# Describing Multivariate Distributions with Nonlinear Variation Using Data Depth [1]

Rima Izem[2], Eynat Rafalin [3] and Diane L. Souvaine[3]

# 1 Introduction

Growth curves of plants and animals, human speech, gene expression signals, and medical images or 3-dimensional shapes of cancer tumors, are all real life examples of high dimensional multivariate data referred to as functional data [80, 81, 26, 33]. Variability in these data sets could be representative of respectively environmental or genetic variation of growth patterns for plants or animals, accent variation for pronunciation of words, genetic differentiation in cell tissues, and morphological variation of cancer tumors. Effectively describing variation in these data set could lead respectively to better understanding of patterns of growth important in agriculture, to development of better voice recognition tools, and to the discovery of better disease diagnosis tools.

Several statistical methods such as principal component analysis and analysis of variance are often effective in analyzing variation in the data when the space of variation is linear [7]. However, describing variability is much more difficult when the data varies along nonlinear modes [80, 81, 26, 55]. Simple examples of nonlinear variation in functional data are horizontal shift of curves of common shape [81, 55], frequency change of acoustic signals of common shape [53], or lighting change in images of the same object [86].

On one hand, current methods dealing with nonlinear variation in high dimensional data such as principal curves [39], manifold learning methods [86], or kernel methods [10] are mainly concerned with dimensionality reduction, not quantification of the variation. On the other hand, the rediscovered measures of center and spread in manifolds, the Fréchet mean and Fréchet variance [70], used in shape analysis [32, 11] and curve analysis [54, 53], are not robust measures and do not give a full description of the variability. More precisely, as for the usual sample mean and variance, the sample Fréchet mean and variance are not robust to outliers or skewness in the data. Although the usual Euclidean sample variance is a matrix of dimension $k$ when the data lies in a $k$ dimensional Euclidean space, the Fréchet variance is only a scalar, even when the data lies in a $k$ dimensional manifold.

Data depth methods [83, 69, 68, 94, 95], such as half-space depth [49, 89] or simplicial depth [67], are particularly attractive descriptions of the distribution of multivariate data because they are nonparametric, robust and geometrically motivated. They use local geometry of the distribution to define its global properties, such as contours. However, classical data depth definitions and methods are ill suited for the description of multivariate distributions when these distributions lie in a non-convex or a smaller dimensional manifold.

This project presents novel data depth functions that are capable of describing variability in multivariate data when the space of variation is a manifold or the result of nonlinear variation in the data. These newly proposed statistical concepts show significant potential for quantifying properties of the probability distribution, for reducing dimensionality, for decomposing variability and for inference. We propose to estimate our measures of depth from the data using proximity graphs, which are graphs that capture the geometry of the point set. The statistical challenges are: first, in proving that our intuitive definitions are meaningful quantification of variation for a large class of probability distributions, and are equivalent to usual depth methods for convex Euclidean spaces; second, in characterizing the distribution of our sample estimates and deriving the bias and efficiency of these estimates for small samples. The computational challenges are: first, in creating new efficient and flexible algorithms to estimate the proposed sample depth, especially for high dimensional data where existing algorithms

[1] This in the text of a proposal submitted to the National Science Foundation in response to Program Solicitation NSF 05-622 *Mathematical Sciences: Innovations at the Interface with the Physical and Computer Sciences and Engineering, MSPA-MCS*

[2] Department of Statistics, Harvard University, Cambridge, MA, 02138

[3] Department of Computer Science, Tufts University, Medford, MA 02155

and evaluation of complexity are currently scarce; second, in developing new and improved algorithms for manifold learning and boundary detection for high dimensional data that are of independent interest. This interdisciplinary project is data driven and arose from a common research interest of a statistician, and two computer scientists. Partial progress has already been achieved, and successful completion will not only advance the statistics and computer science fields but also bring fresh insights to scientists analyzing real data.

This research will progress in several directions. The first stage will focus on characterization of the class of distributions for which our definition coincides with other classical definitions of data depth such as half-space depth, and simplicial depth. Practical methods of estimating this depth function based on proximity graphs will be developed in the second stage. The third stage will focus on creating statistical analysis and quantification methods, specifically methods detecting outliers or skewness in a sample, for dimensionality reduction, for description of variation, and for inference. Our algorithms will build on existing methods for manifold learning and representation of high-dimensional manifolds and suggest new methods for partial manifold learning. As part of our research we will quantify the required sampling conditions for the estimators to approximate the underlying distributions and suggest additional methods that will improve the behavior of the suggested estimators. To finalize this research, an experimental study will be conducted with varying constructions of graphs on point sets and applied for analysis of real life data sets.

**Intellectual merit** derives from the development of these methodologies for the quantification of variation of multivariate data on manifolds, through a successful interdisciplinary collaboration between statisticians and computer scientists. **Broader Impacts** are twofold: first, providing fresh insights for data analysis through application of these statistical measures to real-world data sets; second, training of future statisticians and computer scientists and promoting interdisciplinary collaboration between the two groups. To this effect, a graduate student seminar on the topic of this proposal will be offered collaboratively at Tufts and Harvard, and results of this project will be integrated into undergraduate classes.

**Research Team:** Principal Investigator Izem is an expert on functional data analysis, which is the analysis of multivariate data of high dimension, such as sets of curves, images, and shapes. Her current collaborations are with colleagues in Biology, Engineering, Economics, Astrophysics and Anthropology working in developing novel statistical methods for analysis of variation in high dimensional data such as growth curves, acoustic signals, and spatio-temporal data [55, 60, 56, 54, 53]. Principal Investigator Souvaine and Co-Investigator Rafalin are experts in the field of computational geometry and have a history of bringing their theoretical discoveries in geometry and algorithms to practical fruition. In the last few years Souvaine and Rafalin have concentrated on the computational study of *data-depth*, a statistical analysis method that is geometrical in nature and provides an attractive alternative to classical statistics [78, 16, 73, 15, 75, 77, 76, 51]. Consultant Herrera provides an expertise in differential topology and Riemannian geometry [42, 44, 41, 43, 45, 47, 46]. Discussions at the Radcliffe Institute for Advanced Study initiated this collaboration and preliminary results have already been achieved

**The structure of this document:** Section 2 gives an overview of data depth methods and some motivating examples of nonlinear variations in multivariate data. It shows how classical methods fail to describe the variation and to represent the geometry of the sample distribution when these distributions lie in non-convex subspaces or in curved surfaces or manifolds. Section 3 and 4 describe our major contributions. In Section 3, more meaningful definitions of data depth which respect the geometry of distributions with nonlinear variation are presented. It also shows how the *desirable properties for data depth functions* can be generalized for distributions with nonlinear variations. Section 4 describes a paradigm using proximity graph, to approximate these depth functions from a finite sample. Section 5 describes the computational and experimental framework of this proposal. Sections 6, 7, and 8 include summary of proposed research, incorporation of this research in graduate and undergraduate training and results of prior support. Finally, the appendix (Section 9) defines the necessary terminology in differential geometry.

# 2 Motivation: Statistical Analysis of Nonlinear Variation

This section illustrates multivariate data with nonlinear variation by simple examples and shows that classical data depth methods are ill suited for describing the distribution of these data. This failure motivates our proposal of new definitions of data depth which respect the geometry of the space of variation. The proposed methods are especially useful for describing and quantifying high dimensional multivariate distributions.

In general, we say that variation in multivariate data is *nonlinear* if it lies in a Euclidean space but varies along nonlinear directions or if it lies in a lower dimensional manifold. We start with an overview of usual data depth methods in Subsection 2.1. Next we show two examples of multivariate data with nonlinear variations: simple bivariate data in Subsection 2.2 and high dimensional data in Subsection 2.3. We discuss why usual methods of data depth fail to describe the *true* variability in both examples and propose more meaningful data depth concepts, which would better respect the geometry of the space of variation. Finally, we survey statistical methods applied to data with nonlinear variations in Subsection 2.4.

## 2.1 Data Depth for Analysis of Data in Euclidean Space

*Data depth* is a statistical analysis method that assigns a numeric value to a point, corresponding to its centrality relative to $F$, a probability distribution in $\mathbb{R}^d$ or relative to a given data cloud.
Figure 1 shows the *depth contours* formed by applying the half-space depth function to a set of 500 normally distributed data points. Depth contours [89] are nested regions of increasing depth that serve as a topological map of the data. The $j$th depth contour consists of all those points in space of depth $\geq j$. The half-space depth function [49, 89] (in the literature sometimes called *location depth* or *Tukey depth*) is one of the most well studied depth functions both by the statistical and computational communities. It is attractive as it attains all four desirable properties of depth functions (see Subsection 3.2).

**Definition 2.1.** *The* **half-space depth** *of a point $x$ relative to a set of points $\mathcal{S} = \{X_1, ..., X_n\}$ is the minimum number of points of $\mathcal{S}$ lying in any closed half-space determined by a line through $x$.*

For a point $x_1$ that is shallow in the set there is a half-plane passing through $x_1$ that contains a small number of points of $S$ in its interior, while for a point $x_2$ that lies in the center of the set, every half-plane contains a relatively large number of points. Half-space depth contours in the sample case are nested, connected, convex and compact [95, 24].

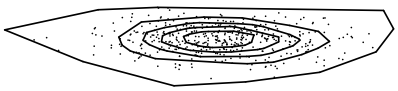


Figure 1: $20\%, 40\%, 60\%, 80\%$ and $100\%$ contours approximated using the convex hull of the $20\%, 40\%, 60\%, 80\%$ and $100\%$ deepest points for a data set consisting of 500 points, normally distributed, width scaled by 5.

