Principal Manifolds of Middles: A Framework and Estimation Procedure Using Mixture Densities

Kun Meng\textsuperscript{1,*} and Ani Eloyan\textsuperscript{1}

\textsuperscript{1}Department of Biostatistics, Brown University School of Public Health, Providence, RI 02903, USA
\textsuperscript{*}Corresponding Author: e-mail: kun.meng@brown.edu, Address: 121 S. Main St, Providence, RI 02903, USA.

September 5, 2018

Abstract

Principal manifolds are used to represent high-dimensional data in a low-dimensional space. They are high-dimensional generalizations of principal curves and surfaces. The existing methods for fitting principal manifolds have several shortcomings: model bias, heavy computational burden, sensitivity to outliers, and difficulty of use in applications. We propose a novel method for modeling principal manifolds that addresses these limitations. It is based on minimization of penalized mean squared error functionals, providing a nonlinear summary of the data points in Euclidean spaces. We introduce the framework in the context of principal manifolds of middles and develop an estimate by proposing a high-dimensional mixture density estimation procedure. The Sobolev embedding theorem guarantees the regularity of the derived manifolds and analytical expressions of the embedding maps are obtained. The algorithm is computationally efficient and robust to outliers. We used simulation studies to illustrate the comparative performance of the proposed method in low-dimensions and found that it performs better than competitors. In addition, we analyze computed tomography images of lung cancer tumors focusing on two important clinical questions - estimation of the tumor surface and identification of tumor interior classifier. We used the obtained analytic expressions of embedding maps to construct a tumor interior classifier.

Keywords: projection index, self-consistency, Sobolev spaces, tumor interior classifier
1 Introduction

Consider a dataset of $n$ observations for two random variables $X$ and $Y$. It is common to visualize the joint behavior of $X$ and $Y$ in a scatter plot as presented in Figure 1. The type of summary measure used to quantify this joint behavior depends on the goal of the analysis. If the goal is to find a linear rule for modeling the response, say $Y$, using the realizations of explanatory variable $X$, $E(Y|X)$ can be modeled as a linear function of $X$. The linear regression procedure is equivalent to finding a line minimizing the sum of squared vertical deviations shown in Figure 1. The result of linear regression for a given dataset is very sensitive to the choice of explanatory variable. The difference between two linear regression results for the same dataset with different choices of explanatory variables is shown in Figure 1. In many situations, we do not have a preferred variable that we wish to label as explanatory variable. Then simply assigning the roles of explanatory variable could lead to a poor summary. An alternative to linear regression is to summarize the data points by a linear manifold that treats $X$ and $Y$ symmetrically. For example, linear principal component analysis (PCA) (Jolliffe (1986)) estimates a line that minimizes the squared orthogonal deviations as shown in Figure 1.

As an analogue to the generalization from linear regression to nonlinear regression, which allows nonlinearity of the manifolds minimizing the sum of squared vertical deviations, a nonlinear PCA was proposed by Hastie (1984) and Hastie and Stuetzle (1989) (and referred to as the HS algorithm hereafter) by allowing nonlinearity of the manifolds minimizing the sum of squared orthogonal deviations from data points. Suppose $\{z_i\}_{i=1}^I$ is a set of observations from a random $D$-vector, the first $d$ linear principal components, with $d < D$, are given by $\arg\inf_{L \in \mathcal{L}} \sum_{i=1}^I \|z_i - \Pi_L(z_i)\|_2^2$, where $\mathcal{L}$ denotes the collection of all $d$ dimensional planes in $\mathbb{R}^D$ and $\Pi_L(z_i)$ denotes the orthogonal projection of $z_i$ onto plane $L$. Linear PCA can be generalized by extending $\mathcal{L}$ to a larger collection including manifolds with nonzero curvature. Motivated by the generalization of linear PCA allowing nonlinearity, we propose a framework for estimation of principal manifolds in high dimensions. The principal curves and surfaces are special cases of the proposed method. As shown by Hastie and Stuetzle (1989), their proposed principal curves and surfaces have model bias. To overcome the bias, Tibshirani (1992) proposed a new definition of principal curves based on a mixture model. Duchamp et al. (1996) studied the differential geometric properties of the principal curve and analyzed their first and second variations. Yue et al. (2016) proposed an effective algorithm for numerically deriving principal surfaces. The concepts of principal curve and surface fitting have been widely used in various fields of applications since their introduction. For example, Chen et al. (2004) applied the principal-curve algorithm to model freeway traffic streams. Yue et al. (2016) used principal surfaces to parameterize manifold-like...
white matter tracts in the brain using diffusion tensor imaging data.

Principal curves and surfaces provide a natural geometric framework for nonlinear dimension reduction. Martínez-Morales (2004) considered the principal curve framework from a differential manifold viewpoint, presenting principal curves as a special case of manifold fitting. Martínez-Morales (2004) further generalized the framework into principal embedding and introduced harmonic energy as a regularizing term for determining a local minimum of principal embedding. However, this work did not provide a practical algorithm for constructing a principal embedding. Dai and Müller (2017) proposed Riemannian functional principal component analysis to serve as an intrinsic principal component analysis of Riemannian manifold-valued functional data. Many other frameworks exist for dimension reduction depending on various assumptions of data structure and modeling assumptions. Sammon’s mapping (Sammon, 1969) is one of the first nonlinear dimension reduction techniques. The algorithm maps a high dimensional dataset to a lower dimensional space by preserving the higher dimensional structure of inter-point distances in the lower dimensional projection. Curvilinear distance analysis proposed by Demartines and Hérault (1997) trains a self-organizing
neural network to fit the manifold and seeks to preserve geodesic distances in its embedding. Gaussian process latent variable models (Lawrence, 2005) find a lower dimensional nonlinear embedding of high dimensional data using a Gaussian process. Local tangent space alignment (Zhang and Zha, 2004) computes the tangent space at every point by computing the first $d$ principal components in each local neighborhood followed by optimization to find an embedding that aligns the tangent spaces. Diffeomorphic dimensionality reduction (Walder and Schölkopf, 2009) learns a smooth diffeomorphic mapping which transports the data onto a lower dimensional linear subspace. While these and other methods have been proposed for dimension reduction, they often lack in explicit definition of modeling assumptions, have high computational burden, or can only be used in limited applications.

From the differential geometry viewpoint, curves and surfaces are 1 and 2 dimensional manifolds, respectively. In this article, we develop a framework of principal manifolds based on the assumption that the estimated lower dimensional projections are in a Sobolev space. The proposed framework is a higher dimensional generalization of principal curves and surfaces. Based on the Sobolev embedding theorem (Rudin (1991)), the smoothness of the principal manifolds in this framework is guaranteed. To avoid the model bias in the generalization of the HS estimate, reduce computational burden, and obtain estimates robust to outliers, we propose middles of random vectors. Our developed principal manifold estimation algorithm is based on modeling the middle of a random vector using a high dimensional mixture density estimation procedure. The proposed algorithm provides the analytic expressions of the derived smooth manifolds, which are not provided by the existing procedures.

The proposed algorithm can be applied in several fields of applications (e.g. medical imaging, 3D printing and engineering) where high dimensional objects are observed with useful low-dimensional features. In this paper, we consider a problem in radiation therapy for patients with cancer tumors. Computed tomography (CT) images are collected for cancer patients routinely for disease diagnosis, surgical planning, and treatment. One of the important questions in analyzing the CT scans is the delineation of the tumor surface. Often a radiologist marks several voxels (i.e. three-dimensional pixels) on the periphery of the tumor, followed by the use of an algorithm to estimate the tumor surface based on these vertices. Clearly, the selection of peripheral vertices by the radiologist is a time consuming process. We show that our proposed algorithm can provide a high quality estimate of the tumor surface by using relatively few vertices. Secondly, for radiation therapy planning it is important to identify whether a voxel is an interior point of the tumor. Using the proposed algorithm, we develop a classifier identifying the interior points of a tumor. This classifier can help radiation therapists identify the target of ionizing radiation and spare the region
outside of a tumor from receiving doses above specified tolerance levels. Even though outside of the scope of this paper, our proposed algorithm can potentially be used to 3D print artificial human organs using biomaterials.

The rest of the article is organized as follows. In Section 2, we discuss some useful results on high dimensional geometry, which will be used throughout the article. The framework of principal manifolds is proposed in Section 3 where the relationship between principal manifolds, principal curves and surfaces, and self-consistency is discussed from functional viewpoint. To avoid model bias and high computational burden, we propose middles of random vectors in Section 3.2. In Section 4, we present the principal manifold algorithm to numerically derive principal manifolds from datasets. Simulation studies presented in Section 5 illustrate the performance of the proposed algorithm compared to other existing methods. In Section 6, tumor imaging datasets are analyzed using the proposed algorithm and a classifier identifying the interior region of tumors is developed. Finally, Section 7 concludes the article.

2 Geometry in the High Dimensional Space

Throughout this article, we use the positive integers $d$ and $D$ to denote the dimensions of interest, with $d < D$, such that $D$ denotes the dimension of the space where we observe the data and $d$ denotes the dimension the low-dimensional space of the target fitted manifold. The (trivial) manifolds can be defined as follows (see Chern (1951) and Milnor (1997) for more details).

**Definition 2.1.** Suppose $f$ is a bijective map from $\mathbb{R}^d$ to a subset of $\mathbb{R}^D$ such that both $f$ and its inverse $f^{-1}$ are continuous, then $M^d_f := \{f(t) : t \in \mathbb{R}^d\}$ is called a $d$-dimensional manifold determined by $f$. Furthermore, $M^d_f$ is a manifold embedded into $\mathbb{R}^D$ and $f$ is referred to as the embedding map of $M^d_f$.

From topological viewpoint, the trivial manifolds defined above are homeomorphic to $\mathbb{R}^d$. However, some surfaces of interest may not be homeomorphic to $\mathbb{R}^d$. The existence of such surfaces gives the motivation for extending the above definition to non-trivial manifolds defined as follows.

**Definition 2.2.** Let $M^d$ be a subset of $\mathbb{R}^D$. If there exists a family of $d$-dimensional trivial manifolds embedded into $\mathbb{R}^D$, say $\{M^d_f\}_{f \in \mathcal{F}}$, where $\mathcal{F}$ defines a family of embedding maps, such that $M^d = \bigcup_{f \in \mathcal{F}} M^d_f$, then $M^d$ is called a non-trivial $d$ dimensional manifold. $M^d$ is often denoted by $M^d_{\mathcal{F}}$ and $\mathcal{F}$ is called an embedding family of $M^d_{\mathcal{F}}$.

Since to fit a non-trivial manifold, it suffices to fit several trivial manifolds, we mainly consider trivial manifold fitting in this article. An approach to non-trivial manifold fitting will be discussed in Section 6.
2.1 Projection Indices

A key concept in defining and deriving principal manifolds is the projection index. Suppose \( x \in \mathbb{R}^D \) and \( M_f^d \) is a manifold determined by \( f \). Intuitively, the projection index of \( x \) onto \( M_f^d \) is the parameter \( t \) in \( \mathbb{R}^d \) such that \( f(t) \) is closest to \( x \). Formally, the projection index is \( \pi_f(x) := \arg\inf_{t \in \mathbb{R}^d} \|x - f(t)\|_{\mathbb{R}^D}^2 \). The left panel of Figure 2 illustrates a projection index for a 1-dimensional manifold \( M^1_f \). However, this formal definition is not rigorous as there might exist more than one \( t \) such that \( \|x - f(t)\|_{\mathbb{R}^D}^2 = \inf_{t' \in \mathbb{R}^d} \|x - f(t')\|_{\mathbb{R}^D}^2 \) resulting in ambiguity in the choice of \( t \) minimizing the distance. For example, in the right panel of Figure 2, \( x \) is located at the center of a semi-circle \( f \) implying \( \|x - f(t_1)\|_{\mathbb{R}^D} = \|x - f(t_2)\|_{\mathbb{R}^D} = \|x - f(t_3)\|_{\mathbb{R}^D} \) for three points in the space denoted as \( t_1, t_2 \) and \( t_3 \). Hence, the points \( t_1, t_2 \) and \( t_3 \) all minimize \( \|x - f(t)\|_{\mathbb{R}^D} \). For \( d = 1 \), a rigorous definition of \( \pi_f \) is given by Hastie and Stuetzle (1989) as \( \pi_f = \sup \{ t : \|x - f(t)\|_{\mathbb{R}^D}^2 = \inf_{t'} \|x - f(t')\|_{\mathbb{R}^D}^2 \} \).

