01 Apr 2018

A Multi-Step Nonlinear Dimension-Reduction Approach with Applications to Bigdata

R. Krishnan

V. A. Samaranayake
Missouri University of Science and Technology, vsam@mst.edu

Jagannathan Sarangapani
Missouri University of Science and Technology, sarangap@mst.edu

Recommended Citation

R. Krishnan et al., "A Multi-Step Nonlinear Dimension-Reduction Approach with Applications to Bigdata," *Procedia Computer Science*, vol. 144, pp. 81-88, Elsevier, Apr 2018.
The definitive version is available at https://doi.org/10.1016/j.procs.2018.10.507

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 144 (2018) 81–88

Procedia Computer Science

www.elsevier.com/locate/procedia

INNS Conference on Big Data and Deep Learning 2018

A Multi-step Nonlinear Dimension-reduction Approach with Applications to Bigdata

R. Krishnan[a,*], V.A. Samaranayake[b], S. Jagannathan[a]

[a] *Department of Electrical and Computer Engineering, Missouri University of Science and Technology, USA*
[b] *Department of Mathematics and Statistics, Missouri University of Science and Technology, USA*

Abstract

In this paper, a multi-step dimension-reduction approach is proposed for addressing nonlinear relationships within attributes. In this work, the attributes in the data are first organized into groups. In each group, the dimensions are reduced via a parametric mapping that takes into account nonlinear relationships. Mapping parameters are estimated using a low rank singular value decomposition (SVD) of distance covariance. Subsequently, the attributes are reorganized into groups based on the magnitude of their respective singular values. The group-wise organization and the subsequent reduction process is performed for multiple steps until a singular value-based user-defined criterion is satisfied. Simulation analysis is utilized to investigate the performance with five big data-sets.

Keywords:

1. Introduction

Typically, bigdata sets are accompanied by a large number of incomplete and redundant dimensions where [5] complications such as nonlinear relationships, noise, spurious correlations and incidental endogeneity are common. Within this domain, classification is an important problem and the challenges from big data result in increasing prediction errors [5].

To mitigate the issue, it is common to extract relevant features using dimension-reduction approaches such as principal component analysis (PCA), factor analysis, etc [14]. However, when these methodologies [14] are utilized in the presence of nonlinearities[14], imperfect estimation is observed because correlation matrices cannot measure nonlinear relationships [9].

On the other hand, to handle nonlinear relationships, several dimension-reduction techniques such as Isomap [2], LLE [12], Hessian LLE [6], Laplacian eigenmaps [3] and kernel PCA [13] have been introduced. Due to the properties

[*] Corresponding author.
E-mail address: krm9c@mst.edu

10.1016/j.procs.2018.10.507

of high dimensional spaces [7], it can be challenging to use these methodologies because they aim to discover the geometric structure of the data.

In the above-mentioned dimension-reduction approaches [10, 2, 12, 6, 3, 13], a one-step mapping for dimension reduction is common. A one-step approach is however susceptible to improper estimation due to noise and redundancies and can incur large computational cost. Therefore, multi-step dimension reduction approaches are preferable to address the challenges mentioned earlier[7] and motivated by the challenges, one such approach (NDR) is presented in this paper.

In this paper, dimensions are reduced while considering nonlinear relationships using distance covariance. However, a distance covariance matrix characterizing pairwise dependencies can be rank-deficient if the number of sample points n are significantly smaller than dimensions/attributes p in the data [9]. To mitigate this problem, group-wise organization of the attributes is introduced, where the size of the group is kept smaller than the number of observations.

The p attributes are initially organized at random into groups. Next, nonlinear relationships are measured in each group using distance covariance [9]. For each group, the features are transformed with a parametric map constructed using the singular value decomposition (SVD) of distance covariance. The cardinality of the transformed space is set by a user-defined parameter determining the amount of information to be captured in each group.

The new set of attributes are then reorganized into groups but, now, the magnitude of the singular values is utilized to enable organization. In the proposed approach, the group-wise organization is based on SVD results, in contrast with [1], where application specific relationships could be incorporated independent of data analysis.