The concept of *data depth* [83, 69, 68] has been developed over the last decade as a method of nonparametric multivariate data analysis. Other examples include simplicial depth [67], Mahalanobis depth [71], Oja depth [74], projection depth [90] and the convex hull peeling depth [9, 27]. Data depth is "robust against the possibility of one or several unannounced outliers that may occur in the data and yield reasonable results even if several unannounced outliers occur in the data" [82]. In addition, data depth enables one to quantify and compare statistical attributes of high dimensional data sets. "The value of a depth function for a specific point may vary with the notion of data depth but for each notion of depth a larger value always implies a deeper (or more central) point with respect to the probability distribution" [69].

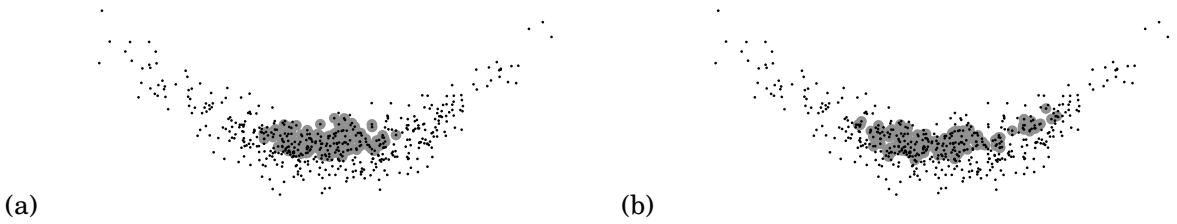
Statisticians have developed a systematic approach for defining quantitative distributional characteristics and inference methods based on the concept of data depth [69]. This work also introduces several simple graphs to display the characteristics of the distribution visually. They suggest a natural generalization of the quantiles by using *level contours*, as follows: The *$p$th central region*, $C_p$, is the smallest region enclosed by depth contours to amass probability $p$ and its boundary is denoted $Q_F(p)$.

This methodology when applied to the sample case can be viewed as a multivariate generalization of the univariate rank methods, where the ranking is center-outward and is induced by depth. *Rank*

*statistics* are based on the ordering of observations, where the order *reflects extremeness, contiguity, variability or the effect of external contamination and provides parameter estimation method* [9]. Other methods of multivariate ranking exist [9, 48], but cannot be easily extended to distributions with non-linear variations.

## 2.2 Example 1: *Curve* Shape Bivariate Distribution

We see in Figure 2(a) an example of a bivariate distribution where the space of variation resembles the shape of a parabola or a banana. Although the data in this case lie in a two dimensional Euclidean space, the distribution of the data is not along linear axes, and the space of variation is non-convex. Figure 2(a) demonstrates why the classical concept of data depth does not give meaningful results for distributions with non-linear variations. We would expect the deepest or most central points to lie close to the medial axis of the banana. However, the half-space depth calculates the deepest points as points that are close to the center of the convex hull of the shape, points that are on the inner boundary of the banana. We define in Section 3 data depth measures which better respect the geometry of the data. These data depth measures use the geodesic distance, the shorter length of paths between two points along the support of the distribution, rather than the Euclidean distance, as measure of spread in the data. We propose in Section 4 a method for estimating this revised depth from a sample. The results of this new method are shown in Figure 2(b).



(a) (b)

Figure 2: Distribution following a banana shape. *Left (a):* The 30% deepest points computed according to the half-space depth do not follow the medial axis of the data and rather, are close to the center of the convex hull of the set. *Right (b):* The $30\%$ deepest points computed according the the Delaunay-based proximity depth describe better the geometry of the space of variation. These subplots were generated using the *Depth Explorer* software [50, 51].

## 2.3 Example 2: Growth Curves

The growth curves in Figure 3 show the change of height over time for different genotypes of the same species of plant[4]. It is an example of a large family of curves collected in the biological sciences, called reaction norm curves, which measure the change of a trait as a function of the environment for different genotypes (see [59] and [54] for other examples). Each growth curve represents a family and the variation in these curves represents the genetic variation in the population. Data depth methods applied to these curves would allow biologists to understand the genetic variation in the population, in particular to define the center of the distribution (curve of highest depth) as well as rank all curves with respect to this distribution.

This set of curves is one example of functional data of common shape. Although the height was measured at 6 time points, the height varies continuously over time. Functional Data Analysis (FDA) considers the infinite-dimensional curve, rather than the 6-dimensional vector, as the statistical entity of interest. Earlier models described the growth curves in a logistic parametric form, such that

$$y_{ij} = f_i(t_j) + \epsilon_{i,j}, \text{ and } f_i(t) = \frac{K_i}{1 + (K_i/N_i - 1)\exp(-r_i t)};$$

[4]Data kindly provided by John Stinchcombe, Biology Department at the University of Toronto, and Johanna Schmitt, Biology Department at Brown University

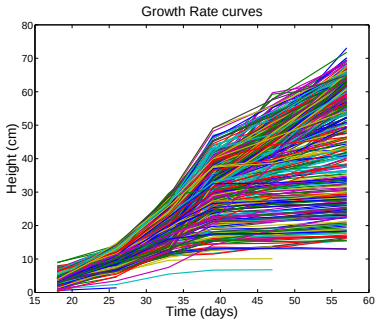



Figure 3: Growth curves, showing the change in height over time for 49 plant genotypes. Each curve represents the mean height of a genotype at 6 time points. The variation in this data represents the genetic variation in the population.

where $y_{ij}$ is the measured height of family $i$ at time $t_j$, $f_i(t_j)$ is the value of the height function at time $t_j$ and $\epsilon_{i,j}$ is the additive noise due to measurement error. The three parameters $N_i$, $K_i$ and $r_i$ are respectively the initial height, final height and growth rate of family $i$.

We could think of the distribution of these curves in different ways: first, since the height of the plant was collected at 6 time points, as variation in a 6-dimensional space; second, as variation of functional data in an infinite-dimensional space; third, as variation in the 3-dimensional parametric space. More precisely, if the data is generated by the parametric logistic curves, variation in the growth curves could be represented by variation in the population of the three parameters of growth rate $r$, final height $K$ and initial height $N$. Figures 4(a) and 4(b) show particular examples of logistic curves, their 3-d space of variation (by considering only 3 measurements, so that we could visualize the space of variation), and the corresponding parametric variation. Figure 4(a) (resp. Figure 4(b)) shows the variation when the parameter $r$ varies (resp. $r$ and $K$ vary), keeping $K$ and $N$ fixed (resp. $N$ fixed).

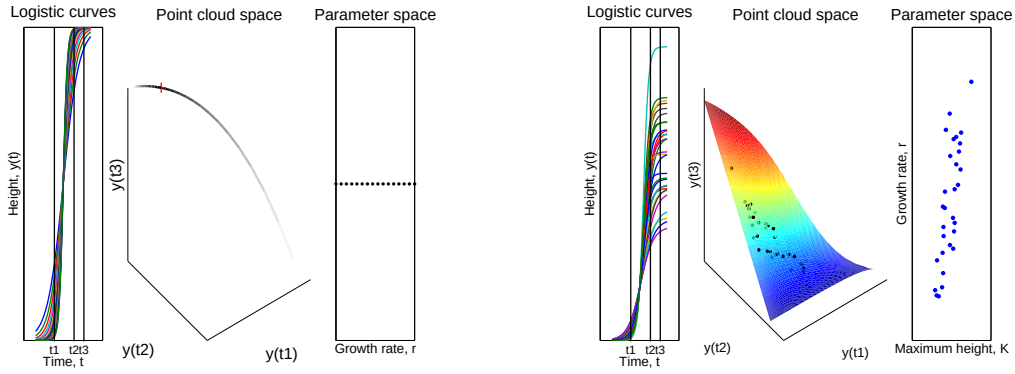
We see in Figures 4(a) and 4(b) that variations in the parameters $K$ and $r$ result in nonlinear variation in the height. Because of this nonlinearity, describing the distribution by classical data depth methods would produce distinctly different answers for the three different ways of describing data variation. Given that the parametrization of variation of the data is generally unknown and that different parametrization could produce equivalent variation of the data, we focus here on data depth methods on the original scale of the data rather than on any parametrization. An additional advantage of this approach is that the dimensions in the original view of functional data are on comparable scales, whereas the parameters may lie on different scales. For example, the dimensions in the original data for logistic curves represent heights and lie on the same scale, but under parametrization, the initial height $N$ and final height $K$ share a scale, but the growth rate $r$ has a different scale.

Although the data in Figure 4 (a) lies in a one-dimensional curve, classical data depth methods applied to this curve would find that all point have equal depth, so they would fail to identify a center or spread of the distribution. For example, according to the half-space depth, for every data point we could find a half-plane passing through that point that has zero points on one side, and thus the depth of all points will be zero. We would like to define a data depth concept that would be invariant or consistent with increasing dimensionality of the data. Moreover, we would like to define a correspondence between the depth in the original data and a possible parametrization of the variation. For example, the shades of gray in the second subplot in Figure 4(a) correspond to contours that we believe better reflect the distribution in the data set.

## 2.4 Nonlinear Variation in Statistics

The novelty of the proposed work is that it builds both on statistical methods for describing nonlinear variation and some existing algorithms for dimensionality reduction to proceed further in understanding and quantifying variability in a more robust way. This subsection surveys statistical methods used for describing variability in non-linear spaces of variation.