If \( d > 1 \), then \( \mathbb{R}^d \) is only a partially ordered set, whereas \( \mathbb{R}^1 \) is a totally ordered set (Schmidt (2011)), hence the definition of projection index for \( d = 1 \) does not apply in higher dimensions. We introduce a general definition of a projection index for a high-dimensional space with the desirable property of uniqueness. For any \( x \in \mathbb{R}^D \), let \( S^f_x := \arg\inf_{t \in \mathbb{R}^d} \|x - f(t)\|_{\mathbb{R}^D} \) define the collection of all points \( t \in \mathbb{R}^d \) such that \( f(t) \) is closest to \( x \).

**Figure 2:** Illustration of a projection index onto a 1-dimensional manifold \( M^1_f \) (left). The plot on the right shows three possible values of \( t \) that minimize the distance between the center of the semi-circle and the values of the function at those points.

**Lemma 2.1.** If \( f : \mathbb{R}^d \to \mathbb{R}^D \) is continuous, then \( S^f_x \) is non-empty and compact for all \( x \in \mathbb{R}^D \).

The proof of Lemma 2.1 is given in Appendix 9.4.

**Definition 2.3.** Let \( M_f^d \) denote a manifold determined by a continuous map \( f : \mathbb{R}^d \to \mathbb{R}^D \). For each \( x \in \mathbb{R}^D \), if there exists a probability measure \( \nu_x \) defined on \( S^f_x \), then for \( (M_f^d, S^f_x, \nu_x) \), the projection index is defined by \( \pi_f(x) := \left( \int_{S^f_x} t_1 \nu_x(dt), \cdots, \int_{S^f_x} t_d \nu_x(dt) \right) \).
The following lemma ensures that the projection defined above is unique.

**Lemma 2.2.** The projection $\pi_f$ is well defined if $f$ is a continuous function.

**Proof.** From Lemma 2.1, the continuity of $f$ implies the compactness of $S^f_x$. Then $m_i = \sup_{t \in S^f_x} |t_i| < \infty$. Hence $\pi_f$ is well defined.

If $S^f_x$ contains only one point $t^*_x$, then the probability measure on $S^f_x$ can only be the point mass at $t^*_x$ and $\pi_f(x) = t^*_x$. In the special case when $d = 1$, $T_x := \sup \left\{ t : t \in S^f_x \right\} \in S^f_x$. Let $\delta_{(T_x)}$ denote the point mass at $T_x$, then the triple $(M^f_x, S^f_x, \delta_{(T_x)})$ implies the projection index defined by [Hastie and Stuetzle (1989)] implying that the proposed projection index is a generalization of the distance based projection that is the bases of the HS algorithm.

## 3 Principal Manifolds in High Dimensions

To guarantee the smoothness of principal manifolds, we restrict the components of $f$, which determine a manifold, to Sobolev spaces [Adams and Fournier (2003)]. The following notations on Sobolev spaces will be used throughout this article. While we provide the notations for a general $s < \frac{d}{2}$, we only use the special case where $s = 0$.

(i) $\tilde{H}^s(\mathbb{R}^d)$ denotes a homogeneous Sobolev space defined by

$$\tilde{H}^s(\mathbb{R}^d) = \left\{ u \in \mathcal{S}'(\mathbb{R}^d) : \hat{u} \in L^1_{\text{loc}}(\mathbb{R}^d), \int_{\mathbb{R}^d} \|\xi\|^{2s} \left| \hat{u}(\xi) \right|^2 d\xi < \infty \right\},$$

where $L^1_{\text{loc}}(\mathbb{R}^d)$ denotes the collection of all locally integrable functions in $\mathbb{R}^d$, $\mathcal{S}'(\mathbb{R}^d)$ is the collection of all tempered distributions, and $\hat{u}$ denotes the Fourier transform of $u$.

(ii) $\nabla^{-2}\tilde{H}^s(\mathbb{R}^d)$ denotes the collection of distributions such that the norm of their second derivative belongs to a homogeneous Sobolev space, i.e. $\nabla^{-2}\tilde{H}^s(\mathbb{R}^d) := \left\{ u \in \mathcal{S}'(\mathbb{R}^d) : \|\nabla^2 u\|_{\mathbb{R}^{d \times d}} \in \tilde{H}^s(\mathbb{R}^d) \right\}$, where $\mathcal{S}'(\mathbb{R}^d)$ is the collection of all distributions, $\nabla^2 u = \left( \frac{\partial^2 u}{\partial t_i \partial t_j} \right)_{1 \leq i, j \leq d}$ is the Hessian matrix of $u$ and $\|\nabla^2 u\|_{\mathbb{R}^{d \times d}}^2 = \sum_{i, j=1}^d \left| \frac{\partial^2 u}{\partial t_i \partial t_j} \right|^2$.

(iii) $\nabla^{-2}\tilde{H}^s(\mathbb{R}^d \to \mathbb{R}^D)$ denotes the collection of $\mathbb{R}^d \to \mathbb{R}^D$ maps such that each of their components is in $\nabla^{-2}\tilde{H}^s(\mathbb{R}^d)$.

(iv) $C(\mathbb{R}^d \to \mathbb{R}^D)$ denotes the collection of continuous $\mathbb{R}^d \to \mathbb{R}^D$ maps. For simplicity, we denote $C(\mathbb{R}^d \to \mathbb{R}^D) \cap \nabla^{-2}\tilde{H}^s(\mathbb{R}^d \to \mathbb{R}^D)$ by $C \cap \nabla^{-2}\tilde{H}^s(\mathbb{R}^d \to \mathbb{R}^D)$. 

7
Motivated by Hastie and Stuetzle (1989) and Martínez-Morales (2004), we propose the following definition of a functional of manifolds measuring the balance between data fit and smoothness from variational viewpoint. The principal manifold is then defined as the minima of this functional.

**Definition 3.1.** Suppose $X$ is a random $D$-vector, its distribution is determined by probability measure $\mathbb{P}$ and has finite second moments, $\mathbb{E}_\mathbb{P}$ is the expectation with respect to probability measure $\mathbb{P}$, $\mathcal{C} \subset C \cap \nabla^{-2}\hat{H}^0(\mathbb{R}^d \to \mathbb{R}^D)$, and $w \in [0, +\infty]$. The penalized mean squared error functionals for $X$ are given by

$$M_{w,\mathbb{P}}(f, g) := \mathbb{E}_\mathbb{P} \left\| X - f (\pi_g(X)) \right\|_{\mathbb{R}^D}^2 + w \left\| \nabla^2 f \right\|_{L^2(\mathbb{R}^d)}^2, \quad M_{w,\mathbb{P}}(f) := M_{w,\mathbb{P}}(f, f), \quad f, g \in \mathcal{C},$$

where $\left\| \nabla^2 f \right\|_{L^2(\mathbb{R}^d)}^2 = \int_{\mathbb{R}^d} \sum_{i=1}^D \left\| \nabla^2 f^l(t) \right\|_{\mathbb{R}^d}^2 dt$ and $f^l$ is the $l$th component of $f$. A manifold $M_d^f$ determined by $f^* \in \mathcal{C}$ is called a $(\mathcal{C}, w)$ principal manifold for $X$ if $f^* = \arg \inf_{f \in \mathcal{C}} M_{w,\mathbb{P}}(f)$, where $w$ is called the smoothing parameter.

Since $\mathcal{C} \subset \nabla^{-2}\hat{H}^0(\mathbb{R}^d \to \mathbb{R}^D), \left\| \nabla^2 f \right\|_{L^2(\mathbb{R}^d)}^2$ is finite. As we are using functions in Sobolev spaces to develop the framework of principal manifolds, the discussion heavily depends on the Fourier transform defined on $\mathbb{R}^d$ rather than a unit interval of $\mathbb{R}^d$. Hence, the principal manifolds in this article are parameterized on $\mathbb{R}^d$ rather than the unit interval as in previous work. The first term of (3.1) measures the fit to the data and is a mean squared error term, while $\left\| \nabla^2 f \right\|_{L^2(\mathbb{R}^d)}^2$ in the second term describes the curvature of $M_d^f$. In other words, the second term in (3.1) penalizes the curvature of the fitted manifold. $w$ establishes a tradeoff between the two. Intuitively, as $w$ varies from 0 to $+\infty$, the corresponding principal manifold varies from rough to extremely smooth (flat), and $w \in (0, \infty)$ indexes the class of manifolds of varying smoothness between the two extremes. To formalize this claim, we denote by $F(\cdot, w) := \arg \inf_{f \in \mathcal{C}} M_{w,\mathbb{P}}(f)$, for all $w \in [0, +\infty]$, the set of principal manifolds of $f$ as a function of $w$. The map $F : \mathbb{R}^d \times [0, +\infty] \to \mathbb{R}^D$ is a homotopy between two special cases: (i) A $(\mathcal{C}, +\infty)$ principal manifold for $X$ is uniquely determined by the mean and first $d$ linear principal components of $X$ (as shown in Theorem 3.1); (ii) A $(\mathcal{C}, 0)$ principal manifold for $X$ is a principal manifold of self-consistent type in some specific settings (Theorem 3.3).

**Theorem 3.1.** Suppose $X$ is a random $D$-vector, $v_1, v_2, \ldots, v_D$ and $\lambda_1, \lambda_2, \ldots, \lambda_D$ are eigenvectors and eigenvalues of the variance of $X$, respectively. $v_i$ corresponds to $\lambda_i$ for $i = 1, 2, \ldots, D$ and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_D$. Let $\mathcal{L}$ be the collection of all $d$-variate linear functions and $\mathcal{C} \subset C \cap \nabla^{-2}\hat{H}^0(\mathbb{R}^d \to \mathbb{R}^D)$. Then both $(\mathcal{C}, +\infty)$ and $(\mathcal{L}, w)$ principal manifolds are linear manifolds $\mathbb{E}X + \text{Span}\{v_1, v_2, \ldots, v_d\} := \left\{ \mathbb{E}X + \sum_{i=1}^d \alpha_i v_i : \alpha_i \in \mathbb{R}^1, i = 1, 2, \ldots, d \right\}$.

The proof of Theorem 3.1 is shown in Appendix 9.5.
To estimate the principal manifolds given in Definition 3.1, we implement the following iteration steps with an initial value $f_0$.

$$f_{(n+1)} = \arg \inf_{f \in \mathcal{C}} M_{w, P}(f, f_{(n)}).$$  \hfill (3.2)

$f_0$ is usually obtained by estimating the first $d$ principal components. The following theorem shows that the HS principal-curve algorithm given by

$$f_{(n+1)} := T f_{(n)}, \text{ where } T f_{(n)}(t) := \mathbb{E} \left( X | \pi f_{(n)} = t \right), \quad t \in \mathbb{R}^1, \quad n \in \mathbb{N}$$  \hfill (3.3)

is a special case of (3.2) with $d = 1, w = 0$.

**Theorem 3.2.** Suppose $\mathcal{C} \subset \nabla^{-2} \mathcal{H}^0(\mathbb{R}^1 \to \mathbb{R}^D), f_{(n)} \in \mathcal{C}$ and $T f_{(n)}$ is defined by (3.3). If $T f_{(n)} \in \mathcal{C}$, then $T f_{(n)} = \arg \inf_{f \in \mathcal{C}} M_{0, P}(f, f_{(n)})$.

The proof of Theorem 3.2 is presented in Appendix 9.8.

