PCA-Based Out-of-Sample Extension for Dimensionality Reduction

Yariv Aizenbud Amit Bermanis Amir Averbuch

November 5, 2013

Abstract

Dimensionality reduction methods are very common in the field of high dimensional data analysis, where the classical analysis methods are inadequate. Typically, algorithms for dimensionality reduction are computationally expensive. Therefore, their applications to process data warehouses are impractical. It is visible even more when the data is accumulated non-stop. In this paper, an out-of-sample extension scheme for dimensionality reduction is presented. We propose an algorithm which performs an out-of-sample extension to newly-arrived multidimensional data points. Unlike other extension algorithms, such as the Nyström algorithm, the proposed algorithm uses the intrinsic geometry of the data and the properties of dimensionality reduction map. We prove that the error of the proposed algorithm is bounded. Additionally to the out-of-sample extension, the algorithm provides a residual for any new data point that tells us the abnormality degree of this data point.

1 Introduction

Analysis of high-dimensional data warehouses is of great interest since it may illuminate the underlying phenomena. To cope with big high-dimensional data, it is often assumed that there are some (unobservable) dependencies between the parameters (sensed, streamed, computed, achieved) of the multidimensional data points (we will call them data points). Mathematically, it means that the data is sampled from a low-dimensional manifold that is embedded in a high dimensional ambient space. Dimensionality reduction

methods, which rely on the presence of a manifold, map the data into a low-dimensional space while preserving certain properties from the data. Particularly, these methods preserve local linear (or nearly linear) structures of the data in addition to other properties.

Kernel-based methods characterize a broad class of dimensionality reduction methods. The kernel encapsulates a measure of mutual affinities (or similarities) between data points. Spectral analysis of the associated kernel matrix via singular values decomposition (SVD) obtains an embedding of the data into an Euclidean space, where hidden structures become more visible. The dimensionality of the embedding space is affected by the spectrum's decay rate. Two classical kernel-based methods for dimensionality reduction are principal component analysis (PCA) [11, 12] and multidimensional scaling (MDS) [7, 14]. PCA projects the data points onto a space spanned by the significant singular vectors of the data's covariance matrix. The MDS does the same but it uses the Gram matrix of the inner products between data points. Both embeddings preserve significant directions of change in the data, i.e. its directional variances. Other kernel methods use the same mechanism with different kernels (Gram matrices). Examples of kernel methods are diffusion maps (DM) [5], local linear embedding (LLE) [19], Laplacian eigenmaps [2], Hessian eigenmaps [10] and local tangent space alignment [22, 23].

From a practical point of view, kernel methods have a significant computational drawback: spectral analysis of the kernel matrix becomes impractical for large datasets due to high computational complexity required to manipulate a kernel matrix. They are also global which is disadvantage. Furthermore, in many applications, the process of data analysis is dynamic where the data accumulates over time and the embedding has to be modified once in a while to produce a new kernel matrix that has to be analyzed. Additionally, processing kernel matrix in memory becomes impractical when the datasets are huge. Another obstacle for achieving efficient computational methods is the fact that spectral methods are global. A general solution scheme that embeds a subset of the source data is usually referred to as a training dataset. Then, the embedding is extended to any new out-of-sample data point. The Nyström method [1, 9, 17], which is widely used in integral equations solvers, has become very popular as an out-of-sample extension method for dimensionality reduction. For a review of spectral clustering and Nyström extension see Section 2 in [20]. Geometric Harmonics (GH) [6] is another out-of-sample extension method. It uses the Nyström extension of eigenfunctions of the kernel defined on the data. In order to avoid numerical instabilities, it uses only the significant eigenvalues,

therefore, inconsistencies with the in-sample data might occur. This problem, additionally to the fixed interpolation distance problem, is treated in [18], where a multiscale interpolation scheme is introduced. Another multiscale approach that aims to solve the aforementioned limitations was recently introduced in [3].

All these methods use a kernel matrix (or, perhaps, its low rank approximation) as an interpolation matrix. This mechanism is strongly related to a variety of isotropic interpolation methods that employ radial basis functions (RBF). Such methods are used for scattered data approximation, where the data lies in a metric space. More details about RBF and scattered data approximation can be found in [4] and [21], respectively.

In this paper, we employ the manifold assumption to establish an anisotropic out-of-sample extension. We suggest a new anisotropic interpolation scheme that assigns for each data point a likelihood neighborhood. This likelihood is based on geometric features of the dimensionality reduction map by using a local Principal Component Analysis (PCA) of the map's image. Incorporation of such neighborhood information produces a linear system for finding the out-of-sample extension for this data point. This method also provides an abnormality measure for the newly-mapped data point.

The paper has the following structure: Section 2 introduces the problem and the needed definitions. Section 3 establishes the geometric-based stochastic linear system, on which the interpolant is based. Three different interpolants are presented, each considers different geometric considerations. In Section 4, an analysis of the interpolation's error is presented for the case of Lipschitz mappings. Computational complexity analysis of the scheme is presented in Section 5, and experimental results for both synthetic data and real-life data are presented in Section 6.

2 Problem Setup

Let $\mathcal{M}$ be a compact low-dimensional manifold of intrinsic dimension m that lies in a high-dimensional ambient space $\mathbb{R}^n$ $(m < n)$, whose Euclidean metric is denoted by $\|\cdot\|$. Let ψ be a smooth, Lipschitz dimensionality reducing function defined on $\mathcal{M}$, i.e. $\psi : \mathcal{M} \to \mathcal{N} \subset \mathbb{R}^d$ $(m < d < n)$, where $\mathcal{N}$ is a m-dimensional manifold. Let $M = \{x_1, \ldots, x_p\} \subset \mathcal{M}$ be a finite training dataset, sufficiently dense sampled from $\mathcal{M}$, whose image under ψ, $\psi(M) = \{\psi(x_1), \ldots, \psi(x_p)\}$, was already computed. Given an out-of-sample data point $x \in \mathcal{M} \backslash M$, we aim to embed it into $\mathbb{R}^d$, while preserving some local

properties of ψ. The embedding of x into $\mathbb{R}^d$, denoted by $\hat{\psi}(x)$, is referred to as the extension of ψ to x.

