Research Article

A Manifold-Based Dimension Reduction Algorithm Framework for Noisy Data Using Graph Sampling and Spectral Graph

Tao Yang,1,2 Dongmei Fu,1,2 and Jintao Meng1

1School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, China
2Beijing Engineering Research Center of Industrial Spectrum Imaging, School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, China

Correspondence should be addressed to Tao Yang; yangtao@ustb.edu.cn

Received 10 June 2020; Revised 16 July 2020; Accepted 3 August 2020; Published 29 September 2020

Academic Editor: Marcin Mrugalski

Copyright © 2020 Tao Yang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This paper proposes a new manifold-based dimension reduction algorithm framework. It can deal with the dimension reduction problem of data with noise and give the dimension reduction results with the deviation values caused by noise interference. Commonly used manifold learning methods are sensitive to noise in the data. Mean computation, a denoising method, is an important step in data preprocessing but leads to a loss of local structural information. In addition, it is difficult to measure the accuracy of the dimension reduction of noisy data. Hence, manifold learning methods often transform the data into an approximately smooth manifold structure; however, practical data from the physical world may not meet the requirements. The proposed framework follows the idea of the localization of manifolds and uses graph sampling to determine some local anchor points from the given data. Subsequently, the specific range of localities is determined using graph spectral analysis, and the density within each local range is estimated to obtain the distribution parameters. Then, manifold-based dimension reduction with distribution parameters is established, and the deviation values in each local range are measured and further extended to all data. Thus, our proposed framework gives a measurement method for deviation caused by noise.

1. Introduction

Manifold learning is used for nonlinear dimension reduction of natural and general data. These data are often assumed to be nonlinear manifolds embedded in low-dimensional space [1]. When the data are ideally located on a smooth manifold, manifold-based methods such as ISOMAP [2], local linear embedding (LLE) [3], Laplace embedding (LE) [4], and local tangent space alignment (LTSA) [5] can give effective and accurate low-dimensional structures. However, the performance of these methods decreases when the noise interferes with the data and does not allow it to be accurately embedded on the manifold. In practice, noise from the natural environment, sensors, and human intervention interfere with most of the data collected from general backgrounds and objects. Thus, the dimension reduction of noisy data is imperative. This is highly significant for data mining applications in various fields.

Manifold-based dimension reduction of noisy data is challenging. Traditionally, Hein and Maier [6] and Gong et al. [7] reduce the noise impact through an error penalty factor, resulting in excessive smoothing and even deformation of the data. Current works mainly focus on the combination of denoising preprocessing and manifold dimension reduction. Zhao et al. [8] propose a semisupervised local multimanifold framework by linear embedding which incorporates the neighborhood reconstruction error to preserve local topology structures between both labeled and unlabeled samples. Kim and Lee [9] construct a metric function to calculate the distance between the data points and the expected smooth manifold and then project the data onto this smooth manifold. A similar work refers to
Rajagopal et al. [10]. In Jin and Bachmann’s work [11], the data are split into multiple regions and the dimension in each region is reduced to obtain principal components, connected to reduce the noise interference. Zhang et al. [12] use sparse-constrained manifold mapping to weaken the noise. Zhang et al. [13] propose a marginal Isomap for manifold learning which computes the shortest path on graphs to better separate interclass samples, and the authors in [14] further provide the semisupervised marginal manifold visualization. Hao et al. [15] project the data onto a local tangent plane and achieve denoising by averaging the projected coordinates, but the original local structure information is not retained simultaneously. Zhang et al. [16] propose a robust unsupervised nonnegative adaptive feature extraction algorithm which preserves the manifold structures by joint weights sharing in low-dimensional representation and error approximation method.

Although the above manifold learning methods can suppress noise to a certain extent, the original data are possibly overprocessed or artificially intervened, and the effectiveness of dimension reduction becomes difficult to evaluate. Yan et al. [17], which is based on graph spectral of graph theory and the frequency domain of noise mostly belonging to high frequency bands, realize the screening of noise points by the characteristics of the data itself and do not affect the distribution of the original data; however, it is still a combination of denoising and dimension reduction and does not provide the possible low-dimensional space of the original data. Graph theory is also used in uncertain nonlinear systems with environmental noises, and the edge-based adaptive algorithm is proposed to solve the distributed optimal consensus problem [18], while the form of noises is needless to be focused. Li and Lian [19] consider the frequency of persistent dwell time switching in switching Markov jump systems to determine the rule of control while not specific to the manifold-based method.