### 3.1 Principal Manifolds of self-consistent type

The concept of principal curves was defined on the basis of self-consistency. In this context a principal curve approximating a high-dimensional scatter plot is such that at each point $t$, the function $f(t)$ is the average of the points projected to that point. Next, we propose a generalization of this property in high dimensions.

Again, we define the generalization in $\mathbb{R}^d$, while the previous work used the unit interval, as $\mathbb{R}^d$ and the unit interval are equivalent up to a reparameterization.

**Definition 3.2.** Suppose $X$ is a random $D$-vector, its distribution is determined by probability measure $\mathbb{P}$ and has finite second moments, and $\mathcal{C} \subset C(\mathbb{R}^d \to \mathbb{R}^D)$. $f \in \mathcal{C}$ satisfies the self-consistency condition if

$$\mathbb{E}_\mathbb{P} (X | \pi f(X) = t) = f(t).$$  \hfill (3.4)

If there exists an $f \in \mathcal{C}$ satisfying the self-consistency condition (3.4) then the manifold $M_f^d$ determined by $f$ is called a $\mathcal{C}$ principal manifold of self-consistent type for $X$.

The definition of principal curves (Hastie and Stuetzle 1989) is a special case of principal manifolds of self-consistent type given by Definition 3.2 when $d = 1$, implying that principal curves and manifolds share the same intuitive motivation. For any $t$ in the parameter space $\mathbb{R}^d$, $f(t)$ is the local expectation over points.
such that $f(t)$ is their closest point on $M_f$. When $w = 0$, principal manifolds and principal manifolds of self-consistent type are equivalent for $d = 1$ as shown in the following theorem.

**Theorem 3.3.** Suppose $\mathcal{C} \subset \nabla^{-2} \mathcal{H}(\mathbb{R}^1 \to \mathbb{R}^D)$ and $\mathbb{P}$ is a probability measure with a bounded support, then $(\mathcal{C}, 0)$ principal manifolds defined in Definition 3.1 are $\mathcal{C}$ principal manifolds of self-consistent type.

The proof of Theorem 3.3 is given in Appendix 9.6.

### 3.2 Principal Manifolds for The Middles of Random Vectors

While the random $D$-vector $X$ can have any distribution $\mathbb{P}$, in practice, $\mathbb{P}$ is usually set to be the empirical distribution estimated using the data. This approach may be problematic depending on the practical question of interest. For instance, an algorithm based on the use of the empirical distribution can be computationally demanding limiting its use when the sample size is large. In addition, the principal manifolds given in 3.1 share the same model bias as the HS principal curves. In order to solve the two problems and motivated by Tibshirani (1992), we give a new definition of principal manifolds in this subsection by introducing the concept of *the middles of a random vector*.

In order to define the middle of a random $D$-vector $X$, we assume that each realization of $X$ is generated in two stages: step 1, a realization of a latent random $D$-vector $T$, say $t$, is generated according to some probability measure $p_T(dt)$; step 2, a realization of $X$ is generated according to the conditional distribution given $T = t$, which is determined by a transition probability $p_{X|T}(t, dx)$, with $\mathbb{E}(X|T = t) = t$. For any $t$, if $p_{X|T}(t, dx)$ gives a distribution whose variance is isotropic in all directions, then $t$ is in the center of the realizations of $X|T = t \sim p_{X|T}(t, dx)$.

**Definition 3.3.** Suppose $X$ is a random $D$-vector with finite first moments. If there exists a random $D$-vector $T$ such that $\mathbb{E}(X|T = t) = t$ for all $t$ in the support of $T$, then $T$ is called a middle of $X$.

Figure 3 shows an intuitive illustration of the geometric relationship between a random 2-vector $(X, Y)$ and a middle of it defined by $T$. It can be observed that $T$ is located in the middle of the realizations $(X, Y)$.

**Definition 3.4.** Suppose $X$ and $T$ are random $D$-vectors and $T$ is a middle of $X$. The distribution of $T$ is represented by probability measure $p_T(dt)$. Then, for a collection $\mathcal{C} \subset C \cap \nabla^{-2} \mathcal{H}(\mathbb{R}^d \to \mathbb{R}^D)$, a $(\mathcal{C}, w)$ principal manifold for $T$ denoted by $M_{f^*}$ with $f^* = \arg\inf_{f \in \mathcal{C}} M_{w, p_T}(f)$, is called a $(\mathcal{C}, w)$ principal manifold for a middle of $X$.

In applications, middles of random vectors are not observable. For a random $D$-vector $X$, we propose to construct a sequence of random $D$-vectors $\{X_N\}_{N \in \mathbb{N}}$ such that each $X_N$ has a known middle $T_N$. If $X_N$
Figure 3: Example of a middle of a random 2-vector \((X, Y)\). The blue dots indicate simulated data points, while the red line corresponds to a middle for the random vector.

converges to \(X\) as \(N \to \infty\), then the principal manifold of the middle \(T_N\) of \(X_N\) with an appropriately large \(\hat{N}\) will be used as an estimate for the principal manifold of \(X\) with reduced model bias. For an absolutely continuous random \(D\)-vector \(X\), the following theorem, which is a generalization of Theorem 2.1 in Eloyan and Ghosh (2011), provides a construction of a sequence \(X_N\) and guarantees uniform convergence of \(X_N\) to \(X\).

**Theorem 3.4.** Suppose the following conditions are satisfied.

(i) \(p \in C(\mathbb{R}^D)\) is a probability density function, where \(C(\mathbb{R}^D)\) is the collection of all uniformly continuous and bounded functions defined on \(\mathbb{R}^D\), and \(\text{supp}(p) := \{x \in \mathbb{R}^D : p(x) > 0\}\).

(ii) For each \(N \in \mathbb{N}\), there exists a finite set \(M_N = \{\mu_{j,N} \in \mathbb{R}^D : j = 1, 2, \cdots, N\}\).

(iii) There exists a partition of \(\text{supp}(p)\), say \(\{A_{j,N}\}_{j=1}^N\), with \(\bigcup_{j=1}^N A_{j,N} = \text{supp}(p)\) and \(A_{i,N} \cap A_{j,N} = \emptyset\) whenever \(i \neq j\), such that \(A_{j,N} \cap M_N = \{\mu_{j,N}\}\) and

\[
d_N := \sup_{j=1,2,\cdots,N} \sup_{x_1,x_2 \in A_{j,N}} \|x_1 - x_2\|_{\mathbb{R}^D} \to 0, \quad N \to \infty.
\]

(iv) \(\psi\) is a strictly positive function on \(\mathbb{R}^D\) with \(\int \psi = 1\) and \(\psi \in C(\mathbb{R}^D)\).

(v) \(\sigma_N > 0\) and \(\sigma_N \to 0\) as \(N \to \infty\).

(vi) The triple \((\psi, d_N, \sigma_N)\) satisfies

\[
\sup_{\mu_1, \mu_2 : ||\mu_1 - \mu_2\|_{\mathbb{R}^D} \leq d_N} \left\{ \left\| \frac{1}{\sigma_N} \psi \left( \frac{x - \mu_1}{\sigma_N} \right) - \frac{1}{\sigma_N} \psi \left( \frac{x - \mu_2}{\sigma_N} \right) \right\|_{L^\infty(\mathbb{R}^D)} \right\} \to 0, \quad N \to \infty.
\]
Let $\Theta_N := \{ \theta \in [0,1]^N : \sum_{j=1}^N \theta_j = 1 \}$ and

$$p_N(x|\theta_N) := \sum_{j=1}^N \theta_{j,N} \frac{1}{\sigma_N} \psi \left( \frac{x - \mu_{j,N}}{\sigma_N} \right),$$

(3.5)

where $\theta_N = (\theta_{1,N}, \theta_{2,N}, \ldots, \theta_{N,N}) \in \Theta_N$. Then $\inf_{\theta_N \in \Theta_N} \| p_N(\cdot|\theta_N) - p \|_{L^2(\mathbb{R}^D)} \to 0$ as $N \to \infty$.

The proof of this theorem is given in Appendix 9.7. The parameters $\{M_N\}_{N \in \mathbb{N}}$, $\sigma_N$, $N$ and $\theta_N$ can be estimated by the high dimensional mixture density estimation (HDMDE) procedure presented in Appendix 9.2 which is a generalization of the 1 dimensional MDE procedure presented by Eloyan and Ghosh (2011).

In the rest of this article, the kernel function $\psi$ is set to be the Gaussian kernel on $\mathbb{R}^D$.

From Theorem 3.4, the sequence of random vectors $X_N$ whose distribution is determined by the probability measure $p_N(x|\theta_N) dx$ converges to $X$ in a uniform sense. It is straightforward that the random $D$-vector $T_N$ following

$$p_{T_N}(dt) = \sum_{j=1}^N \theta_{j,N} \delta_{\{\mu_{j,N}\}}(dt)$$

(3.6)

is a middle of $X_N$. Using the asymptotic hypothesis testing procedure within the HDMDE algorithm we can estimate $\hat{N}$ such that the distance between $p_N(x|\theta_N)$ and $p_{N+1}(x|\theta_{N+1})$ is smaller than a desired value (see Appendix 9.2 for more details). The $(\mathcal{C}, w)$ principal manifold for $T_N$ is given by

$$M_{w,p_{T_N}(dt)}(f,g) := \sum_{j=1}^N \theta_{j,N} \| \mu_{j,N} - f(\pi_g(\mu_{j,N})) \|_{L^2(\mathbb{R}^D)}^2 + w \| \nabla^2 f \|_{L^2(\mathbb{R}^D)}^2, \; f,g \in \mathcal{C}.$$  

The manifold $M_{f^*}$ with $f^* = \arg\inf_{f \in \mathcal{C}} M_{w,p_{T_N}(dt)}(f,f)$ is an estimate of the principal manifold of $X$ with reduced model bias. Since $\hat{N}$, estimated by an asymptotic hypotheses testing within the HDMDE procedure, is usually much smaller than the sample size, this approach reduces computational burden often significantly.

A suggestion for choosing $w$ is

$$w = \kappa \times \overline{\theta_{1,N}} \times \sigma_N,$$

(3.7)

where $\overline{\theta_{1,N}} = \frac{1}{N} \sum_{j=1}^N \theta_{j,N}$. (3.7) will be used through the rest of this article. The rationale for (3.7) is given as follows.

Motivation 1: The more curvature the target manifold has, the smaller the value of $w$. An increase in the curvature of the target manifold tends to result in a larger $\hat{N}$. From the HDMDE procedure, a larger value of $\hat{N}$ implies a smaller value of $\overline{\theta_{1,N}}$. Hence, we set $w$ proportional to $\overline{\theta_{1,N}}$ to penalize for the curvature in the function.
**Motivation 2:** The higher the random noise in the data, the smaller $w$ should be to keep the curvature of the derived manifold comparatively low and resist noise. In applications, if a given data set has a low signal-to-noise ratio the estimated value of $\sigma_N$ tends to be higher. Hence, it is reasonable to set $w$ proportional to $\sigma_N$ to penalize for possible curvature resulting from the increased random noise in the data.

### 4 A Principal Manifold Estimation Algorithm

Based on the proposed HDMDE algorithm and the iteration scheme (3.2) we propose a principal manifold estimation (PME) algorithm to estimate the principal manifold of a random vector with reduced model bias. As we proposed in Section 3.2, $X_N$ converges to $X$ and we can use $X_N$’s middle $T_N$ to estimate the principal manifold of $X$. Since the distribution of $T_N$ is given by $p_{T_N}(dt)$ in (3.6), from (3.2), the principal manifold of $T_N$ can be derived by the following iteration.

$$f_{(n+1)} = \arg \inf_{f \in C} \left\{ \sum_{j=1}^{N} \theta_{j,N} \left( \mu_{j,N} - f \left( \pi f_{(n)} (\mu_{j,N}) \right) \right)^2_{R^D} + w \| \nabla^2 f \|_{L^2(R^d)}^2 \right\}. \tag{4.1}$$

In this section, we give an analytic formula for deriving $f_{(n+1)}$ from $f_{(n)}$ within $C$ when $d \leq 3$. In applications, we usually require the derived manifolds to be continuous and to some degree smooth. For example, the surface of a tumor, which can be estimated by the proposed algorithm, is naturally expected to be smooth. Hence, we propose to choose $C$ such that the choice guarantees the smoothness of the derived manifolds. In the meantime, $C$ is expected to contain most of the $R^d \to R^D$ maps we use in applications. When $d \leq 3$, the Sobolev space $\nabla^{-2}H^s(R^d)$ is an appropriate choice for $C$ in the minimization above. It is widely used in the partial differential equations in recent decades to guarantee the smoothness of functions. From the viewpoint of functional analysis, $\nabla^{-2}H^0(R^d)$ is a very large function space and can accommodate many applications.