In FDA and image analysis, nonlinear variations of curves around a common shape, such as registration or horizontal shift, have been recognized for some time, see Chapter 5 in [81]. However, most existing FDA methods of analysis of variance which account for nonlinearity in the data, have treated nonlinearity somewhat as a nuisance. Although nonlinear variations are sometimes important variation in the data [56], these methods remove the nonlinear variation to estimate the common curve




(a) (b)

Figure 4: Two examples of high dimensional logistic curves variation. Curve space (left); space of variation of this data shown in 3-d (middle), where coordinates of each point $i$ represent the heights $(y_i(t_1), y_i(t_2), y_i(t_3))$ at times $t_1$, $t_2$ and $t_3$) and the parameter space (right). *Subplot(a)* left: logistic curves with varying growth rate; middle: the space of variation is a one dimensional curve, the shades of the curves from lighter to darker represent smaller to larger depth (for our ideal measure of depth) for this sample; right: the parameters $r$ generating the curves are equally spaced. *Subplot (b)* left: logistic curves with varying final height $K$ and growth rate $r$; middle: the space of variation is supported in a 2-d manifold; right: the parameter space generating the data, $r$ versus $K$.

better, see [92, 93], and then apply conventional linear techniques [84, 81], they do not account for nonlinear components in the quantification of variation.

Fréchet [70] conducted some of the earliest work on generalizing the concept of mean and variance to distributions in manifolds rather than Euclidean spaces. These concepts were rediscovered for analysis of nonlinear variation in curves [56, 54, 53], and statistical shape analysis [63, 64, 11, 32]. The Fréchet mean in the manifold coincides with the usual mean when the support of the distribution is Euclidean with the usual metric. However, the Fréchet variance is a number, not a matrix, which represents the total variation in the data, even when the data lie in a $k$ ($k \geq 1$) dimensional manifold. Although the Fréchet variance addresses the problem of nonlinearity, it does not fully describe the variation. As in Fréchet's mean and variance, we propose to use a metric in the manifold to define the depth of a point; the data depth methods proposed in Section 3, however, provide a more nuanced understanding of the variability in the data.

Hastie and Stuetzle's [39] pioneering, and highly cited [21, 22, 85], work on principal curves relies on finding principal directions of variations that are nonlinear, as opposed to linear principal component. Principal curves are the curves in the middle of the data which are formally defined as the curves which satisfy the self-consistency principle. Many other manifold learning methods have been recently developed, mostly in computer science, to analyze variation of data in a nonlinear manifold using local geometric properties [86, 25, 20, 19, 23, 36, 18] and kernel methods [10]. The principal curve and manifold learning tools have been mainly used for dimensionality reduction, and not for quantification of variability along principal directions. As in manifold learning methods, the methodology we propose will use proximity graphs to learn the geometry of the support. In addition, we address the quantification of the variability using proximity graphs.

# 3 Depth Functions for Distributions with Nonlinear Variation

Our goal is to define medians and depth functions for data and distributions with nonlinear variation. This method will make it possible to quantify statistical properties of the probability distribution, useful for dimensionality reduction, variability decomposition and inference.

We propose two meaningful measures of data depth in Subsection 3.1. These definitions provide a

measure of depth which respects the geometry of the distribution. We propose also general desirable properties for data depth functions for distributions in Subsection 3.2. The method we propose to approximate this depth function, from a sample, will be presented in Section 4. Readers not familiar with differential geometry definitions of *Riemannian manifold*, *Tangent space*, *geodesics* and *log map* are encouraged to read the Appendix for reference.

## 3.1 Depth Functions

We propose two different ways of defining depth in manifolds, they both respect the geometry of the support of the distribution. The first definition links the depth measure in a manifold to the depth measure in the *tangent space*, which is a Euclidean space. The second definition is related to the concept of minimizing the $L_1$ distance to define the median or point of highest depth. These depth measures are broadly defined for a large class of distributions supported on a complete connected *Riemannian manifold* or submanifold $\Omega$. Our definitions of depth use *geodesic* distances $d(x, y)$, or the shortest distance of paths along the support $\Omega$ between two points $x$ and $y$. [5] In particular, if $\Omega$ is $\mathbb{R}^k$ ($k \leq d$), or any convex set in $\mathbb{R}^k$, then this distance is the usual Euclidean distance.

### 3.1.1 Definition 1: Generalized Depth in Manifolds as Depth in *Tangent Space*

This depth definition links any classical depth function, defined in Euclidean spaces, with definition of depth in a manifold through the *tangent spaces* to the manifold and the *log map* (as defined in Appendix).

**Definition 3.1.** *The generalized [D] depth of a point $x$ in the manifold is its [D] depth in its tangent space $T_x\Omega$, where all other points in the manifold were transformed to the tangent space using the isometry log map $Log_x$ and [D] is any depth function defined with respect to Euclidean space.*

Since the *tangent space* $T_x\Omega$ is a Euclidean space, any measure of depth [D] can be used in this Euclidean space, such as half-space depth or simplicial depth.

If the support $\Omega$ is a convex subspace of a Euclidean space $\mathbb{R}^k$, the log map at any point is the identity function. So, the advantage of this data depth is that the measure coincides exactly with the usual definition of depth in Euclidean spaces. On the other hand, if the support $\Omega$ is a non-convex or a nonlinear manifold this depth measure respects the geometry of the manifold.

### 3.1.2 Definition 2: Depth in Manifolds Using $L_1$ Medians

This depth definition links the $L1 - median$ in a Euclidean space with the $L1 - median$ in a manifold. The depth of a point is then defined from center outward relative to the median.

**Definition 3.2.** *Let $F$ be a distribution supported in $\Omega$ such that $\int_\Omega d(x, y)dF(y) < \infty$ for all $x \in \Omega$. A point $m \in \Omega$ is called a median of $F$ if $\int_\Omega d(m, y)dF(y) \leq \int_\Omega d(x, y)dF(y), \forall x$. Moreover, the generalized $L_1$ depth of a point $x$ in the support is the ratio $D(x) = \frac{\int_\Omega d(y,m)dF(y)}{\int_\Omega d(x,y)dF(y)}$.*

Under this definition, the depth of a point decreases as we move away from a *median* along a geodesic. This is a new definition of depth. We will investigate the conditions for which this definition will coincide with other depth measures in convex Euclidean spaces.

## 3.2 Desirable Properties of Depth Functions

Desirable properties of depth functions and depth contours, as a tool to analyze and evaluate depth functions, were suggested and often cited statistical papers [67, 94, 95, 40]. In particular, Serfling and Zuo [94] formulated a general definition of desirable properties of depth functions to *serve most effectively as a tool providing center outward ordering of points* (affine invariance; maximality at center; monotonicity relative to deepest point; and vanishing at infinity). They evaluated several depth functions according to these properties. As shown in [50] these desirable properties are meaningful and give the expected results if the underlying distribution is *unimodal* and the ordering of the points is center-outward.

[5] This distance is well defined for connected complete *Riemannian manifolds* or submanifolds, see Appendix 9.

We will revise some of these properties, such as affine invariance, and monotonicity relative to deepest point, to account for distributions in manifolds or non-convex spaces. Zuo and Serfling use these properties to prove important convergence properties of depth contours. We plan to extend their work, and prove comparable results for data with nonlinear variations.

**Affine Invariance:** A depth function is said to be *affine invariant* if its coordinate system can change without affecting the relative geometry of the function, i.e. the geometry remains consistent with rotation, scale, or translation of the axes. For such a function, the choice of scale of coordinate system will not affect the relative depth between two points, nor the depth contours. Thus, for experimental data with several dimensions/variables, even if the relative proportions between the dimensions are not known, the output of the analysis will remain unchanged.

We consider that the relative scale is important and that some affine transformations can change dramatically local properties related to scale, as shown in Figure 3.2. Instead of letting the data depth be invariant to all affine transformations, we propose to modify this property as follows:
*Invariance under expansion of the space:* The data depth should be invariant to any transformation of the distribution by $H$ such that $H : x \in \Omega \mapsto H(x) \in \Upsilon$ and $d_\Upsilon(H(x), H(y)) = \alpha * d_\Omega(x, y)$ for some scalar $\alpha$, where $d_\Omega$ (resp. $d_\Upsilon$) is the geodesic distance in $\Omega$ (resp. $\Upsilon$).

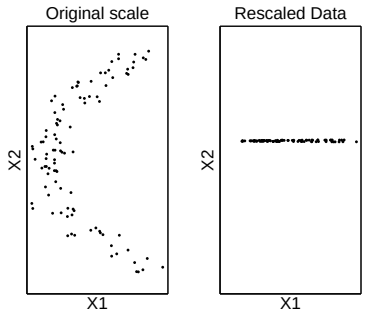



Figure 5: Distribution following a banana shape after an affine transformation that shrinks the variable $X_2$ while leaving the variable $X_1$ the same. *Left*: original data set. *Right*: data set after affine transformation. This transformation does not preserve the relative scale and changed the local properties of the distributions.

**Maximality at Center:** The *maximality at center* property states that for a distribution having a uniquely defined center (e.g. a point of symmetry with respect to some notion of symmetry), the depth function should attain maximum value at the center. By changing the definition of *symmetry* to account for symmetry on a manifold, we consider the property of maximality at the center to remain desirable for manifolds.

**Monotonicity Relative to Deepest Point:** For the *monotonicity property* to hold, as a point $x$ moves away from the 'deepest point' (the point at which the depth function attains maximum value) along any fixed ray through the center, the depth at $x$ should decrease monotonically.