The group-wise organization and reduction procedure is repeated until a stopping criterion is satisfied. A user-defined information loss criterion based on singular values is therefore defined. Due to the proposed criterion, NDR can estimate the number of dimensions to which to reduce the data to, in contrast with traditional approaches [10, 2, 12, 6, 3]. Furthermore, it is generally difficult to specify the number of dimensions that must be extracted from the data [10, 2, 12, 6, 3] because information about their usefulness is unknown and requires trial and error. In contrast, specifying the percentage of information to be kept during dimension reduction, as done with NDR, is a more natural choice.

Finally, the performance of NDR is demonstrated on five big data sets using standard classification methods. The contributions of this paper therefore include: (1) development of a multi step dimension-reduction approach using distance covariance while handling nonlinear relationships and noisy dimensions; (2) design of a generic group-wise organization process with explicit stopping criterion to control information loss and perform organization; (3) demonstration of the performance using five data-sets in the problem of classification and fault diagnostics.

The rest of the paper is organized as follows. Motivation are established in Section II. NDR is described in Section III. Simulation results are outlined in Section IV while Section V provides the conclusions for the paper. The preliminaries are discussed next.

2. Preliminaries and Distance Covariance

Let $\mathbb{R}$ to represent the set of real numbers and denote matrices and vectors by boldface. Let a sample of data be denoted as $\boldsymbol{\mathcal{Y}} \in \mathbb{R}^{n \times p}$, where n represents the number of sample points and p is total number of attributes and consider a data-point $\boldsymbol{y} \in \boldsymbol{\mathcal{Y}}$. The objective of this work is to predict the category for $\boldsymbol{y}$ by transforming $\boldsymbol{x}$ into $\boldsymbol{y}$ using $\phi(.)$, where $\arg\max(\boldsymbol{y})$ indicates the category for $\boldsymbol{x}$ such that

$$p(\boldsymbol{y} \in \boldsymbol{\Psi}_k | \boldsymbol{\mathcal{X}}) = \phi(\boldsymbol{x}) + \varepsilon. \tag{1}$$

where $p(\boldsymbol{y} \in \boldsymbol{\Psi}_k | \boldsymbol{\mathcal{X}})$ is the vector of $\mathbb{R}^{\mathcal{F} \times 1}$ and ε is the approximation error. In fault diagnostics problems, prediction involves detecting whether $\boldsymbol{y}$ belongs to the healthy case or is at one of the faults $\mathcal{F}$. Statistically speaking, one can consider that the data-points in $\boldsymbol{\mathcal{Y}}$ coming from category k belong to population $\boldsymbol{\Psi}_k$ where $k = 1, 2, 3, \cdots, \mathcal{F}$. One may observe that the objective is to determine to which population does $\boldsymbol{y}$ belongs and $p(\boldsymbol{y} \in \boldsymbol{\Psi}_k | \boldsymbol{\mathcal{X}})$ measures the probability of $\boldsymbol{y}$ belonging to each k.

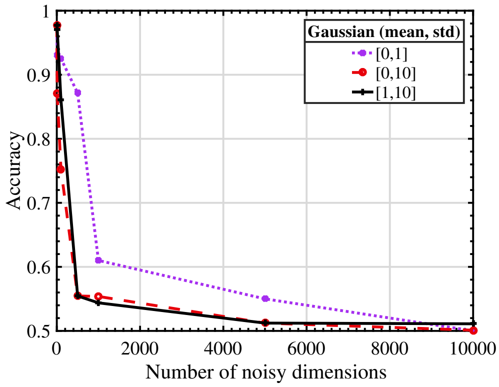

Fig. 1: Change in accuracy with increase in the number of noisy dimensions.

Algorithm 1 A Nonlinear dimension-reduction (NDR)

```
Input: $\boldsymbol{\mathcal{X}}$, $\theta$
Output:Reduced dimension $\boldsymbol{\mathcal{X}}$
Let $\boldsymbol{\mathcal{X}}^{(\mathbf{1})} = \boldsymbol{\mathcal{X}}$
Standardize $\boldsymbol{\mathcal{X}}^{(1)}$
for Each step starting from $i = 1$
    if $\frac{1}{T^{(i)}} \sum_{t=0}^{T^{(i)}} \sum_{l=0}^{\kappa} \lambda_l \geq \theta$
        Stop the dimension-reduction procedure
    end if
    if i>1
        Create groups according to Eq (5)
    else
        Generate groupings at random
    end if
    for Every group at step $i$
        Calculate DC matrix
        Evaluate $\kappa$
        Evaluate low rank approximation [16]
    end for
end for
Train the regression parameter [4]
```