#### 4.1 Results on Functions from Sobolev Spaces

To give the desired analytic formula deriving $f_{(n+1)}$ from $f_{(n)}$, we present some notations and results related to functions in Sobolev spaces. For all $u \in \nabla^{-2}H^s(R^d)$, we define semi-norms by $\|u\|_{2,s} := \left( \int_{R^d} \| \xi \|_{R^d}^2 \| \mathcal{F} (\nabla^2 u) (\xi) \|_{R^{d \times d}}^2 d\xi \right)^{1/2}$, where $\mathcal{F}$ denotes the Fourier transform. For any $s \in R$, a non-homogeneous Sobolev space $H^s(R^d)$ is defined by

$$H^s(R^d) := \left\{ u \in \mathcal{S}'(R^d) : \hat{u} \in L^1_{\text{loc}}(R^d), \int_{R^d} \left( 1 + \| \xi \|_{R^d}^2 \right)^s |\hat{u}(\xi)|^2 d\xi < \infty \right\}.$$
Suppose \( \Omega \) is an open, bounded subset of \( \mathbb{R}^d \), then \( H^s(\Omega) \) is the restriction set of tempered distributions (generalized functions) in \( H^s(\mathbb{R}^d) \) to \( \Omega \). \( H^s_{\text{loc}}(\mathbb{R}^d) \) is the set of distributions on \( \mathbb{R}^d \) whose restriction to any bounded open set \( \Omega \) is in \( H^s(\Omega) \). The following result guarantees the regularity of functions in \( \nabla^{-2} \hat{H}^s(\mathbb{R}^d) \).

**Theorem 4.1.** If \(-2 - \frac{d}{2} < s < \frac{d}{2}\), then \( \nabla^{-2} \hat{H}^s(\mathbb{R}^d) \subset H^{2+s}_{\text{loc}}(\mathbb{R}^d) \subset C^k(\mathbb{R}^d) \) for \( k < 2 + s - \frac{d}{2} \).

**Proof.** It straightforward that \( \nabla^{-2} \hat{H}^s(\mathbb{R}^d) \subset H^{2+s}_{\text{loc}}(\mathbb{R}^d) \). From Sobolev embedding theorem (Theorem 7.25, Rudin (1991)), \( H^{2+s}_{\text{loc}}(\mathbb{R}^d) \subset C^k(\mathbb{R}^d) \) for \( k < 2 + s - \frac{d}{2} \). \( \square \)

The next result provides an analytic formula for solving a minimization problem within \( \nabla^{-2} \hat{H}^s(\mathbb{R}^d) \).

**Theorem 4.2.** [Duchon (1977)] Suppose \( A \) is a finite subset of \( \mathbb{R}^d \), \( P_d[t_1, t_2, \cdots, t_d] \) denotes the linear space of polynomials on \( \mathbb{R}^d \) such that the degree is less than 2, and any polynomial in \( P_d[t_1, t_2, \cdots, t_d] \) is uniquely determined by its values on \( A \). Then there exists exactly one function of the form

\[
\sigma(t) = \sum_{a \in A} s_a \eta_{4+2s-d}(\|t-a\|_{\mathbb{R}^d}) + p(t)
\]

with \( p \in P_d \) and \( \sum_{a \in A} s_a q(a) = 0 \), \( \forall q \in P_d \), taking specific values on \( A \), where \( \eta_\lambda(t) = |t|^\lambda \log(|t|) \) if \( \lambda \) is an even and positive integer and \( \eta_\lambda(t) = |t|^\lambda \) otherwise. In addition, if \( f \) is another function taking defined values on \( A \), then one has \( \|f\|_{2,s} \geq \|\sigma\|_{2,s} \).

Recall that here we only use the special case of \( s = 0 \). Theorem 4.2 gives the analytic expression of a minima within \( \nabla^{-2} \hat{H}^0(\mathbb{R}^d) \), rather than \( C \cap \nabla^{-2} \hat{H}^s(\mathbb{R}^d) \). Theorem 4.1 shows that if \( f \in \nabla^{-2} \hat{H}^0(\mathbb{R}^d \rightarrow \mathbb{R}^D) \), (i) \( d = 1, 2 \), \( f \in C^1(\mathbb{R}^d \rightarrow \mathbb{R}^D) \); (ii) \( d = 3 \), \( f \in C(\mathbb{R}^d \rightarrow \mathbb{R}^D) \); (iii) \( d > 3 \), there is no guarantee for the continuity of \( f \). Based on these results for \( d \leq 3 \) we obtain

\[
C \cap \nabla^{-2} \hat{H}^0(\mathbb{R}^d \rightarrow \mathbb{R}^D) = \nabla^{-2} \hat{H}^0(\mathbb{R}^d \rightarrow \mathbb{R}^D),
\]

Hence, we assume \( d \leq 3 \) in the rest of this section to derive the desired formula using Theorem 4.2.

**4.2 An Analytic Formula for Minimization**

Suppose \( f_{(n)} \) is given in the \( n^{th} \) step of the iteration scheme. From Theorem 4.2, we can present the elements of the \( D \)-dimensional function (4.1) as follows.

\[
f_{(n+1)}^l(t) = \sum_{j=1}^{N} s_j^l \times \eta_{4-d} \left( \|t - \pi_{f_{(n)}}(\mu_{j,N})\|_{\mathbb{R}^d} \right) + \sum_{k=1}^{d+1} \alpha_k^l \times p_k(t), \quad l = 1, 2, \cdots, D,
\]

(4.2)
where \( \eta_{d-4} \) is given by

\[
\eta_{d-4}(t) = \begin{cases} 
|t|^{4-d} \log(|t|), & d = 2, t \neq 0 \\
0, & d = 2, t = 0, \\
|t|^{4-d}, & \text{otherwise,}
\end{cases}
\]

\( t = (t_1, t_2, \cdots, t_d) \in \mathbb{R}^d, l = 1, 2, \cdots, D, \) the \( p_1, p_2, \cdots, p_{d+1} \) are linearly independent polynomials spanning \( P_2[t_1, t_2, \cdots, t_d] \), \( \alpha^l = (\alpha_1^l, \alpha_2^l, \cdots, \alpha_{d+1}^l)^T \) and \( s^l = (s_1^l, s_2^l, \cdots, s_N^l)^T \) are unknown parameter vectors subject to the constraints \( T^T s^l = 0 \) for \( l = 1, 2, \cdots, D \) and \( T \) is an \( N \times (d + 1) \) matrix with \( T_{ij} = p_j(\pi_{j(n)}(\mu_i, N)) \). Define \( E_{ij} = \eta_{d-4}\left(\|\pi_{f(n)}(\mu_i, N) - \pi_{f(n)}(\mu_j, N)\|_{\mathbb{R}^d}\right) \), \( W = \text{diag}\{\theta_1, \theta_2, \cdots, \theta_N\} \) and \( \mu^l = (\mu_{1,N}^l, \mu_{2,N}^l, \cdots, \mu_{N,N}^l)^T \) for \( l = 1, 2, \cdots, D \), then the minimization (4.1) is equivalent to:

\[
\inf_{s^l \in \mathbb{R}^N, \alpha^l \in \mathbb{R}^{d+1}, T^T s^l = 0, l = 1, 2, \cdots, D} \left\{ \|W^{1/2} \left( \mu^l - Es^l - T\alpha^l \right) \|^2_{\mathbb{R}^l} + w(s^l)^T Es^l \right\}.
\]

By the method of Lagrangian multiplier, the minimization above is equivalent to solving the following linear equations,

\[
A := \begin{pmatrix} 2EWE + 2wE & 2EWT \\ 2TT^WE & 2TT^WT \\ T^T & 0 & 0 \end{pmatrix}, \quad A \begin{pmatrix} s^l \\ \alpha^l \\ \lambda^l \end{pmatrix} = \begin{pmatrix} 2EW\mu^l \\ 2TTW\mu^l \\ 0 \end{pmatrix}, \quad l = 1, 2, \cdots, D,
\]

where \( \lambda^l \)'s are the Lagrangian multipliers. \( s^l \) and \( \alpha^l \) can be obtained by calculating the (generalized) inverse of \( A \). By plugging the estimated values of \( s^l \) and \( \alpha^l \) into (4.2), we can obtain \( f_{(n+1)} = (f_{(n+1)}^1, f_{(n+1)}^2, \cdots, f_{(n+1)}^D)^T \).

As a result, (4.2)(4.3)(4.4) combined provide the desired analytic formula.

### 4.3 Principal Manifold Estimation Algorithm

Before presenting the steps of the iterative algorithm for estimating the principal manifolds we propose a measure for performance of the estimation procedure. Specifically, we use the mean squared distance for this purpose following Hastie and Stuetzle (1989), where mean squared distance was used to measure the performance of principal curves. In addition, the motivation of linear PCA is to minimize the mean squared distance of the data points from the fitted lines.

**Definition 4.1.** Suppose \( \{x_i\}_{i=1}^I \subset \mathbb{R}^D \) is a dataset of interest and \( M_f^d \) is a manifold determined by a
continuous map \( f : \mathbb{R}^d \to \mathbb{R}^D \). The mean squared distance (MSD) from \( \{x_i\}_{i=1}^I \) to \( M_f^d \) is defined by

\[
MSD(f) = MSD \left( \{x_i\}_{i=1}^I, f \right) := \frac{1}{I} \sum_{i=1}^I \| x_i - f(\pi_f(x_i)) \|^2_{\mathbb{R}^D}.
\]

The following list presents the parameters that need to be set or initialized in the proposed algorithm.

(i) \( \{x_i\}_{i=1}^I \subset \mathbb{R}^D \): Dataset of interest;

(ii) \( N_0 \): Initial number of \( \mu_{j,N} \)'s in HDMDE;

(iii) \( \alpha \): Confidence level of the hypothesis test determining the number of \( \mu_{j,N} \)'s in HDMDE;

(iv) \( \kappa \): Factor in (3.7) determining the curvature of the derived manifold;

(v) \( \epsilon \): Threshold terminating iteration;

(vi) \( n_{\text{max}} \): Upper limit of the number of steps in the iteration.

The PME algorithm

(i) Use HDMDE and inputs \( N_0, \alpha \) and \( \{x_i\}_{i=1}^I \) to estimate \( \hat{N}, \hat{\mu}_{j,N} \) and \( \hat{\theta}_{j,N} \);

(ii) Calculate \( w \) by (3.7).

(iii) Calculate the mean and the first \( d \) principal components of the dataset, defined by \( \bar{x} = \frac{1}{I} \sum_{i=1}^I x_i \) and \( v_1, v_2, \cdots, v_d \). Set the initial embedding map \( f_0(t_1, t_2, \cdots, t_d) = \bar{x} + \sum_{i=1}^d t_i v_i \) and counting variable \( n = 0 \), calculate \( MSD(0) \);

(iv) Use the analytic formula in Section 4 to derive \( f_1 \) from \( f_0 \) and calculate \( MSD(1) \).

while(\( n \leq n_{\text{max}} \) and \( |MSD(n+1)/MSD(n) - 1| < \epsilon \) ) do:

(i) \( n \leftarrow n + 1 \);

(ii) Use the analytic formula in Section 4 to derive \( f_{(n+1)} \) from \( f_{(n)} \) and calculate \( MSD_{(n+1)} \),

where \( MSD_{(n)} = MSD(\{x_i\}_{i=1}^I, f_{(n)}) \).

