_0^\partial(x)h^m} \sum_{i=1}^N K(h, x, X_i)$$

is a consistent estimator of $f(x)$. Here

$$m_0^\partial(x) = \int_{\mathbb{R}^{m-1}} \int_{-\infty}^{b_x/h} K(h, x, x + hz_\perp + hz_\parallel \eta_x) dz_\parallel dz_\perp$$

where $z_\perp \perp \eta_x$ and z_\parallel is a scalar. Moreover, $f_{h,N}^\partial(x)$ has bias

$$(7) \quad \mathbb{E} [f_{h,N}^\partial(x) - f(x)] = hm_1^\partial(x)\eta_x \cdot \nabla f(x) + \mathcal{O}(h^2)$$

and variance

$$\text{var} (f_{h,N}^\partial(x) - f(x)) = \frac{h^{-m}}{N} \frac{m_0^{2,\partial}(x)}{m_0^\partial(x)} f(x) + \mathcal{O}(1/N)$$

The proof of Theorem 4.1 is in Appendix B. Intuitively Theorem 4.1 says that finding a consistent estimator of $f(x)$ for points near the boundary requires correcting the zeroth moment m_0 . For interior points, the zeroth moment is the integral of the kernel over the entire tangent space, but for boundary points, the integral only extends to the boundary. Since we choose an orientation with η_x pointing towards the boundary (for boundary points

η_x is the unit normal vector), the integral over z_{\parallel} extends infinitely in the negative direction (into the interior of the manifold) but only up to b_x/h in the positive direction (toward the boundary). One should think of $hz_{\parallel}\eta_x$ being a tangent vector which extends up to b_x , which explains why z_{\parallel} extends to b_x/h . Finally, notice that for $dM(x) \gg h$, the zeroth moment $m_0^{\partial}(x)$ reduces to the zeroth moment $m_0(x)$ for manifolds without boundary up to an error of higher order in h due to the decay of the kernel. This shows that the estimator of Corollary 3.4 is consistent for all interior points. However, for a fixed h the bias will be significantly larger using the estimator of Corollary 3.4 than for the estimator of Theorem 4.1 for points with $b_x \leq h$.

For general local kernels, the formula for $m_0^{\partial}(x)$ can be very difficult to evaluate near the boundary. One solution is to apply an asymptotic expansion in b_x/h , for example,

$$\begin{aligned} m_0^{\partial}(x) &= \int_{\mathbb{R}^{m-1}} \int_{-\infty}^0 K(h, x, x + hz_{\perp} + hz_{\parallel}\eta_x) dz_{\parallel} dz_{\perp} \\ &\quad + \frac{b_x}{h} \int_{\mathbb{R}^{m-1}} K(h, x, x + hz_{\perp}) dz_{\perp} + \mathcal{O}\left(\left(\frac{b_x}{h}\right)^2\right). \end{aligned}$$

However, working with these asymptotic expansions is very complicated. Moreover, the asymptotic expansion suggests a fundamental connection between $m_0^{\partial}(x)$ and the standard zeroth moment $m_0(x)$ for an $(m-1)$ -dimensional manifold. Exploiting this connection requires a kernel which can convert the vector sum $hz_{\perp} + hz_{\parallel}\eta_x$ into a product. Of course, the only kernel which can make this separation exactly is the exponential kernel,

$$K(h, x, y) = \pi^{-m/2} \exp\left(-\frac{\|x - y\|^2}{h^2}\right)$$

(in general it is also possible to have h depend on x) where we have,

$$K(h, x, x + hz_{\perp} + hz_{\parallel}\eta_x) = \pi^{-(m-1)/2} \exp(-\|z_{\perp}\|^2) \pi^{-1/2} \exp(-z_{\parallel}^2).$$

This property dramatically simplifies KDE for manifolds with boundary, as shown by the following explicit computation,

$$\begin{aligned} m_0^{\partial}(x) &= \pi^{-(m-1)/2} \int_{\mathbb{R}^{m-1}} \exp(-\|z_{\perp}\|^2) dz_{\perp} \pi^{-1/2} \int_{-\infty}^{b_x/h} \exp(-z_{\parallel}^2) dz_{\parallel} \\ &= \pi^{-1/2} \int_{-\infty}^0 \exp(-z_{\parallel}^2) dz_{\parallel} + \pi^{-1/2} \int_0^{b_x/h} \exp(-z_{\parallel}^2) dz_{\parallel} \\ (8) \quad &= \frac{1}{2}(1 + \text{erf}(b_x/h)) \end{aligned}$$

Due to this simplification, we advocate the exponential kernel for KDE on manifolds with boundary.

Making use of Theorem 4.1 with $m_0^\partial(x)$ from (8) requires estimating the distance to the boundary. The next theorem shows how to use (6) together with Theorem 4.1 to estimate b_x .

THEOREM 4.2 (Boundary Direction Estimation). *Under the same hypotheses as Theorem 4.1, $\mu_{h,N}(x)$ has expectation*

$$\mathbb{E}[\mu_{h,N}(x)] = -\eta_x f(x) m_1^\partial(x) + \mathcal{O}(h\nabla f(x), hf(x))$$

where $\eta_x \in T_x \mathcal{M}$ is a unit vector pointing towards the closest boundary point (η_x is the outward pointing normal for $x \in \partial \mathcal{M}$) and

$$m_1^\partial(x) = - \int_{\mathbb{R}^{m-1}} \int_{-\infty}^{b_x/h} K(h, x, x + hz_\perp + hz_\parallel \eta_x) z_\parallel dz_\parallel dz_\perp$$

The proof of Theorem 4.2 is in Appendix B. Notice that the minus sign in the definition of $m_1^\partial(x)$ implies that for most kernels, $m_1^\partial(x) > 0$ (since the integral is heavily weighted toward the negative z_\parallel direction). This choice of minus sign gives the correct impression that $\mu_{h,N}(x)$ points into the interior (the opposite direction of η_x).

Intuitively, Theorem 4.2 follows from the fact that the integrand $K(h, x, x + hz)z$ is odd and the domain of integration is symmetric in every direction $z \perp \eta_x$. The only non-symmetric direction is parallel to η_x due to the boundary. Thus, the integral is zero in every direction except $-\eta_x$, where the minus sign follows from the fact that there are more points in the interior direction than in the boundary direction (since the boundary cuts off the data). Of course, it is possible for a large density gradient to force μ to point in a different direction, which explains the bias term of order $h\nabla f(x)$, but this is a higher order error.

For the Gaussian kernel, we again have an exact expression for the integral

$$(9) \quad \begin{aligned} \mathbb{E}[\mu_{h,N}(x)] &= \eta_x f(x) \pi^{-1/2} \int_{-\infty}^{b_x/h} \exp(-z_\parallel^2) z_\parallel dz_\parallel \\ &= -\eta_x \frac{f(x)}{2\sqrt{\pi}} \exp\left(-\frac{b_x^2}{h^2}\right) \end{aligned}$$

and we will use this expression combined with (8) to find b_x . Since $f(x)$ is unknown, and appears in both $f_{h,N}(x)$ and $\|\mu_{h,N}(x)\|$, the natural quantity to consider is

$$\frac{f_{h,N}(x)}{\sqrt{\pi} \|\mu_{h,N}(x)\|} = (1 + \text{erf}(b_x/h)) e^{b_x^2/h^2}.$$

In order to find b_x we will solve the above expression numerically by setting $c = \frac{f_{h,N}(x)}{\sqrt{\pi} \|\mu_{h,N}(x)\|}$ and defining

$$F(b_x) = (1 + \operatorname{erf}(b_x/h)) e^{b_x^2/h^2} - c,$$

where we note that

$$F'(b_x) = \frac{2}{\sqrt{\pi} h} + 2(1 + \operatorname{erf}(b_x/h)) e^{b_x^2/h^2} \frac{b_x}{h^2},$$

Newton's method can be used to solve $F(b_x) = 0$ for b_x . In fact, using the fact that $1 \leq 1 + \operatorname{erf}(b_x/h) < 2$, a very simple lower bound for b_x is

$$b_x \geq h \sqrt{\max\{0, -\log(c/2)\}}$$

and this can be a useful initial guess for Newton's method.