To learn the map $\phi(.)$, we need representative samples from each $\boldsymbol{\Psi}_k$, therefore consider a dataset $\mathcal{D} = \{\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{T}}\}$ collected from $\mathcal{F}$ categories, where $\boldsymbol{\mathcal{X}}$ represents the observations with $\boldsymbol{\mathcal{T}}$ referring to the corresponding labels. Note that Eq. (1) is a generic structure for a parametric map based classification regimes, for example, neural networks (NN), logistic regression (LR), etc [4].

While classifying data-points in the big data scenario, the number of dimensions in the data can adversely effect the classification efficiency. To observe this effect, consider classification with the MNIST hand-written recognition [8] data-set. A three layer NN is used for classification and noisy dimensions are introduced synthetically. Observe that, when the number of noisy dimensions in the data increase, the accuracy falls drastically as observed from Fig. 1.

In such cases, it is desirable to reduce the number of dimensions prior to classification such that the noisy dimensions are removed. Specifically, noisy dimensions refer to those that are redundant and to remove these dimensions, we require a measure of redundancies among the dimensions. Pearson's correlation and/or covariance is widely used in such applications [4].

The fundamental drawback arises because Pearson's correlation and/or covariance coefficient is not a true characterization of independence when two dimensions are nonlinearly related [9]. A measure of dependency that can characterize independence in the presence of nonlinear relationships was proposed by Professor Szekely [15] and is known as distance covariance (DC). For illustration, let $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ and $p = 2$ such that $\boldsymbol{x} = [\boldsymbol{a} \quad \boldsymbol{b}]$. It follows that $\boldsymbol{a}$ and $\boldsymbol{b}$ are univariate vectors of sample size n each. DC is defined as follows.

Definition 1. *Let pair-wise distances for each element in $\boldsymbol{a}$ be written as $\boldsymbol{A}_{m,l} = \|\boldsymbol{a}_m - \boldsymbol{a}_l\|, \; \forall \quad m,l = 1,2,\cdots n$. Similarly, the pairwise distances for $\boldsymbol{b}$ is denoted as $\boldsymbol{B}_{m,l} = \|\boldsymbol{b}_m - \boldsymbol{b}_l\|, \; \forall \quad m,l = 1,2,\cdots n$. Double-center the distance matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ to get $\tilde{\boldsymbol{A}}$ and $\tilde{\boldsymbol{B}}$ such that*

$$\tilde{\boldsymbol{A}}_{m,l} = \begin{cases} \boldsymbol{A}_{m,l} - \frac{1}{n}\mathbf{1}_n\boldsymbol{A} - \\ \frac{1}{n}\boldsymbol{A}\mathbf{1}_n^T + \frac{1}{n^2}\mathbf{1}_n\boldsymbol{A}\mathbf{1}_n^T & m \neq l \\ 0 & m == l, \end{cases} \quad \tilde{\boldsymbol{B}}_{m,l} = \begin{cases} \boldsymbol{B}_{m,l} - \frac{1}{n}\mathbf{1}_n\boldsymbol{B} - \\ \frac{1}{n}\boldsymbol{B}\mathbf{1}_n^T + \frac{1}{n^2}\mathbf{1}_n\boldsymbol{B}\mathbf{1}_n^T & m \neq l \\ 0 & m == l, \end{cases} \tag{2}$$

where $\mathbf{1}_{(.)}$ denotes the vector of ones with length (.). The sample DC is defined under finiteness of the first order moment for $\boldsymbol{a}$ and $\boldsymbol{b}$ as

$$\nu_n^{\boldsymbol{a},\boldsymbol{b}} = \sqrt{\frac{1}{n^2}\mathbf{1}_n\tilde{\boldsymbol{A}} \odot \tilde{\boldsymbol{B}}\mathbf{1}_n^T}, \tag{3}$$

where Hadamard product is denoted by $\odot$.

The proposed nonlinear dimension-reduction approach based on DC is presented next.