This property clearly will not hold as stated for data from a distribution lying on a non-convex support. Consider a two dimensional manifold which resembles a *swiss roll* (see [86, 25]). Then a fixed ray through the center can cross this manifold several times, where consecutive cuts are not necessarily cutting consecutive regions in the manifold

Rather than moving in a line from the deepest point to another point, we can move through the *geodesics* between two points. The depth along this path should decrease monotonically as we move away from the deepest point. We propose the following:
*Monotonicity Relative to Deepest Point in a Manifold:* For every geodesic path connecting a point $x \in M$ to the 'deepest point' (the point at which the depth function attains maximum value), the depth on the path should decrease monotonically.

**Vanishing at Infinity:** According to the *vanishing at infinity* property the depth of a point $x$ should approach zero as $||x||$ approaches $\infty$, where $||x||$ is the Euclidean norm of $x$. In other words, the depth of point $x$ should approach zero as it moves away from the deepest point to infinity. This property is desirable in order to ensure the boundedness of any depth contour. While this property can remain unchanged it can be relaxed by using the geodesic distance instead of the Euclidean distance, i.e. the depth of point $x$ should approach zero as $d(x, m)$ goes to infinity, where $m$ is a point of highest depth.

### 3.3 Proposed Future Research Related to Depth in Manifolds

We propose to prove formally:

1. That both proposed data depth measures satisfy our defined desirable properties.
2. For which distributions our depth definitions are equivalent to usual depth definitions–we expect these definitions to be equivalent when the distribution is unimodal and its support is convex.
3. For which distributions in manifolds would both of our definitions coincide.
4. Convergence properties based on the desirable properties of depth functions.

## 4 Data Depth for Sample Sets

For most existing depth functions, as $n$ goes to infinity, the depth value of a point $x$ relative to a finite sample converges almost surely to the comparable depth in the continuous case, relative to the underlying model, where the probability distribution is in $\mathbb{R}^d$: $lim_{n\to\infty} D_n(x) = D(x)$, where $D(x)$ is the theoretical data depth of $x$ (see Section 3) and $D_n(x)$ is the estimate of the depth of $x$ generated from sample $\{X_1, \ldots, X_n, x\}$ [24, 67]. Under certain conditions (for example elliptic distributions) and for most well behaved depth functions the depth contours for the finite sample case track the contours of the underlying model [40, 94, 95]

Our goal is to prove comparable results assuming that the underlying distribution is not linear. We will suggest depth functions to estimate depth values and depth contours for non-linear distributions from a finite sample $S$, that are based on the notion of *proximity graphs*. Proximity graphs, graphs in which points that are close to each other under some definition of closeness are connected [57], are the main technique used to capture the structure of a manifold $\Omega$ given a finite sample from $\Omega$ [86, 20, 19, 23, 36, 18]. Thus, the use of proximity graphs as a tool for depth computation is a natural extension of this approach. As part of this research we will define the required sampling conditions, under which the estimators approximate the underlying distribution well. Intuitively, if the underlying distribution lies on a manifold with a small feature size, then a better sampling is required in order to guarantee that the depth values relative to the sample set track the depth values of the underlying model.

Section 4.1 describes our suggested depth functions. Sections 4.2 and 4.3 provide background on proximity graphs and the reasons that proximity graphs are attractive as a tool for studying high-dimensional manifolds. Sections 4.4, 4.5 and 4.6 describe the main research directions: quantifying the required sampling conditions to achieve meaningful approximations of the continuous case; providing graphs for which the path length closely approximates geodesic distance; and suggesting algorithms for detection of boundary points in high dimensional manifolds.

### 4.1 Definition of Depth Relative to a Sample Set

Definitions 4.1 and 4.2 are the sample version of the definitions in Sections 3.1 and 3.2 respectively, and are achieved by estimating the *geodesic* distance by distance along graphs generated from the sample $S$. Definition 4.3 is a definition based on an existing depth measure (the proximity graph depth), that is easier to compute relative to the first two definitions and under certain conditions, approximates them.

**Definition 4.1. Sample Depth of a Manifold as Depth in Tangent Space:** *The sample [D] depth of a point $x$ with respect to a set $S = \{X_1, \ldots X_n, x\}$ is the sample [D] depth of the point $x$ with respect to the set $Log_{n,x}(S) = \{Log_{n,x}(X_1) \ldots Log_{n,x}(X_n)\}$, where [D] is any depth function defined in Euclidean space, $Log_{n,x}$ is the estimated values of the log map at point $x$ using estimates of the geodesic from the sample $S$ (see Appendix for more details on log map).*

At each point of the sample $x$, the estimated values of the *log map* $Log_{n,x}(x_i)$ can be found by embedding the set $S$ on a Euclidean space (estimated *tangent space*) using the matrix of geodesic distances.

**Definition 4.2. Generalized Sample Depth of a Manifolds Using $L_1$ Medians:** *The generalized median in a sample $S = \{X_1, \ldots X_n\}$ is a point $m \in S$ such that $\sum_i d_n(x_i, m) \leq \sum_i d_n(x_i, x), \forall x \in S$. The generalized sample $L_1$ depth of a point $x$ with respect to a set $S = \{X_1, \ldots X_n\}$ is $D(x) = \frac{\sum_i d_n(x_i,m)}{\sum_i d_n(x_i,x)}$.*

**Definition 4.3.** *The **Sample [Proximity graph] Depth** of a point $x$ relative to a point set $S = \{X_1 \dots X_n\}$ is the minimum path length in the [proximity graph] of $S$ that must be traversed in order to travel from $x$ to any point on the* boundary *of $S$.*

The *proximity graph depth* is a new class of depth functions [76, 77, 50], generalized from the concept of *Delaunay depth* [37, 1]. We have shown experimentally [76, 77, 50], that unlike most existing depth functions, the proximity graph depth can distinguish multimodal data sets. Note that in the original definition the proximity graph depth was defined using the minimal number of edges in a path from a point $x$ to the boundary. However, in order to approximate the generalized depth function for the continuous case we chose to use the path length rather than the path size. This will not add to the complexity of computing the depth of a point, but will guarantee certain properties of the depth function.

## 4.2 Proximity Graphs

*Proximity graphs* are graphs in which points close to each other by some definition of closeness are connected [57]. Examples include the *Delaunay triangulation* [34], the family of *$\beta$-skeletons* [61] which include as a special case the *Gabriel graph* [35] and the *relative neighborhood graph* [87], the *$k$-relative neighborhood graphs* [57], *rectangular influence graph* [52], *sphere of influence graph* [88], *$\gamma$-neighborhood graphs* [91], and others. See Figure 6 for examples of the Delaunay triangulation and Gabriel graph of a set of points.

The **Delaunay triangulation** is one of the most commonly used proximity graphs and has many application area. The Delaunay triangulation of a $d$-dimensional point set $S$ is the simplicial decomposition of the convex hull of $S$ such that the $d$-sphere defined by the points of every simplex in the decomposition contains no point $r \in S$ [34]. In a closed Riemannian manifold, the Delaunay triangulation of a set of sites exists and has properties similar to the Euclidean case, as long as the set of sites is sufficiently dense [66, 38].

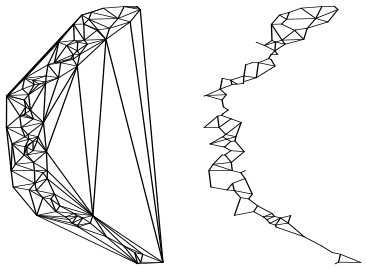


Figure 6: The Delaunay triangulation and Gabriel graph of a distribution following a banana shape. The Delaunay triangulation triangulates the region which is the convex hull of the set while the Gabriel graph is more indicative of the shape of the cloud. However, this may not be true for other point sets. This figure demonstrates that the boundary of a shape cannot be defined simply as the set of outer edges.

## 4.3 Manifold Learning

The **manifold learning problem** is the problem of computing a model of a $k$-dimensional manifold $\Omega$ embedded in $d$-dimensional Euclidean space $\mathbb{R}^d$ only from a finite set $S$ of sample points. The approximated manifold is usually represented as a triangulation or more generally simplicial complexes [28]. [6] This problem has applications in fields such as speech recognition [13] and neural networks. Most existing algorithms are based on the reconstruction of a *proximity graph* on the point set, and refinement of the resulting structure.

In the general *manifold learning problem* the manifold can be a collection of manifolds, not necessarily of the same dimension. In our problem, however, we assume that the probability distribution lies on a connected manifold, and thus we can assume that we are dealing with only one manifold.

Special cases of the manifold learning problem in low dimensions are *curve reconstruction* [29, 3] and *surface reconstruction* [4, 5, 6]. Shape reconstruction in the general case is a difficult problem, and some progress have been achieved in the study of this problem only recently. Current algorithms study restricted classes of non-linear manifolds and use varying types of proximity graphs to capture the shape of the manifold.

[6] A *triangulation* is a partition of a geometric domain into simplicies that meet only at shared faces. A *simplicial complex* $K$ is a finite set of simplicies such that (i) if $\sigma$ is a simplex of $K$ and $\tau$ is a face of $\sigma$ then $\tau$ is a simplex of $K$, and (ii) if $\sigma$ and $\tau$ are simplicies of $K$ then $\sigma \bigcup \tau$ is either empty or a connected face of $\tau$ and $\sigma$.

