Computational Statistics and Data Analysis 107 (2017) 1–17

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Density estimation on manifolds with boundary

Tyrus Berry *, Timothy Sauer

Department of Mathematical Sciences, George Mason University, 4400 University Drive, MS: 3F2 Exploratory Hall, Room 4400 Fairfax, VA 22030, United States

ARTICLE INFO

Article history:
Received 5 January 2016
Received in revised form 23 September 2016
Accepted 27 September 2016
Available online 8 October 2016

Keywords:
Kernel density estimation
Manifold learning
Boundary correction
Geometric prior

ABSTRACT

Density estimation is a crucial component of many machine learning methods, and manifold learning in particular, where geometry is to be constructed from data alone. A significant practical limitation of the current density estimation literature is that methods have not been developed for manifolds with boundary, except in simple cases of linear manifolds where the location of the boundary is assumed to be known. This limitation is overcome by developing a density estimation method for manifolds with boundary that does not require any prior knowledge of the location of the boundary. To construct an appropriate estimator, statistics are introduced that provably approximate the distance and direction of the boundary, which allows us to apply a cut-and-normalize boundary correction. Then, multiple cut-and-normalize estimators are used to build a consistent kernel density estimator that has uniform bias, at interior and boundary points, on manifolds with 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. When lack of "first principles" understanding of a process limits effective parametric modeling, data-driven approaches are valuable. 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*,[1] 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* dimensionality of the ambient space, which is often high.

Recently the geometric prior has received some attention in the density estimation field (Hendriks, 1990; Pelletier, 2005; Kim and Park, 2013; Ozakin and Gray, 2009), although use of these methods remains restricted for several reasons. For example, the methods of Hendriks (1990) and Pelletier (2005) require the structure of the manifold to be known a priori. Recently, 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 (Belkin and Niyogi, 2003; Coifman and Lafon, 2006). These methods have also been extended to a large class of noncompact manifolds (Hein, 2005, 2006; Hein et al., 2005) with natural assumptions on the

* Corresponding author.
E-mail addresses: tberry@gmu.edu (T. Berry), tsauer@gmu.edu (T. Sauer).

[1] We should note that the word "prior" here is used only to indicate an assumption which is implicitly placed on the data (namely that it lies on a manifold) and the geometric prior should not be interpreted as a prior in the sense of Bayesian inference.

http://dx.doi.org/10.1016/j.csda.2016.09.011
0167-9473/

geometry of the manifold. The assumptions introduced in Hein (2005) include all compact manifolds, as well as many non-compact manifolds, such as any linear manifold, which implies that standard kernel density estimation theory on Euclidean spaces is included as a special case. While these manifold learning methods make implicit assumptions on the geometry of the underlying manifold (such as bounded curvature), kernel density estimation requires knowledge of the dimension of the manifold in order to obtain the correct normalization factor. For ease of exposition, we will assume the dimension of the manifold is known, although this is not necessary: In Appendix B we include a practical method of empirically tuning the bandwidth parameter that also estimates the dimension. Following Coifman and Lafon (2006), Hein (2005, 2006) and Hein et al. (2005), in this manuscript we will assume that the data points lie exactly on the manifold embedded in the ambient Euclidean space. In fact, there is evidence that kernel based manifold learning methods are robust to perturbations which are smaller than the bandwidth parameter (Coifman and Lafon, 2006).

The remaining 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 (Jones, 1993; Jones and Foster, 1996; Jones and Zhang, 1999; Chen, 2000; Karunamuni and Alberts, 2005; Malec and Schienle, 2014).

Here 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 briefly review nonparametric density estimation on embedded manifolds. The boundary correction method using BDE is introduced in Section 3, and the results are demonstrated on several illustrative examples. We conclude with a brief discussion in Section 4.

2. Background

Assume one is given N samples $\{X_i\}_{i=1}^N$ (often assumed to be independent) of a probability distribution on $\mathbb{R}^n$ with a density function $f(x)$. The problem of nonparametric density estimation is to find an estimator $f_{h,N}(x)$ that approximates the true density function. A kernel density estimator is typically constructed (Parzen, 1962; Loftsgaarden and Quesenberry et al., 1965; Whittle, 1958) as

$$f_{h,N}(x) = \frac{1}{Nh^n} \sum_{i=1}^{N} K\left(\frac{\|x - X_i\|}{h}\right) \tag{1}$$

where the kernel function is defined via a univariate shape function K and $h \to 0$ as $N \to \infty$. The kernel function must be normalized to integrate to 1 for each h to have a consistent estimator.

The standard KDE formulation (1) assumes that the density is supported on the Euclidean space from which the data is sampled. However, real data may be restricted to lie on a lower dimensional submanifold of this Euclidean space. This assumption, which we call the *geometric prior*, is a potential workaround to the curse of dimensionality for high dimensional data. Since the geometric prior assumes that the density is supported on a submanifold of the ambient 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 (1990), who modernized the Fourier approach of Whittle (1958) using a generalized Fourier analysis on compact Riemannian manifolds without boundary, based on the eigenfunctions of the Laplace–Beltrami operator. The limitation of Hendriks (1990) 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 Pelletier (2005). 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 (Kim and Park, 2013). However, evaluating the kernel of Kim and Park (2013) 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, Rosenberg, 1997