3. Nonlinear Dimension-reduction (NDR)

In bigdata applications, the sheer volume of dimensions is a major issue. As a result, in each step of the dimension reduction procedure, we propose to divide the number of dimensions in the data into groups. Next, we reduce the dimensions in each group while taking into account nonlinear relationships within these dimensions. Finally, we merge the newly extracted dimensions. The above explained split and merge process is repeated until pre-defined condition is satisfied. The overall procedure is detailed in Algorithm. 1 and the details of NDR are as follows.

For illustration, consider a simple case, where the sample $\boldsymbol{X}$ consists of only eight dimensions such that $p = 8$. Denote the data at the first step as $\boldsymbol{X}^{(1)} = \{x_j^{(1)}, j = 1,2,3,\cdots,8\}$ such that $x_j^{(1)}$ denotes the j^{th} coordinate in the data vector $\boldsymbol{X}^{(1)}$. First, the eight attributes are grouped as $(1,2),(3,4),(5,6),(7,8)$ such that $(x_1^{(1)}, x_2^{(1)})$ form group $t = 1$ and $(x_3^{(1)}, x_4^{(1)})$ form group $t = 2$ and so on. After the initial grouping, the proposed methodology is applied in two stages.

Stage 1. For each group, let $\boldsymbol{C}_1^{(1)}$ be the DC matrix estimated for the pair $(x_1^{(1)}, x_2^{(1)})$ such that

$$\boldsymbol{C}_1^{(1)} = \begin{bmatrix} \nu_n^{x_1^{(1)},x_1^{(1)}} & \nu_n^{x_2^{(1)},x_1^{(1)}} \\ \nu_n^{x_1^{(1)},x_2^{(1)}} & \nu_n^{x_2^{(1)},x_2^{(1)}} \end{bmatrix}.$$

As observed, the matrix $\boldsymbol{C}_1^{(1)}$ is symmetric and of size 2 × 2, where group size is two and the elements represent the strength of pairwise relationships. By using distance covariance to measure the relationships, we ensure that even when nonlinear relationships are present within the dimensions, the strength of the relationships is captured.

Under the assumption that the amount of information contained by a dimension is proportional to the magnitude of variance, the objective is to transform the data into a new dimensional space such that the redundancies among the two attributes in the group are minimized while the variance, that is the amount of information to be retained, is maximized. Since singular values of $\boldsymbol{C}_1^{(1)}$ signify the variance of each dimension in the group, the transformation is constructed using the singular vectors of $\boldsymbol{C}_1^{(1)}$ [11].

To this end, the singular values are first normalized such that they lie between zero and one. Next, the size of the projected space is determined depending on a threshold θ set by the practitioner which refers to the percentage of information to be retained in each group. Then, all singular values with magnitude greater than $\theta/100$ are retained and the rest of the dimensions are discarded. Using the selected singular values, a low dimensional approximation corresponding to $\boldsymbol{C}_t^{(1)}$ is computed. To compute the approximation, one procedure is given in [16]. The transformation parameters that best extract $\theta\%$ variation from the group are obtained as a result of this approximation.

With the current example, let $\theta = 95\%$ and assume that in each group, the number of dimensions to be extracted is one. Using the low dimensional approximation of SVD, we transform the data in this group. Next, the above explained reduction procedure is repeated for each of the four groups to achieve the reduced number of dimensions at the first step. In the example illustrated above, the dimension are reduced to four after the first step.

Let $\boldsymbol{P}_t^{(1)}$ be the transformation parameters obtained from the approximation in each group t. The transformation for all the data-points in $\boldsymbol{X}^{(1)}$ is given as

$$\hat{\boldsymbol{X}}^{(2)} = \{\boldsymbol{X}_t^{(1)}\boldsymbol{P}_t^{(1)}, \forall t = 1, 2, 3, \cdots, T^{(1)}\}, \tag{4}$$

where $T^{(1)}$ is the number of groups at step one. Note that $\boldsymbol{X}_t^{(1)}$ denotes the data in each group therefore for the example considered here $\boldsymbol{X}_t^{(1)} \in \mathbb{R}^{n\times 2}$ and after transformation $\hat{\boldsymbol{X}}_t \in \mathbb{R}^{n\times 1}$. Next, data is aggregated from all the groups to achieve $\hat{\boldsymbol{X}}^{(2)} = [\hat{\boldsymbol{X}}_1^{(2)} \quad \hat{\boldsymbol{X}}_2^{(2)} \quad \hat{\boldsymbol{X}}_3^{(2)} \quad \hat{\boldsymbol{X}}_4^{(2)}]$ which is the data available after dimension reduction at the first step.