- *ISOMAP*[86] studies the class of non-linear manifolds which are isometric to a convex domain of Euclidean space (like the *swiss roll* example). It uses two types of proximity graphs, either they connect two points if they are at a distance of no more than some global threshold, or they use the $k$ nearest neighbor graph (a fixed number of nearest neighbors around each point). A variant of this algorithm *C-ISOMAP* [20, 19] was later described to recover mappings of a larger class of conformal embeddings provided that the original sampling density is reasonably uniform.
- *CoconeShape* [23] reconstructs a shape from a sample derived from a smooth manifold embedded in some Euclidean space $\mathbb{R}^d$ and is based on computation of the Delaunay triangulation. An improvement of the algorithm [18] can produce a triangulation $T$ interpolating $S$ such that $\Omega$ and the underlying space of $T$, denoted $|T|$, are homeomorphic and the Hausdorff distance between $\Omega$ and $|T|$ and the distance between their respective normal spaces are provably small. This algorithm requires strict sampling conditions and although theoretically important, is impractical. An algorithm based on the *adaptive neighborhood graph* [36] allows to approximate the geodesic distances between the points in $S$ provided a certain standard sampling condition is fulfilled. However, this algorithm is not guaranteed to produce an approximation that is topologically equivalent and geometrically close to a samples one.

To the best of our knowledge practical algorithms for high dimensional manifold reconstruction are rare (ISOMAP and its relatives). There is need for more practical tools for manifold learning, with guaranteed performance under certain conditions. Our application of data depth computation presents a relaxed manifold learning problem, as we do not require the computed model to be geometrically close to $\Omega$, but rather require that the path lengths be a good approximation of the geodesic distances. We plan to provide a practical implementation to solve this problem, based on the adaptive neighborhood graph, or to suggest a new algorithm, with guaranteed performance.

## 4.4 Boundary and Boundary Detection

For elliptically symmetric distributions, the *boundary* of $S$ is defined as the convex hull of $S$. However, for data lying on a manifold, this definition needs to be modified. For example, consider a banana shape, as in Figure 2, then the boundary of the banana consists of points that are not necessarily on the convex hull of the shape.

The problem of detecting boundary points, although may seem straightforward, is not simple, especially in dimensions higher than 2: although in two dimensions we can trace the edges forming the boundary of the shape (see for example Figure 6), in higher dimensions edges connecting pairs of points form only the skeleton of the shape and they cannot define any boundary to it.

A well known technique for boundary detection in two and three dimensions is based on $\alpha$-shapes [29, 30], a family of polytopes derived from the Delaunay triangulation of the point set. A similar notion was suggested by Veltkamp as $\gamma$-graphs for three dimensional point sets [91]. However, these algorithms are based on construction of the Delaunay triangulation which is computationally inefficient for high-dimensions. As part of our research, we will suggest methods of extracting boundary point, based on construction of a proximity graphs, which are more feasible to compute than Delaunay triangulations. A possible approach is to extend the $\alpha$-shapes methods for other types of proximity graphs. Note that we do not need to output the shape of the boundary, only boundary points. Therefore, it might be possible to propose a more efficient algorithm than existing methods (as a comparison, algorithms for computing the convex hull of a set of points have an exponential dependance on dimension, while algorithm for computing only the extreme points have linear dependance).

## 4.5 Path Length to Approximate Geodesic Distance

The proximity graph depth will approximate the generalized depth function provided that the path length between two points in the graph will approximate the geodesic distance. The dilation (or *spanning ratio*) of a graph defined on $n$ points is the maximal ratio over all pairs of data points $(x, y)$, of the minimum graph distance between $x$ and $y$, over the Euclidean distance between $x$ and $y$. Not all proximity graphs have bounded dilation: while Delaunay triangulation has bounded dilation of approximately 2.42 [58], it has been shown that the family of $\beta$-skeletons have unbounded dilation

[12, 31]. However, the algorithmic complexity of computing the Delaunay triangulation makes it impractical for use in higher dimensions. In addition, we require that the ratio is computed relative to the geodesic path and not the Euclidean path.

Therefore, we will need to investigate the type of proximity graphs that have bounded dilation for geodesic distances and what are the required sampling conditions to correctly approximate the geodesic distance from the graph. One possible candidate is the *adaptive neighborhood graph* [36] that allows to approximate the geodesic distances between the points in $S$ provided a certain standard sampling condition is fulfilled.

### 4.6 Sampling Conditions

Surface reconstruction algorithms assume that the input points satisfy a sampling condition. Assume that the manifold $\Omega$ is compact and smooth. The *medial axis* of $\Omega$ is the closure of the set of points in $\mathbb{R}^d$ that have more than one closet point in $\Omega$. *Local feature size* for $\Omega \subseteq \mathbb{R}^d$ is a continuous function $f : \Omega \to \mathbb{R}$ where $f(x)$ is the distance of $x \in \Omega$ to the medial axis of $\Omega$. Intuitively, $f$ measures how complicated $\Omega$ is in the neighborhood of $x$.

Some algorithms for manifold learning require *uniform sampling*, however, relaxed sampling conditions are possible. A sample $S$ of $\Omega$ is an $\epsilon$ **sample** if for each $x \in \Omega$ there is a sample point $p \in S$ such that $||p - x|| \leq \epsilon f(x)$ [3]. Dey et al. [23] showed that this condition is not always sufficient (for applications of dimension detection) and suggested a stricter sampling condition: A sample $S$ of $\Omega$ is $(\epsilon, \delta)$ **sampling** if for each $x \in \Omega$ there is a sample point $p \in S$ such that $||p - x|| \leq \epsilon f(x)$ and for any $p, q \in S$ $||p - q|| \geq \delta f(p)$ and $\epsilon/\delta$ is a constant. This sampling is very restricted, and is used in [18], but can potentially be relaxed if the manifold dimension is known in advance.

We will investigate the sampling conditions required to guarantee a good approximation of the depth values and depth contours.

### 4.7 Proposed Future Research Related to Depth from Sample

1. Find computationally tractable ways to generate proximity graphs with desirable properties.
2. Show that our sample data depth is a consistent estimate of the distribution data depth defined in Section 3.
3. Derive the distribution of our data depth estimates, and estimate bias and variance for small sample.
4. Study the effect of the sampling conditions on the quality of the results.
5. Suggest algorithms for detection of boundary points from a graph representing a manifold.

## 5 Experimentation and Computational Framework

The experimentation and computational aspects of this project span several directions. The first direction is *programming* our suggested algorithms, specifically the algorithms for computation of depth functions in the sample case, as suggested in Section 4. This work will require the translation of our theoretical definitions to efficient code; the design of appropriate data structures that will be easily extendable for large high dimensional data sets; and the testing of the correctness of the resulting programs, by designing appropriate data sets. The second direction is *generation of data* that will enable to test and validate, through these simulations, our methods and the reliability of our code. The third direction is to apply our algorithms to *real life data* to extract fresh insights.

We will begin by analyzing generated data sets in two and higher dimensions, using two main techniques. In one approach we will define varying types of probability distributions, create samples based on these distributions and compare the statistical attributes computed based on the sample set to the statistical attributes of the original set. In the second approach we will generate data sets from simple elliptical distribution and create data with nonlinear variations by transforming the original data set. This can be done, for example by using a parabolic function to create a banana shape or by defining parameters of variation and projecting the data using these parameters (see Figure 4). Our goal is that the median and, in most cases also the depth contours, in the generated sets will coincide with the median and the depth contours in the original set. For these generated sets the underlying

distribution is not known. Second, we will apply the tested algorithms to real life data sets and try to extract fresh insights from the data using our methods. Examples include growth curves shown in Section 2, face data [86] and speech recognition data.

We expect that in the preliminary phases of this research most code will be written in MATLAB [72], as it provides an easy platform to generate data sets and to quickly write and test code. Experimenting with large high-dimensional data sets or real life data will require faster algorithms and will therefore be coded in C++ or Java. We expect to use existing geometric algorithms as part of the LEDA [65] or CGAL [17] libraries. These software libraries are in common use in the geometry community and allow code to be built based on verified existing modules, shortening programming time, and improving the quality and flexibility of the code.

# 6 Summary of Proposed Research

Our goal is to propose novel data depth functions that would extend data depth concepts to describe variation of multivariate data (such as curves, images or shapes), when the space of variation is a manifold or the result of nonlinear variation in the data. We propose to generate statistical tools and efficient algorithms for the application of our method in data analysis.

As detailed in the preceding pages, the major tasks to pursue are as follows:

- Characterize the class of distributions $F$ for which our definition coincides with other classical definitions of data depth.
- Develop statistical analysis and quantification methods based on this definition.
- Suggest practical methods of estimating this depth function based on proximity graphs.
- Conduct an experimental study with varying constructions of point sets and apply it for analysis of real life data sets.

In the course of our research we expect collaboration opportunities with other researchers at MIT, Stanford and elsewhere.