The low-dimensional space of noisy data will show certain randomness under the interference condition; that is to say, noise changes the existing distribution characteristics because of the low-dimensional coordinates changes, and these characteristics reflect the existence of certain deviation values in the dimension reduction results. Some works, such as that by Zhang et al. [20] propose a joint label prediction to have representations of data which preserved manifold structures explicitly and adaptively, Ye et al. [21] provide a new linear discriminant analysis to solve the problem of insufficient robustness to outliers, and Zhang et al. [22] propose a reliable and robust two-dimensional neighborhood preserving projection method, pay attentions to robust feature extraction problem. The aims of these papers are different from this paper. This paper, aiming at manifold-based dimension reduction of noisy data, solves the problem of measuring the deviation values of low-dimensional data caused by noise and determines the local range of the data.

In the determination of localization, which is an important process in most manifold methods for proper localization, the key is to change the low-dimensional structure to an approximately nonlinear, high-dimensional one. This paper uses graph sampling [23] and spectral graph [24] to provide a solution for noisy data. In our paper, graph sampling is used in the first step of our framework to select useful nodes, and spectral graph is the theory foundation which gives the basis to graph wavelet analysis in our framework. This paper proposes to determine proper local ranges with knowledge of graph sampling and spectral graph and such local ranges are considered to be relatively natural, where the structure, including noise interference, of the original data is retained. The algorithm framework proposed needs all structural information, which is completed by combining graph sampling and spectral graph theory, to measure the deviation values of low-dimensional samples under noise interference.

For measuring the deviation values of the low-dimensional samples, this study uses a maximum likelihood estimation in each local range to construct a normal distribution function adapted to the current local. This function can be used to obtain the maximum deviation of the noisy data in each locality, and the dimension reduction result should be affected by the fluctuation of noise interference in one or more localities. The distance weights are computed to calculate the deviation results caused by noise of the dimension reduction results in our algorithm framework.

As can be seen from the above, the contribution of this paper is as follows:

1. This paper proposes a manifold dimension reduction framework for noisy data.
2. A kind of local ranges determination is provided for complex noisy data by graph sampling and graph spectral theory.
3. This paper proposes a weighted sum method to compute the low-dimensional data deviation values from different local ranges caused by noise.

Furthermore, our framework fits for many kinds of manifold dimension reduction methods and thus can be possibly applied to different research fields.

The remainder of this paper is organized as follows. In next section, the preliminary works are introduced. Afterwards, the section details the proposed algorithm framework, and the following section provides the experimental results of the simulation data and practical data. Finally, a summary is given in the final section.

2. Preliminary Works

The noisy data are denoted as \( x = \{ x_i \mid i = 1, 2, \ldots, N \} \), \( x_i \in \mathbb{R}^D \). Assume that noise obeys the Gaussian distribution \( \delta \sim N(0, \sigma^2) \), and the data \( \tilde{x} \) sampled from the unknown smooth manifold \( \mathcal{M} \) are denoted as \( x = \tilde{x} + \delta \). Based on the data \( x \), an undirected weighted graph \( G = (V, E) \) is constructed, in which \( V \) represents the set of nodes in the graph, and the nodes have a one-to-one correspondence to the original data samples, that is, \( v_i \sim x_i, \forall i \). \( E \) represents the set of edges connecting two nodes in the graph. A weight matrix \( W \) is calculated, and a nonnegative weight is assigned to each edge. In this study, the calculation of weights uses a Gaussian kernel function.
The algorithm framework in this paper is divided into three parts: the determination of local anchor vectors, local range and distribution estimation, and manifold dimension reduction with distribution parameters.

3.1. Local Anchor Vectors. The localization analysis is the first step for reducing the dimension of the manifold. The determination of the local anchor vectors is important for identifying the local position. Graph sampling is similar to image downsampling. Graph sampling can select some data samples to form a reduced data graph based on the polarity of the eigenvectors of the Laplacian matrix, retaining a part of the data samples of interest, so as to determine some key samples for computing the local ranges. These key samples are collectively referred to as “local anchor vectors” in this paper.

Let the local anchor vector set be \( V_1 \); then, \( V_1 \subset V \), implying that the anchor vector set is a part of the graph nodes, and each local anchor vector and its local range represent its nearest neighbors.

This paper adopts a graph sampling method based on the polarity of the components of the largest eigenvector, that is, \( V_1 = \{ \nu_l \in V | \mu_{\text{max}}(\ell) > 0 \} \), with eigenvalues arranged in the ascending order. According to Shuman et al. [27], graph sampling should meet the following conditions: (1) the sampled node set should be approximately half of the total number of nodes in the original graph, that is, \( |V_1| \approx (|V|/2) \); (2) the removed nodes are not connected to edges with high weights; and (3) the computation is efficient and feasible. The polarity of the largest eigenvector sampling method satisfies three conditions, and the local anchor vector set can be obtained for the local range analysis.

3.2. Local Range and Distribution Estimation. After each local anchor vector is determined, it is necessary to further determine its local range and then calculate information such as the deviation values caused by noise interference.