As a result we obtain the analytic expression of an embedding \( \mathbb{R}^d \to \mathbb{R}^D \) map \( \hat{f} \) determining the desired manifold \( M_f^d \) with \( d \leq 3 \), \( D \in \mathbb{N}_+ \), and \( MSD (\{x_i\}_{i=1}^I, f) \).

5 Simulations

In this section, we use two settings \( d = 1, D = 2 \) and \( d = 2, D = 3 \) to illustrate the performance of the proposed PME algorithm by simulation studies and compare the results with existing approaches. The PME algorithm and the analyses herein are implemented in the \( R \) software [R Core Team 2017]. The code is available upon request.
Case 2, Principal Curves: We compare the performance of the proposed PME algorithm with the HS procedure implemented using the function `principal.curve()` in the R package "princurve". We generate 100 samples of size \( n = 1000 \) from four choices of true distribution functions \( f(\cdot) \). For each dataset, we compare the performance of HS and PME algorithms at the 10\(^{th} \) iteration step. For the PME algorithm, we use the following settings of the input variables \( N_0 = 10, \alpha = 0.05, \kappa = 1, \epsilon = 0 \) and \( n_{\text{max}} = 10 \), while we applied the HS algorithm using the default settings in the package. The choices of four true density functions are given as follows.

I. \((\theta + e_1, \sin \theta + e_2)\) with \( \theta \sim N(0, \pi) \) and \( e_1, e_2 \) are independent \( N(0, 0.2) \).

II. \((10 \cos \theta + e_1, 10 \sin \theta + e_2)\) with \( \theta \sim Unif(-0.1\pi, 1.1\pi) \) and \( e_1, e_2 \) are independent \( N(0, 1) \).

III. Suppose \( \rho = 5, \theta \sim N(0, 0.5\pi) \) and \( e_1, e_2 \) are independent \( N(0, 1) \), random 2-vector is

\[
\left( (\rho \theta + e_1) \cos \frac{3\pi}{10} - (\rho \cos \theta + e_2) \sin \frac{3\pi}{10}, (\rho \theta + e_1) \sin \frac{3\pi}{10} + (\rho \cos \theta + e_2) \cos \frac{3\pi}{10} \right). \tag{5.1}
\]

IV. \([5.1]\) with \( \rho = 1.5, \theta \sim N(0, \pi) \) and \( e_1, e_2 \) are independent \( N(0, 1) \).

The results of the simulations are presented in Figure\[4\]. The simulations show that the proposed PME algorithm outperforms the HS algorithm uniformly for these four random vectors. We observe that the PME algorithm performs particularly well when the true underlying function has a high level of curvature such as in cases I and II implying that the algorithm will perform better in brain imaging applications where objects often have high levels of curvature. In addition, the simulation studies indicate that PME has a superior performance in estimating the boundary of the function of interest. Furthermore, the PME algorithm gives the analytic expression of the embedding map \( f \) of the derived principal manifold \( M_d^\ast \). For example, the analytic expression of the embedding map of the function in Figure\[3\] IV is given as follows.

\[
f^l(t) = \sum_{j=1}^{11} s_j^l |t - t_j^o|^3 + \alpha_1^l + \alpha_2^l t, \quad l = 1, 2,
\]

\((s_1^1, s_2^1, \ldots, s_{11}^1) = 10^2 \times (4.45, 2.67, 0.35, -4.52, -4.66, 1.29, -0.17, -2.19, 5.84, 1.14, -3.97),\)

\((s_1^2, s_2^2, \ldots, s_{11}^2) = 10^2 \times (-3.23, -1.77, -0.48, 3.25, 2.79, -0.51, 0.10, 1.89, -4.18, -0.94, 2.89),\)

\((\alpha_1^1, \alpha_2^1) = 10^4 \times (0.91, 2.40),\)

\((\alpha_1^2, \alpha_2^2) = 10^4 \times (-0.60, -1.17),\)

\((t_1^*, t_2^*, \ldots, t_{11}^*) = (-3.55, -6.08, 13.14, 1.52, -8.25 - 11.42, -1.70, 8.71, 3.31, 5.71, -0.70).\)
Figure 4: Comparison of the PME and HS algorithms for four simulation settings. The left panel presents one example simulated dataset for each setting with the estimated HS (black solid line), PME (red dashed line), and a middle of the function (green dot-dashed line). The right panel shows the boxplots of the MSD for each of the two algorithms.
principal surface algorithm (PSA) introduced by Yue et al. (2016). We generate 100 samples of size \( n = 1000 \) from \((\theta_1 + e_1, -\theta_1^2 - \theta_2^2 + e_2, \theta_2 + e_3)\) where \( \theta_1, \theta_2 \) are independent \( \text{Unif}(0, 1) \) and \( e_1, e_2, e_3 \) are independent \( \mathcal{N}(0, 0.1) \) and use both PME and PSA algorithms to fit a principal surface to each of the generated datasets. We compare the performance of the two algorithms using the MSD. The starting values of the functions in both algorithms are obtained as the linear principal components of the dataset. In the PME algorithm we used \( N_0 = 100, \alpha = 0.05, \kappa = 1, \epsilon = 0 \) and \( n_{\text{max}} = 10 \).

Figure 5 shows the results of the 100 simulations. Visually, the principal surfaces estimated by the PME and PSA algorithms using a sample simulated dataset are similar, with the PME estimate indicating slightly more curvature than the PSA estimated surface. The bottom-right panel shows the boxplots, means and standard deviations of MSDs calculated using the estimates by PME and PSA algorithms in the 100 simulations. The MSD values indicate that the PME algorithm performs marginally better than PSA. An intuitive reason for the observed higher curvature of the PME result and smaller MSD values can be found in that HDMDE estimates different weights for the data points while PSA distributes equal weights to all data points. Specifically, the points within neighborhoods of high point density are given a higher weight in PME while the points within low point density neighborhoods are given lower weights. As a result the curve fits the data better at the neighborhoods with high point density that often have higher curvature. For a given dataset, if the data points are evenly distributed on the support of the data, then PME and PSA are likely to perform similar to each other.

Furthermore, PME can give the analytic expression of the derived 2-dimensional principal manifold. The resulting analytic expression can be used in many applications, one of which is shown in Section 6. Throughout Section 5 we set \( \kappa = 1 \). Cross-validation techniques can be used to numerically adjust \( \kappa \) to reduce MDE. Additionally, as shown in Section 4 the proposed PME algorithm can be applied to all pairs \((d, D)\) with \( d \leq 3 \) and \( D \in \mathbb{N}_+ \). The running time of the proposed algorithm implemented in R is similar to the running times of the algorithms used as competitors in all the simulations performed in this section.

6 Tumor Surface and Interior Estimation in Cancer

In this section we consider the important question of surface estimation and tumor segmentation in cancer imaging. Radiation therapy uses ionizing radiation to control or kill cancer cells. Advanced radiotherapy techniques (e.g. proton and charged particle radiotherapy) enable good precision in the dose delivery and spatial distribution of radiation. To avoid harming healthy tissue with unnecessary doses of radiation, identifying the interior region of a tumor is very important in radiation therapy. We analyze a subset of
Figure 5: A comparison of the PME and PSA procedures. The top plots show the estimated principal surfaces by PME (left) and PSA (right) algorithms for one of the simulated datasets presented in the bottom left plot. The bottom right plot shows the boxplots of MSD values for 100 simulated datasets.
n = 8 patients from a publicly available dataset collected for 422 patients with non-small cell lung cancer and available at [http://www.cancerimagingarchive.net/](http://www.cancerimagingarchive.net/) Computed tomography (CT) scans of the tumors within the lung are obtained for each study participant along with masks of the tumor hand segmented by a radiologist. The details on imaging parameters are available on the website and the references provided therein.

The 3D plots of the data points for two study participants are shown in the left column of Figure 6. We use the proposed PME algorithm to estimate the surface of each tumor and develop a classifier identifying points inside the tumor. The classifier can be used to estimate the interior region of the tumor, which is the target of radiation therapy. In developing this classifier, we show the importance of the analytic expression of embedding maps derived by the PME algorithm.

### 6.1 Tumor Interior Classifier

Before presenting the results of the analyses, we give a description of our classifier based on tumor surface fitting. Suppose a tumor $T$, whose boundary surface $\partial T$ is a 2-dimensional non-trivial manifold embedded into $\mathbb{R}^3$, is contained by a cuboid $\Omega$. $\{x_i\}_{i=1}^I$ denotes the voxels (three-dimensional pixels) that are provided by the radiologist as locations on the boundary surface of tumor $T$. The first question to address in automated segmentation of a tumor is the estimation of the tumor surface. While we discuss tumor surface estimation in this section, the proposed method can be implemented in other surface estimation problems such as estimation of the cortical surface of the brain using magnetic resonance imaging, where the surface is much more complex with higher curvature (i.e. including complex structures such as the gyri and sulci). Motivated by the definition of non-trivial manifolds, we propose an estimation approach based on a combination of principal manifolds estimated locally. In other words, we divide $\Omega$ into a collection of cubes defined by $\Omega = \bigcup_{j=1}^J \Omega_j$, with $\Omega_i \cap \Omega_j = \emptyset$, when $i \neq j$ and estimate the surface of the tumor piecewise. For each $j = 1, 2, \cdots, J$, we use the voxels in $\Omega_j^* := \left\{x : \text{dist} (\Omega_j, x) < 0.5 \times \text{diam} (\Omega_j)\right\}$, where $\text{diam} (\Omega_j)$ stands for the diameter of $\Omega_j$, to estimate a 2-dimensional principal manifold $M^2_{\hat{f}_j}$ using the PME algorithm. As a result, we obtain the analytic expression of the embedding map $f_j : \mathbb{R}^2 \rightarrow \mathbb{R}^3$. The estimate of the surface of the tumor is given by

$$\hat{\partial}T := \bigcup_{j=1}^J \left(\Omega_j \cap M^2_{\hat{f}_j}\right). \quad (6.1)$$
Figure 6: Vertices on the tumor surface (colored points) and the overlaid surface estimate (in gray) are presented in the left column. The first three rows show the first tumor and the last three rows show the second tumor. The estimate based on a subsample of the data is presented in the middle column. The interior classifier for the tumor is shown in the right column. Each row presents a different angle of the tumor.
The MSD of the estimate can be computed using $\text{MSD}(\hat{T}, T) := \frac{1}{J} \sum_{j=1}^{J} \text{MSD}(\Omega_j \cap \{x_i\}_{i=1}^{I}, f_j)$ and can be used to measure the performance of the tumor surface estimation procedure. The normal vector field on $M^2_{f_j}$ is given by $n_p = (f_{j1}, f_{j2}, f_{j3})$. Suppose $c^\ast$ is the estimated center of the tumor calculated by $c^\ast := \frac{1}{I} \sum_{i=1}^{I} x_i$, then the estimation formula of the interior region of the tumor is given by

$$\hat{T} := \bigcup_{j=1}^{J} \{ x \in \Omega_j \mid \langle f_j (\pi_{f_j}(x)) - x, n (\pi_{f_j}(x)) \rangle \times \langle f_j (\pi_{f_j}(c^\ast)) - c^\ast, n (\pi_{f_j}(c^\ast)) \rangle \geq 0 \},$$

(6.2)

where $\langle \cdot, \cdot \rangle$ is the inner product in $\mathbb{R}^3$. Since the tumor interior region estimation fully depends on the tumor surface estimation, the performance of tumor interior region estimation can also be measured by MSD.

### 6.2 Results of Tumor Data Analyses

Some practical aspects of analyzing tumor data include the selection of cubes that divide the cuboid containing the data. We discuss this using one example tumor from the dataset contained in the cuboid $\Omega = [-125, -27] \times [-30, 67] \times [-50, 26]$. We can partition $\Omega$ into $\{\Omega_{ijk}\}_{i,j,k}$ by dividing the three edges of $\Omega$. Depending on the distribution of available data points on the tumor surface the choice of the division will have an effect on the estimated manifold. For instance, if the boundary of $\Omega_{ijk}$ has a very large curvature with few data points surface fitting algorithms usually do not perform well as there is not enough data points near the boundary. In applications, one can use various approaches for dividing the cuboid including dividing each edge into equally spaced intervals, manual division, and dividing each edge based on the quantiles of the data points in the corresponding direction. Depending on the application at hand a semi-automatic division using scatterplots as guides may be a better alternative.