# 7 Multidisciplinary Training

This project will impact training of future statisticians and computer scientists and will promote interdisciplinary collaboration between the two groups. A graduate student seminar on the topic of this proposal will be jointly offered by Principal Investigator Izem at Harvard and Principal Investigators Souvaine and Rafalin at Tufts to encourage statisticians and computer scientists to collaborate and share ideas and train graduate students in the area. Results of this project will be integrated into undergraduate classes as well. We believe that many statistical concepts we propose, quantifying center and spread of a distribution, are fairly intuitive and could be illustrated with simple toy examples in an introductory undergraduate class in statistics or more advanced classes such as analysis of variance and multivariate analysis. PI Izem has taught several undergraduate courses in introductory statistics and experimental design and has required students to conduct small research projects as assignments for the class. The data depth concepts discussed in this proposal will serve as a basis for some undergraduate research projects on data analysis or exploration. An exploratory computer-science course that incorporates real life problems was designed by Souvaine and taught by Souvaine and Rafalin and was successful in attracting undergraduate students to computer-science. The course will be expanded to include the results of this research. Souvaine is the PI of an NSF CSEMS project devoted to recruiting/retaining needy members of under-represented groups in computer science, engineering and mathematics, a program that exposes these undergraduates to research results in their first and second years.

# 8 Results of Prior Support

**NSF grant CCF-0431027:** "Impact of Computational Geometry on Depth-Based Statistics"
Period of support: August 15, 2004 - August 14, 2006 Amount: $ 170,402
Principal investigator: Diane L. Souvaine Postdoctoral Associate: Eynat Rafalin

Data depth is a statistical analysis method for multivariate data that requires no prior assumptions on the probability distribution of data and handles outliers. This ongoing project applies computational-geometry techniques to develop more efficient tools for data-depth analysis, including combinatorial and algorithmical results. Results to date include the following. A new depth measure, based on the minimum path length along proximity graph edges to the convex hull of a point set was suggested and evaluated theoretically and experimentally [50, 76, 77]. A novel approach for analyzing depth in multi-modal data sets was presented including experimental results demonstrating that, unlike most depth functions, the proximity graph depth can distinguish multimodal data sets. An interactive software tool for the generation of data sets and visualization of the performance of multiple depth measures was developed to provide experimental evaluation, recognize problems with the existing measures, and to suggest modifications [50, 51]. Two competing notions of depth contours were identified and efficient algorithms for dynamic computation of both versions of half-space depth contours were provided [15, 75]. A new definition of simplicial depth was proposed which remains valid a continuous probability field, but fixes flaws incurred by the current definition in the finite sample case. Tight bounds on the value of the simplicial depth based on the half-space depth were proved, and the computational complexity of the new definition was evaluated [14]. A novel approach for topologically sweeping the complete graph, that is used to compute the simplicial depth median, was presented [79, 75].

# 9 Appendix - Geometry of the Support of a Distribution

This section includes the notation, terminology, and definitions we need in Section 3 to define data depth functions for distributions $F$. Needed concepts are *support of a distribution*, *Riemannian manifold*, and *geodesic distance*. Good resources for an introduction to differential geometry are [8], [62], and [2].

**Support of a distribution:** *The support $\Omega$ of a distribution $F$* is the space in which the distribution lives. For example, the support of the uniform distribution on a compact set $K$ in $\mathbb{R}^d$ is the set $K$. On the other hand, the support of the multivariate normal distribution in $\mathbb{R}^d$ is $\mathbb{R}^d$. More formally, the support $\Omega$ of a distribution $F$ is the intersection of all closed sets in $\mathbb{R}^d$ with $F$ measure 1. The data depth function defined in Subsection 3.1 have distributions with support in smooth spaces in which the shortest distance of paths, along the space, between two points is well defined. This condition is satisfied if the support spaces are complete *Riemannian manifolds*, and the distances that interest us will be geodesic distances.

**k-dimensional differentiable manifold:** A *$k$-dimensional manifold* is locally homeomorphic to $\mathbb{R}^k$, and we call the local homeomorphisms charts. For the manifold to be smooth, we need these charts to be smooth and to patch up together smoothly as well ( to be $C^\infty$ related ). More formally, *$k$-dimensional differentiable manifold* is a set $M$ together with a family $(M_i)_{i\in I}$ of subsets such that: (i) $M = \cup_{i\in I} M_i$; (ii) [Local homeomorphisms] For every $i \in I$, there is an injective map $\phi_i : M_i \in \mathbb{R}^k$ so that $\phi_i(M_i)$ is open in $\mathbb{R}^k$ and $C^\infty$; and (iii) [$C^\infty$ related charts] For $M_i \cap M_j \neq \emptyset$, $\phi_i(M_i \cap M_j)$ is open in $\mathbb{R}^k$, and the composition $\phi_j o \phi_i^{-1} : \phi_i(M_i \cap M_j) \to \phi_j(M_i \cap M_j)$ is $C^\infty$ for arbitrary $i$ and $j$.
Each $\phi_i$ is called a *chart*, and $(M_i, \phi_i)_i$ is called an *atlas*. In addition, we suppose that manifolds satisfy the Haussdorf separation axiom (i.e. every two points $p$ and $q$ in $M$ have disjoint open neighborhoods $U_p$ and $U_q$ in $M$).

**Tangent space:** Given a manifold $M$ in $\mathbb{R}^d$, we can associate a linear subspace of $\mathbb{R}^d$ at each point $p$ of $M$ called the *tangent space*. A tangent space $T_pM$ of $M$ at $p$ is the set of all tangent vectors of $M$ at $p$. A tangent vector at $p$ can be thought of as a directional derivative, so that tangent vectors are tangents to curves lying on the manifold. Formally, a tangent vector at $p$ is an equivalence class of differentiable curves or paths $c : (-\epsilon, \epsilon) \to M$ with $c(0) = p$, where $c_1 \sim c_2 \Leftrightarrow \frac{d}{dt}(\phi o c_1)(0) = \frac{d}{dt}(\phi o c_2)(0)$ for every chart $\phi : M \to \mathbb{R}$ containing $p$. The following Theorem [62] gives the dimensionality of the tangent space $T_pM$ (which is the same as the manifold $M$) as well as a coordinate system in that linear subspace.

**Theorem 9.1.** *The tangent space at $p$ on a $k$-dimensional differentiable manifold is an $k$-dimensional $\mathbb{R}$-vector space and is spanned in appropriately chosen coordinates $x^i, \ldots, x^k$ in a given chart by $\frac{\partial}{\partial x^1}|_p, \ldots, \frac{\partial}{\partial x^k}|_p$, where $\frac{\partial}{\partial x^i}|_p(f) := \frac{\partial (f o \phi^{-1})}{\partial u^i}|_{\phi(p)}, \forall$ differentiable functions $f$ in a chart $\phi$ which contains $p$.*

*For every tangent vector $V$ at $p$, one has* $V = \sum_{i=1}^{n} V(x^i) \frac{\partial}{\partial x^i}|_p$. A *vector field* in the manifold $M$ is a function that smoothly assigns to each point $p$ in $M$ a tangent vector $V$ in $T_pM$.

**Riemannian Geometry:** *Riemannian geometry* will provide the context for an inner product in the manifold. Note that in $\mathbb{R}^k$, we can consider that the usual Euclidean distance is derived from the definition of inner product, so that $||x|| = \sqrt{< x, x >}$. Similarly, a *Riemannian metric* is defined with respect to an inner product in the tangent space. Let $L^2(T_pM; \mathbb{R}) = (T_pM)^*$ be the dual space of $T_pM$, i.e. the space of all bilinear forms from $T_pM \text{x} T_pM$ to $\mathbb{R}$. Let $\{dx^i|_p \otimes dx^j|_p, i, j = 1, \ldots, k\}$ be the dual basis in the dual space $L^2(T_pM; \mathbb{R})$, then a *Riemannian metric* $g$ on $M$ is an association from $g : p \in M \mapsto g_p \in L^2(T_pM; \mathbb{R})$, where $g_p$ defines at every point $p$ an inner product on the tangent space $T_pM$, i.e $g_p$ satisfies the following conditions: (i) Symmetry. $g_p(X, Y) = g_p(Y, X)$ for all $X, Y$; (ii) $g_p$ is positive definiteness. $g_p(X, X) > 0$ for all $X \neq 0$; (iii) Differentiability. $g_p = \sum_{i,j} g_{i,j}(p) dx^i|_p \otimes dx^j|p$ where the coefficients $g_{i,j}$ are differentiable functions. The pair $(M, g)$ is then called a *Riemannian manifold*, and one refers to the *Riemannian metric* as the metric tensor.

**Geodesics:** In Euclidean spaces, the shortest path between two points is a straight line, and the distance between the points is measured as the length of that straight line segment. In Section 3.1, we define the distance between two points in *the support of a distribution* $\Omega$ as the distance of the shortest path between two points along $\Omega$. This quantity is guaranteed to exist and be well defined for complete and connected *Riemannian manifolds*, and the shortest path is called geodesic. If $c : [a, b] \to M$ is a differentiable curve on a *Riemannian manifold* $M$ with endpoints $c(a) = x$ and $c(b) = y$. The length of $c$ can be defined as $L(c) = \int_a^b ||c'(t)||dt$ where $c'(t)$ is a vector in the tangent space $T_{c(t)}M$, reparameterized proportional to arclength, and the norm is given by the *Riemannian metric* at $c(t)$.
The idea of a global minimum of length leads to a definition of distance, not to be confused with the *Riemannian metric*. It is defined as $d(x, y) = \inf\{L(c) : c \text{ differentiable curve between } x \text{ and } y\}$. Any differentiable path $c$ with minimum length $L(c)$ is called *geodesic*[7]. *Geodesics* are guaranteed to exist for complete, connected *Riemannian manifolds*, as described in the Hopf-Rinow Theorem [62].

