LINEAR AND NONLINEAR DIMENSIONALITY REDUCTION FOR FACE RECOGNITION

Weilin Huang and Hujun Yin

School of Electrical and Electronic Engineering, The University of Manchester
Manchester, M60 1QD, UK
weilin.huang@postgrad.manchester.ac.uk, h.yin@manchester.ac.uk

ABSTRACT

Principal component analysis (PCA) has long been a simple, efficient technique for dimensionality reduction. However, many nonlinear methods such as local linear embedding and curvilinear component analysis have been proposed for increasingly complex nonlinear data recently. In this paper, we investigate and compare linear PCA and various nonlinear methods for face recognition. Results drawn from experiments on real-world face databases show that both linear and nonlinear methods yield similar performance and differences in classification rate are insignificant to conclude which method is always superior. A nonlinearity measure is derived to quantify the degree of nonlinearity of a data set in the reduced subspace. It can be used to indicate the effectiveness of nonlinear or linear dimensionality reduction.

***Index Terms*—**PCA, dimensionality reduction, nonlinear manifold, face recognition

1. INTRODUCTION

With the fast increasing quantity and complexity of data in an information-rich age, it becomes difficult, challenging or even impossible for engineers or analysts to deal with raw data directly. Data preprocessing thus becomes an important and emerging topic in many data-driven applications such as image processing and bioinformatics. Dimensionality reduction provides an efficient way for data abstraction and representation as well as feature extraction. It aims to detect intrinsic structures of data and to extract a reduced number of variables (dimensions) that capture and retain the main features of the high-dimensional data. For instance, images contain a large number of pixel values and are presented as high-dimensional arrays. Operating directly on these arrays is inefficient and may lead to high computational and storage demands. Reducing dimensionality has thus become a desired and key technique in many image processing applications.

PCA is a primary technique and is regarded as the theoretical foundation of many dimension reduction techniques. It seeks a linear projection that best fits a data set in the least-square sense and has been widely used for feature extraction in pattern classification due to its computational and analytical simplicity [1]. Eigenface [2] is a famous application of PCA for face recognition. However, the linearity of PCA limits its power for complex data sets as it is unable to capture nonlinear structure of the data defined by beyond second order statistics. Several nonlinear techniques have been proposed recently for instance, kernel PCA [3], LLE [4], Isomap [5] and CCA [6]. A review can be found in [7].

There has been previous work on applying these linear and nonlinear projections for face recognition [8-11] and performances vary with training/test schemes, preprocessing methods and choices of classifier. Thus an objective comparison is desirable for evaluating these methods and their effectiveness. This paper investigates these methods on their capabilities for dimension reduction of face data. Comparisons are conducted on the same training/test scheme and choice of classifiers. Then a quantitative nonlinearity measure is proposed for analyzing the complexity of the data, which further supports the findings.

Section 2 describes the algorithms concerned. Experimental results are then shown in Section 3, followed by discussions and a nonlinearity analysis in Section 4. Section 5 concludes the paper.

2. ALGORITHM REVIEW

2.1. Dimensionality reduction methods

PCA is a classical linear projection aiming at finding orthogonal principal directions of a data set by solving an eigenvalue problem. While discarding a (large) number of minor components, a (small) number of principal components are retained on a linear, low-dimensional subspace, also known as *eigenface* in face recognition. The data projection from a high-dimensional space to the feature subspace is described as below,

$$y_{ik} = \mathbf{v}_k^T \mathbf{X}_i \tag{1}$$

where $\mathbf{X}_i$, i=1,2,…, N, is the i-th n-dimensional sample (image), $\mathbf{v}_k$ (k=1,2,…,d, $d \le n$) is the k-th eigenvector corresponding to the k-th largest eigenvalue of the covariance matrix of the data set, and y_{ik} is the k-th component of the projected data in the d-dimensional subspace. Raw images are projected onto the feature subspace first, and classification is conducted in the reduced space.