The proposed extension scheme is based on local geometric properties of ψ in the neighborhood of x, denoted by $N(x)$. Specifically, the influence of a neighbor $x_j \in N(x)$ on the value of $\hat{\psi}(x)$ depends on its distance from x and on the geometry of the image $\psi(N(x))$ of $N(x)$ under ψ. This approach is implemented by considering $\hat{\psi}(x)$ as a random variable with mean $\mathbb{E}\hat{\psi}(x) = \psi(x_j)$ and a variance $\mathbb{V}\hat{\psi}(x)$ that depends both on the distance of x from x_j and on some geometric properties of $\psi(N(x))$ that will be discussed in details in Section 3.2. Mathematically,

$$\hat{\psi}(x) = \psi(x_j) + \omega_j, \tag{2.1}$$

where ω_j is a random variable with mean $\mathbb{E}\omega_j = 0$ and variance $\mathbb{V}\omega_j = \sigma_j$ that, as previously mentioned, depends on the local geometry of ψ in the neighborhood of x. Thus, we get $|N(x)|$ equations for $\hat{\psi}(x)$, one for each neighbor $x_j \in N(x)$. The optimal solution is achieved by the generalized least squares approach described in Section 2.1.

2.1 Generalized Least Squares (GLS)

In this section, we briefly describe the GLS approach that will be utilized to evaluate $\hat{\psi}(x)$. In general, the GLS addresses the problem of a linear regression that assumes either independence or common variance of the random variables. Thus, if $y = (y_1, \ldots, y_k)^T$ are random variables, corresponding to k data points in $\mathbb{R}^d$, the addressed regression problem is

$$y = X\beta + \mu, \tag{2.2}$$

where X is an $k \times d$ matrix that stores the data points as its rows, and $\mu \in \mathbb{R}^k$ is an error vector. Respectively to the aforementioned assumption, the $k \times k$ conditional covariance matrix of the error term $W = \mathbb{V}\{\mu|X\}$ is not necessarily scalar or diagonal. The GLS solution to Eq. 2.2 is

$$\hat{\beta} = (X^T W^{-1} X)^{-1} X^T W^{-1} y. \tag{2.3}$$

The Mahalanobis distance between two random vectors v_1 and v_2 of the same distribution with conditional covariance matrix W is

$$\|v_1 - v_2\|_W \triangleq \sqrt{(v_1 - v_2)^T W^{-1} (v_1 - v_2)}. \tag{2.4}$$

Observation 2.1. *The Mahalanobis distance in Eq. 2.4 measures the similarity between v_1 and v_2 in respect to W. If the random variables are independent, then W is diagonal. Then, it is more affected by low variance random variables and less by high variance ones.*

The GLS solution from Eq. 2.3 minimizes the squared Mahalanobis distance between y and the estimator $X\beta$, i.e.

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^d} \|y - X\beta\|_W . \tag{2.5}$$

Further details concerning GLS can be found for example in [13].

In our case, for a fixed out-of-sample data point $x \in \mathcal{M} \backslash M$ and its $k = |N(x)|$ neighbors, a linear system of k equations, each of the form of Eq. 2.1, has to be solved for $\hat{\psi}(x)$. Without loss of generality, we assume that $N(x) = \{x_1, \ldots, x_k\}$. The matrix formulation for such a system is

$$J\hat{\psi}(x) = \Psi + \Omega, \tag{2.6}$$

where $J = [I_d, \ldots, I_d]^T$ is the $kd \times d$ identity blocks matrix, Ω is a kd-long vector, whose j-th section is the d-long constant vector $(\omega_j, \ldots, \omega_j)^T$, and Ψ is a kd-long vector, whose j-th section is the d-long vector $\psi(x_j)$. The vector Ψ encapsulates the images of $N(x)$ under ψ, i.e. the neighborhood of $\hat{\psi}(x)$ in $\mathcal{N}$. The corresponding covariance matrix is the $kd \times kd$ blocks diagonal matrix W,

$$W = diag(w_1, \ldots, w_k), \tag{2.7}$$

whose j-th diagonal element is $w_j = \sigma_j^2 I_d$. Therefore, due to Eq. 2.3, the GLS solution to Eq. 2.6 is

$$\hat{\psi}(x) \triangleq (J^T W^{-1} J)^{-1} J^T W^{-1} \Psi, \tag{2.8}$$

and it minimizes the Mahalanobis distance

$$m(x) \triangleq \|J\hat{\psi}(x) - \Psi\|_W \tag{2.9}$$

that measures the similarity (with respect to W) between $\hat{\psi}(x)$ and its neighbors $\{\psi(x_1), \ldots, \psi(x_k)\}$ in $\mathcal{N}$, which are encapsulated in Ψ. Once W is defined as an invertible covariance matrix, $\hat{\psi}(x)$, as defined in Eq. 2.8, is well posed. The definition of W depends on the definition of w_j for any $j = 1, \ldots, k$, which can be chosen subjected to the similarity properties to be preserved by ψ. These properties are discussed in Section 3. Once Eq. 2.8 is solved for

$\hat{\psi}(x)$, the Mahalanobis distance from Eq. 2.9 provides a measure for the disagreement of the out-of-sample extension of ψ to x with the surrounding geometry. Thus, large value of $m(x)$ indicates that x resides outside of $\mathcal{M}$ and thus, in data analysis terminology, it can be considered as an anomaly.