The spectral graph theory is used to further analyze the local ranges. A spectral graph is a special type of spectral analysis. The spectral analysis itself is based on the frequency domain where a signal is characterized by its spectral coefficient or spectral energy. According to [28], stable data distribution often exhibits low-frequency characteristics, while noise may appear in the full frequency band and mainly cover high frequencies. Therefore, the local range of a local anchor vector should be located in the low-frequency band of a signal on the anchor vector.

This study uses the spectral graph wavelet method to obtain the local range. This method is more flexible and has fewer parameters than the common k-nearest neighbor algorithm. If a signal is applied to an anchor vector, it will propagate and extend to the area around. Owing to distance and noise, the signal strength will be attenuated. When determining the local range using the spectral graph wavelet, the range in which the signal is attenuated to a certain degree and the bandpass characteristic of the wavelet function will also affect the speed of attenuation.
Let a signal function $\delta_n \in \mathbb{R}^N$ act on each node of the graph with $\delta_n(n) = 1$ and $\delta_n(m) = 0$, $n \neq m$. For a certain node $n$, its wavelet coefficients $\psi_{\delta,n}(n) = T_g \delta_n(n)$. Suppose the filter function is $g$. Then,

$$
\psi_{\delta,n} = \sum_{\ell=1}^N g(\lambda_\ell) \delta_n(\ell) u_{\ell} = \sum_{\ell=1}^N g(\lambda_\ell) \left( \sum_n \delta_n(m) u_{\ell}(m) \right) u_{\ell} = \sum_{\ell=1}^N g(\lambda_\ell) u_{\ell}(n) u_{\ell}.
$$

(7)

The filter function will control the span of signal spreading, and it can take out the area where the local range determination method using an anchor vector energy is, the better the signal spread in this area will be.

### Complexity

$$
\delta
$$

According to maximum likelihood and the expression of the parameters, it is necessary to estimate the distribution. The distribution can also be approximately Gaussian. To determine the local range of each anchor vector. On the Gaussian noise premise, the deviation values of low-dimensional manifolds under noise interference. A maximum likelihood estimation is used in each local range to construct a normal distribution function adapted to the current local. In this study, manifold mapping functions are established for the local mean and deviation results, to obtain the manifold dimension reduction results with distribution parameters and then to calculate the deviations between the deviation dimension reduction result and the mean dimension reduction result.

Similarly, using the original data to establish a manifold mapping function and considering that each data belongs to the range of one or more local anchor vectors, the dimension reduction result should also be affected by the fluctuation of noise interference in one or more localities. In this study, the distance weights are used to calculate the weighted sum of the deviation results of the dimension reduction results of each data sample.

For each anchor vector and its local range, the mean and covariance were estimated. In reality, the mean of all anchor vectors constitutes a type of denoising result. This mean result will form an approximately smooth manifold. Therefore, the manifold dimension reduction of the mean values is used as the benchmark of the deviation measure. The framework proposed in this study can use any manifold dimension reduction method to establish a mapping function. Assume that the selected manifold dimension reduction method is expressed as $\mathcal{F}$; then, $\mathcal{F}(x): \mathbb{R}^D \rightarrow \mathbb{R}^d$ and $d \leq D$ represent reduction from dimension $D$ to dimension $d$. When the mapping function is obtained by mean values, it is denoted as $\mathcal{F}_\mu$. In addition, it is necessary to establish the dimension reduction results of the deviation value caused by covariance. Taking the anchor vector $v_p$ and its local range as an example, since $\mu_p \in \mathbb{R}^D$ and $\Sigma_p \in \mathbb{R}^{D \times D}$ for a certain dimension $m$, we can calculate the deviation values because of the variance parameter in dimension $m$, that is,

$$
{b}_p^m = \mu_p^m - a_m^m,
$$

$$
{m}_p^b = \mu_p^m + a_m^m,
$$

(10)

where $\mu_p^m$ is the $m$-th element in the mean vector $\mu_p$ and $\sigma_p^m$ is the $m$-th diagonal element in the covariance $\Sigma_p$. By traversing all the $D$ dimensions of the data, the deviation values can be obtained. The deviation values are divided into positive and negative values because of the direction of the variance and are denoted as $[b_p^m, b_p^m]$.