Finally, using the estimated value for b_x we can evaluate $m_0(x) = \frac{1}{2}(1 + \operatorname{erf}(b_x/h))$ and use the KDE formula in Theorem 4.1 with this $m_0(x)$ to which yields a consistent estimator of $f_{h,N}(x)$ on manifolds with boundary.

EXAMPLE 4.3 (KDE on a Disc). In this example we verify the above expansions for data sampled on the disk $D^2 = \{(r \cos \theta, r \sin \theta) \in \mathbb{R}^2 : r \leq 1\}$ according to the density $f(r, \theta) = \frac{2}{3\pi}(2 - r^2)$. In order to generate samples $x_i = (r_i, \theta_i)$ from the density f , we use the rejection sampling method. We first generate points on the disc sampled according to the uniform density $f_0(r, \theta) = \operatorname{vol}(D^2)^{-1} = \pi^{-1}$ by generating 12500 uniformly random points in $[-1, 1]^2$ and then eliminating points with distance to the origin greater than 1. Next we set $M = \max_{r,\theta}\{f(r, \theta)/f_0(r, \theta)\} = 4/3$ and for each uniformly sampled point \tilde{x}_i on D^2 , we draw a uniformly random variable $\xi_i \in [0, 1]$ and we reject the i -th point if $\xi_i \geq \frac{f(\tilde{x}_i)}{M f_0(\tilde{x}_i)} = 1 - r^2/2$ and otherwise we accept the point as a sample x_i of f . In this experiment there were $N = 7316$ points remaining after restricting to the unit disc and rejection sampling.

Using the data x_i , we first evaluate the standard KDE formula (without boundary correction)

$$f_{h,N}(x_i) = \frac{1}{Nh^2} \sum_{j=1}^N K(h, x_i, x_j)$$

and

$$\mu_{h,N}(x_i) = \frac{1}{Nh^3} \sum_{j=1}^N K(h, x_i, x_j)(x_j - x_i)$$

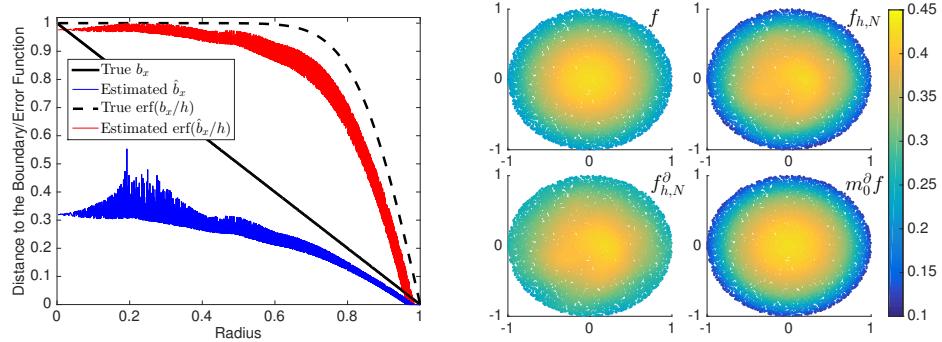


FIG 3. Verifying the estimation of the distance to the boundary with $h = 0.2$ on the disk data set. Left: The true distance to the boundary (black, solid curve) is compared to the recovered distance (blue) is shown for each point as a function of the radius, we also show the value of $\text{erf}(b_x/h)$ for both the true distance (black, dashed curve) and the recovered distance (red). Right: The true density f is compared to the standard KDE $f_{h,N}$ and the boundary correction $f_{h,N}^{\partial}$ as well as the theoretically derived standard KDE result $m_0^{\partial}f$.

on each data point. In this example we use the standard Gaussian kernel described above. In order to correct the KDE on the boundary, we first estimate the distance to the boundary using the strategy outlined above, and the results of this estimate are shown in Figure 3 (top, left). We then compute $m_0^{\partial}(x)$ which allows us to compute the boundary correction $f_{h,N}^{\partial}(x)$ and in Figure 3 (top, right) we compare this to the standard KDE $f_{h,N}(x)$ as well as the true density f and the theoretically derived large data limit $m_0^{\partial}(x)f(x)$ of the standard KDE. These quantities are also compared qualitatively on the disc in Figure 3, which clearly shows the underestimation of the standard KDE on the boundary, which also agrees with the theoretically derived standard KDE result which is $m_0^{\partial}(x)f(x)$. In contrast, the boundary correction $f_{h,N}^{\partial}$ slightly overestimates the true density on the boundary, due to $h = 0.2$ being quite large in this example.

4.2. The Cut and Normalize Method. The weakness of the previous approach is that the estimate of b_x may not be very accurate, especially for points far from the boundary. Of course, since the function $\text{erf}(b_x/h)$ saturates for b_x sufficiently large, this somewhat ameliorates the problem of underestimating b_x . However, it would be preferable in terms of bias to have an exact value for b_x . In fact, the kernel weighted average $\mu_{h,N}(x)$ makes this possible. Notice that the unit vector in the direction of $-\mu_{h,N}(x)$ is an

estimate of η_x , namely

$$\hat{\eta}_x \equiv \frac{-\mu_{h,N}(x)}{||\mu_{h,N}(x)||} = \eta_x + \mathcal{O}(h).$$

Since $\mu_{h,N}(x)$ tells us the direction of the boundary, we can protect against underestimation of b_x by actually cutting of the kernel at the estimated distance to the boundary.

Given an estimate \hat{b}_x of b_x , the cut-and-normalize method only includes samples X_i such that $X_i \cdot \hat{\eta}_x \leq \hat{b}_x$, which gives us the following estimator,

$$f_{h,N}^c(x) \equiv \frac{1}{N(1 + \text{erf}(\hat{b}_x/h))h^m} \sum_{X_i \cdot \hat{\eta}_x \leq \hat{b}_x} K(h, x, X_i)$$

which is a consistent estimator for any $0 < \hat{b}_x < b_x$. Of course, this cut-and-normalize method has several potential downsides. The first is that by not including the maximum possible number of points, we have increased the variance of our estimator. The second is that for points in the interior, the cut-and-normalize method may eliminate the symmetry of the region of integration, leading to increased bias for interior points. However, as long as the estimate b_x is larger than h for points that are far from the boundary, the effect of cutting the domain outside of h will be negligible (see proof of Theorem A.1 for details). In our empirical investigations, we have found that the error introduced by the cut-and-normalize method is very small compared to the error of using an incorrect estimate of b_x direction in $m_0^\partial(x)$. In Figure 4 we apply the cut-and-normalize method to Example 4.3 and show that for interior points, the method produces results that are comparable to the standard KDE. This should be compared with Figure 3 which simply renormalizes using the estimated distance to the boundary without cutting. Figure 3 does not match the standard KDE for interior points.

4.3. Higher-Order Boundary Correction. The above method obtains an asymptotically unbiased estimate of the sampling density at all points of the manifold, including the boundary. However, the bias in the interior of the manifold is $\mathcal{O}(h^2)$, which is significantly smaller than for points very near the boundary, where the bias is $\mathcal{O}(h)$. In order to obtain a uniform rate of convergence at all points, we need to eliminate the order- h term $hm_1^\partial(x)\eta_x \cdot \nabla f(x)$ appearing in the bias of Theorem 4.1.