3 Geometric-based covariance matrix

As mentioned in Section 2.1, the GLS solution minimizes the Mahalanobis distances between $\hat{\psi}(x)$ and its neighbors according to the stored information in W. Thus, if the variances are determined subject to some feature, then $\hat{\psi}(x)$, which is defined in Eq. 2.8, is the closest point in $\mathcal{N}$ to its neighbors with respect to this feature. In this section, W is constructed such that the resulted out-of-sample extension $\hat{\psi}(x)$ agrees with the principal direction of its neighborhood in $\mathcal{N}$. The neighborhood $N_\varepsilon(x)$ can be defined variously. In this paper,

$$N_\varepsilon(x) \triangleq \{y \in M : \|x - y\| \leq \varepsilon\},$$

is used for some positive ε, which ensures locality of the extension scheme. The parameter ε should be fixed according to the sampling density of $\mathcal{M}$, such that $|N(x)| \geq d$. This restriction enables to detect the principal directions of the image of $\psi(N_\varepsilon(x))$ in $\mathcal{N}$.

In the rest of this section, W is constructed. The first construction, presented in Section 3.1, provides a simple mechanism to control the influence rate as a function of its distance from x of a data point $x_j \in N_\varepsilon(x)$ on the value of $\hat{\psi}(x)$. The second construction for W, presented in Section 3.2, incorporates information regarding the principal variance directions of $N_\varepsilon(x)$ such that the result from the sample extension $\hat{\psi}(x)$ "agrees" with these directions.

3.1 Distance based covariance matrix W

Although the definition of $N(x)$ localizes the scheme, it is reasonable to require that data points in $N(x)$, which are distant from x will affect less than close ones. For this purpose, an "affection weight"

$$\lambda_j \triangleq \frac{1}{\|x - x_j\|} \tag{3.1}$$

is assigned to each data point $x_j \in N(x)$. Of course, any other decreasing function of the distance between x and x_j can be utilized. Then, by defining the variance σ_j to be

proportional to the distance $\|x - x_j\|$ such as $\sigma_j \triangleq \lambda_j^{-1}$, we get a diagonal matrix W, whose j-th diagonal element is

$$w_j \triangleq \lambda_j^2 I_d. \tag{3.2}$$

Thus, due to Observation 2.1, close data points in $N_\varepsilon(x)$ affect $\hat{\psi}(x)$ more than those that are far.

3.2 Tangent space based covariance matrix W

In this section, we present a covariance matrix that encapsulates geometric information concerning the manifold $\mathcal{N}$. The covariance matrix W is a set such that the resulted extension obeys the Lipschitz property of ψ.

Let $\mathcal{T}_j$ be the tangent space of $\mathcal{N}$ in $\psi(x_j)$ and let $\mathcal{P}_j$ be the orthogonal projection on $\mathcal{T}_j$. We denote the tangential component of ω_j by $\omega_j^t = \mathcal{P}_j\omega_j$ and its orthogonal complement by $\omega_j^o = (\mathcal{I} - \mathcal{P}_j)\omega_j$, where $\mathcal{I}$ is the identity transformation. Proposition 3.1 quantifies the tangential and the perpendicular components of ω_j from Eq. 2.1, as functions of the curvature of $\mathcal{N}$ in x_j and $\|x - x_j\|$.

Proposition 3.1. *Let $\|x - x_j\| \leq r$ and assume that the curvature of $\mathcal{N}$ in x_j is bounded by a constant c_j. If ψ is assumed to be Lipschitz with constant k, then $\omega_j^t \leq kr$ and $\omega_j^o \leq (c_j kr)^2$.*

Proof. Without loss of generality, we assume that $\psi(x_j) = 0 \in \mathbb{R}^d$ and $\mathcal{T}_j = \mathbb{R}^m$. We denote the graph of the manifold $\mathcal{N}$ in the neighborhood of 0 by the function $f : \mathcal{T}_j \to \mathbb{R}^{d-m}$, where the data points of $\mathcal{N}$ are $(z, f(z))$, $z \in \mathcal{T}_j$. Thus, we get $f(0) = 0$ and $\frac{\partial f}{\partial z}(0) = 0$. Let $x \in \mathcal{M}$ be a data point in the neighborhood of x_j and let $\psi(x) = (z_x, f(z_x))$. Namely, $z_x = \mathcal{P}_j\psi(x)$ and $f(z_x) = (\mathcal{I} - \mathcal{P}_j)\psi(x)$. Then, the Taylor expansion of $f(z_x)$ around 0 yields $f(z_x) = f(0) + \frac{\partial f}{\partial z}(0)(z_x) + O(\|z_x\|^2)$. Since ψ is assumed to be Lipschitz with constant k we get $\|z_x\| = \|\mathcal{P}_j(\psi(x) - \psi(x_j))\| \leq \|\psi(x) - \psi(x_j)\| \leq kr$. Thus, we get that $\|\omega_j^t\| = \|\mathcal{P}_j(\psi(x) - \psi(x_j))\| \leq kr$ and $\|\omega_j^o\| \leq (c_j kr)^2$. □

Proposition 3.1 actually provides a relation between the tangential and perpendicular components of ω_j from Eq. 2.1. Thus, Ω from Eq. 2.6 is a kd-long vector, whose j-th section is a d-long vector $(\omega_j^t, \ldots, \omega_j^t, \omega_j^o, \ldots, \omega_j^o)^T$, where its first m entries are the tangential weights, and the rest $d-m$ entries are the perpendicular ones. The corresponding

covariance matrix is the $kd \times kd$ blocks diagonal matrix W, whose j-th diagonal element is the $d \times d$ diagonal matrix

$$w_j = \begin{pmatrix} \lambda_j^2 & & & & & \\ & \ddots & & & & \\ & & \lambda_j^2 & & & \\ & & & (c\lambda_j)^4 & & \\ & & & & \ddots & \\ & & & & & (c\lambda_j)^4 \end{pmatrix},$$

where the first m diagonal elements are λ_j^2 (see Eq. 3.1) and the rest $d-m$ elements are $(c\lambda_j)^4$. The solution then is given by Eq. 2.6, while J and Ψ remain the same.