For data points in each $\Omega_{ijk}$, we use our proposed PME algorithm with input $N_0 = 8$, i.e. 8% of the vertices in the dataset are used for estimation, $\alpha = 0.05$, $\kappa = 1$, $\epsilon = 0$ and $n_{\max} = 5$ to obtain an embedding map $f_{ijk}$. The estimated tumor surface by (6.1) is presented in the left column of Figure 6 as the grey translucent surface. Figure 6 visually shows that the surface estimate fits the data points well. (6.2) provides the formula for estimation of the interior region. We use the tumor interior classifier to classify the points in point collection

$$\left\{ \left( -125 + \frac{-27 + 125}{20} \xi, -30 + \frac{67 + 30}{20} \eta, -50 + \frac{26 + 50}{20} \zeta \right) \right\}^{30}_{\xi, \eta, \zeta = 0}.$$

The points identified as inside of the tumor are shown in the right column of Figure 6. The MSD of the
estimated tumor surface is 0.463.

One of the issues in tumor segmentation by hand is the time consuming and costly nature of hand segmentation by a radiologist or a trained technician. An automatic algorithm that can be applied to estimate the tumor surface by using a significantly smaller number of vertices at random locations of the tumor surface identified by a radiologist can be very useful in reducing the time of hand segmentation. We show that the proposed PME algorithm can be used to estimate the tumor surface and interior regions using only a small randomly selected portion of the vertices without increasing the MSD significantly. We randomly select 25% of the vertices from the tumor dataset and apply the PME algorithm input $N_0 = 30\%$ to estimate the tumor surface. The new MSD is 0.487, which is calculated using the new surface estimation and all points in the original data set. Reducing the sample size by 75\% only increases MSD by 5\%. The last three rows of Figure 6 show the performance of surface fitting and interior classifier for another example dataset in our set. Similarly, we find that using a random sample of 25\% of the data the obtained surface increases the MSD by only 4\%.

We evaluated the performance of the two algorithms measured by the MSD based on the data for $n = 8$ patients. As we mentioned in Section 5, we expect the two methods to perform similar to each other as the data points are mostly evenly distributed along a curve within each slice in these datasets. To compare the algorithms directly for each dataset, we computed the logarithm of the ratio of MSDs, i.e. $\log\left(\frac{MSD_{PME}}{MSD_{PSA}}\right)$, for each participant. The median log-ratio of the MSDs estimated by the PSA and the proposed PME is -0.127 (mean = -0.008, sd = 0.27) implying that PME still performs marginally better than PSA. Although PME and PSA perform well for the surface fitting, only PME provides the analytical expression of the estimated surface, which can be used to derive the tumor interior classifier using (6.2).

7 Conclusions

In this paper, we constructed a mathematically rigorous framework of principal manifolds based on function spaces of Sobolev type and proposed an iteration scheme for numerically deriving principal manifolds. Both the principal manifolds and iteration scheme are generalizations of the HS principal curves defined by Hastie (1984) and Hastie and Stuetzle (1989). To reduce computational burden and eliminate the model bias shared by HS principal curves and surfaces and principal manifolds, we proposed middles of random vectors and the approximating principal manifolds for the middle of a random vector. A high dimensional mixture density based algorithm was proposed to model the middles of the principal manifolds. Finally, we proposed the PME algorithm to numerically derive approximating principal manifolds of middles of random vectors. This
algorithm estimates $d$ dimensional manifolds with $d \leq 3$. In addition to reducing computational burden and eliminating model bias, the proposed PME algorithm guarantees the smoothness of derived principal manifolds and is resistant to outliers. Furthermore, the PME algorithm provides the analytic expression of the embedding maps of the derived manifolds.

We used simulation studies to compare the performance of the proposed algorithm with existing methods in two settings: 1) when $d = 1, D = 2$ we compare the proposed algorithm to the HS algorithm for estimating principal curves, 2) when $d = 2, D = 3$ we compare the proposed algorithm with the PSA algorithm for estimating principal surfaces. The studies illustrate that the proposed method performs equivalently or uniformly better compared with the HS and PSA algorithms in the sense of minimizing the MSD. The proposed PME algorithm provides analytic expressions of derived curves and surfaces in addition to the results that can be obtained by the HS or PSA approaches. By applying our PME algorithm to analyze CT images for patients with cancer, we showed that the algorithm can not only estimate the boundary surfaces of tumors from CT scans, but also give a classifier to identify spatial points inside the tumors that can be the target of radiation therapy. The surface fitting and classifier proposed in our paper can be applied in surface estimation problems in brain imaging and in 3D printing.

8 Acknowledgements

The authors would like to thank Dr. Chen Yue for providing the R function for implementing PSA. The project described was supported by Grant Number 5P20GM103645 from the National Institute of General Medical Sciences.

References

R. A. Adams and J. J. Fournier. *Sobolev spaces*, volume 140. Academic press, 2003.

D. Chen, J. Zhang, S. Tang, and J. Wang. Freeway traffic stream modeling based on principal curves and its analysis. *IEEE Transactions on Intelligent Transportation Systems*, 5(4):246–258, 2004.

S.-s. Chern. *Topics in differential geometry*. The Institute for Advanced Study, 1951.

X. Dai and H.-G. Müller. Principal component analysis for functional data on riemannian manifolds and spheres. *arXiv preprint arXiv:1705.06226*, 2017.
P. Demartines and J. Hérault. Curvilinear component analysis: A self-organizing neural network for non-linear mapping of data sets. *IEEE Transactions on neural networks*, 8(1):148–154, 1997.

T. Duchamp, W. Stuetzle, et al. Extremal properties of principal curves in the plane. *The Annals of Statistics*, 24(4):1511–1520, 1996.

J. Duchon. Splines minimizing rotation-invariant semi-norms in sobolev spaces. *Constructive theory of functions of several variables*, pages 85–100, 1977.

R. Durrett. *Probability: theory and examples*. Cambridge university press, 2010.

A. Eloyan and S. K. Ghosh. Smooth density estimation with moment constraints using mixture distributions. *Journal of nonparametric statistics*, 23(2):513–531, 2011.

T. Hastie. Principal curves and surfaces. Technical report, STANFORD UNIV CA LAB FOR COMPUTATIONALSTATISTICS, 1984.

T. Hastie and W. Stuetzle. Principal curves. *Journal of the American Statistical Association*, 84(406):502–516, 1989.

I. T. Jolliffe. Principal component analysis and factor analysis. In *Principal component analysis*, pages 115–128. Springer, 1986.

N. Lawrence. Probabilistic non-linear principal component analysis with gaussian process latent variable models. *Journal of machine learning research*, 6(Nov):1783–1816, 2005.

J. L. Martínez-Morales. Geometric data fitting. In *Abstract and Applied Analysis*, volume 2004, pages 831–880. Hindawi Publishing Corporation, 2004.

J. W. Milnor. *Topology from the differentiable viewpoint*. Princeton University Press, 1997.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL [https://www.R-project.org/](https://www.R-project.org/).

W. Rudin. Functional analysis. international series in pure and applied mathematics, 1991.

J. W. Sammon. A nonlinear mapping for data structure analysis. *IEEE Transactions on computers*, 100(5):401–409, 1969.

G. Schmidt. *Relational mathematics*, volume 132. Cambridge University Press, 2011.
9 Appendix

9.1 A Generalized Framework

Principal manifolds in Definition 3.1 are given by minimization

\[
\arg\inf_{f \in \nabla^{-2} \hat{H}^0(\mathbb{R}^d \to \mathbb{R}^D)} \left\{ \mathbb{E}_\mathbb{P} \| X - f(\pi_f(X)) \|^2_{\mathbb{R}^D} + w \| \nabla^2 f \|^2_{L^2(\mathbb{R}^D)} \right\}.
\]  

(9.1)

In this subsection, we cast (9.1) into a generalized framework. Before giving this generalization, we recall the definition of semi-Hilbert spaces as follows.

**Definition 9.1.** Let \( \mathcal{H} \) be a linear space and \( \| \cdot \| \) be a semi-norm defined on \( \mathcal{H} \). \( N := \{ x \in \mathcal{H} : \| x \| = 0 \} \), then \( N \) is a linear subspace of \( \mathcal{H} \), called the null of \( \| \cdot \| \). Define \( \| [x] \| \) on the quotient space \( \mathcal{H}/N \) by \( \| [x] \| := \| x \| \) for all \( x \in \mathcal{H}/N \). It is straightforward that \( \| [x] \| \) does not depend on the choice of \( x \) in \([x]\). If \((\mathcal{H}/N, \| [\cdot] \|)\) is a Hilbert space, then \((\mathcal{H}, \| \cdot \|)\) is called a semi-Hilbert space with null \( N \).

It is straightforward to verify that \( \left( \nabla^{-2} \hat{H}^0(\mathbb{R}^d \to \mathbb{R}^D), \| \cdot \|_{2,0} \right) \) is a semi-Hilbert space with null space \( P_2 \) and \( \| x - y \|^2_{\mathbb{R}^D} \) is a loss function of \((x, y) \in \mathbb{R}^D \times \mathbb{R}^D \). If we generalize \( \left( \nabla^{-2} \hat{H}^0(\mathbb{R}^d \to \mathbb{R}^D), \| \cdot \|_{2,0} \right) \) to be a general semi-Hilbert space \((\mathcal{H}, \| \cdot \|)\) with null space \( N \) and \( \| x - y \|^2_{\mathbb{R}^D} \) to be a general loss function \( L : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}_+^1 \), Definition 3.1 is generalized as follows.

**Definition 9.2.** Suppose \( X \) is a random \( D \)-vector and its distribution is determined by probability measure \( \mathbb{P} \) and has finite second moments, \( \mathbb{E}_\mathbb{P} \) is the expectation with respect to probability measure \( \mathbb{P} \), \((\mathcal{H}, \| \cdot \|)\) is a semi-Hilbert space of functions on \( \mathbb{R}^d \). \( \mathcal{H}(\mathbb{R}^d \to \mathbb{R}^D) \) denotes the collection of all \( \mathbb{R}^d \to \mathbb{R}^D \) maps such
that each of their components is in $\mathcal{H}$. Let $L : \mathbb{R}^D \times \mathbb{R}^D \rightarrow \mathbb{R}_+^1$ is a loss function. For any $\mathcal{C} \subset C(\mathbb{R}^d \rightarrow \mathbb{R}^D) \cap \mathcal{H}(\mathbb{R}^d \rightarrow \mathbb{R}^D)$ and $w \in [0, +\infty]$, the penalized risk functionals for $X$ and $L$ are

$$M_{w,P,L}(f,g) := \mathbb{E}_x L(X,f(\pi_g(X))) + w\|f\|^2, \quad M_{w,P,L}(f) := M_{w,P,L}(f,f), \quad f,g \in \mathcal{C}.$$  

A manifold $M^d_{f^*}$ determined by $f^* \in \mathcal{C}$ is called a $(\mathcal{C},w)$ principal manifold for $X$ and $L$ if $f^* = \arg\inf_{f \in \mathcal{C}} M_{w,P,L}(f)$, where $w$ is called the smoothing parameter and $\|f\|^2$ is called the penalty functional.

The iteration $f_{(n+1)} := \arg\inf_{f \in \mathcal{C}} M_{w,P,L}(f,f_{(n)})$ can be applied to numerically derive the principal manifolds in Definition 9.2.