To construct a higher-order kernel we will use Richardson extrapolation, which is a general method of combining estimates from multiple values of h to form a higher order method. Its use is common in the kernel density

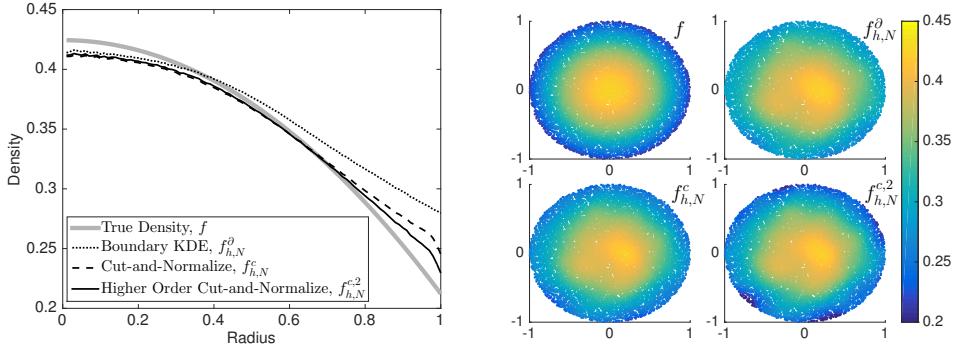


FIG 4. Comparison of the boundary correction, cut-and-normalize method, and higher order cut-and-normalize method on the disk of Example 4.3. Left: Average value of each density estimation method as a function of radius after repeating the experiment of Example 4.3 10 times with independent data sets which shows the bias (variance error is averaged out). Notice that the higher order cut-and-normalize method $f_{h,N}^{c,2}$ has similar bias on the boundary and the interior, which is order $h^2 = 0.04$ in each case. Right: A single realization of the experiment in Example 4.3 showing the true density and all three density estimates (note that the color scale is different than in Figure 3 to better show the differences in these estimates).

estimation literature [28, 14, 18]. Our goal is to cancel the bias term (7)

$$hm_1^{\partial}(x)\eta_x \cdot \nabla f(x) = h(1 + \text{erf}(b_x/h))e^{-b_x^2/h^2}\eta_x \cdot \nabla f(x)$$

using a linear combination of two KDE formulas with different values of h . Consider the bias for bandwidths h and $2h$:

$$(10) \quad \mathbb{E}[f_{h,N}^c(x)] = f(x) + h(1 + \text{erf}(b_x/h))e^{-b_x^2/h^2}\eta_x \cdot \nabla f(x) + \mathcal{O}(h^2)$$

$$(11) \quad \mathbb{E}[f_{2h,N}^c(x)] = f(x) + 2h(1 + \text{erf}(b_x/(2h)))e^{-b_x^2/(4h^2)}\eta_x \cdot \nabla f(x) + \mathcal{O}(h^2).$$

Set $C = \frac{(1 + \text{erf}(b_x/(2h)))e^{-b_x^2/(4h^2)}}{(1 + \text{erf}(b_x/h))e^{-b_x^2/(h^2)}}$ and define the second-order cut-and-normalize density estimator as

$$f_{h,N}^{c,2}(x) \equiv \frac{2Cf_{h,N}^c(x) - f_{2h,N}^c(x)}{2C - 1}.$$

The order- h term of the bias cancels, so that

$$\mathbb{E}[f_{h,N}^{c,2}(x)] = f(x) + \mathcal{O}(h^2),$$

which is the same asymptotic bias as the standard KDE in Corollary 3.4 for embedded manifolds without boundary. It is also interesting to note that as b_x becomes larger than h , the higher-order formula reduces to $f_{2h,N}^c$. This shows that this kernel is only “higher-order” on the boundary, and in fact is the same order as the standard KDE on the interior, so in fact $f_{h,N}^{c,2}(x)$ has a bias which is order- h^2 on the boundary and the interior. The higher order cut-and-normalize method KDE is implemented in the examples below and show bias that is significantly reduced compared to the naive cut-and-normalize method.

We first consider an example on a noncompact manifold with boundary, namely a Gaussian distribution restricted to a half-plane. The manifold in this case is the entire half-plane, which is a simple linear manifold with infinite injectivity radius (see note in Appendix B) and $R(x,y) = 1$ for all pairs of points. This means that the half-plane is a uniformly tangible manifold and so we can estimate the density effectively at each point of a sample set.

EXAMPLE 4.4 (Gaussian in the Half-Plane). We generated 20000 points from a standard 2-dimensional Gaussian and then rejected all the points with first coordinate less than zero. Setting $h = \sqrt{0.06}$, the standard KDE formula $f_{h,N}$ and the BDE $\mu_{h,N}$ were computed. Then the cut-and-normalize estimator $f_{h,N}^c$ and the second-order cut-and-normalize estimator $f_{h,N}^{c,2}$ were calculated as in the flowchart of Figure 2. These estimates are compared in Figure 5. Notice that the standard KDE moves the mode of the distribution into the right half-plane, whereas both cut-and-normalize methods yield a mode very close to zero. Of course, the input to the algorithm are the data points only; no information about the manifold is assumed known.

The next example demonstrates the benefits of the higher-order boundary correction on a portion of a hemisphere with an oscillating boundary (see Figure 6). This manifold is particularly difficult for density estimation due to the large curvature of the boundary. For a point in the middle of one of the arms, there are two boundaries which are equidistant apart. Of course, in the limit of very small h , these points will not be able to see either boundary, but for large h this can lead to significant bias.

EXAMPLE 4.5 (Hemisphere with Oscillating Boundary). To generate this data set, we began by sampling 50000 points uniformly from $[-1, 1]^2$ in the plane, and keep only the points with

$$r \leq \sin(6(\theta - \frac{\pi}{12}))/8 + \frac{3}{4},$$

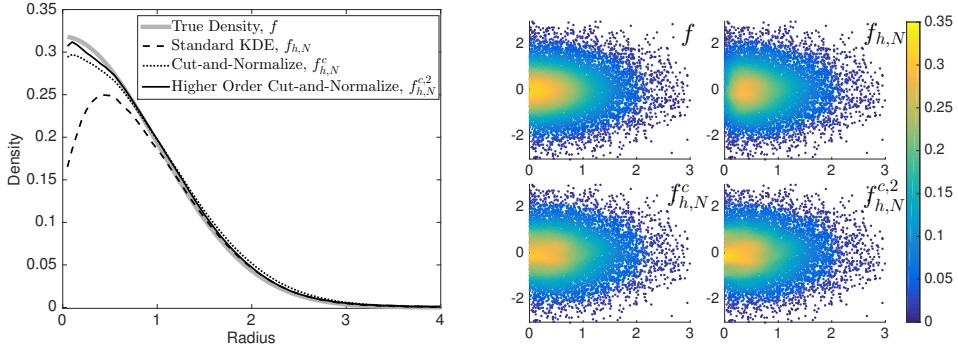


FIG 5. Comparison of the standard KDE, cut-and-normalize method, and higher order cut-and-normalize method on the Gaussian restricted to the half-plane. Left: Average value of each density estimation method as a function of radius after repeating the experiment 10 times with independent data sets which shows the bias (variance error is averaged out). Right: A single realization of the experiment showing the true density and all three density estimates.

which gives a subset of the disk of radius $7/8$ with an oscillating boundary. A z -coordinate on the unit sphere is assigned to each point by setting $z = \sqrt{1 - x^2 + y^2}$. The volume form is given by $dV = \det(D\mathcal{H}^\top D\mathcal{H})^{1/2}$ where $\mathcal{H} : (x, y) \mapsto (x, y, \sqrt{1 - x^2 - y^2})$ which is $dV = (1 - x^2 - y^2)^{-1/2}$. Thus, by mapping uniformly sampled points from the disk onto the hemisphere, the sampling measure of the data at this point is proportional to $dV^{-1} = \sqrt{1 - x^2 - y^2}$.

To normalize the distribution, this function dV^{-1} is integrated against the volume form dV , and in polar coordinates $r = \sqrt{x^2 + y^2}$ the integral is

$$\int_0^{2\pi} \int_0^{\sin(6\theta - \pi/2)/8+3/4} r dr d\theta = \frac{73\pi}{128}.$$

The initial density is $q(r) = \frac{128}{73\pi} \sqrt{1 - r^2}$. This density is largest in the interior, and the density gradient helps to insure that $\mu_{h,N}$ points in the correct direction (into the interior of the manifold). In order to make the problem more challenging we will change the sampling density to be proportional to $f(r) = (1 - r^2)^{-1/2}$ which concentrates more density at the boundary. We will create this sampling density by rejection sampling the initial density. We first compute the normalization factor of the new density by integrating it against the volume form

$$\alpha = \int_0^{2\pi} \int_0^{\sin(6\theta - \pi/2)/8+3/4} \frac{r}{1 - r^2} dr d\theta \approx 2.81893.$$

The new density will be $f(r) = (1 - r^2)^{-1/2}/\alpha$. In order to perform rejection sampling, note that the ratio

$$f(r)/q(r) = \frac{73\pi}{128\alpha(1 - r^2)}$$

has maximum value $M = f(7/8)/q(7/8)$ since $r = 7/8$ is the maximum radius on the oscillating boundary. For each point sampled from q a uniform random number ξ is drawn; the point is accepted as a sample of f if and only if $M\xi < f/q$. After implementing this process in the realization shown in Figure 6, the remaining 10422 points were independent samples of the density f on the hemisphere with oscillating boundary. Using this data set with $h = \sqrt{0.02}$, we computed the standard KDE for each data point, estimated the distance to the boundary, and computed the cut-and-normalize and higher-order cut-and-normalize estimates of the density.