3.2.1 Tangent Space Approximation

In real life applications, in order to use the tangent space to $\mathcal{N}$ in $\psi(x_j)$, it has to be first approximated using its neighboring data. In this section we approximate the tangent space and the principle directions of the manifold $\mathcal{N}$ at $\psi(x_j)$. Then, these approximations are incorporated in the construction of W by assigning heavy weights to the tangential direction and less significant weights to the perpendicular ones. The principle directions of the data and the variance of each direction are the eigenvectors and the eigenvalues of the covariance matrix of a data points $\psi(x_j)$, respectively. This covariance matrix is also known as the PCA matrix. A multi-scale version of the local PCA algorithm is analyzed in [15] and it can be used for our analysis. It is important to take at least as many data points as the dimensionality of $\mathcal{N}$.

The covariance matrix of a data point $\psi(x_j)$ is computed in the following way: for simplicity of calculations, we take as the neighboring set of $\psi(x_j)$ the set $\psi(N_{\varepsilon_1}(x)) = \psi(x_1), \ldots, \psi(x_k)$. Then, we form the $k \times d$ matrix X, whose rows constitue the aforementioned set:

$$X = \begin{pmatrix} -\psi(x_1)- \\ \vdots \\ -\psi(x_k)- \end{pmatrix}.$$

Accordingly, we define

$$cov\left(\psi(x_j)\right) \triangleq \left(\frac{1}{\varepsilon_1^2}\right)\left(\frac{1}{k}\right) XX^t, \quad \text{for all } i = 1, \ldots, k. \tag{3.3}$$

Since we take the same set of data points then for all i, j we have $cov(\psi(x_j)) = cov(\psi(x_i))$. To make the calculation and the stability issues easier we add the $(c \cdot \lambda_i)^4$ component to all the diagonal components, and consequently we define:

$$w_j \triangleq \left(\lambda_j^{-2} cov(\psi(x_j)) + \begin{pmatrix} (c \cdot \lambda_j)^{-4} & 0 & \dots & 0 \\ 0 & (c \cdot \lambda_j)^{-4} & \dots & 0 \\ & \vdots & \ddots & \\ 0 & \dots & 0 & (c \cdot \lambda_j)^{-4} \end{pmatrix} \right)^{-1} . \quad (3.4)$$

It is invertible since w_i is positive definite. We notice that it was possible to add the $(c \cdot \lambda_i)^{-4}$ weight component only to the least significant directions of the covariance matrix by computing the SVD [15] of the covariance matrix. It does not improve the accuracy significantly and it becomes more computationally complex . W is a block diagonal matrix with the same structure as appears in Eq. 2.7.

Another option is to make different estimations for the tangent space in different data points by using different sets of data points in the covariance matrix computation. While this estimation should be more accurate, it is more computationally expensive.

4 Bounding the error of the out-of-sample extension

In this section, we prove that the error of out-of-sample extension is bounded, in both cases of distance-based weights from Eq. 3.2 or the tangential-based approximation from Eq. 3.4. This means that for any function $\psi : \mathcal{M} \to \mathcal{N}$, which agrees on a given set of data points and satisfies certain conditions, the out-of-sample extension $\hat{\psi}(x)$ of a data point x is close to $\psi(x)$.

First we prove the consistency of the algorithm. In other words, the out-of-sample extension of data points, which coverage to an already known data point, will converge to its already known image.

Lemma 4.1. *Assume $x \in \mathcal{M}$. If $x \to x_j \in M$ then $\hat{\psi}(x) \to \psi(x_j)$.*

An intuition for the proof of Lemma 4.1 is that the distance from the data point $x_i \in M$ is inversely proportional to the weight of the equation $y = \psi(x_i)$ in Eq. 2.6, therefore, when $x \to x_i$ the distance tends to 0 and the weight tends to ∞. Notice that when $x = x_i$ then, according to Eq. 3.1, $\lambda = \infty$ and the out-of-sample extension is undefined.

Definition 4.2. *The dataset $M \subset \mathcal{M}$ is called a δ-net of the manifold $\mathcal{M}$ if for any data point $x \in \mathcal{M}$ there is $\tilde{x} \in M$ such that $\|x - \tilde{x}\| \leq \delta$.*

Theorem 4.1. *Assume that M is a δ-net of $\mathcal{M}$. Let $\psi : \mathcal{M} \to \mathcal{N}$ be a Lipschitz function with a constant K. If $\varepsilon_1 = \delta$ and $\hat{\psi}(x)$ is computed using the weights in Eq. 3.2, then $\|\hat{\psi}(x) - \psi(x)\| \leq 3K\delta$.*