### 9.2 High Dimensional Mixture Density Estimation Method

In this subsection, we propose a high dimensional generalization of mixture density estimation method proposed by Eloyan and Ghosh (2011). In Theorem 3.4, for any given $N \in \mathbb{N}$, we assume that the parameters in (3.5) defined by $M_N = \{\mu_{j,N}\}_{j=1}^N$, $\theta_{j,N}$’s and $\sigma_N$, are known. In applications, the data usually informs the choice of these parameters. The choice of $N$ is also an important issue as a small $N$ might not be sufficient for obtaining a good approximation and a large $N$ might result in redundancy and produce computational burden.

#### 9.2.1 Suggested Choices for $\mu_{j,N}$’s and $\sigma_N$

Suppose $N$ is given. Using k-means clustering, we divide the dataset into $N$ clusters and $M_N = \{\mu_{j,N}\}_{j=1}^N$ is set to be the collection of the centers of the $N$ clusters.

The $j^{th}$ cluster is denoted by $\{\xi_k^{(j)}\}_{k=1}^K$ with

$$\xi_k^{(j)} = (\xi_k^{(j),1}, \xi_k^{(j),2}, \ldots, \xi_k^{(j),D})^T, \quad \xi^{(j),l} := (\xi_1^{(j),l}, \xi_2^{(j),l}, \ldots, \xi_K^{(j),l})^T, \quad l = 1, 2, \ldots, D.$$  

Set $Q(\xi^{(j),\mu}) = \frac{1}{K-1}(\xi^{(j),\mu})^T(I_{K \times K} - \frac{1}{K} J_{K \times K})\xi^{(j),\mu}$ where $J_{K \times K}$ is a $K \times K$ matrix with elements equal to 1 and $(\sigma^{(j)})^2 := \frac{1}{D} \sum_{\mu=1}^D Q(\xi^{(j),\mu})$, then $\sigma_N := \sqrt{\frac{1}{N} \sum_{j=1}^N \sigma^{(j)}^2}$. The rationale for this estimation is that $\mathbb{E}Q(\xi^{(j),l}) = \sigma^2$ for $l = 1, 2, \ldots, D$ if $\xi_1^{(j)}, \xi_2^{(j)}, \ldots, \xi_K^{(j)}$ independently follow a multivariate normal distribution with variance $\sigma^2 I_{D \times D}$. 


9.2.2 Mixture Weight Estimation Using the EM Algorithm

For a given $N$ and $M_N = \{\mu_{j,N}\}_{j=1}^N$, $\sigma_N$ computed using the approach presented in Section 9.2.1, we propose an approach to estimate $\theta_{j,N}$ using a constrained EM-algorithm. For simplicity of notations, we denote $\theta_{j,N}$, $\mu_{j,N}$ and $\theta_N$ by $\theta_j$, $\mu_j$ and $\theta$, respectively. Suppose $Z_i$ are unobserved random variables satisfying

$$X_i|\theta, Z_i = z_i \sim \frac{1}{\sigma_N} \psi \left( \frac{x_i - \mu_{z_i}}{\sigma_N} \right) dx_i, \quad Z_i|\theta \sim \theta z_i \sum_{j=1}^N \delta_{(j)}(dz_i), \quad i = 1, 2, \ldots, I, \quad z_i = 1, 2, \ldots, N,$$

then

$$(X_i, Z_i)|\theta \sim \theta z_i \frac{1}{\sigma_N} \psi \left( \frac{x_i - \mu_{z_i}}{\sigma_N} \right) \left( dx_i \times \sum_{j=1}^N \delta_{(j)}(dz_i) \right),$$

$$Z_i = j|\theta, X_i = x_i \sim \frac{\theta_j}{\sum_{j'=1}^N \theta_{j'}} \frac{1}{\sigma_N} \psi \left( \frac{x_i - \mu_{j'}}{\sigma_N} \right) =: w_{ij}(\theta).$$

The complete likelihood of the joint random variables $\{(X_i, Z_i)\}_{i=1}^I$ with respect to the product probability measure $\prod_{i=1}^I \left\{ dx_i \times \sum_{j=1}^N \delta_{(j)}(dz_i) \right\}$ is $L_C(\theta|x, z) = \prod_{i=1}^I \theta_{z_i} \frac{1}{\sigma_N} \psi \left( \frac{x_i - \mu_{z_i}}{\sigma_N} \right)$. For a computed $\theta^{(k)}$, the expectation step of EM algorithm gives

$$Q(\theta|\theta^{(k)}) = \mathbb{E} \left( \log L_C(\theta|X, Z)|X = x, \theta^{(k)} \right)$$

$$= \sum_{i=1}^I \sum_{j=1}^N w_{ij}(\theta^{(k)}) \log \left( \frac{1}{\sigma_N} \psi \left( \frac{x_i - \mu_j}{\sigma_N} \right) \right) + \sum_{i=1}^I \sum_{j=1}^N w_{ij}(\theta^{(k)}) \log \theta_j.$$

Suppose the approximation $p_N$ and the dataset $\{x_i\}_{i=1}^I \subset \mathbb{R}^D$ share the same mean, then we have the constraint $\mathbb{E}_{p_N} X = \sum_{j=1}^N \theta_j \times \int_{\mathbb{R}^D} \frac{x}{\sigma_N} \psi \left( \frac{x - \mu_j}{\sigma_N} \right) dx = \sum_{j=1}^N \theta_j \mu_j = \bar{x} = \frac{1}{I} \sum_{i=1}^I x_i$. As the weights of the density should add to one, we obtain another constraint given by $\sum_{j=1}^N \theta_j = 1$. By combining these constraints we build the Lagrangian

$$Q_\lambda(\theta|\theta^{(k)}) = Q(\theta|\theta^{(k)}) + \lambda_1 \left( 1 - \sum_{j=1}^N \theta_j \right) + \lambda_2^T \left( \bar{x} - \sum_{j=1}^N \theta_j \mu_j \right),$$

$$p_N \left( \frac{x}{\sigma_N} \psi \left( \frac{x - \mu_j}{\sigma_N} \right) \right) \approx \frac{1}{\sigma_N} \psi \left( \frac{x - \mu_j}{\sigma_N} \right) \frac{1}{\sum_{j=1}^N \frac{1}{\sigma_N} \psi \left( \frac{x - \mu_j}{\sigma_N} \right)}$$
where $\lambda_1 \in \mathbb{R}^1$ and $\lambda_2 \in \mathbb{R}^D$. By taking derivatives, we obtain

$$
\frac{\partial Q_\lambda}{\partial \theta_j} = \frac{1}{\theta_j} \sum_{i=1}^I w_{ij}(\theta^{(k)}) - \lambda_1 - \lambda_2^T \mu_j = 0, \quad j = 1, 2, \ldots, N,
$$

$$
\frac{\partial Q_\lambda}{\partial \lambda_1} = 1 - \sum_{j=1}^N \theta_j = 0, \quad \frac{\partial Q_\lambda}{\partial \lambda_2} = \bar{x} - \sum_{j=1}^N \theta_j \mu_j = 0.
$$

These equations result in the following system

$$
\theta_j = \sum_{i=1}^I \frac{w_{ij}(\theta^{(k)})}{\lambda_1 + \lambda_2^T \mu_j}, \quad j = 1, 2, \ldots, N,
$$

$$
\sum_{j=1}^N \left( \sum_{i=1}^I \frac{w_{ij}(\theta^{(k)})}{\lambda_1 + \lambda_2^T \mu_j} \right) = 1, \quad \sum_{j=1}^N \left( \sum_{i=1}^I \frac{w_{ij}(\theta^{(k)})}{\lambda_1 + \lambda_2^T \mu_j} \right) \mu_j = \bar{x}. \quad (9.2)
$$

Then the solution to (9.2) denoted by a triple $(\theta_j^{(k+1)}, \hat{\lambda}_1, \hat{\lambda}_2)$ is given by

$$
\theta_j^{(k+1)} = \sum_{i=1}^I \frac{w_{ij}(\theta^{(k)})}{\lambda_1 + \lambda_2^T \mu_j}, \quad j = 1, 2, \ldots, N,
$$

$$
(\hat{\lambda}_1, \hat{\lambda}_2) = \arg \inf_{\lambda_1 \in \mathbb{R}, \lambda_2 \in \mathbb{R}^D} \left\{ \left[ \sum_{j=1}^N \left( \sum_{i=1}^I \frac{w_{ij}(\theta^{(k)})}{\lambda_1 + \lambda_2^T \mu_j} \right) \right]^{-1} + \left[ \sum_{j=1}^N \left( \sum_{i=1}^I \frac{w_{ij}(\theta^{(k)})}{\lambda_1 + \lambda_2^T \mu_j} \right) \mu_j - \bar{x} \right]^2 \right\}. \quad (9.3)
$$

Using the iteration defined by (9.3), we obtain $(\theta_j^{(k)})_{k \in \mathbb{N}}$. For a threshold $\varepsilon$, an estimate for $\theta_j$ is given by $\hat{\theta}_j := \theta_j^{(k^*)}$ with $k^* := \inf \left\{ k \in \mathbb{N} : |\theta_j^{(k)} - \theta_j^{(k+1)}| < \varepsilon \right\}$. Then $p_N(x) := \sum_{j=1}^N \hat{\theta}_j \times \frac{1}{\sigma_N} \psi \left( \frac{x - \mu_j}{\sigma_N} \right)$ gives the desired estimation of $p$ if $N$ is given.

### 9.2.3 Choice for $N$

The criterion for choosing $N$ is based on an asymptotic hypotheses testing procedure. Suppose the desired confidence level is $\alpha \in (0, 1)$ and $z_{1-\alpha/2}$ denotes the $(1 - \alpha/2)$ upper quantile of the standard normal density. For the data set $\{x_i\}_{i=1}^I$, let

$$
\hat{\Delta}_{i,N} := \log \frac{p_{N+1}(x_i|\hat{\theta}_{N+1})}{p_N(x_i|\hat{\theta}_N)}, \quad \Sigma_{i,N} := \frac{1}{I} \sum_{i=1}^I \hat{\Delta}_{i,N}, \quad S_{i,N}^2 = \frac{\sum_{i=1}^I \left( \hat{\Delta}_{i,N} - \Sigma_{i,N} \right)^2}{I-1}, \quad Z_{i,N} := \sqrt{\frac{\Sigma_{i,N}}{S_{i,N}}},
$$

where $\hat{\theta}_N = (\hat{\theta}_{1,N}, \hat{\theta}_{2,N}, \ldots, \hat{\theta}_{N,N})$ and $\hat{\theta}_{N+1} = (\hat{\theta}_{1,N+1}, \hat{\theta}_{2,N+1}, \ldots, \hat{\theta}_{N+1,N+1})$ are derived using constrained EM-algorithm developed in the previous subsection. Let $N_0$ be the initial guess for $N$, then our proposed estimation of $N$ is given by $N_\alpha := \inf \left\{ N \geq N_0 : -z_{1-\alpha/2} \leq Z_{i,N} \leq z_{1-\alpha/2} \right\}$. The rational of this estimation...
procedure is given in Section 4 of Eloyan and Ghosh (2011).

9.3 Remarks

9.3.1 A Remark on Theorem 3.4

A triple \((\psi, d_N, \sigma_N)\) satisfying condition (vi) of Theorem 3.4 exists. We give an example as follows. Let \(\psi(x) = C_D \exp\{-\|x\|_{\mathbb{R}^D}\}\) with \(C_D^{-1} = \frac{2\pi^{D/2}D^{-1}(D)}{\Gamma(D/2)}\). It is straightforward to verify that \(\int \psi = 1\) and \(\psi \in C(\mathbb{R}^D)\). For all \(x \in \mathbb{R}^D\),

\[
\Delta_N(x) := \left| \frac{1}{\sigma_N} \psi \left( \frac{x - \mu_1}{\sigma_N} \right) - \frac{1}{\sigma_N} \psi \left( \frac{x - \mu_2}{\sigma_N} \right) \right| \leq \frac{C_D}{\sigma_N^D} \exp \left( \frac{\|x - \mu_2\|_{\mathbb{R}^D} - \|x - \mu_1\|_{\mathbb{R}^D}}{\sigma_N} \right) - 1.
\]

Set \(\sigma_N = \frac{1}{N^D}\) and \(d_N = \frac{1}{N^{D+1}}\), then \(|x - \mu_2|_{\mathbb{R}^D} - |x - \mu_1|_{\mathbb{R}^D} \leq N\|\mu_1 - \mu_2\|_{\mathbb{R}^D} \leq Nd_N \leq 9\) whenever \(\|\mu_1 - \mu_2\|_{\mathbb{R}^D} \leq d_N\). Then \(|\psi - 1| \leq 1000|x|\) whenever \(|x| \leq 9\) and

\[
\Delta_N(x) \leq C_D \times N^D \times 1000 \left| \frac{x - \mu_2 - x - \mu_1}{\sigma_N} \right| \leq \frac{1000C_D}{N} \to 0, \quad N \to \infty.
\]

9.3.2 A Remark on Theorem 4.2

From the functional viewpoint, \(\eta_{4+2s-d}\) is an explicit representation of a reproducing kernel of \(\nabla^{-2} \hat{H}^s(\mathbb{R}^d)\) as a semi-Hilbert subspace of \(H^2 \cap \mathbb{R}^d\).