Since the information contained in a dimension is proportional to the singular values, an average of singular values across all the group provides an idea about the amount of information that is retained at a step. To control loss of information, this average of singular values, verified after each step of dimension reduction can be used as a stopping criterion. Therefore, the dimension reduction procedure is stopped when the average information retained at each step, becomes less than the information threshold set by the practitioner. For illustration, let the average of the singular values after the first step be 0.98, which implies that 98% of the information is retained. Since the threshold is 95%, information loss condition is not satisfied yet and therefore stage two of the dimension reduction procedure is initiated.

Stage 2. In this stage, the first objective is to group the attributes. Therefore, denote the set of data-points from the first step in the dimension reduction procedure as $\boldsymbol{X}^{(2)} = [x_1^{(2)} \quad \cdots x_4^{(2)}]$. Note that the dimensions are ordered according to the normalized singular values.

Next, these new dimensions are grouped according to the strength of singular values. In other words, the dimensions with smaller singular values are grouped with dimensions indicating large singular values. By doing this, it is ensured that the dimensions with large variance do not come in one group. Thus, for every $t \in \{1, 2, \cdots, T^{(2)}\}$ the groups are given as

$$\boldsymbol{X}_t^{(2)} = \cup_{l=1}^{\eta^{(1)}} \hat{\boldsymbol{X}}_j^{(2)} : j = i \times T^{(2)}, j < \eta^{(1)} \forall \quad t = 1, 2, \cdots, T^{(2)}, \tag{5}$$

where $\eta^{(1)}$ is the number of attributes at a particular step. Using the grouping mechanism described in Eq. (5), two groups can be constructed for our example. The first attribute belongs to the same group as the third attribute. Furthermore, the second and the fourth attribute form the second group.

The two stages described above constitute a single step in NDR. The two stages of the proposed procedure are continued till the singular value condition fails. For the example considered here, the dimension reduction process is continued for two steps because at the end of the second transformation, the average of the singular values became 94% which is less that 95%. Finally, at the end of the two steps, the overall procedure has reduced the number of dimensions from eight to two.

The dimension-reduction process, need not progress for two steps. The idea is to choose θ; such that it is feasible to perform classification with the remaining dimensions. In a general case, there may be I steps is the dimension-reduction procedure. The data after I transformation may be written as $\boldsymbol{X}^{(I)} = \Phi(\boldsymbol{X}^{(1)})$, where Φ is the dimension reducing transformation.

With this transformation, the final step in the proposed methodology is the process of regression/ classification on the dimension-reduced data where it is desirable to estimate the conditional probability $p(\boldsymbol{y} \in \boldsymbol{\Psi}_k|\boldsymbol{\mathcal{X}})$. As any parametric map can be used for classification, traditional logistic regression is considered in this paper [4]. To this end, every data-point $\boldsymbol{y} \in \boldsymbol{\mathcal{Y}}$ is tested by performing the dimension-reducing transformation on the observations using the transformation parameters obtained while transforming $\boldsymbol{\mathcal{X}}$. A typical classification process with softmax regression is used to determine the probabilities in order to determine the fault/class [4].

Testing can be performed batch-wise or observation-wise. For batch-wise prediction , one would like to take the average across the batch. Next section presents the application of NDR.

4. Results and Discussions

A total of five data-sets are used for analysis in this study and are tabulated in Table 1. Eighty percent of the data is randomly selected for training and the rest is used for testing. Fault diagnostics performance of the methodology will be shown first using the sensorless drive diagnostics data-set. The performance of NDR on classification methods is shown next. All results are averaged over 1000 runs and the hyper parameters for all classification approaches are shown in Table 2.