Proof. We denote by $N_\delta(x) = \{x_1, ...x_k\}$ the set of data points in the $\varepsilon_1 = \delta$ neighborhood of x. It is easy to see that all the data points of $\psi(x_i)$ are inside a ball $B \subset \mathcal{N}$ of radius $K\varepsilon_1$. Therefore, the out-of-sample extension y is also in this ball, namely

$$\left\|\hat{\psi}(x) - \psi(x_i)\right\| < 2K\varepsilon_1. \tag{4.1}$$

Since ψ is a Lipschitz function and $\|x - x_i\| < \varepsilon_1$, we have

$$\|\psi(x_i) - \psi(x)\| < K\varepsilon_1. \tag{4.2}$$

By combining Eqs. (4.1) and (4.2), we get

$$\left\|\hat{\psi}(x) - \psi(x)\right\| \leq \left\|\hat{\psi}(x) - \psi(x_i)\right\| + \|\psi(x_i) - \psi(x)\| \leq 3K\varepsilon_1.$$

□

Next, we show an identical result for the case where the weights from Eq. 3.4 are utilized to construct the covariance matrix W. Moreover, the approximations of the tangent spaces converge to the correct tangent space as ε_1 tends to 0.

Theorem 4.3. *Let M be a δ-net of $\mathcal{M}$ and let $\psi : \mathcal{M} \to \mathcal{N}$ be a Lipschitz function with a constant K. If $\varepsilon_1 = \delta$ and $\hat{\psi}(x)$ is computed using the weights in Eq. 3.4, then $\|\hat{\psi}(x) - \psi(x)\| \leq 3K\delta$.*

Proof. Let $N_\delta(x) = \{x_1, ...x_k\}$ be the δ neighborhood of x. Then, the weight matrix becomes

$$W = \begin{pmatrix} w_1 & 0 & \dots & 0 \\ 0 & w_2 & \dots & 0 \\ & \vdots & \ddots & \\ 0 & \dots & 0 & w_k \end{pmatrix}.$$

By using Eq. (2.8), we get

$$\hat{\psi}(x) = \left(\sum w_i\right)^{-1} \sum w_i \psi(x_i), \tag{4.3}$$

where the matrices w_i are defined in Eq. 3.4. The structure of the w_i matrices allows us to find a basis in which all the matrices w_i become diagonal. Let us denote the diagonal form of w_i by D_i. Then $D_i = T w_i T^{-1}$ where T is a transformation matrix. We can rewrite Eq. (4.3) to become

$$\begin{aligned} T\hat{\psi}(y) &= T\left(\sum w_i\right)^{-1} T^{-1} T \sum w_i \psi(x_i) \\ &= \left(\sum T w_i T^{-1}\right)^{-1} \sum T w_i T^{-1} T \psi(x_i) \\ &= \left(\sum D_i\right)^{-1} \sum D_i T \psi(x_i). \end{aligned}$$

Since all D_i are diagonal, we get a weighted average of the data points $\psi(x_i)$ in the new basis, which is known to be in convex hull. Thus, it is located inside a ball that contains all the data points. It means that $\hat{\psi}(x)$ is inside a ball of radius $K\varepsilon_1$ that contains all $\psi(x_i)$. Therefore,

$$\begin{aligned} \|\hat{\psi}(x) - \psi(x)\| &= \|\hat{\psi}(x) - \psi(x_i) + \psi(x_i) - \psi(x)\| \\ &\leq \|\hat{\psi}(x) - \psi(x_i)\| + \|\psi(x_i) - \psi(x)\| \\ &\leq 2K\varepsilon_1 + K\varepsilon_1 = 3K\varepsilon_1 = 3K\delta. \end{aligned}$$

□

5 Out-of-sample extension complexity

Recall that the dataset M consists of p data points, and assume that the number of data points in the neighborhood of x is k. The covariance matrix of a data point $\psi(x_j)$ from Eq. 3.3 is also computed once for each data point in M, considering each of its k neighbors. The complexity of the neighborhood computation is $O(p)$ operations. Then, the covariance matrix is computed in $O(dk^2)$ operations. Thus, the complexity of this pre-computational phase is $O(p \cdot (p + dk^2)) = O(p^2)$ operations. For each data point, we multiply vectors of of size $d \cdot k$ by matrices of size $k \times d \cdot k$ or $k \times k$. Thus, the out-of-sample extension complexity is $O(k^2 \times d^2)$ operations.

6 Experimental results

6.1 Example I: Data points on a sphere

The function $\psi : [0, \pi] \times [0, \pi] \to \mathbb{R}^3$ maps the spherical coordinates (ϕ, θ) into a 3-D sphere of radius 1. More specifically, $\psi(\phi, \theta) = (\sin(\phi)\cos(\theta), \sin(\phi)\sin(\theta), \cos(\phi))$. We generate

900 data points angularly equally distributed where we have 30 data points on each axis as a training dataset. We generate 100 random data points for which we compute the out-of-sample extension. The results from the application of the algorithm using weights as defined in Eq. 3.2, are shown in Fig. 6.1. In Fig. 6.2 we can see three different results for the out-of-sample extension using different weights as presented in Section 3. In Table 6.1, we show how the results get better for more advanced weight algorithm. We display an accurate error mean for algorithm. We also show the improvement of the results when we take 2500 data points angularly equally distributed with 50 on each axis:

Weight computation	Color in Fig. 6.2	Mean error for 900 data points	Mean error for 2500 data points
Weights as in Eq. 3.2	Yellow	$1.04 \cdot 10^{-2}$	$6.01 \cdot 10^{-3}$
Weights as in Eq. 3.4	Red	$8.08 \cdot 10^{-3}$	$4.45 \cdot 10^{-3}$
Weights as in Eq. 3.4 but with different estimations for the tangent space at each data point	Black	$6.14 \cdot 10^{-3}$	$3.17 \cdot 10^{-3}$

Table 6.1: The mean error performances of the algorithms for different number of data points