9.4 Proof of Lemma 2.1

To finalize the proof Lemma 2.1, we provide the following result along with its proof.

Lemma 9.1. Suppose \(M_f^d\) is a \(d\)-dimensional manifold determined by a continuous map \(f : \mathbb{R}^d \to \mathbb{R}^D\) and \(x \in \mathbb{R}^D\). Then \(B_{M_f^d}(x, r) := \{t \in \mathbb{R}^d : \|x - f(t)\|_{\mathbb{R}^D} \leq r\}\) is a compact subset of \(\mathbb{R}^d\) for all \(r > 0\).

Proof: Clearly \(B_{M_f^d}(x, r) = f^{-1} \left( \overline{B(x, r)} \cap M_f^d \right) \) and \(\overline{B(x, r)} = \{x' \in \mathbb{R}^D : \|x' - x\|_{\mathbb{R}^D} \leq r\}\) is a compact subset of \(\mathbb{R}^D\). The desired result follows from the fact that \(\overline{B(x, r)} \cap M_f^d\) is a compact subset of \(M_f^d\) with respect to the topology of \(M_f^d\) and \(f\) is a homeomorphism. \(\square\)

Proof of Lemma 2.1. Let \(r = \inf_{t' \in \mathbb{R}^d} \|x - f(t')\|_{\mathbb{R}^D}\), then \(B_{M_f^d}(x, 2r) \neq \emptyset\) and \(\inf_{t' \in \mathbb{R}^d} \|x - f(t')\|_{\mathbb{R}^D} = \inf_{t' \in B_{M_f^d}(x, 2r)} \|x - f(t')\|_{\mathbb{R}^D}\). From the previous lemma, we know that \(B_{M_f^d}(x, 2r)\) is compact. So there exists \(t^* \in B_{M_f^d}(x, 2r)\) such that \(\|x - f(t^*)\|_{\mathbb{R}^D} = \inf_{t' \in B_{M_f^d}(x, 2r)} \|x - f(t')\|_{\mathbb{R}^D}\), which implies \(t^* \in S_{x}^f\). So \(S_{x}^f\) is non-empty. From the previous lemma, \(S_{x}^f = \arg \inf_{t \in \mathbb{R}^d} \|x - f(t)\|_{\mathbb{R}^D} = B_{M_f^d}(x, r)\) is compact. \(\square\)
9.5 Proof for Theorem 3.1

Since \(M_{+x, \mathcal{G}}(f) < +\infty\) if and only if each component of \(f\) is a linear function, it suffices to show (ii). Without loss of generality, we assume that \(\mathbb{E}_{\mathcal{P}} X = 0\). Since \(\nabla^2 f = 0\) for all \(f \in \mathcal{L}\), we have

\[
\inf_{f \in \mathcal{L}} M_{w, \mathcal{L}}(f) = \inf_{P \in \mathcal{P}, \text{tr}(P) = d, a \in \mathbb{R}^D} \mathbb{E}_{\mathcal{P}} \|X - (a + PX)\|^2_{\mathbb{R}^D} = \inf_{P \in \mathcal{P}, \text{tr}(P) = d} \text{tr}((I - P)\text{Var}_{\mathcal{P}}(X)),
\]

(9.4)

where \(\mathcal{P}\) denotes the collection of all \(D \times D\) projection matrices and \(\text{Var}_{\mathcal{P}}\) denotes the variance with respect to \(\mathcal{P}\). Suppose \(\{\lambda_i\}_{i=1}^l\) are the eigenvalues of \(\text{Var}_{\mathcal{P}}(X)\) with \(\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_D\) and \(v_i\) is the eigenvector corresponding to \(\lambda_i\). Let \(\tilde{P}\) denotes the projection matrix onto the \(d\) dimensional subspace \(\text{Span}\{v_1, v_2, \cdots, v_d\} = \left\{ \sum_{i=1}^d t_i v_i | t_1, t_2, \cdots, t_d \in \mathbb{R}^1 \right\}\). Then (9.4) implies \(\inf_{f \in \mathcal{L}} M_{w, \mathcal{L}}(f) = \mathbb{E}_{\mathcal{P}} \|X - \tilde{P}X\|^2_{\mathbb{R}^D}\). Hence, the \((\mathcal{L}, w)\) principal manifold for \(X \sim \mathcal{P}\) is the \(d\)-dimensional hyperplane \(\text{Span}\{v_1, v_2, \cdots, v_d\}\).

\(\square\)

9.6 Proof of Theorem 3.3

Let \(f^* = \arg \inf_{f \in \mathcal{G}} M_{0, \mathcal{P}}(f)\), then \(f^*\) is a critical point of \(M_{0, \mathcal{P}}\) in \(\mathcal{G}\), i.e. \(\frac{d}{dt}\big|_{t=0} M_{0, \mathcal{P}}(f^* + tg) = 0\) for all \(g \in \mathcal{G}\). Using the same method as in the proof of Proposition 4 in Hastie and Stuetzle (1989), we can show that \(f^*\) satisfying the self-consistency (3.4) is equivalent to the fact that \(f^*\) is a critical point of \(M_{0, \mathcal{P}}\). Then \(f^*\) gives a principal manifold (curve) of self-consistent type.

The method used in the proof for Proposition 4 in Hastie and Stuetzle (1989) requires the boundedness of \(\|f^* (\pi_{f^*}(X))\|_{\mathbb{R}^D}, \|g (\pi_{f^*}(X))\|_{\mathbb{R}^D}, \|f^* (\pi_{f^*}(X))\|_{\mathbb{R}^D}\) and \(\|g (\pi_{f^*}(X))\|_{\mathbb{R}^D}\), which enables us to use Lebesgue’s dominated convergence theorem. With the fact that \(\pi_{f^*}(x)\) is continuous at \(t = 0\) for almost every \(x\) in the support of \(\mathcal{P}\), which is essentially shown in the proof of Lemma 4.3.1 in Hastie (1984), compactness of the support of \(\mathcal{P}\) and the embedding \(\nabla^{-2} \mathcal{H}^0(\mathbb{R}^1 \to \mathbb{R}^D) \subset C^1(\mathbb{R}^d \to \mathbb{R}^D)\) from Sobolev embedding theorem imply the necessary boundedness.

\(\square\)

9.7 Proof of Theorem 3.4

To prove Theorem 3.4, we need the following lemma.

Lemma 9.2. Suppose \(p \in C(\overline{\mathbb{R}^D})\) and \(\psi\) is a non-negative function on \(\mathbb{R}^D\) with \(\int \psi = 1\). Then

\[
\left\| p \ast \frac{1}{\sigma^D} \psi \left( \frac{x}{\sigma} \right) - p \right\|_{L^\infty(\mathbb{R}^D)} \to 0, \quad \sigma \to 0,
\]

where \(p \ast \frac{1}{\sigma^D} \psi \left( \frac{x}{\sigma} \right) = \int_{\mathbb{R}^D} p(\mu) \times \frac{1}{\sigma^D} \psi \left( \frac{x - \mu}{\sigma} \right) d\mu\).
Define $\hat{N}$ from Theorem 5.1.8 of Durrett (2010), we have \[ \arg \inf \frac{1}{\sigma_N} \psi \left( \frac{x - \mu_j, N}{\sigma_N} \right) \leq \frac{1}{\sigma_N} \psi \left( \frac{x - \mu_j, N}{\sigma_N} \right). \]

Since $\hat{N}$, \[ \hat{N} = \left( \hat{1}, \hat{2}, \cdots, \hat{N}, \right) \in \Theta_N \text{ as } \int p = 1 \text{ and } \bigcup_{j=1}^{N} A_{j,N} = \text{supp}(p). \]

Then \[ p_N(x|\hat{N}) = \sum_{j=1}^{N} \int_{A_{j,N}} p(\mu) d\mu \times \frac{1}{\sigma_N} \psi \left( \frac{x - \mu_j, N}{\sigma_N} \right) = \sum_{j=1}^{N} \int_{A_{j,N}} p(\mu) \times \frac{1}{\sigma_N} \left( \psi \left( \frac{x - \mu_j, N}{\sigma_N} \right) - \psi \left( \frac{x - \mu_j, N}{\sigma_N} \right) \right) d\mu \]

Since $\|\mu - \mu_j, N\|_{\mathbb{R}^D} \leq d_N$ whenever $\mu \in A_{j,N}$,

\[ I \leq \sup_{\|\mu_1, \mu_2\|_{\mathbb{R}^D} \leq d_N} \left\| \frac{1}{\sigma_N} \psi \left( \frac{x - \mu_1}{\sigma_N} \right) - \frac{1}{\sigma_N} \psi \left( \frac{x - \mu_2}{\sigma_N} \right) \right\|_{L^\infty(\mathbb{R}^D)} \to 0, \quad N \to \infty. \]

Since $II = p \ast \frac{1}{\sigma_N} \psi \left( \frac{x}{\sigma_N} \right)$ and $\sigma_N \to 0$, from Lemma 9.2, we have $\|p \ast \frac{1}{\sigma_N} \psi \left( \frac{x}{\sigma_N} \right) - p\|_{L^\infty(\mathbb{R}^D)} \to 0$ as $N \to \infty$. Then the desired result follows. \[ \square \]

9.8 Proof for Theorem 3.2

From Theorem 5.1.8 of Durrett (2010), we have $\arg \inf_{Y \in L^2(\mathbb{P})} E \|X - Y\|_{\mathbb{R}^D}^2 = E \left( X|\pi_f(\omega)(X) \right)$, where $\sigma \left( \pi_f(\omega)(X) \right)$ is the $\sigma$-algebra generated by $\pi_f(\omega)(X)$ and \[ L^2(\mathbb{P})|_{\sigma \left( \pi_f(\omega)(X) \right)} := \{ Y \in L^2(\mathbb{P}) : Y \in \sigma \left( \pi_f(\omega)(X) \right) \}. \]

Let $\mathcal{M}$ denotes the collection of all measurable maps from $\mathbb{R}^1$ to $\mathbb{R}^D$. It is straightforward that

\[ E \left( X - E \left( X|\pi_f(\omega)(X) \right) \right)^2 = \inf_{Y \in L^2(\mathbb{P})} \inf_{f \in \mathcal{M}} E \left( X - f \left( \pi_f(\omega)(X) \right) \right)^2. \]
Define $f_{(n+1)}(t) := \mathbb{E}\left(X|\pi f_{(n)}(X) = t\right) = Tf_{(n)}(t) \in \mathcal{C}$, then $f_{(n+1)}(\pi f_{(n)}(X)) = \mathbb{E}\left(X|\pi f_{(n)}(X)\right)$ and

$$f_{(n+1)} = \arg \inf_{f \in \mathcal{C}} \mathbb{E}\left\| X - f(\pi f_{(n)}(X)) \right\|^2_{\mathbb{R}^D} = \arg \inf_{f \in \mathcal{C}} M_{0,\mathbb{P}}(f, f_{(n)}) = Tf_{(n)},$$

which is the desired result. $\blacksquare$