4.1. Fault Diagnostics

The sensorless drive diagnostics data-set presents an 11-fault estimation problem involving 48 dimensions. The proposed dimension-reduction approach is applied in conjunction with various classification methodologies. The resulting accuracies are shown at the first row in Table 3. Based on these results, NDR appears to capture information accurately as observed from the high accuracies for different classification methods. Suboptimal results are observed for DT, RF, Ada-boost and Naive-Bayes while using this data-set.

Next, different dimension-reduction approaches are compared and the results are summarized in Table 4 [1]. Better accuracies are observed while using NDR relative to other dimension-reduction methods. There is a 24 % improvement in accuracy over PCA with the sensorless drive diagnostics data-set.

In all of these tests, the data are reduced from 48 dimensions to 25. Since the sensorless drive data-set contains nonlinear relationships, PCA is expected to fail as seen from the results. In the next section, performance of NDR on classification is studied.

4.2. Classification

The results are summarized in Table 3. NDR is observed to achieve reasonable performance for all the data-sets that includes linear and nonlinear relationships.

NDR addresses the data-sets with $p < n$ well, examples include the Arcene, Gisette and Dexter data-sets. Accuracies beyond 81 % are observed in each case. The best case accuracy for the high dimensional data-sets are observed while using NDR. The best accuracies are highlighted in bold in Table 3. Twenty five dimensions are extracted for each of the dimension reduction approaches and θ is kept at 0.75 for NDR, which yields 25 dimensions.

Impressive performance for NDR is observed even when traditional methods are seen to achieve sub-optimal performance. Moreover, even when, the number of observations are fewer than the number of dimensions, optimal performance is observed.

5. Conclusions

In this paper, a multi-step dimension-reduction approach was proposed to address nonlinear relationships and noisy dimensions. Due to the use of an information-loss criterion, NDR can dynamically determine the number of

[1] Acronyms for different dimension-reduction techniques are **PCA** (Principal Component Analysis), **ISOMAP** (Isometric Mapping), **FastICA** (Independent Component Analysis), **LLE** (Locally Linear Embedding) and **NDR** (Nonlinear dimension-reduction).

Table 1: Summary descriptions of the different data-sets used in this paper

Data-set	**Dimensions**	**Data points**	**Classes**
Sensorless [8]	48	78000	11
CIFAR-10 [8]	3072	50000	10
Gisette [8]	5000	6000	2
Arcene [8]	10000	100	2
Dexter[8]	20000	300	2

Table 2: Hyper-parameters for different methodologies in this paper.

Method	**Hyper-parameters**
KNearest Neighbors (KNN)	3 neighbors
Support Vector Machine (SVM)	C = 0.025
Kernel SVM	RBF kernel, gamma = 2, C=1
Desicion Trees (DT)	max depth =5
Random Forest (RF)	max depth=5 estimators=10, max features=1
Shallow Neural Network (SNN)	lr = 0.01, 1 hidden layer, 500 neurons
AdaBoost	number of estimators=50, learning rate=1.0
Naive-Bayes	No Priors
Quadratic Discriminant Analysis (QDA)	No Priors
Logistic Regression (LR)	Random initialization, learning rate = 0.001, 1000 iterations

Table 3: Accuracies for the various data-sets with different methods using NDR as the dimension-reduction methodology with type one error in parentheses. These results are highlighted in bold. The value of θ is chosen as 0.95 for all of the data-sets.

Data-sets ↓	**KNN**	**SVM**	**LDA**	**DT**	**RF**	**SNN**	**AdaBoost**	**Naive-Bayes**	**QDA**	**LR-NDR**
Sensorless	0.88(0.008)	0.91 (0.01)	0.60(0.016)	0.50(0.013)	0.95(0.021)	0.47(0.004)	0.74 (0)	0.83 (0.056)	0.84(0.069)	**0.93(0.004)**
CIFAR-10	0.29(0.080)	0.44(0.047)	0.37(0.056)	0.26(0.048)	0.21(0.090)	0.50(0.036)	0.51(0.077)	0.36 (0.074)	0.44(0.021)	**0.66(0.036)**
Arcene	0.83 (0)	0.84 (0)	0.55 (0)	0.54 (0)	0.83 (0)	0.72 (0)	0.74 (0)	0.51 (0)	0.76 (0)	**0.87 (0)**
Dexter	0.48 (0)	0.54 (0)	0.54 (0)	0.77 (0)	0.53 (0)	0.57 (0)	0.69 (0)	0.52 (0)	0.59 (0)	**0.81(0.015)**
Gisette	0.94(0.031)	0.93(0.012)	0.91(0.052)	0.92(0.049)	0.74(0.074)	**0.98**(0.015)	0.97(0.019)	0.74 (0.256)	0.63(0.126)	**0.99(0.015)**

Table 4: Average classification accuracy (**Avg. Class. Acc.**) rates and computational times (**Comp. Time**) for various dimensions-reduction techniques (**Dim-red**) and classifiers using the sensorless drive diagnostics data-set.