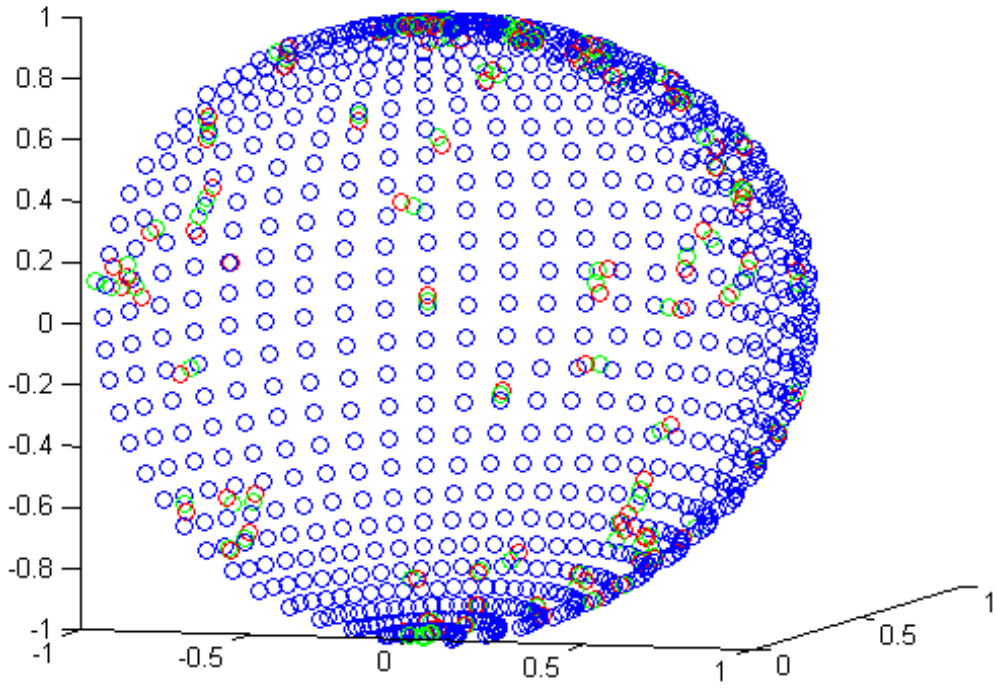

Figure 6.1: An illustration of the out-of-sample extension algorithm on a sphere. Blue denotes the original data set, green denotes the correct images and red denotes the out-of-sample extension calculated using the algorithm with weights as presented in Eq 3.2

.

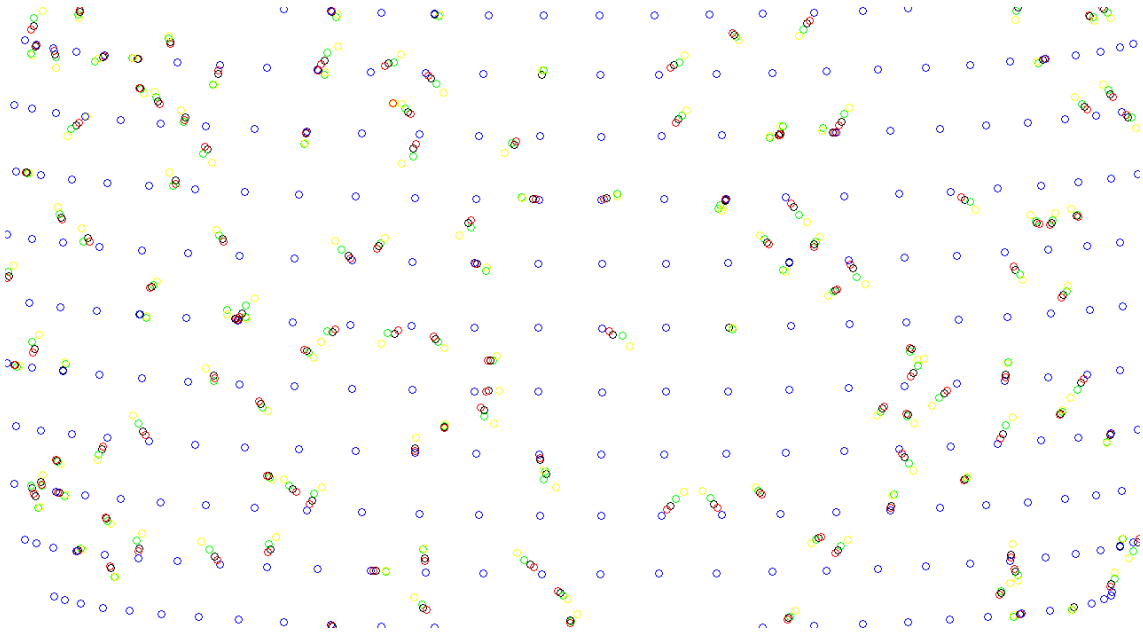

Figure 6.2: An illustration of the algorithms in Table 6.1 on a sphere. Blue denotes the original data set, green denotes the correct images and yellow denotes the out-of-sample extension computed using the algorithm with weights as presented in Eq 3.2. Red denotes the out-of-sample extension computed using the algorithm with weights as presented in Eq 3.4. Black denotes the out-of-sample extension computed using the weights as presented in Eq 3.4 with different estimations for the tangent space at each data point.

6.2 Example II: Dimensionality reduction example

DARPA datasets [16] from 1998 and 1999 are used here to find anomalies in them. All the activities and non-activities are labeled and published. These datasets contain different types of cyber attacks that we consider as anomalies.