We iterate through all anchor vectors to obtain all local deviation value sets $[b_p^m, b_p^m]$ and use the negative and positive deviations to obtain the dimension reduction mapping functions $\mathcal{F}_{b_p}^m$ and $\mathcal{F}_{b_p}^m$, respectively. For now, we
obtain the manifold dimension reduction with distribution parameters, that is, \(\mathcal{F}_\mu, \mathcal{F}_{\nu_r}, \mathcal{F}_{\nu_{r'}}\). In order to measure the low-dimensional deviation values caused by noise, the framework in this study requires the following two calculations:

1. **Calculation of deviation value in each dimension of the low-dimensional space.** We use \((\mathcal{F}_\mu, \mathcal{F}_{\nu_r}, \mathcal{F}_{\nu_{r'}}), j \in \{1, 2, \ldots, d\}\) to denote the dimension reduction with distribution parameters. For any \(j\)-th dimension, the formula for calculating the deviation is as follows:

\[
T_v^{j} = \frac{\sum_{r} D_r \nu_r^{j} \nu_{r'}^{j}}{D_v^{j}}, \quad v_p \in V_1, \quad v_q \in V_1.
\]  

(11)

The above formula represents the deviation calculation of the \(j\)-th dimension within the local range of the anchor vector \(v_p\).

2. **Distance calculation of low-dimensional space data.** The computation formula is as follows:

\[
T_v^{j} = \frac{\sum_{r} D_r \nu_r^{j} \nu_{r'}^{j}}{D_v^{j}}, \quad v_p \in V_1, \quad v_q \in V_1.
\]  

(12)

The Euclidean distance calculation in the above formula represents the deviation calculation of the data distance within the local range of the anchor vector \(v_p\).

In order to further achieve the dimension reduction of the overall noisy data and measure the interference effect of noise, based on obtaining the manifold dimension reduction with distribution parameters of the anchor vectors and their local deviation values, the framework in this study provides a distance-weighted method for dimension reduction values of the original noisy data.

The reason of using distance-weighted method is that any one data sample \(x_i, \forall i\) may be located in the local range of one or more anchor vectors, and noise interferences represented by each anchor vector and its locality will affect the current sample \(x_i\). In this study, the ratio of the distance between the current sample \(x_i\) and the local mean of each anchor vector is used as a weight. The deviation value of \(x_i\) comes from the weighted sum of the deviation results of the dimensions of each anchor vector. Specifically, suppose the data sample \(x_i\) is in the localities of \(R\) anchor vectors, and these anchor vectors are expressed as \(V_R = \{v_{1, V_i}, v_{2, V_i}, \ldots, v_{R, V_i}\}\). Then, we calculate the distances from the local mean values of these anchor vectors, that is,

\[
D_r = \left| x_i - \mu_{r, V_i} \right|, \quad r \in R,
\]

\[
D_{V_i} = \sum_r D_r .
\]  

(13)

The deviation values of each dimension of the low-dimensional local of anchor vectors in \(V_R\) are denoted as \(T_{v_p, v_q}^{j}, T_{v_p, v_q}^{j}, j \in \{1, 2, \ldots, d\}\); then, the deviation values of each dimension in the low-dimensional space of \(x_i\) are

\[
T_{v_p}^{j} = \sum_r D_r \nu_r^{j} \nu_{r'}^{j}.
\]  

(14)

For the entire data \(X\), the dimension reduction result is \(\mathcal{F}_X\), and the deviation values in each dimension of (11) are introduced to obtain the deviation values of the entire low-dimensional entire data with noise interference. For a certain dimension \(j\), it is expressed as follows:

\[
\mathcal{F}_X^{j} = \mathcal{F}_X^{j} + T_{v_p}^{j},
\]

\[
\mathcal{F}_X^{j} = \mathcal{F}_X^{j} + T_{v_p}^{j}.
\]  

(15)

We iterate overall low-dimension \(d\) and obtain \(\mathcal{F}_{X,v_p}\) and \(\mathcal{F}_{X,v_q}\). Thus, the deviation values of the entire data distance in the low-dimensional space can be obtained from the Euclidean distance calculation in (12).

The above deviation value calculation for each dimension in the low-dimensional space and for the overall data in the low-dimensional space is shown in Figure 1.

In Figure 1 and according to (14), we can obtain the computing of \(w_1\) and \(w_2\) in the following:

\[
w_1 = \frac{|x_i - \mu_{v_p}^{j}|}{|x_i - \mu_{v_q}^{j}| + |x_i - \mu_{v_q}^{j}|},
\]

\[
w_2 = \frac{|x_i - \mu_{v_p}^{j}|}{|x_i - \mu_{v_q}^{j}| + |x_i - \mu_{v_q}^{j}|}.
\]  

(16)

Through the weights, the different local ranges can simultaneously affect the final deviation values.

Summarize the above three parts of algorithm to have the schematic illustration of Figure 2 in the following.

The computing process of the proposed framework is as follows.

Input (1) data samples with noise \(X = \{x_i\}_{i=1}^N, x_i \in \mathbb{R}^D\); (2) functions \(\delta_i, i = 1, 2, \ldots, N\); (3) filter functions \(g_e\) and (4) a manifold dimension reduction algorithm \(\mathcal{F}: \mathbb{R}^D \rightarrow \mathbb{R}^d\).

**Procedure.**

Step 1. Construct a graph \(G = (V, E)\) according to the data \(X\), where the weight matrix \(W\) is calculated by (1).