**Exponential and log map:** When the geodesic $c$ exists in the interval $[0, 1]$, such that $c(0) = p$ and $c'(0) = v$, the Riemannian *exponential map* at the point $p$ of a manifold, denoted $Exp_p : T_pM \to M$, is defined as $Exp_p(v) = c(1)$. If $M$ is a complete manifold, the *exponential map* is defined for all vectors $v \in T_pM$, and the mapping $Exp_p$ is a diffeomorphism in some neighborhood $U \subset T_pM$ containing 0. The inverse of the exponential map is called the *log map* and denoted by $Log_p : Exp_p(U) \to T_pM$. The Riemannian *log map* $Log_p$ is an isometry, it preserves the length of geodesic curves from $p$ to other points in the manifold.

[7]Geodesics may not be unique, in the case of the sphere, there are infinitely many geodesics between two points. On the other hand, geodesics may not exist, for example for the plane without the origin (i.e. $\mathbb{R}^2 - \{(0, 0)\}$) there is no geodesic between the two points $(-1, 0)$ and $(1, 0)$.

# References


[1] M. Abellanas, M. Claverol, and F. Hurtado. Point set stratification and Delaunay depth, 2005. ACM Computing Research Repository, cs.CG/0505017.

[2] S.-I. Amari, O. E. Barndorff-Nielsen, R. E. Kass, S. L. Lauritzen, and C. R. Rao. *Differential geometry in statistical inference*. Institute of Mathematical Statistics Lecture Notes—Monograph Series, 10. Institute of Mathematical Statistics, Hayward, CA, 1987.

[3] N. Amenta, M. Bern, and D. Eppstein. The crust and the $\beta$-skeleton: Combinatorial curve reconstruction. *Graphical models and image processing: GMIP*, 60(2):125–135, 1998.

[4] N. Amenta, M. Bern, and M. Kamvysselis. A new Voronoi-based surface reconstruction algorithm. In *SIGGRAPH '98: Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, pages 415–421, New York, NY, USA, 1998. ACM Press.

[5] N. Amenta, S. Choi, T. K. Dey, and N. Leekha. A simple algorithm for homeomorphic surface reconstruction. *Internat. J. Comput. Geom. Appl.*, 12(1-2):125–141, 2002.

[6] N. Amenta, S. Choi, and R. K. Kolluri. The power crust, unions of balls, and the medial axis transform. *Comput. Geom.*, 19(2-3):127–153, 2001. Combinatorial curves and surfaces.

[7] T. W. Anderson. *An introduction to multivariate statistical analysis*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, third edition, 2003.

[8] L. Auslander and R. E. MacKenzie. *Introduction to differentiable manifolds*. Dover Publications Inc., New York, 1977. Corrected reprinting.

[9] V. Barnett. The ordering of multivariate data. *J. Roy. Statist. Soc. Ser. A*, 139(3):318–355, 1976.

[10] A. J. S. Bernhard Schlkopf. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press, 2002.

[11] R. Bhattacharya and V. Patrangenaru. Large sample theory of intrinsic and extrinsic sample means on manifolds. *Annals of Statistics*, 31:1–29, 2003.

[12] P. Bose, L. Devroye, W. Evans, and D. Kirkpatrick. On the spanning ratio of Gabriel graphs and $\beta$-skeletons. In *LATIN 2002: Theoretical informatics (Cancun)*, volume 2286 of *Lecture Notes in Comput. Sci.*, pages 479–493. Springer, Berlin, 2002.

[13] C. Bregler and S. M. Omohundro. Nonlinear manifold learning for visual speech recognition. In *Proc. of 5th International Conference on Computer Vision*, pages 494– 499, 1995.

[14] M. Burr, E. Rafalin, and D. L. Souvaine. Simplicial depth: An improved definition, analysis, and efficiency for the sample case. DIMACS technical report 2003-28, DIMACS, `http://dimacs.rutgers.edu/TechnicalReports/ abstracts/2003/2003-28.html` , 2003. Also appeared as *TUFTS-CS Technical Report 2003-01*, Tufts University, September, 2003 and in *Proc. 15th Canad. Conf. Comput. Geom., 2004*.

[15] M. Burr, E. Rafalin, and D. L. Souvaine. Dynamic update of half-space depth contours. *Tufts CS Technical Report 2006-1*, Tufts University, 2006. Abstract appeared in *Proceedings of the 14th Annual Fall Workshop on Computational Geometry*, MIT, 2004, pages 3-4.

[16] M. Burr, E. Rafalin, and D. L. Souvaine. Simplicial depth: An improved definition, analysis and efficiency for the sample case. In R. Liu, editor, *Data Depth: Robust Multivariate Analysis, Computational Geometry and Applications*, DIMACS book series. AMS, 2006.

[17] CGAL. Computational geometry algorithms library. `www.cgal.org`.

[18] S.-W. Cheng, T. K. Dey, and E. A. Ramos. Manifold reconstruction from point samples. In *Proc. 16th Symp. Discrete Algorithms*, pages 1018–1027. ACM and SIAM, Jan 2005.

[19] V. de Silva and J. B. Tenenbaum. Local versus global methods for nonlinear dimensionality reduction. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems*, volume 15, pages 705–712. MIT Press, Cambridge, 2003.

[20] V. de Silva and J. B. Tenenbaum. Unsupervised learning of curved manifolds. In *Nonlinear estimation and classification (Berkeley, CA, 2001)*, volume 171 of *Lecture Notes in Statist.*, pages 453–465. Springer, New York, 2003.

[21] P. Delicado. Another look at principal curves and surfaces. *J. Multivariate Anal.*, 77(1):84–116, 2001.

[22] P. Delicado and M. Huerta. Principal curves of oriented points: theoretical and computational improvements. *Comput. Statist.*, 18(2):293–315, 2003. Euroworkshop on Statistical Modelling (Bernried, 2001).

[23] T. K. Dey, J. Giesen, S. Goswami, and W. Zhao. Shape dimension and approximation from samples. *Discrete Comput. Geom.*, 29(3):419–434, 2003.

[24] D. L. Donoho and M. Gasko. Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *Ann. Statist.*, 20(4):1803–1827, 1992.

[25] D. L. Donoho and C. Grimes. Hessian eigenmaps: locally linear embedding techniques for high-dimensional data. *Proc. Natl. Acad. Sci. USA*, 100(10):5591–5596 (electronic), 2003.

[26] I. L. Dryden and K. V. Mardia. *Statistical shape analysis*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons Ltd., Chichester, 1998.

[27] W. Eddy. Convex hull peeling. In H. Caussinus, editor, *COMPSTAT*, pages 42–47. Physica-Verlag, Wien, 1982.

[28] H. Edelsbrunner. Modeling with simplicial complexes (topology, geometry, and algorithms). In *Proc. 6th Canadian Conf. Comput. Geom.*, pages 36–44, 1994.

[29] H. Edelsbrunner, D. G. Kirkpatrick, and R. Seidel. On the shape of a set of points in the plane. *IEEE Trans. Inform. Theory*, 29(4):551–559, 1983.

[30] H. Edelsbrunner and E. P. Mucke. Three-dimensional alpha shapes. *ACM Trans. Graphics*, 13:43–72, 1994.

[31] D. Eppstein. Beta-skeletons have unbounded dilation. *Comput. Geom.*, 23(1):43–52, 2002.

[32] P. Fletcher. *Statistical Variability in Nonlinear Spaces: Application to Shape Analysis and DT-MRI*. Ph.D. thesis, Dept. Comput. Sci., University of North Carolina at Chapel Hill, Chapel Hill, NC, Aug. 2004.

[33] P. Fletcher, C. Lu, S. Pizer, and S. Joshi. Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE TRANSACTIONS ON MEDICAL IMAGING*, 23(8):995–1005, 2004.

[34] S. Fortune. Voronoi diagrams and Delaunay triangulations. In *Handbook of discrete and computational geometry*, CRC Press Ser. Discrete Math. Appl., pages 377–388. CRC Press, Inc., Boca Raton, FL, USA, 1997.

[35] K. Gabriel and R. Sokal. A new statistical approach to geographic variation analysis. *Systematic Zoology*, 18:259–278, 1969.

[36] J. Giesen and U. Wagner. Shape dimension and intrinsic metric from samples of manifolds. *Discrete Comput. Geom.*, 32(2):245–267, 2004.

[37] P. J. Green and R. Sibson. Computing Dirichlet tessellations in the plane. *The Computer Journal*, 21(2):168–173, 1978.

[38] C. I. Grima and A. Márquez. *Computational geometry on surfaces - Performing computational geometry on the cylinder, the sphere, the torus, and the cone*. Kluwer Academic Publishers, Dordrecht, 2001.

[39] T. Hastie and W. Stuetzle. Principal curves. *J. Amer. Statist. Assoc.*, 84(406):502–516, 1989.

[40] X. He and G. Wang. Convergence of depth contours for multivariate datasets. *Ann. Statist.*, 25(2):495–504, 1997.

[41] H. Herrera. Gromov invariants of $S^2$-bundles over 4-manifolds. In *Proceedings of the 1999 Georgia Topology Conference (Athens, GA)*, volume 124, pages 327–345, 2002.

[42] H. Herrera and R. Herrera. Classification of positive quaternion-Kähler 12-manifolds. *C. R. Math. Acad. Sci. Paris*, 334(1):43–46, 2002.

[43] H. Herrera and R. Herrera. $\hat{A}$-genus on non-spin manifolds with $S^1$ actions and the classification of positive quaternion-Kähler 12-manifolds. *J. Differential Geom.*, 61(3):341–364, 2002.

