# Manifold Learning Using Kernel Density Estimation and Local Principal Components Analysis

**Kitty Mohammed** KITTYMOHAMMED1985@GMAIL.COM
*Department of Statistics*
*University of Washington*
*Seattle, WA 98195-4322 USA*

**Hariharan Narayanan** HARIHARAN.NARAYANAN@TIFR.RES.IN
*School of Technology and Computer Science*
*Tata Institute of Fundamental Research*
*Mumbai, Maharashtra 400 005 India*

## Abstract

We consider the problem of recovering a $d-$dimensional manifold $\mathcal{M} \subset \mathbb{R}^n$ when provided with noiseless samples from $\mathcal{M}$. There are many algorithms (e.g., Isomap) that are used in practice to fit manifolds and thus reduce the dimensionality of a given data set. Ideally, the estimate $\mathcal{M}_{\text{put}}$ of $\mathcal{M}$ should be an actual manifold of a certain smoothness; furthermore, $\mathcal{M}_{\text{put}}$ should be arbitrarily close to $\mathcal{M}$ in Hausdorff distance given a large enough sample. Generally speaking, existing manifold learning algorithms do not meet these criteria. ( ) have developed an algorithm whose output is provably a manifold. The key idea is to define an approximate squared-distance function (asdf) to $\mathcal{M}$. Then, $\mathcal{M}_{\text{put}}$ is given by the set of points where the gradient of the asdf is orthogonal to the subspace spanned by the largest $n-d$ eigenvectors of the Hessian of the asdf. As long as the asdf meets certain regularity conditions, $\mathcal{M}_{\text{put}}$ is a manifold that is arbitrarily close in Hausdorff distance to $\mathcal{M}$. In this paper, we define two asdfs that can be calculated from the data and show that they meet the required regularity conditions. The first asdf is based on kernel density estimation, and the second is based on estimation of tangent spaces using local principal components analysis.

**Keywords:** manifold learning, KDE, local PCA, ridges

## 1. Introduction

It is often the case that high-dimensional data sets have lower-dimensional structure taking the form of a manifold. Manifold learning consists of algorithms that take a high-dimensional data set as input and output a fit of the manifold structure. Many of these algorithms (such as Isomap, Laplacian eigenmaps, locally linear embedding, etc.) are used in practice and have a theoretical literature supporting them. ( ) give a concise overview of these methods.

A drawback of most manifold learning algorithms is that if we are given data from a manifold, their output is not an actual manifold that is close to the original manifold. ( ) develop an algorithm whose output is provably a manifold of certain smoothness. They start by defining an approximate squared-distance function (asdf) from the data in a manner that uses exhaustive search, utilizing the data only indirectly. Thus, a very large number of potential asdfs are examined before an approximately optimal one is chosen. In this paper, we do away with the exhaustive search, albeit in the specific case of noiseless data that is sampled uniformly from a manifold.

( ) prove a key theorem that states that as long as we are able to define an asdf meeting certain general conditions, their algorithm outputs a set that is a manifold with bounded smoothness and Hausdorff distance to the original manifold. We demonstrate two different methods of estimating the true manifold via asdfs that can be calculated from the data. The two asdfs in our paper are based on 1) kernel density estimation, and 2) approximating the manifold using tangent planes which are in turn approximated with local principal components analysis (PCA).

( ) learn manifolds by forming a kernel density estimator (KDE) from the data points and finding its $d$-dimensional ridges. We give a more precise definition later, but a ridge is essentially a higher-dimensional analog of the mode and is related to the output set from the algorithm of ( ).
( ) give a practical method for finding the ridges through a variant of gradient descent where the descent is constrained to the subspace spanned by the largest eigenvectors of the Hessian of the KDE. We state their algorithm in Section of our paper and use it to produce simulation results. Although they only apply subspace-constrained gradient descent to find ridges of the KDE, the method is more general and can be used to find ridges of both of our asdfs.

### 1.1 Related work

Manifold learning has existed as an area of statistics and machine learning since the early 2000s. Some classical manifold learning algorithms are Isomap ( , ), locally linear embedding ( , ), and Laplacian eigenmaps ( , ). Many of these early algorithms rely on spectral graph theory and start off by constructing a graph which is then used to produce a lower-dimensional embedding of the data set. The theoretical guarantees are centered around proving that asymptotically, certain values such as the geodesic distance can be approximated to arbitrary precision.

More recently, there have been quite a few papers combining ridge estimation with manifold learning (including the work of , ). Some early results on ridge estimation are due to ( ), ( ), and ( ). Ridge sets can be constructed to estimate a probability density or an embedded submanifold. Theoretical guarantees in this setting have been given by ( ), ( ), ( ), and ( ). Of these, the most relevant results for us are from ( ). They prove that as the sample size goes to infinity, their ridge set gets arbitrarily close to an underlying manifold in Hausdorff distance. ( ) also define a procedure related to ridge estimation methods that can be used to estimate an underlying manifold. For our purposes, the major advances of their work are twofold. First, their method is general; as long as a function meets a few conditions, it can be used to define an estimator that can be made arbitrarily close to an underlying manifold in Hausdorff distance. Furthermore, they show that this