Step 2. Calculate the Laplacian matrix \(L\) of the graph \(G\) and obtain the eigenvector matrix \(U\) and eigenvalue matrix \(\Lambda\) by matrix decomposition.

Step 3. Determine the anchor vector set \(V_1\) according to the polarity of the largest eigenvector.

Step 4. For each \(v_i \in V_1\), do the following:

Apply a signal function \(\delta_i\) on \(v_i\). Calculate the spectral graph wavelet coefficient of the signal function according to (7). Determine the local range of the anchor vector \(v_i\) according to (8).
Obtain the distribution parameters of the anchor vector locality according to (9).

Step 5. Learn to obtain the manifold dimension reduction result \( \mathcal{F}_\mu \) using all mean vectors of the anchor vector set \( V_1 \).

Step 6. Use all covariances of the anchor vector set \( V_1 \) and (10) to obtain all deviation value vectors and learn to obtain the manifold dimension reduction results \( \mathcal{F}_{p} \), \( \mathcal{F}_{b} \).

Step 7. Learn to obtain the manifold dimension reduction result \( \mathcal{F}_X \) for all data \( X \).

Step 8. Obtain the deviation value of each dimension of the low-dimensional space according to (11).

Step 9. For each \( x_i \in X \), do the following:

According to (13), calculate the distance ratio between \( x_i \) and the local mean values of the anchor vectors consisting of the local ranges that contain \( x_i \).

Calculate the deviation values \( \mathcal{F}_{b_{x_i}^+}, \mathcal{F}_{b_{x_i}^-} \) of \( \mathcal{F}_{x_i} \) according to (14) and (15).

Output. Low-dimensional data samples. \( (\mathcal{F}_x, \mathcal{F}_{b_{x_i}^+}, \mathcal{F}_{b_{x_i}^-}) \).

4. Experiments

For the proposed algorithm framework, two simulated datasets and two image datasets are selected for algorithm implementation and simulation.
4.1. Swiss Roll Dataset. This three-dimensional dataset has a total of 1000 data samples. Gaussian noise with 0 mean and a standard deviation of 0.5 is added to each sample point, as shown in Figure 3.

We construct the graph $G$ and calculate the Laplacian matrix $L$ and its eigenvalue decomposition result $L = UΛU^*$. The polarity of the largest eigenvector is used to determine the set of anchor vectors for local analysis, as shown in Figure 4.

Next, the graph spectral wavelet method determines the local range of each anchor vector. The filtering function selected in this study is $g(x) = \exp\left(-\frac{rx}{\lambda_{\text{max}}}\right)$, where $\lambda_{\text{max}}$ represents the largest eigenvalue of the Laplacian matrix. This function has low-pass characteristics. The parameter $r$ can control the low-pass frequency domain width and therefore change the ranges of local area. In practice, the filter function with its parameter should be tested to choose proper ones. For the current simulation process, $r = 30$ is selected. A signal function $δ$ is applied to each anchor vector, and the wavelet coefficients of the nodes in the graph are calculated so that the local range of each anchor vector is determined by the total energy ratio of the wavelet coefficients. Four anchor vectors are randomly selected as visualization examples, as shown in Figure 5.

For each local range, we estimate the distribution parameters, that is, the mean and covariance parameters. According to (9), the local mean vector and deviation values of each anchor vector local range in the three-dimensional space can be plotted, as shown in Figure 6.

The mean vectors are used to learn to determine the dimension reduction $\mathbf{F}_{\mu}$, and the deviations are used to learn to determine the dimension reduction $\mathbf{F}_{\bm{\delta}}, \mathbf{F}_{\bm{\beta}}$. In order to visualize the deviation value obtained by the algorithm framework shown in (16), the positive $Z$-axis value in the three-dimensional space is used to represent the deviation value because of the existence of the positive standard deviation of the distribution parameters, and the negative value $Z$-axis value is used to represent the one because of the existence of the negative standard deviation of the distribution parameters, that is,

$$T_{\mu,\bm{\delta}} = \|\mathbf{F}_{\mu} - \mathbf{F}_{\bm{\delta}}\|,$$

$$T_{\mu,\bm{\beta}} = \|\mathbf{F}_{\mu} - \mathbf{F}_{\bm{\beta}}\|.$$  

(17)

For now, we can obtain the dimension reduction results of the mean vectors and deviation values, as shown in Figure 7. This algorithm framework can introduce any type of manifold dimension reduction method. In the experiment, four commonly used methods are selected: LE, ISOMAP, LLE, and LTSA.

As can be seen from Figure 7, the low-dimensional results obtained by different manifold learning methods and the deviation values caused by noise interference are different.