The density estimates are compared visually in Figure 6. We also repeated this experiment 10 times and computed the average of each of the estimates on the positive x-axis and the positive y-axis (which correspond to the shortest and longest radii, respectively) and these curves are compared to the true density in Figure 6. Despite the gradient of the density increasing in the direction of the boundary, the $\mu_{h,N}$ computation still appears to have pointed into the interior as evidenced by the significant improvement of the cut-and-normalize method over the standard KDE. This example also showed the largest difference between the cut-and-normalize method and the higher order cut-and-normalize method, possibly due to the large gradient at the boundary making the order h term quite large. The complexity of the boundary in this example illustrates the advantage of our method, which does not require any prior knowledge of the boundary.

5. Discussion. Our goal in this manuscript was to overcome the limitations which make KDE impractical for large data sets lying on manifolds embedded in \mathbb{R}^n . These limitation include the need to know the structure of the manifold as in [23, 17], the restriction to compact manifolds in [7, 4], and the restriction in all previous work to manifolds without boundary. In fact, the need to know the structure of the manifold was alleviated in the work of [7], and our new proofs extend the results of [7] to the larger class of tangible manifolds and to the larger class of local kernels first introduced in [4]. These advancements do not affect the algorithm for KDE which is implicit to [7, 4] and are purely advancements in the theory.

The more practical advancement is the generalization of the ‘cut-and-normalize’ strategy for boundary correction [10, 18] to manifolds, especially

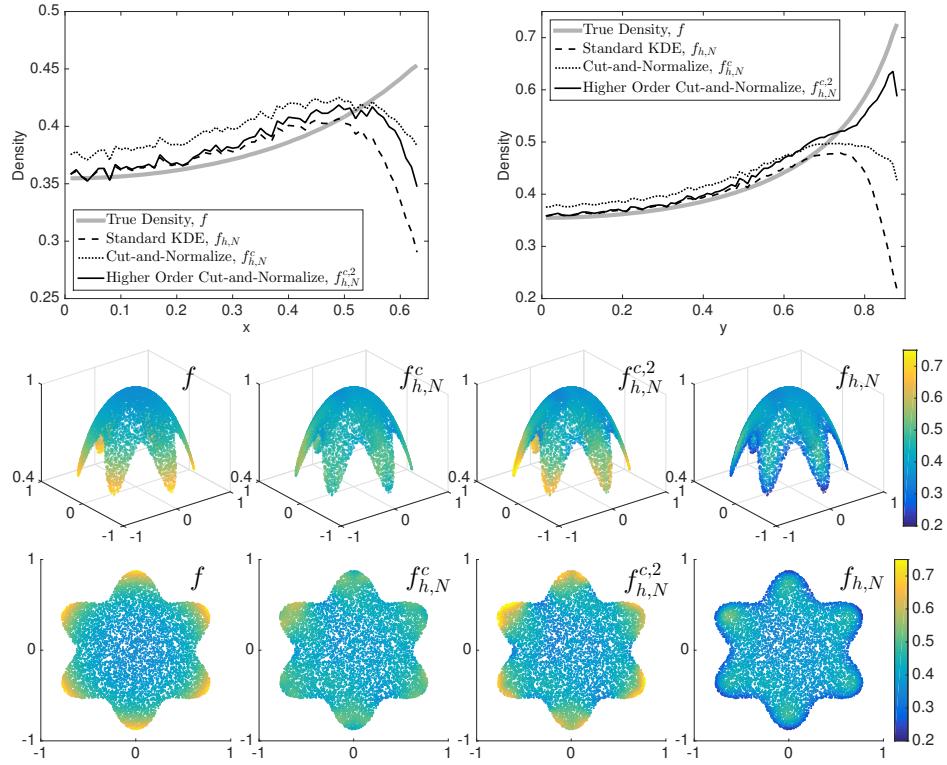


FIG 6. Comparison of the standard KDE, cut-and-normalize method, and higher order cut-and-normalize method on the hemisphere with oscillating boundary. Top, Left: True density compared to estimates on the positive x -axis. Top, Right: True density compared to estimates on the positive y -axis. Below we visualize the various estimates in 3-dimensions and 2-dimensions.

when we cannot assume we know the location of the boundary. In Section 4 we showed that the key to extending the cut-and-normalize strategy was estimating the distance and direction of the boundary and then deriving the correct normalization factor. Another practical consequence of this theory is that the exponential kernel has a significant advantage over other kernels, which is that the boundary normalization factor has a very simple form independent of the dimension of the manifold. Finally, using the distance and direction information derived here, the various boundary correction methods of [29, 16, 13, 14, 27, 18, 11, 6, 15, 20] can now be extended to manifolds and to the case where the boundary is unknown.

APPENDIX A: KERNEL DENSITY ESTIMATION WITH LOCAL KERNELS

The intrinsic geometry of a prototypical kernel is only affected by the second moment, which is essentially the projection of $A(x)$ onto $T_x\mathcal{M}$. More formally, let $\mathcal{I}(x) : T_x\mathbb{R}^n \rightarrow T_x\mathcal{M} \cong \mathbb{R}^m$ be the projection from the tangent space of the ambient space to the tangent space of \mathcal{M} at x . (We can think of $\mathcal{I}(x)$ as the derivative of the embedding map $\iota : \mathcal{M} \rightarrow \mathbb{R}^n$, which for an embedded manifold is just the identity on \mathcal{M} which implies that $\mathcal{I}(x)\mathcal{I}(x)^\top = I_{m \times m}$.) In order to find the moments of K_A we first define the projection of $A(x)$ onto the tangent space as the symmetric positive definite matrix

$$\hat{A}(x) = \mathcal{I}(x)A(x)\mathcal{I}(x)^\top.$$

Then the zeroth moment of the prototypical kernel K_A is

$$m_0(x) = \det(\hat{A}(x))^{1/2} = |\hat{A}(x)|^{1/2}$$

and the second moment is the $m \times m$ matrix valued function

$$C(x) = m_0(x)\hat{A}(x).$$

For a general local kernel we can define $\hat{A} \equiv C(x)/m_0(x)$ to be the second moment normalized by the zeroth moment. We note that the prototypical kernels in [4] included a mean shift of order h , but following [33] we will always use a mean-zero kernel for density estimation so that the kernel is symmetric in x and y .

Following [4] we can now define the intrinsic geometry of a local kernel on a manifold \mathcal{M} . Let g be the Riemannian metric which $\mathcal{M} \subset \mathbb{R}^n$ inherits from the ambient space, namely

$$g_x(u, v) = \left\langle \mathcal{I}(x)^\top u, \mathcal{I}(x)^\top v \right\rangle_{\mathbb{R}^n} = u^\top \mathcal{I}(x) \mathcal{I}(x)^\top v = u^\top v$$

and define the intrinsic geometry \hat{g} by

$$\hat{g}_x(u, v) = g_x(\hat{A}(x)^{-1/2}u, \hat{A}(x)^{-1/2}v) = u^\top \hat{A}(x)^{-1}v.$$

Notice that the zeroth moment of the local kernel $m_0(x) = \sqrt{|\hat{A}(x)|}$ relates the volume form of the intrinsic geometry

$$d\hat{V}(x) = \sqrt{|\hat{A}(x)|} dV(x) = m_0(x) dV(x)$$

to the volume form dV of the geometry g that \mathcal{M} inherits from the ambient space.

We will see below that a kernel density estimate that uses a local kernel will estimate the density relative to the volume form of the intrinsic geometry. In particular, when $|\hat{A}(x)| = 1$, the estimated density will be relative to the volume form which \mathcal{M} inherits from the ambient space. Moreover, we will find that the bias of the estimator also depends on the intrinsic geometry.