We use this dataset to evaluate the performance of the out-of-sample extension using weights as presented in Eq 3.4 and the Mahalanobis distance from Eq. 2.9. The experiment done by following the example in [8]. We use the same data and the same mapping that was developed in [8]. Diffusion Maps (DM) [5] reduced the dimensionality from $\mathbb{R}^{14}$ to $\mathbb{R}^5$ so that $\psi : \mathbb{R}^{14} \to \mathbb{R}^5$, and all the 1321 data points are used as our training dataset. We show in Fig. 6.3 the image of these data points under ψ. The normal behavior manifold in the embedded space in Fig. 6.3 is the "horseshoe" and we can see a few data points, which are considered as anomalous, are the labeled attacks. Then, a newly arrived data points are assigned coordinates in the embedded space via

the application of Nyström extension as can be seen in the left image in Fig. 6.4. It is also done by using the algorithm in [3]. Data point #51, which is a newly arrived data point, is classified as an anomaly that can be seen as an outlier on the left side of the normal ("horseshoe") manifold.

We apply our out-of-sample extension algorithm using weights from Eq 3.4, to the same set of newly arrived data points. The results are shown on the right image in Fig. 6.4. To find anomalies, we compute the Mahalanobis distance of the extension using Eq. (2.9) for each of the newly arrived data points. We see that data point #51 emerged as having a much higher residual error (2.7210^{-7}) than the other data points whose average residual error is 6.2110^{-10}. All the Mahalanobis distance values are shown in Fig. 6.5.

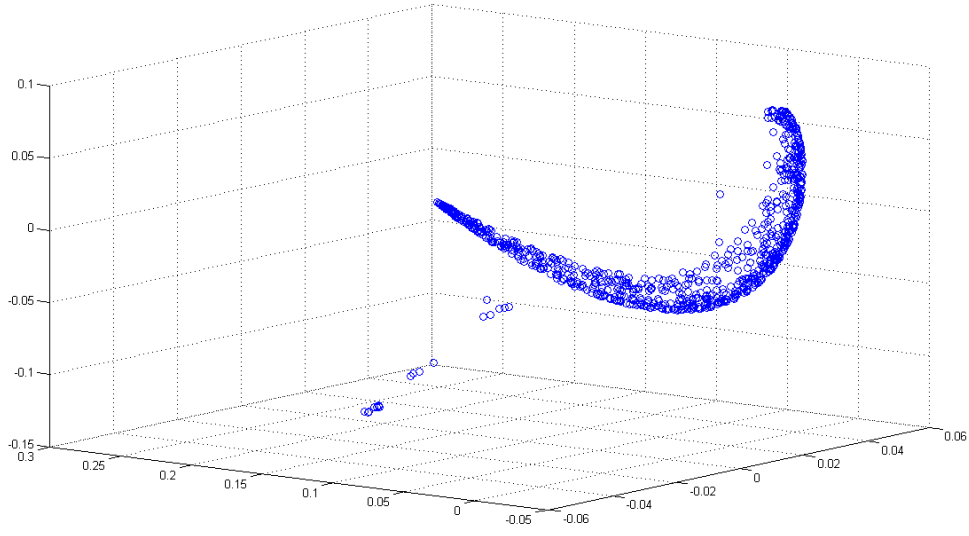

Figure 6.3: The first three coordinates of the data points after the embedding into $\mathbb{R}^5$.

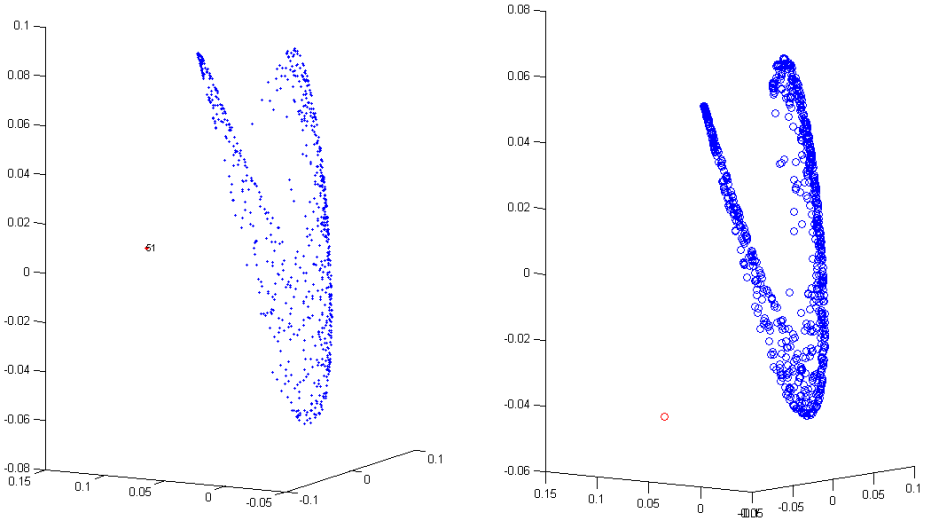

Figure 6.4: Out-of-sample extension is computed for a new day. In the left side, the out-of-sample extension is computed via the Nyström extension algorithm. Data point #51 is known to be anomalous. In the right image, the output of the algorithm using weights from Eq 3.4, is presented by the red data point which is data point # 51.