We use manifold learning to perform dimension reduction on the entire noisy data and the proposed distance weighting method to calculate the deviation values of the noise interference in the dimension reduction result. In order to visualize the deviation results, the positive and negative $Z$-axes are still selected to denote the deviation values in low-dimensional space, as shown in Figure 8.

As can be seen from Figure 8, the dimension reduction results of deviation values with the distribution parameters exhibit certain morphological characteristics, which are different for different dimension reduction methods.

4.2. S-Shaped Dataset. This dataset has a total of 1000 data samples. Gaussian noise with 0 mean and a standard deviation of 0.5 is added to each sample point, as shown in Figure 9.

According to the anchor vector set and the local range determination method given in the framework, the anchor point set is shown in Figure 10 and the local ranges represented by the wavelet coefficients of six randomly selected anchor vectors are shown as examples in Figure 11.

After determining the anchor vector and local range, the distribution parameters are estimated and the deviation can
be obtained based on the positive and negative directions of the standard deviation, as shown in Figure 12.

Similar to the Swiss roll dataset experiment, the mean vector and the deviation values in Figure 12 are used to calculate the deviation values caused by noise interference. The four commonly used manifold methods: LE, ISOMAP, LLE, and LTSA reduce it to two-dimension and use the positive and negative directions of the $Z$-axis to indicate the calculated deviation values, as shown in Figure 13.

Then, manifold dimension reduction processing is performed on the original noisy data, and deviation values are calculated for noise interference in the dimension reduction results. Similarly, select the positive and negative directions of the $Z$-axis to indicate the deviation values in low-dimensional space.

Figure 14 also shows that the interference of noise on the manifold dimension reduction results varies with the method and has different characteristics.

4.3. MNIST Dataset. The MNIST dataset is a set of handwritten digital grayscale images of size $28 \times 28$ in pixels. The purpose of the experiment in this section is to test the effectiveness of the proposed algorithm framework on datasets with classification labels. Therefore, 2000 images of “0” and “1” are arbitrarily selected as two-class samples, and 1500 images of “0,” “1,” and “2” are arbitrarily selected as three-class samples. For visualization, each image data is vectorized into $28 \times 28 = 768$ vector sample, and the manifold dimension reduction method is used for two-dimensional reduction, and the positive and negative $Z$-axes of the three-dimensional space are used to characterize the deviation values caused by noise interference in low-dimensional space.

4.3.1. Two-Class Samples. Gaussian noise is added to the original 2000 vectors, with 0 mean and a standard deviation of 0.5. The proposed algorithm framework uses LE, ISOMAP, LLE,
and LTSA, the commonly used manifold dimension reduction methods, and the parameter in the filter function \( g \) is \( \tau = 35 \). Figure 15 is an example of the original image. Figure 16 shows the results of the manifold dimension reduction results and the deviation values in the positive and negative standard deviation directions in low-dimensional space.

As seen in Figure 16, when reducing to two dimensions, the distance between the different classes is obvious, the class labels can be well distinguished, and the deviation values of the dimension reduction results with distribution parameters also have obvious class discrimination. The difference can also indicate the change in the noise interference in different localities of the dataset.

4.3.2. Three-Class Samples. Gaussian noise is added to the original 1500 vectors, with 0 mean and a standard deviation of 0.5. Other conditions are similar to the two-class experiment. Figure 17 is an example of the original image. Figure 18 shows the results of the dimension reductions and their deviations.

The results in Figure 18 also have obvious class discrimination, and the noise interference difference of each sample point can be expressed by the difference between the manifold dimension reduction results and the distribution parameters.

4.4. Fashion-MNIST Dataset. The fashion-MNIST dataset belongs to an extended version of MNIST dataset and includes 28 × 28 grayscale image sets of ten category labels such as items “T-shirt,” “jeans,” and “sweater.”

4.4.1. Two-Class Samples. We arbitrarily select “T-shirts” and “boots” as two-class samples. The processing method is the same as in the MNIST dataset experiment. Gaussian
Figure 8: Dimension reduction and deviation value results for entire noisy Swiss roll data.

(a) ISOMAP method dimension reduction of entire data
   - The deviation values (negative direction)
   - The deviation values (positive direction)

(b) LE method dimension reduction of entire data
   - The deviation values (negative direction)
   - The deviation values (positive direction)

(c) LLE method dimension reduction of entire data
   - The deviation values (negative direction)
   - The deviation values (positive direction)

(d) LSTA method dimension reduction of entire data
   - The deviation values (negative direction)
   - The deviation values (positive direction)

Figure 9: Visualization of noisy S-shaped dataset.

- Original noisy S-shape data
noise with mean of 0 and standard deviation of 0.5 is introduced. Each class is randomly composed of 1,000 data samples. The proposed algorithm framework is used to process samples, and the parameter of the filter function $g$ is $\tau = 25$. Figure 19 is an example of the original image. Figure 20 shows the result of the manifold dimension reduction and deviation values of the dimension reduction results with the distribution parameters.