THEOREM A.1 (KDE on Tangible Manifolds). *Let \tilde{f} be a density supported on an m -dimensional tangible manifold $\mathcal{M} \subset \mathbb{R}^n$ without boundary. Let $\tilde{f} = f dV = \hat{f} d\hat{V}$ where dV is the volume form on \mathcal{M} inherited from the embedding and $d\hat{V} = m_0(x)^{-1} dV$ is the volume form of the intrinsic geometry. We assume that $f \in \mathcal{C}^4(\mathcal{M})$ is bounded above by a polynomial. Let $K : [0, \infty) \times \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ be a local kernel with zeroth moment $m_0(x)$ and second moment $C(x)$ and zero first and third moments. If X_i are independent samples of f then*

$$\hat{f}_{h,N}(x) = \frac{1}{Nh^m} \sum_{i=1}^N K(h, x, X_i)$$

is a consistent estimator of $\hat{f}(x) = m_0(x)f(x)$ with bias

$$\mathbb{E} [\hat{f}_{h,N}(x) - \hat{f}(x)] = \frac{h^2}{2} \left(\sum_{i,j=1}^m C(x)_{ij} H(f)(x)_{ij} + f(x)\omega(x) \right) + \mathcal{O}(h^4)$$

where $H(f) = \nabla \nabla f$ is the Hessian of f and

$$\omega(x) = \frac{1}{3} \sum_{ij} C(x)_{ij} (R(x)_{ij} + \tilde{R}(x)_{ij})$$

depends on the intrinsic Ricci curvature $R(x)_{ij}$ and the extrinsic curvature of the embedding via $\tilde{R}(x)_{ij}$. The variance of the estimator is

$$\text{var} (\hat{f}_{h,N}(x) - \hat{f}(x)) = \frac{h^{-m}}{N} m_0^2(x) f(x) + \mathcal{O}(1/N)$$

where $m_0^2(x)$ is the zeroth moment of K^2 .

PROOF. We first compute the expectation

$$\mathbb{E} \left[\frac{1}{Nh^m} \sum_{i=1}^N K(h, x, X_i) \right] = \frac{1}{h^m} \int_{\mathcal{M}} K(h, x, y) f(y) dV(y)$$

by splitting the integral over \mathcal{M} into two disjoint regions. Assume that $h < \text{inj}(x)$ which implies that for some $\gamma \in (0, 1)$ we have $h^\gamma < \text{inj}(x)$ (we will explain the need for h^γ below). Since h^γ is less than the injectivity radius, for any $s \in T_x\mathcal{M}$ with $\|s\| < h^\gamma$ we can map s to \mathcal{M} diffeomorphically via $\exp_x(s) \in \mathcal{M}$. We will split the manifold into the image of this ball $\exp_x(B_{h^\gamma}(x))$ and the complement $\mathcal{M} \cap \exp_x(B_{h^\gamma}(x))^c$. We first show that the integral over the complement is small. Since $K(h, x, x + hz)p(z)$ is integrable for any polynomial p , taking p to be $z^{\ell+\kappa}$ where κ is the degree of the polynomial upper bound of f we have $K(h, x, x + hz) < \|z\|^{-\ell-\kappa}$ and therefore $|K(h, x, x + hz)f(x + hz)| < a\|z\|^{-\ell}$ for some constant a , where ℓ was arbitrary. Making the change of variables $y = x + hz$ we find that $z \in \tilde{\mathcal{M}} \cap \exp_0(B_{h^{\gamma-1}}(0))^c$ where $\tilde{\mathcal{M}}$ is translated so that $z = 0$ corresponds to the point $x \in \mathcal{M}$, and $dV(y) = h^m dV(z)$ so we have

$$\frac{1}{h^m} \left| \int_{\mathcal{M} \cap \exp_x(B_{h^\gamma}(x))^c} K(h, x, y)f(y) dV(y) \right| \leq \int_{\tilde{\mathcal{M}} \cap \exp_x(B_{h^{\gamma-1}}(0))^c} a\|z\|^{-\ell} dV(z).$$

Notice that the decay of the kernel is in the ambient space distance $\|z\|$, whereas the region $\exp_x(B_{h^\gamma}(0))^c$ only guarantees that the geodesic distance from 0 to z is large. In order for this integral to be small, we now need the guarantee that large geodesic distance implies large Euclidean distance, which is exactly our assumption that $R(x, y) > 0$. Since x is tangible, let $R(x, y) > c$, we then have, $\|z\| > cd_g(0, z) > h^{\gamma-1}$, so

$$\int_{\tilde{\mathcal{M}} \cap \exp_x(B_{h^{\gamma-1}}(0))^c} a\|z\|^{-\ell} dV(z) \leq ac^{-\ell} \int_{\tilde{\mathcal{M}} \cap \|z\| > h^{\gamma-1}} \|z\|^{-\ell} dV(z).$$

We can bound the previous integral by the integral over all $\|z\| > h^{\gamma-1}$ in \mathbb{R}^n , and switching to polar coordinates we find

$$\begin{aligned} ac^{-\ell} \int_{\tilde{\mathcal{M}} \cap \|z\| > h^{\gamma-1}} \|z\|^{-\ell} dV(z) &= aV_n c^{-\ell} c^{-\ell} \int_{h^{\gamma-1}}^{\infty} r^{-\ell} r^n dr \\ &\leq \frac{aV_n}{\ell - n - 1} c^{-\ell} h^{(\gamma-1)(-\ell+n+1)} \end{aligned}$$

for $\ell > n + 2$ where V_n is the volume of the unit n -ball. Since ℓ was arbitrary and $\gamma - 1 < 0$, we can bound this integral by $\mathcal{O}(h^k)$ for any k .

Having established that the integral outside the image of the ball $B_{h^\gamma}(x)$ is small to an arbitrarily high order in h , we now consider the integral inside the ball,

$$\frac{1}{h^m} \int_{\exp_x(B_{h^\gamma}(x))} K(h, x, y)f(y) dV(y).$$

Since h^γ is less than the injectivity radius, we can write the integral in terms of geodesic normal coordinates $s = \exp_x^{-1}(y)$ based at x . In these coordinates we have the following expansion of the Riemannian metric,

$$g_{ij} = \delta_{ij} - \frac{1}{3} \sum_{k,l} R_{ikjl} s_k s_l + P_3(s) + \mathcal{O}(s^4)$$

where R_{ikjl} is the Riemannian curvature tensor and $P_3(s)$ is a homogeneous polynomial of degree 3 in the components of s . This yields the expansion of the volume form in geodesic coordinates based at x ,

$$dV(y) = \sqrt{|g(s)|} ds = \left(1 - \frac{1}{6} \sum_{i,j} R_{ij} s_i s_j + P_3(s) + \mathcal{O}(s^4) \right) ds$$

where $R_{ij} = \sum_k R_{ikjk}$ is the Ricci curvature. Let $y = \exp_x(s)$ and let γ be geodesic curve with $\gamma(0) = x$ and $\gamma(\|s\|) = y$ parametrized by arclength so that $\|\gamma'(t)\| = 1$ for all t . We can expand γ in t as

$$\gamma(t) = \gamma(0) + t\gamma'(0) + t^2\gamma''(0)/2 + P_3(t) + \mathcal{O}(t^4)$$

Notice that $\gamma'(0)$ is a unit vector in the direction of s , so that $\|s\|\gamma'(0) = s$. Moreover, since the γ is a geodesic, it satisfies the geodesic equation $\nabla_{\gamma'}\gamma' = 0$ on \mathcal{M} , which says that $\gamma''(0)$ is orthogonal to the tangent plane $T_x\mathcal{M}$ so that $\gamma''(0) = (\nabla_{\gamma'}\gamma')^\perp = \Pi(\gamma'(0), \gamma'(0))$ where Π is the second fundamental form. This gives us the expansion,

$$y = \gamma(\|s\|) = x + s + \Pi(s, s)/2 + \text{III}(s, s, s)/6 + \mathcal{O}(s^4)$$

where $\Pi(s, s)$ is the second fundamental form which is a bilinear form defined by $D^2 \exp_x(s)|_{s=0}$ and $\text{III}(s, s, s)$ is the trilinear form defined by $D^3 \exp_x(s)|_{s=0}$. We note that $\text{III}(s, s, s)$ is a universal polynomial in terms of the extrinsic curvature tensor and its derivatives [25]. We also expand the kernel $K(h, x, \exp_x(s))$ in the last component centered around $x + s$ as

$$K(h, x, \exp_x(s)) = K(h, x, x+s) + K_y(h, x, x+s) (\Pi(s, s)/2 + \text{III}(s, s, s)/6) + \mathcal{O}(s^4)$$

and we expand the density $\tilde{f}(s) = f(\exp_x(s))$ around $s = 0$ as

$$f(\exp_x(s)) = f(x) + \sum_{i=1}^m s_i \frac{\partial \tilde{f}}{\partial s_i} + \frac{1}{2} \sum_{i,j=1}^m s_i s_j \frac{\partial^2 \tilde{f}}{\partial s_i \partial s_j} + P_3(s) + \mathcal{O}(s^4).$$