Kernel-PCA [3] is an extension of PCA to nonlinear mapping. Input data is projected onto a high-dimensional feature space (F) by using a hypothetic nonlinear function, $\Phi(X)$. Then the standard PCA is performed on the projected space via a kernel function, $k(X,Y)=(\Phi(X),\Phi(Y))$. The covariance matrix in projected space $\mathbf{K}$ is computed in via the kernel function, Eq. (2), and the projection of an image $\mathbf{X}_t$ to the feature subspace of F is shown in Eq. (3).

$$K_{ij} := k(\mathbf{X}_i, \mathbf{X}_j) = (\Phi(\mathbf{X}_i) \cdot \Phi(\mathbf{X}_j)) \tag{2}$$

$$(\mathbf{v}_k^f \cdot \Phi(\mathbf{X}_t)) = \mathbf{v}_k^T \mathbf{T} \tag{3}$$

where $\mathbf{v}^f_k$ and $\mathbf{v}_k$ (k=1,2,…,d) are the k-th eigenvector in F space and the k-th eigenvector of the covariance matrix $\mathbf{K}$, respectively. $\mathbf{T}$ is the projection of $\mathbf{X}_t$ on all training images, T_i:=$k(\mathbf{X}_i,\mathbf{X}_t)$ =$(\Phi(\mathbf{X}_i)\bullet\Phi(\mathbf{X}_t))$, i=1,2,…,N), $\mathbf{X}_i$ is the i-th training image. Two commonly used kernel functions are *polynomial* and *radial basis*, referred to as KPCA1 and KPCA2 in our experiments, respectively.

LLE [4] is a local PCA method and is able to map high-dimensional nonlinear data onto a single global coordinate system of lower dimensional subspace. The neighborhood is preserved in the embedding by minimizing cost functions in input space and output spaces respectively. The weights ω_{ij} of each input data $\mathbf{X}_i$ from its neighbors $\mathbf{X}_j$ are computed by minimizing the cost function, $E(\omega)$, and the output vector $\mathbf{Y}_i$ reconstructed from ω_{ij} by minimizing the embedding cost function, $E(Y)$,

$$E(\omega)=\sum_i\left\|\mathbf{X}_i-\sum_j\omega_{ij}\mathbf{X}_j\right\|^2,\ E(Y)=\sum_i\left\|\mathbf{Y}_i-\sum_j\omega_{ij}\mathbf{Y}_j\right\|^2 \quad (4)$$

The optimal weights ω_{ij} are found by solving a least-square problem of the cost function in Eq. (4), while the embedding vectors $\mathbf{Y}_i$ in Eq. (4) are solved as an eigenvalue problem [4].

Isomap [5] seeks an underlying manifold structure of a data set by computing the geodesic manifold distances between all pairs of data points. It first defines a neighborhood graph over all data points by connecting each point to all its neighbors in the input space. Then it estimates geodesic distances of all pairs of points by computing the shortest path distances in the neighborhood graph (e.g. using Floyd's algorithm). Then classical multidimensional scaling is applied to the distance matrix to construct the embedding that best preserves the intrinsic geometry structure of the data.

CCA [6] is another method for nonlinear mapping. It detects the intrinsic geometric properties of data by preserving local distance relationships via minimizing an error function defined as,

$$E=\frac{1}{2}\sum_i\sum_{j\neq i}(I_{ij}-O_{ij})^2\eta(O_{ij},\beta_y) \quad (5)$$

where I_{ij} and O_{ij} are the Euclidean distances between points i and j in n-D input space and d-D output space respectively. $\eta(O_{ij},\beta_y)$ is a monotonically decreasing neighborhood function respecting to the distance in the projected space and is used for preserving local topology and maintaining shorter distances than longer ones.

2.2. Classifiers

For classification, four common classifiers were used. The nearest-neighbor (NN) simply classifies a test sample by finding the most similar example in the training set and returning the class of that example. In soft k-NN classifier [12], each principal component outputs a confidence value, which gives the degree of support for the component in every face representation, and then the final decision is given by considering all of these confidence values. The linear discriminant analysis (LDA) [13], a widely used linear classifier, tries to find a linear projection of the data set that minimizes within-class scatter and maximizes between-class separation. The support vector machine (SVM) [14] is a nonlinear classifier which separates data sets by constructing hyperplanes that maximize the margins between the data sets.

3. EXPERIMENTS AND RESULTS

3.1. Experiments

In all implementations, raw face images were first preprocessed by one of the dimension reduction methods, and then classification was performed with one of the classifiers. Two publicly available real-world face databases, Olivetti Research Laboratory (ORL) and Yale, were used. ORL face database consists of 40 subjects, 10 different images for each subject. Images are of the same size of 92×112 and vary in term of lighting conditions, facial expressions or facial details. Yale database contains 165 face images of 15 subjects with size of 243 ×320. Each subject has 11 images with variation in both expression and lighting conditions.