Dim-red	**KNN**	SVM	LDA	DT	RF	SNN	AdaBoost	Naive-Bayes	QDA
PCA	0.52	0.52	0.52	0.45	0.45	0.52	0.18	0.45	0.53
ISOMAP	0.67	0.46	0.43	0.32	0.39	0.58	0.18	0.42	0.49
LLE	0.25	0.08	0.15	0.18	0.18	0.08	0.18	0.19	0.19
KPCA	0.54	0.53	0.52	0.45	0.44	**0.58**	0.28	0.45	0.53
NDR	**0.89**	**0.91**	**0.60**	**0.50**	**0.95**	**0.47**	**0.74**	**0.83**	**0.84**

dimensions to be extracted in contrast with standard dimension-reduction procedures. Due to the use of distance covariance, NDR was able to capture nonlinear relationships from the data while exhibiting 24% improvement over PCA for the sensorless drive diagnostics data-set.

Acknowledgments

The authors would like to thank Dr. Elizabeth Bowles, Research Associate, Biological Sciences Department, for her help in improving the written flow. This research was supported in part by NSF I/UCRC award IIP 1134721 and Intelligent Systems Center.

References

[1] Adragni, K.P., Al-Najjar, E., Martin, S., Popuri, S.K., Raim, A.M., 2016. Group-wise sufficient dimension reduction with principal fitted components. Computational Statistics 31, 923–941.

[2] Balasubramanian, M., Schwartz, E.L., 2002. The isomap algorithm and topological stability. Science 295, 7–7.

[3] Belkin, M., Niyogi, P., 2003. Laplacian eigenmaps for dimensionality reduction and data representation. Neural computation 15, 1373–1396.

[4] Bishop, C.M., 2006. Pattern recognition and machine learning. springer.

[5] Clarke, R., Ressom, H.W., Wang, A., Xuan, J., Liu, M.C., Gehan, E.A., Wang, Y., 2008. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. Nature Reviews Cancer 8, 37–49.

[6] Donoho, D.L., Grimes, C., 2003. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. Proceedings of the National Academy of Sciences 100, 5591–5596.

[7] Fan, J., Han, F., Liu, H., 2014. Challenges of big data analysis. National science review 1, 293–314.

[8] Guyon, I., Gunn, S., Ben-Hur, A., Dror, G., 2005. Result analysis of the nips 2003 feature selection challenge, in: Advances in neural information processing systems, pp. 545–552.

[9] Huo, X., SzÃ©kely, G.J., 2016. Fast computing for distance covariance. Technometrics 58, 435–447. URL: http://dx.doi.org/10.1080/00401706.2015.1054435, doi:10.1080/00401706.2015.1054435, arXiv:http://dx.doi.org/10.1080/00401706.2015.1054435.

[10] Johnson, R.A., Wichern, D.W., 1992. Applied multivariate statistical analysis. volume 4. Prentice hall Englewood Cliffs, NJ.

[11] Jolliffe, I., 2002. Principal component analysis. Wiley Online Library.

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

[13] Schölkopf, B., Smola, A., Müller, K.R., 1997. Kernel principal component analysis, in: International Conference on Artificial Neural Networks, Springer. pp. 583–588.

[14] Sorzano, C.O.S., Vargas, J., Montano, A.P., 2014. A survey of dimensionality reduction techniques. arXiv preprint arXiv:1403.2877 .

[15] Székely, G.J., Rizzo, M.L., Bakirov, N.K., et al., 2007. Measuring and testing dependence by correlation of distances. The annals of statistics 35, 2769–2794.

[16] Witten, D.M., Tibshirani, R., Hastie, T., 2009. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics 10, 515–534.