Multiplying the above expansions to find $K(h, x, y)f(y) dV(y)$ in exponential coordinates and eliminating any odd order terms due to integration over a symmetric domain, we find

$$\begin{aligned} & \frac{1}{h^m} \int_{\exp_x(B_{h^\gamma}(x))} K(h, x, y)f(y) dV(y) \\ &= \frac{1}{h^m} \int_{||s|| < h^\gamma} K(h, x, x + s)f(x) + K(h, x, x + s) \frac{1}{2} \sum_{i,j=1}^m s_i s_j \frac{\partial^2 \tilde{f}}{\partial s_i \partial s_j} \\ & \quad + f(x)K_y(h, x, x + s)(\text{II}(s, s)/2 + \text{III}(s, s, s)/6) \\ & \quad + f(x)K(h, x, x + s) \frac{1}{6} \sum_{i,j=1}^m R_{ij} s_i s_j + (K + K_y)\mathcal{O}(s^4) ds. \end{aligned}$$

Rescaling $s \mapsto hs$ we find

$$\begin{aligned} &= \int_{||s|| < h^{\gamma-1}} K(h, x, x + hs)f(x) + \frac{h^2}{2} K(h, x, x + hs) \sum_{i,j=1}^m s_i s_j \frac{\partial^2 \tilde{f}}{\partial s_i \partial s_j} \\ & \quad + \frac{h^2}{2} f(x)K_y(h, x, x + hs)(\text{II}(s, s) + h\text{III}(s, s, s)/3) \\ & \quad + \frac{h^2}{6} f(x)K(h, x, x + hs) \sum_{i,j=1}^m R_{ij} s_i s_j + (K + K_y)\mathcal{O}(h^4 s^4) ds. \end{aligned}$$

Notice that $D_s K(h, x, x + hs) = hK_y(h, x, x + hs)$, so in fact, the term

$$\frac{h^2}{2} f(x)K_y(h, x, x + hs)\text{II}(s, s) = \frac{h}{2} f(x)D_s(K(h, x, x + hs))\text{II}(s, s)$$

is actually order h . Moreover, using integration by parts, the boundary terms are order h^k for any k by decay of K , and we find the integrand $K(h, x, x + hs)D_s(\text{II}(s, s))$ is linear in s and integrates to zero by symmetry. Applying integration by parts to the term containing $\text{III}(s, s, s)$, we find an order h^2 term and we define the bilinear form $D_s\text{III}(s, s, s)$ so that this term in the expansion becomes

$$\frac{h^2}{6} f(x) \int_{||s|| < h^{\gamma-1}} K(h, x, x + hs)D_s\text{III}(s, s, s) ds.$$

We denote the (s_i, s_j) -entry of the bilinear form $D_s\text{III}(s, s, s)$ by $\tilde{R}(x)_{ij}$ which appears in the definition of $\omega(x)$ in the statement of the theorem.

Extending the integral $\|s\| < h^{\gamma-1}$ to then entire tangent plane (which results in an error less than any polynomial in h as shown above) and using the definition of the moments of the kernel we have

$$\begin{aligned} & \frac{1}{h^m} \int_{\mathcal{M}} K(h, x, y) f(y) dV(y) \\ &= m_0(x) f(x) + \frac{h^2}{2} \left(\sum_{i,j=1}^m C_{ij} H(f)(x)_{ij} + f(x) \omega(x) \right) + \mathcal{O}(h^4) \end{aligned}$$

which verifies the bias formula above. Finally, by independence we have

$$\begin{aligned} & \mathbb{E} \left[\left(\frac{1}{Nh^m} \sum_{i=1}^N K(h, x, X_i) - \hat{f}(x) \right)^2 \right] \\ &= \frac{1}{N^2 h^{2m}} \mathbb{E} \left[\left(\sum_{i=1}^N (K(h, x, X_i) - h^m \hat{f}(x)) \right)^2 \right] \\ &= \frac{1}{Nh^{2m}} \mathbb{E} \left[(K(h, x, X_i) - h^m \hat{f}(x))^2 \right] \\ &= \frac{1}{N} \mathbb{E} \left[h^{-2m} K(h, x, X_i)^2 - 2h^{-m} K(h, x, X_i) \hat{f}(x) + \hat{f}(x)^2 \right] \\ &= \frac{h^{-m}}{N} \mathbb{E} \left[h^{-m} K(h, x, X_i)^2 \right] - \hat{f}(x)^2/N + \mathcal{O}(h^2/N) \\ &= \frac{h^{-m}}{N} m_0^2(x) f(x) + \mathcal{O}(1/N) \end{aligned}$$

which verifies the variance formula. \square

Notice that the previous theorem can be extended to local kernels with nonzero third moments with the bias term correct up to order- h^3 instead of order- h^4 . Similarly, the theorem also applies to density functions $f \in \mathcal{C}^3(\mathcal{M})$ with bias term correct up to order- h^3 .

Since $H(f)(x)_{ij} = \frac{\partial^2 \hat{f}(0)}{\partial s_i \partial s_j}$, by setting $\hat{A}(x) = C(x)/m_0(x)$ and changing variables to $\hat{s} = \hat{A}^{-1/2}(x)s$ so that $\frac{\partial \hat{s}_i}{\partial s_j} = \hat{A}_{ij}^{-1/2}$ it was shown in [4] (Lemma 4.2) that

$$\sum_{i,j=1}^m C_{ij} H(f)(x)_{ij} = m_0(x) (\Delta_{\hat{g}} f(x) + \kappa(x) \cdot \nabla f(x))$$

So if we are interested in estimating the density $f(x)$ relative to the volume form dV inherited from the ambient space, we have the following corollary.

COROLLARY A.2. *Under the same assumptions as Theorem A.1,*

$$f_{h,N}(x) \equiv \frac{1}{Nh^m m_0(x)} \sum_{i=1}^N K(h, x, X_i)$$

is a consistent estimator of $f(x)$ with bias,

$$\mathbb{E}[f_{h,N}(x) - f(x)] = h^2(\Delta_{\hat{g}} f(x) + \kappa(x) \cdot \nabla f(x) + \omega(x) f(x)) + \mathcal{O}(h^4)$$

and variance,

$$\text{var}(f_{h,N}(x) - f(x)) = \frac{h^{-m}}{N} \frac{m_0^2(x)}{m_0(x)} f(x) + \mathcal{O}(1/N)$$

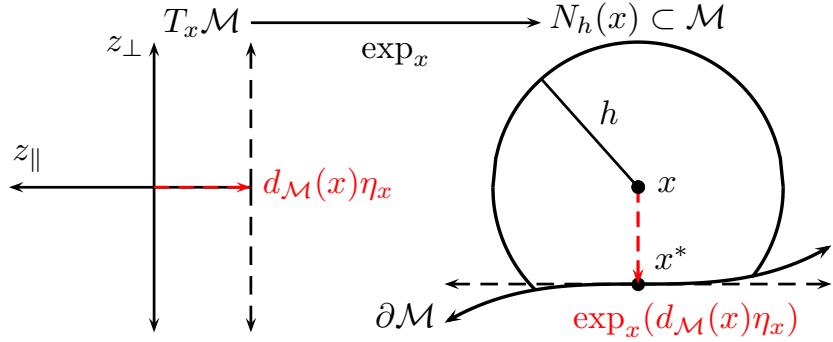
APPENDIX B: PROOFS FOR MANIFOLDS WITH BOUNDARY

In order to extend the definition of a tangible manifold to include manifolds with boundary, notice that for manifolds with boundary, we consider the tangent space for points on the boundary to be the half space. So we consider the injectivity radius to be the largest ball such that the exponential map is well defined on the intersection of the ball and the half space. Similarly for points near the boundary, we consider the tangent space to be a cut space which is cut at b_x in the direction η_x . These definitions allow points on or near the boundary to still have large injectivity radii.