As can be seen in Figure 20, the dimension reduction results have obvious class discrimination, and the deviations by noise interference are nonlinear and have class discrimination.
Figure 13: Dimension reduction of the mean vectors and deviation values in low-dimensional space of each anchor vector.

(a) ISOMAP method of dimension reduction of the mean vectors
   - The deviation values (negative direction)
   - The deviation values (positive direction)

(b) LE method of dimension reduction of the mean vectors
   - The deviation values (negative direction)
   - The deviation values (positive direction)

(c) LLE method of dimension reduction of the mean vectors
   - The deviation values (negative direction)
   - The deviation values (positive direction)

(d) LTSA method of dimension reduction of the mean vectors
   - The deviation values (negative direction)
   - The deviation values (positive direction)

Figure 14: Continued.
Figure 14: Dimension reduction and deviation value results for entire noisy S-shaped dataset.

Figure 15: Example of original "0" and "1" images and images with noise.

Figure 16: Continued.
Figure 16: Manifold dimension reduction results of “0” and “1” and results of deviation values.

Figure 17: Example of original “0,” “1,” and “2” images and images with noise.

Figure 18: Continued.
Figure 18: Manifold dimension reduction results for “0,” “1,” and “2” and results of deviation values.

Figure 19: Examples of original “T-shirt” and “boots” and images with added noise.

Figure 20: Continued.
**Figure 21:** Examples of original “T-shirt,” “pants,” “boots,” and examples with added noise.

**Figure 20:** Manifold dimension reduction results for two-class samples and results of deviation values.

**Figure 22:** Continued.
4.4.2. Three-Class Samples. We arbitrarily select the three-class samples of "T-shirt," "pants," and "boots." The setting of the experiment is the same as that of the two-class process. Gaussian noise with mean of 0 and standard deviation of 0.5 is introduced. We randomly take 500 data samples from each class for dimension reduction processing. Figure 21 is an example of the original image. The experimental results are shown in Figure 22.

As can be seen in Figure 22, the dimension reduction result visually retains the class discrimination, and the deviation values from the noise interference are not only nonlinear but also have a visual class discrimination.

5. Conclusion

This paper presents a manifold-based dimension reduction algorithm framework capable of processing noisy data. Considering the manifold localization and the distribution characteristics brought about by noise interference, we propose a method for determining local anchor vectors using graph sampling and a method for determining the local ranges of anchor vectors based on spectral graph wavelet. In each local range of an anchor vector, maximum likelihood estimation is used to estimate the distribution parameters, and a distance-weighted deviation value calculation method for dimension reduction results with distribution parameters is proposed. Among these, dimension reduction can be adopted as the currently used manifold learning methods.

As seen from the simulation of noisy data, the proposed framework can fit for other kinds of dimension reduction methods and in this way to extend the functionalities of those methods to take noisy data as input to be applied in more complex situations.

This paper will further research on various manifold dimension reduction calculations with other types of distribution parameters, optimization of filter functions, and studying of quantitative evaluation methods.

Data Availability

The paper used the public MNIST and fashion-MNIST dataset.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (no. 61903029), National Key R&D Program of China (no. 2017YFB0702100), and National Material Environmental Corrosion Platform.

References

[1] B. Mikhail, N. Partha, and S. Vikas, “Manifold regularization: a geometric framework for learning from labeled and unlabeled examples,” *Journal of Machine Learning Research*, vol. 7, no. 1, pp. 2399–2434, 2006.
[2] J. B. Tenenbaum, V. de Silva, and J. Langford, “A global geometric framework for nonlinear dimensionality reduction,” *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.
[3] S. T. Roweis and L. Saul, “Nonlinear dimensionality reduction by locally linear embedding,” *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.
18 Complexity

[4] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," Neural Computation, vol. 15, no. 6, pp. 1373–1396, 2003.

[5] Z. Zhang and H. Zha, "Principal manifolds and nonlinear dimensionality reduction via tangent space alignment," SIAM Journal on Scientific Computing, vol. 26, no. 1, pp. 313–338, 2004.

[6] M. Hein and M. Maier, "Manifold denoising," in Proceedings of the 19th International Conference on Neural Information Processing Systems, pp. 561–568, Vancouver, Canada, December 2006.

[7] D. Gong, F. Sha, and G. Medioni, "Locally linear denoising on image manifolds," in Proceedings of the 13th International Conference on Artificial Intelligence and Statistics, pp. 265–272, Sardinia, Italy, May 2010.

[8] Y. Zhao, Z. Zhang, J. Qin, and L. Zhang, "Semi-supervised local multi-manifold Isomap by linear embedding for feature extraction," Pattern Recognition, vol. 76, pp. 662–678, 2018.