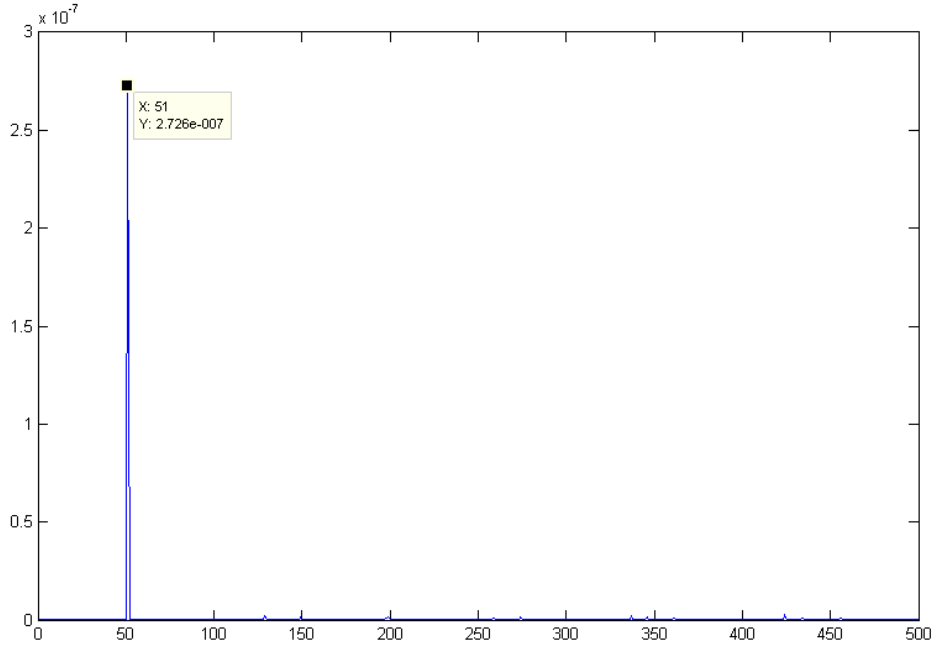

Figure 6.5: The Mahalanobis distance values. We see that data point #51 has the highest value. Therefore, it is classified as an anomalous data point.

7 Conclusions

In this paper, we presented an efficient out-of-sample extension (interpolation) scheme for dimensionality reduction maps, which are widely used in the field of data analysis. The computational cost of such maps is high. Therefore, once such a map was computed over a training set, an efficient extension scheme is needed. The presented scheme is based on the manifold assumption, which is widely used in the field of dimensionality reduction. It provides an optimal solution of a stochastic geometric-based linear equations system that is determined by local PCA of the embedded data. Moreover, the scheme enables the detection of abnormal data points. The interpolation error was analyzed, assuming that the original map is Lipschitz. The scheme was applied to both synthetic and real-life data and provided good results by mapping data from the manifold to the image manifold and by detection of the correct abnormal data points.

Acknowledgments

This research was partially supported by the Israel Science Foundation (Grant No. 1041/10), the Ministry of Science & Technology (Grant No. 3-9096) and by the Binational Science Foundation (BSF Grant No. 2012282).

References

[1] C.T.H. Baker. *The Numerical Treatment of Integral Equations*. Oxford: Clarendon Press, 1977.

[2] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15:1373–1396, 2003.

[3] A. Bermanis, A. Averbuch, and R.R. Coifman. Multiscale data sampling and function extension. *Applied and Computational Harmonic Analysis*, 34:15–29, 2013.

[4] M. D. Buhmann. Radial basis functions: Theory and implementations. Cambridge University Press, 2003.

[5] R.R. Coifman and S. Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006.

[6] R.R. Coifman and S. Lafon. Geometric harmonics: A novel tool for multiscale out-of-sample extension of empirical functions. *Applied and Computational Harmonic Analysis*, 21(1):31–52, 2006.

[7] T. Cox and M. Cox. *Multidimensional Scaling*. Chapman and Hall, London, UK, 1994.

[8] G. David. *Anomaly Detection and Classification via Diffusion Processes in Hyper-Networks*. PhD thesis, School of Computer Science, Tel Aviv University, March 2009.

[9] L.M. Delves and J. Walsh. Numerical solution of integral equations. Clarendon, Oxford, 1974.

[10] D.L. Donoho and C. Grimes. Hessian eigenmaps: New locally linear embedding techniques for high dimensional data. *Proceedings of the National Academy of Sciences of the United States of America*, 100:5591–5596, May 2003.

[11] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24, 1933.

[12] I.T. Jolliffe. *Principal Component Analysis*. Springer, New York, NY, 1986.

[13] T. Kariya and H. Kurata. *Generalized least squares*. Wiley, 2004.

[14] J.B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29:1–27, 1964.

[15] A.V. Little, J. Lee, Yoon-Mo Jung, and M. Maggioni. Estimation of intrinsic dimensionality of samples from noisy low-dimensional manifolds in high dimensions with multiscale svd. In *Statistical Signal Processing, 2009. SSP '09. IEEE/SP 15th Workshop on*, pages 85 –88, 31 2009-sept. 3 2009.

[16] Lincoln Laboratory MIT. Darpa intrusion detection evaluation data sets. `http://www.ll.mit.edu/IST/ideval/data/data_index.html`, 1999.

[17] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery. *Numerical Recipes in C*. Cambridge Univ. Press, 2nd edition, 1992.

[18] N. Rabin and R.R. Coifman. Heterogeneous datasets representation and learning using diffusion maps and laplacian pyramids. *Proceedings of the 12th SIAM International Conference on Data Mining, SDM12.*, 2012.

[19] S.T. Roweis and L.K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

[20] J. Schuetter and T. Shi. Multi-sample data spectroscopic clustering of large datasets using Nyström extension. *Journal of Computational and Graphical Statistics*, pages 531–542, 2011.

[21] H. Wendland. Scattered data approximation. Cambridge University Press, 2005.

[22] G. Yang, X. Xu, and J. Zhang. Manifold alignment via local tangent space alignment. *International Conference on Computer Science and Software Engineering*, December 2008.

[23] Z. Zhang and H. Zha. Principal manifolds and nonlinear dimension reduction via local tangent space alignment. *SIAM Journal on Scientific Computing*, pages 313–338, 2004.