[44] H. Herrera and R. Herrera. A result on the $\widehat{A}$ and elliptic genera on non-spin manifolds with circle actions. *C. R. Math. Acad. Sci. Paris*, 335(4):371–374, 2002.

[45] H. Herrera and R. Herrera. Generalized elliptic genus and cobordism class of nonspin real Grassmannians. *Ann. Global Anal. Geom.*, 24(4):323–335, 2003.

[46] H. Herrera and R. Herrera. Elliptic genera on non-spin Riemannian symmetric spaces with $b_2 = 0$. *J. Geom. Phys.*, 49(2):197–205, 2004.

[47] H. Herrera and R. Herrera. The signature and the elliptic genus of $\pi_2$-finite manifolds with circle actions. *Topology Appl.*, 136(1-3):251–259, 2004.

[48] T. P. Hettmansperger. *Statistical inference based on ranks*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, 1984.

[49] J. Hodges. A bivariate sign test. *The Annals of Mathematical Statistics*, 26:523–527, 1955.

[50] J. Hugg, E. Rafalin, K. Seyboth, and D. Souvaine. An experimental study of old and new depth measures. In *Proceedings of the 8th International Workshop on Algorithm Engineering and Experiments (ALENEX)*, 2006.

[51] J. Hugg, E. Rafalin, and D. Souvaine. Depth Explorer - a software tool for analysis of depth measures. *Tufts CS Technical Report 2005-6*, Tufts University, Nov. 2005. Abstract appeared in *Proceedings of the 15th Annual Fall Workshop on Computational Geometry*, UPenn, 2005, pages 9-10.

[52] M. Ichino and J. Sklansky. The relative neighborhood graph for mixed feature variables. *Pattern Recognition*, 18(2):161–167, 1985.

[53] R. Izem. *Analysis of Nonlinear Variation in Functional Data*. Ph.D. thesis, Dept. Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC, May 2004.

[54] R. Izem and J. G. Kingsolver. Variation in continuous reaction norms: quantifying directions of biological interest. *The American Naturalist*, 166(8):277–289, 2005.

[55] R. Izem and J. S. Marron. Quantifying nonlinear modes of variation of curves. In *Proceedings of the International Statistical Institute 54th session [CD-ROM], August 13-20. Berlin, Germany.*, 2003. Invited talk.

[56] R. Izem, J. S. Marron, and J. G. Kingsolver. Analyzing nonlinear variation of thermal performance curves. In *Proceedings of the American Statistical Association,[CD-ROM], August 8-12, Toronto, Canada.*, 2005. winner of the Best Student paper Award in the Graphics and Computing Section of the ASA.

[57] J. W. Jaromczyk and G. T. Toussaint. Relative neighborhood graphs and their relatives. *Proc. IEEE*, 80(9):1502–1517, sep 1992.

[58] J. M. Keil and C. A. Gutwin. Classes of graphs which approximate the complete Euclidean graph. *Discrete Comput. Geom.*, 7(1):13–28, 1992.

[59] J. Kingsolver, R. Gomulkiewicz, and P. Carter. Variation, selection and evolution of function-valued traits. *Genetica*, 112-113:87–104, 2001.

[60] J. G. Kingsolver, R. Izem, and G. Ragland. Plasticity of size and growth in fluctuating thermal environments: Comparing reaction norms and performance curves. *Integrative and Comparative Biology*, 44(6):450–460, 2004.

[61] D. G. Kirkpatrick and J. D. Radke. A framework for computational morphology. In G. Toussaint, editor, *Computational geometry*, pages 217–248. North-Holland, 1985.

[62] W. Kühnel. *Differential geometry*, volume 16 of *Student Mathematical Library*. American Mathematical Society, Providence, RI, 2002. Curves—surfaces—manifolds, Translated from the 1999 German original by Bruce Hunt.

[63] A. Kume and H. Le. Estimating Fréchet means in Bookstein's shape space. *Adv. in Appl. Probab.*, 32(3):663–674, 2000.

[64] H. Le. Locating Fréchet means with application to shape spaces. *Adv. in Appl. Probab.*, 33(2):324–338, 2001.

[65] LEDA. Library of efficient data structures and algorithms. `http://www.algorithmic-solutions.com/enleda.htm`.

[66] G. Leibon and D. Letscher. Delaunay triangulations and Voronoi diagrams for Riemannian manifolds. In *SCG '00: Proceedings of the sixteenth annual symposium on Computational geometry*, pages 341–349, New York, NY, USA, 2000. ACM Press.

[67] R. Liu. On a notion of data depth based on random simplices. *The Annals of Statistics*, 18:405–414, 1990.

[68] R. Liu. Data depth: center-outward ordering of multivariate data and nonparametric multivariate statistics. In M. Akritas and D. Politis, editors, *Recent Advances and Trends in Nonparametric Statistics*, pages 155–168. Elsevier Science, 2003.

[69] R. Liu, J. Parelius, and K. Singh. Multivariate analysis by data depth: descriptive statistics, graphics and inference. *The Annals of Statistics*, 27:783–858, 1999.

[70] F. M. Les éléments aléatoires de nature quelconque dans un espace distancié. *Ann. Inst. H. Poincaré*, 10:215–310, 1948.

[71] P. C. Mahalanobis. On the generalized distance in statistics. In *Proceedings of the National Academy India*, volume 12, pages 49–55, 1936.

[72] T. MathWorks. Matlab (r). `www.mathworks.com`.

[73] K. Miller, S. Ramaswami, P. Rousseeuw, T. Sellarés, D. Souvaine, I. Streinu, and A. Struyf. Efficient computation of location depth contours by methods of combinatorial geometry. *Statistics and Computing*, 13(2):153–162, 2003.

[74] H. Oja. Descriptive statistics for multivariate distributions. *Statistics and Probability Letters*, 1:327–332, 1983.

[75] E. Rafalin. *Algorithms and Analysis of Depth Functions using Computational Geometry*. Ph.D. thesis, Dept. Comput. Sci., Tufts University, Medford, MA, May 2005.

[76] E. Rafalin, K. Seyboth, and D. Souvaine. Path length in proximity graphs as a data depth measure. *Tufts CS Technical Report 2005-5*, Tufts University, Nov. 2005. Abstract appeared in *Proceedings of the 15th Annual Fall Workshop on Computational Geometry*, UPenn, 2005, pages 11-12.

[77] E. Rafalin, K. Seyboth, and D. Souvaine. Proximity graph depth as a new analysis tool of multimodal high-dimensional data, 2006. Submitted for publication.

[78] E. Rafalin and D. Souvaine. Computational geometry and statistical depth measures. In M. Hubert, G. Pison, A. Struyf, and S. V. Aelst, editors, *Theory and Applications of Recent Robust Methods*, Statistics for Industry and Technology, pages 283–292. Birkhauser, Basel, 2004.

[79] E. Rafalin and D. Souvaine. Topologically sweeping the complete graph, 2006. Submitted for publication.

[80] J. O. Ramsay and B. W. Silverman. *Functional data analysis: methods and case studies*. Springer Series in Statistics. Springer, 2002.

[81] J. O. Ramsay and B. W. Silverman. *Functional data analysis, 2nd edition*. Springer Series in Statistics. Springer, New York, second edition, 2005.

[82] P. J. Rousseeuw. Introduction to positive-breakdown methods. In *Robust Inference*, volume 15 of *Handbook of Statist.*, pages 101–121. North-Holland, Amsterdam, 1997.

[83] P. J. Rousseeuw. Introduction to positive-breakdown methods. In J. E. Goodman and J. O'Rourke, editors, *Handbook of discrete and computational geometry*, Discrete Mathematics and its Applications (Boca Raton), pages xviii+1539. Chapman & Hall/CRC, Boca Raton, FL, second edition, 2004.

[84] B. W. Silverman. Incorporating parametric effects into functional principal components analysis. *J. Roy. Statist. Soc. Ser. B*, 57(4):673–689, 1995.

[85] T. Tarpey and B. Flury. Self-consistency: a fundamental concept in statistics. *Statist. Sci.*, 11(3):229–243, 1996.

[86] J. B. Tenenbaum, V. d. Silva, and J. C. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500):2319–2323, 2000.

[87] G. Toussaint. The relative neighborhood graph of a finite planar set. *Pattern Recognition*, 12:261–268, 1980.

[88] G. T. Toussaint. A graph theoretical primal sketch. In G. T. Toussaint, editor, *Computational morphology*, pages 229–260. North-Holland, 1988.

[89] J. W. Tukey. Mathematics and the picturing of data. In *Proc. of the Int. Cong. of Math. (Vancouver, B. C., 1974), Vol. 2*, pages 523–531. Canad. Math. Congress, Montreal, Que., 1975.

[90] D. E. Tyler. Finite sample breakdown points of projection based multivariate location and scatter statistics. *Ann. Statist.*, 22(2):1024–1044, 1994.

[91] R. C. Veltkamp. The $\gamma$-neighborhood graph. *Comput. Geom.*, 1(4):227–246, 1992.

[92] K. Wang and T. Gasser. Alignment of curves by dynamic time warping. *Ann. Statist.*, 25(3):1251–1276, 1997.

[93] K. Wang and T. Gasser. Synchronizing sample curves nonparametrically. *Ann. Statist.*, 27(2):439–460, 1999.

[94] Y. Zuo and R. Serfling. General notions of statistical depth function. *Ann. Statist.*, 28(2):461–482, 2000.

[95] Y. Zuo and R. Serfling. Structural properties and convergence results for contours of sample statistical depth functions. *Ann. Statist.*, 28(2):483–499, 2000.