PROOF OF THEOREM 4.1. The key to this theorem is deriving the new normalization factor

$$m_0^\partial(x) = \int_{\mathbb{R}^{m-1}} \int_{-\infty}^{b_x/h} K(h, x, x + hz_\perp + hz_\parallel \eta_x) dz_\parallel dz_\perp.$$

To understand this formula, let x^* be a point on the boundary which minimizes the geodesic distance, $d(x, x^*) = b_x$ (since a boundary is always closed such a point always exists although it may not be unique). If $\|x - x^*\| > h$ then the boundary is far enough away that it will have a negligible effect on m_0 since in the proof of Theorem A.1 we bound the integral outside the ball $N_h(x)$. Thus, we restrict our attention to points with $b_x < h$ and we assume the h is sufficiently small that x^* is unique (notice that this will depend on the curvature of the boundary). We define $\eta_x \in T_x \mathcal{M}$ to be the unit vector which points towards x^* , meaning that $\exp_x(b_x \eta_x) = x^*$ and if x lies exactly on the boundary we define η_x to be the outward pointing unit normal vector.



We can now decompose the exponential coordinates in the tangent space $B_{h^\gamma}(x) \subset T_x \mathcal{M}$ into vectors s_{\parallel} which are parallel to η_x and vectors s_{\perp} which are perpendicular to η_x . All vectors perpendicular to η_x can extend up to length h^γ , whereas vectors parallel to η_x can extend up to length h^γ in the direction $-\eta_x$ (away from the boundary), but only up to length b_x in the direction η_x (towards the boundary). With this decomposition, the coefficient of $f(x)$ from the expansion in the proof of Theorem A.1 becomes

$$m_0^\partial(x) = h^{-m} \int_{[-h^\gamma, h^\gamma]^{m-1}} \int_{-h^\gamma}^{b_x} K(h, x, x + s_{\perp} + s_{\parallel}) ds_{\parallel} ds_{\perp}$$

and this is the leading order term. Making the change of variables $s = hz$, and recalling from the proof of Theorem A.1 that the integral is negligible beyond $h^{\gamma-1}$, we can extend the integral over z_{\perp} to all of $\mathbb{R}^{m-1} \subset T_x \mathcal{M}$. On the other hand, the integral over z_{\parallel} cannot be extended to all of $\mathbb{R} \subset T_x \mathcal{M}$, but only to the half-line $(-\infty, b_x/h]$ so that the zeroth moment becomes

$$m_0^\partial(x) = \int_{\mathbb{R}^{m-1}} \int_{-\infty}^{b_x/h} K(h, x, x + hz_{\perp} + hz_{\parallel}\eta_x) dz_{\parallel} dz_{\perp}.$$

Since $m_0^\partial(x)$ is the coefficient of $f(x)$ in the expansion of the standard KDE formula, replacing $m_0(x)$ with $m_0^\partial(x)$ in the standard KDE formula yields $f_{h,N}^\partial(x)$ which is a consistent estimator of $f(x)$.

In order to establish the bias of this estimator, notice that the next term of the expansion in Theorem A.1 is

$$\sum_{i=1}^m \frac{\partial \tilde{f}}{\partial s_i} h^{-m} \int_{\|s\| < h^\gamma} K(h, x, x + s) s_i ds$$

which integrates to zero for x sufficiently far from the boundary due to the symmetry of the domain of integration. However, for points near the

boundary, this integral will not be zero. Instead, this term integrates to zero for every $s \perp \eta_x$ since the domain is symmetric in those directions, so we have $s = s_{\parallel} \eta_x$ and the integral becomes

$$\eta_x \cdot \nabla f(x) h^{-m} \int_{-[h^\gamma, h^\gamma]^{m-1}} \int_{-h^\gamma}^{b_x} K(h, x, x + s_{\perp} + s_{\parallel} \eta_x) s_{\parallel} ds_{\parallel} ds_{\perp}.$$

Notice that we have rewritten the partial derivatives with respect to the geodesic normal coordinates in terms of the gradient operator by inserting the metric g_{ij} (which becomes the dot product) and the inverse metric g^{jk} (which joins with the partial derivatives to become the gradient operator), namely

$$\sum_{i=1}^m (\eta_x)_i \frac{\partial \tilde{f}(0)}{\partial s_i} = \sum_{i,j,k} (\eta_x)_i g_{ij} g^{jk} \frac{\partial \tilde{f}(0)}{\partial s_k} = \sum_{i,j} (\eta_x)_i g_{ij} (\nabla f(x))_j = \eta_x \cdot \nabla f(x).$$

Changing variables to $s = hz$ as above, we find the bias to be $m_1^\partial(x) \eta_x \cdot \nabla f(x)$ where

$$m_1^\partial(x) = \int_{\mathbb{R}^{m-1}} \int_{-\infty}^{b_x/h} K(h, x, x + hz_{\perp} + hz_{\parallel} \eta_x) z_{\parallel} dz_{\parallel} dz_{\perp}$$

Finally, the derivation of the variance follows exactly as in Theorem A.1 with

$$m_0^{2,\partial}(x) = \int_{\mathbb{R}^{m-1}} \int_{-\infty}^{b_x/h} K(h, x, x + hz_{\perp} + hz_{\parallel} \eta_x)^2 dz_{\parallel} dz_{\perp}.$$

□

PROOF OF THEOREM 4.2. The definition

$$\mu_{h,N}(x) \equiv \frac{1}{Nh^{m+1}} \sum_{i=1}^N K(h, x, X_i)(X_i - x)$$

implies a formula for the expectation:

$$\mathbb{E}[\mu_{h,N}(x)] = \frac{1}{h^{m+1}} \int_{\mathcal{M}} K(h, x, y)(y - x) f(y) dV(y).$$

Following the same argument as in the proof of Theorem A.1 we can restrict this integral to the image of the ball $\|s\| < h^\gamma$ under the exponential map,

and then change variables to the geodesic normal coordinates $s \in T_x\mathcal{M}$ with $y = \exp_x(s)$, which yields

$$\mathbb{E}[\mu_{h,N}(x)] = \frac{1}{h^{m+1}} \int_{\|s\| < h^\gamma} K(h, x, \exp_x(s))(\exp_x(s) - x)f(\exp_x(s))dV(\exp_x(s)).$$

Applying the asymptotic expansions from the proof of Theorem A.1, we find

$$\begin{aligned} \mathbb{E}[\mu_{h,N}(x)] &= \frac{1}{h^{m+1}} \int_{\|s\| < h^\gamma} K(h, x, x + s)f(x)s \\ &\quad + K(h, x, x + s)s(s \cdot \nabla f(x) + f(x)D_s \Pi(s, s)/2) \\ &\quad + \mathcal{O}(s^3 K(h, x, x + s))ds. \end{aligned}$$

Following the proof of Theorem 4.1 we decompose $s = s_\perp \oplus s_{\parallel}\eta_x$ and note that the first term of the integral is zero in every direction except $s = s_{\parallel}\eta_x$ which leads to

$$\begin{aligned} \mathbb{E}[\mu_{h,N}(x)] &= \frac{1}{h^{m+1}} \int_{[-h^\gamma, h^\gamma]^{m-1}} \int_{-h^\gamma}^{b_x} K(h, x, x + s_\perp + s_{\parallel})f(x)s_{\parallel}\eta_x \\ &\quad + K(h, x, x + s)s(s \cdot \nabla f(x) + f(x)D_s \Pi(s, s)/2) \\ &\quad + \mathcal{O}(s^3 K(h, x, x + s))ds_{\parallel}ds_\perp. \end{aligned}$$

Changing variables to $s = hz$ we have

$$\begin{aligned} \mathbb{E}[\mu_{h,N}(x)] &= \int_{[-h^{\gamma-1}, h^{\gamma-1}]^{m-1}} \int_{-h^{\gamma-1}}^{b_x/h} K(h, x, x + hs_\perp + hs_{\parallel})f(x)z_{\parallel}\eta_x \\ &\quad + hK(h, x, x + hz)z(z \cdot \nabla f(x) + f(x)D_z \Pi(z, z)/2) \\ &\quad + \mathcal{O}(h^2 z^3 K(h, x, x + hz))dz_{\parallel}dz_\perp \end{aligned}$$

and extending the integrals to \mathbb{R}^{m-1} and $(-\infty, b_x/h)$ respectively (following Theorem A.1) we have

$$\begin{aligned} \mathbb{E}[\mu_{h,N}(x)] &= \eta_x f(x) \int_{\mathbb{R}^{m-1}} \int_{-\infty}^{b_x/h} K(h, x, x + hs_\perp + hs_{\parallel})z_{\parallel} dz_{\parallel} dz_\perp \\ &\quad + \mathcal{O}(h\nabla f(x), h f(x)) \\ &= -\eta_x f(x)m_1^\partial(x) + \mathcal{O}(h, \nabla f(x), h f(x)) \end{aligned}$$

where we recall that the definition of the integral $m_1^\partial(x)$ incorporates a minus sign. \square