[9] K. Kim and J. Lee, "Nonlinear dynamic projection for noise reduction of dispersed manifolds," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 11, pp. 2303–2309, 2014.

[10] A. Rajagopal, R. B. Hamilton, and F. Scalzo, "Noise reduction in intracranial pressure signal using causal shape manifolds," Biomedical Signal Processing and Control, vol. 28, pp. 19–26, 2016.

[11] C. Jin and C. M. Bachmann, "Modeling and mitigating noise in graph and manifold representations of hyperspectral imagery," in Proceedings of the Algorithms and Technologies for Multispectral, Hyperspectral, and Ultrastructural Imagery XXI, Baltimore, MA, USA, May 2015.

[12] Z. Zhang, Y. Zhang, F. Li, M. Zhao, L. Zhang, and S. Yan, "Discriminative sparse flexible manifold embedding with novel graph for robust visual representation and label propagation," Pattern Recognition, vol. 61, no. 1, pp. 492–510, 2017.

[13] Z. Zhang, T. Chow, and M. Zhao, "M-Isomap: orthogonal constrained marginal Isomap for nonlinear dimensionality reduction," IEEE Transactions on Cybernetics, vol. 43, no. 1, pp. 180–191, 2013.

[14] Z. Zhang, T. W. S. Chow, and M. Zhao, "Trace ratio optimization-based semi-supervised nonlinear dimensionality reduction for marginal manifold visualization," IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 5, pp. 1148–1161, 2013.

[15] Z. Hao, J. Liu, S. Ma, X. Jin, and X. Lian, "Noise-removal method for manifold learning," in Proceedings of the International Conference on Intelligent Computing, Networked Control, and Their Engineering Applications, pp. 191–200, Nanjing, China, September 2017.

[16] Y. Zhang, Z. Zhang, S. Li et al., "Unsupervised nonnegative adaptive feature extraction for data representation," IEEE Transactions on Knowledge and Data Engineering, vol. 31, no. 12, pp. 2423–2440, 2019.

[17] J. Yan, H. Sun, H. Chen, N. Junejo, and E. Cheng, "Resonance-based time-frequency manifold for feature extraction of ship-radiated noise," Sensors, vol. 18, no. 4, pp. 936–957, 2018.

[18] D. Wang, Z. Wang, and C. Wen, "Distributed optimal consensus control for a class of uncertain nonlinear multiagent networks with disturbance rejection using adaptive technique," IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 99, 2019.

[19] S. Li and J. Lian, "Almost sure stability of Markov jump systems with persistent dwell time switching," IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 99, 2020.

[20] Z. Zhang, Y. Zhang, G. Liu, J. Tang, S. Yan, and M. Wang, "Joint label prediction based semi-supervised adaptive concept factorization for robust data representation," IEEE Transactions on Knowledge and Data Engineering, vol. 32, no. 5, pp. 952–970, 2020.

[21] Q. Ye, L. Fu, Z. Zhang, H. Zhao, and M. Naiem, "Lp- and ls-distance based robust linear discriminant analysis," Neural Networks, vol. 105, pp. 393–404, 2018.

[22] Z. Zhang, F. Li, M. Zhao, L. Zhang, and S. Yan, "Robust neighborhood preserving projection by nuclear/l2,1-norm regularization for image feature extraction," IEEE Transactions on Image Processing, vol. 26, no. 4, pp. 1607–1622, 2017.

[23] S. K. Narang and A. Ortega, "Downsampling graphs using spectral theory," in Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4208–4211, Prague, Czech Republic, May 2011.

[24] D. K. Hammond, P. Vandergheynst, and R. Gribonval, "The spectral graph wavelet transform: fundamental theory and fast computation," Vertex-Frequency Analysis of Graph Signals, Springer, Berlin, Germany, pp. 141–175, 2019.

[25] S. T. Ali, J. P. Antoine, and J. P. Gazeau, "Wavelets on manifolds," Coherent States, Wavelets, and Their Generalizations, Springer, Berlin, Germany, pp. 457–493, 2014.

[26] S. Deutsch, A. Ortega, and G. Medioni, "Graph-based manifold frequency analysis for denoising," 2016, http://arxiv.org/abs/1611.09510.

[27] D. I. Shuman, M. J. Faraji, and P. Vandergheynst, "A multiscalar pyramid transform for graph signals," IEEE Transactions on Signal Processing, vol. 64, no. 8, pp. 2119–2134, 2016.

[28] S. Deutsch, A. Ortega, and G. Medioni, "Manifold denoising based on spectral graph wavelets," in Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing, Shanghai, China, March 2016.

[29] B. H. Wellenhof and H. Moritz, "Introduction to spectral analysis," Mathematical and Numerical Techniques in Physical Geodesy, vol. 7, pp. 157–259, 2006.