APPENDIX C: DIMENSION ESTIMATION

Notice that the definition of K_A requires the intrinsic dimension m of the manifold. Interestingly, the dimension is not required in [4] to find the Laplace-Beltrami operator of the intrinsic geometry, and in [4] the factor $\pi^{-m/2}$ is not included in the definition of a prototypical kernel. However, in order to find a properly normalized density one must know the intrinsic dimension, and so in this paper we include the normalization factor $\pi^{-m/2}$ in the definition of the kernel for convenience. There are many methods of identifying the dimension from the data, we advocate a method which was introduced in [8] and further refined in [3, 2] which simultaneously determines the dimension and tunes the bandwidth parameter h . The method of [3] uses the fact that when h is well tuned, the unnormalized kernel sum $\frac{1}{N} \sum_{i=1}^N K(h, x, x_i)$ is proportional to h^m as shown in Theorem A.1. By varying h one can estimate the scaling law $m = \frac{d \log D(h)}{d \log h}$, and when h is well tuned this scaling law will be stable under small changes in h .

In order to simultaneously estimate the dimension m and tune the bandwidth h , we first generate a grid of h values, h_j (typically a logarithmic scale is used, such as $h_j = 1.1^j$ for $j = -20, \dots, 0, \dots, 20$). We then evaluate the sum

$$S(x, h_j) = \frac{1}{N} \sum_{i=1}^N K(h_j, x, x_i)$$

which should be proportional to h^m when $h = h_j$ is well tuned. Motivated by this, we compute the scaling law at each h_j by

$$\dim(x, h_j) = \frac{\log(S(x, h_{j+1})) - \log(S(x, h_j))}{\log(h_{j+1}) - \log(h_j)} \approx \frac{d \log S}{d \log h}(h_j)$$

which gives us an approximate dimension for each value of h_j . In [3] they advocated taking value of h_j which maximizes the dimension, however in [2] they showed that the extrinsic curvature can lead to overestimation. Instead, [2] advocates looking for persistent values of dimension, which intuitively means one should look for values of the dimension such that the curve $\dim(x, h_j)$ is flat for a large range of values of h_j . One method is to approximate derivatives of $\dim(x, h_j)$ with respect to h_j and attempt maximize dim while minimizing the derivatives.

Notice that the above method finds a dimension at a single point x . To estimate a single dimension for an entire data set, one can define $S(h_j)$ to be the average value of S over the entire data set and apply the same procedure.

REFERENCES

- [1] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.
- [2] Tyrus Berry and John Harlim. Iterated diffusion maps for feature identification. *Submitted to Journal of Applied and Computational Harmonic Analysis*, 2015.
- [3] Tyrus Berry and John Harlim. Variable bandwidth diffusion kernels. *Applied and Computational Harmonic Analysis*, 2015.
- [4] Tyrus Berry and Timothy Sauer. Local kernels and the geometric structure of data. *Applied and Computational Harmonic Analysis*, 2015.
- [5] Theophilos Cacoullos. Estimation of a multivariate density. *Annals of the Institute of Statistical Mathematics*, 18(1):179–189, 1966.
- [6] SongXi Chen. Probability density function estimation using gamma kernels. *Annals of the Institute of Statistical Mathematics*, 52(3):471–480, 2000.
- [7] R. Coifman and S. Lafon. Diffusion maps. *Appl. Comput. Harmon. Anal.*, 21:5–30, 2006.
- [8] Ronald R Coifman, Yoel Shkolnisky, Fred J Sigworth, and Amit Singer. Graph laplacian tomography from unknown random projections. *Image Processing, IEEE Transactions on*, 17(10):1891–1899, 2008.
- [9] Vassiliy A Epanechnikov. Non-parametric estimation of a multivariate probability density. *Theory of Probability & Its Applications*, 14(1):153–158, 1969.
- [10] Theo Gasser and Hans-Georg Müller. *Kernel estimation of regression functions*. Springer, 1979.
- [11] Peter Hall and Byeong U Park. New methods for bias correction at endpoints and boundaries. *Annals of Statistics*, pages 1460–1479, 2002.
- [12] Harrie Hendriks. Nonparametric estimation of a probability density on a riemannian manifold using fourier expansions. *The Annals of Statistics*, pages 832–849, 1990.
- [13] M.C. Jones. Simple boundary correction for kernel density estimation. *Statistics and Computing*, 3(3):135–146, 1993.
- [14] MC Jones and PJ Foster. A simple nonnegative boundary correction method for kernel density estimation. *Statistica Sinica*, 6(4):1005–1013, 1996.
- [15] Rhoana J Karunamuni and Tom Alberts. On boundary correction in kernel density estimation. *Statistical Methodology*, 2(3):191–212, 2005.
- [16] Rhoana J Karunamuni and Shunpu Zhang. Some improvements on a boundary corrected kernel density estimator. *Statistics & Probability Letters*, 78(5):499–507, 2008.
- [17] Yoon Tae Kim and Hyun Suk Park. Geometric structures arising from kernel density estimation on riemannian manifolds. *Journal of Multivariate Analysis*, 114:112–126, 2013.
- [18] Cha Kyung-Joon and William R Schucany. Nonparametric kernel regression estimation near endpoints. *Journal of statistical planning and inference*, 66(2):289–304, 1998.
- [19] Don O Loftsgaarden, Charles P Quesenberry, et al. A nonparametric estimate of a multivariate density function. *The Annals of Mathematical Statistics*, 36(3):1049–1051, 1965.
- [20] Peter Malec and Melanie Schienle. Nonparametric kernel density estimation near the boundary. *Computational Statistics & Data Analysis*, 72:57–76, 2014.
- [21] Arkadas Ozakin and Alexander G Gray. Submanifold density estimation. In *Advances in Neural Information Processing Systems*, pages 1375–1382, 2009.
- [22] Emanuel Parzen. On estimation of a probability density function and mode. *The*

- Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.
- [23] Bruno Pelletier. Kernel density estimation on riemannian manifolds. *Statistics & probability letters*, 73(3):297–304, 2005.
 - [24] Dimitris N Politis and Joseph P Romano. Multivariate density estimation with general flat-top kernels of infinite order. *Journal of Multivariate Analysis*, 68(1):1–25, 1999.
 - [25] S. Rosenberg. *The Laplacian on a Riemannian manifold*. Cambridge University Press, 1997.
 - [26] Murray Rosenblatt. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27(3):832–837, 1956.
 - [27] M. C. Jones S. Zhang, R. J. Karunamuni. An improved estimator of the density function at the boundary. *Journal of the American Statistical Association*, 94(448):1231–1241, 1999.
 - [28] WR Schucany and John P Sommers. Improvement of kernel type density estimators. *Journal of the American Statistical Association*, 72(358):420–423, 1977.
 - [29] Eugene F Schuster. Incorporating support constraints into nonparametric estimators of densities. *Communications in Statistics-Theory and methods*, 14(5):1123–1136, 1985.
 - [30] Amit Singer and Ronald R. Coifman. Non-linear independent component analysis with diffusion maps. *Applied and Computational Harmonic Analysis*, 25(2):226 – 239, 2008.
 - [31] D. G. Terrell and D. W. Scott. Variable kernel density estimation. *Annals of Statistics*, 20:1236–1265, 1992.
 - [32] George R Terrell and David W Scott. On improving convergence rates for nonnegative kernel density estimators. *The Annals of Statistics*, pages 1160–1163, 1980.
 - [33] P. Whittle. On the smoothing of probability density functions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 20(2):334–343, 1958.

E-MAIL: tberry@gmu.edu

E-MAIL: tsauer@gmu.edu