3.2. Results

For an objective comparison, the performances of six dimension reduction methods were investigated on the same classifier in each experiment, and the results presented here are the typical performances of each method in the same reduced dimension (60 in this case). For ORL data, the number of training images varied from three to six per subject and the remaining seven to four were used for testing. The results are the average of ten independent implementations with different randomly chosen training/test images. The same selections of training/test images were used by all the methods to ensure unbiased comparisons. The results of a projection followed by a classifier are shown in Table 1.

For Yale database, we trained on ten faces and tested on the remaining one of a subject each time. In each test, test faces had the same facial expression or lighting condition. Eleven implementations were conducted through the entire database corresponding to eleven different facial expressions or lighting conditions. The results shown in Table 2 are the averages of these eleven independent implementations.

Table 1. Classification rates of dimesion reduction methods followed by various classifiers (ORL)

No. of training faces	*Classification rates (%)*					
	PCA	KPCA1	KPCA2	LLE	Isomap	CCA
	NN classifier					
3	86.75	86.46	87.75	88.78	85.79	87.75
4	91.92	91.33	92.83	92.51	91.46	91.83
5	94.35	94.25	94.20	94.60	93.15	94.45
6	96.44	96.44	95.94	96.07	95.87	96.04
	soft k-NN classifier					
3	86.75	84.50	88.25	88.71	85.86	87.32
4	91.31	90.58	90.92	92.75	91.21	91.54
5	93.85	93.55	93.00	94.50	93.10	94.15
6	95.74	95.50	94.13	96.06	95.25	96.19
	LDA classifier					
3	90.64	88.93	89.93	90.29	86.54	87.85
4	94.92	94.00	94.04	93.25	91.63	92.79
5	96.20	95.85	95.05	96.25	93.10	94.60
6	96.88	96.69	96.81	97.31	95.69	95.96
	SVM					
3	90.21	88.32	86.75	91.07	86.61	89.71
4	94.79	93.83	91.42	94.46	91.08	94.46
5	96.70	96.35	94.20	95.65	93.30	96.25
6	97.75	97.81	96.26	97.13	96.19	97.06

Table 2. Classification rates of dimesion reduction methods followed by various classifiers (Yale)

Classifier	*Classification rates (%)*					
	PCA	KPCA1	KPCA2	LLE	Isomap	CCA
NN	84.85	81.82	84.85	86.06	83.03	84.85
Soft k-NN	83.03	83.03	84.24	84.85	83.64	83.64
LDA	89.70	85.45	87.27	86.06	81.82	85.45
SVM	88.48	87.27	83.64	86.06	84.24	88.48

For ORL database, the results show that with more training samples the classification rate increases in all methods as expected. LLE slightly outperforms the others with improvement of less than

1% however in most implementations with both NN and soft k-NN classifiers; while PCA has the best performance with LDA and SVM classifiers, but the improvement again is limited. The results on Yale database show a similar pattern. That is, all dimension reduction methods followed by the same classifier give similar performance and PCA followed by LDA classifier has the highest rate in all implementations. Thus nonlinear methods do not always or significantly outperform PCA in reducing dimensionality of face data, and the differences in classification rate between them are not significant enough to single out a particular method. In order to quantify the effect of nonlinearity, a further analysis on complex data structure of face data is conducted next.

4. NONLINEARITY ANALYSIS

There has been previous work showing that nonlinear methods are more powerful than linear PCA for capturing nonlinear structure of high-dimensional data [3-7]. However, most experiments with these nonlinear projections were conducted on artificial data sets such as S-curve and swissroll that distribute on assumed nonlinear manifolds [3-7]. The structures of real-world data sets can be far more different. The question is, are real-world data sets (e.g. faces) distributed nonlinearly or having a rather linear distribution, and how to quantify their degree of nonlinearity?

It is obvious that a data set containing N data points is linear in an N-1 or higher dimensional space or subspace. In both ORL and Yale face databases, the dimension of face data is much greater than the number of images. Thus face data are linear in their input spaces. For example, a training set of 200 faces in ORL is linear in 199 or higher dimensional manifold (subspace). However, are they still linear in a further reduced space, such as in 60-dimensional manifold as used in the experiments? If not, how nonlinear are they? A PCA-based quantitative measure of nonlinearity is derived to address these questions as follows.

4.1. Quantitative measure of nonlinearity

The most common measures of linearity or nonlinearity are based on the residuals of linear and nonlinear regressions [15-17]. The larger the difference of residuals between linear and nonlinear regressions, the more nonlinear the data set is. In our measure, we assume that a nonlinear regression fits “perfectly” the data and there is no or “negligible” residual of the nonlinear regression. The nonlinearity rate (NLR) can be computed as

$$NLR = \frac{1}{N\gamma}\sum_{i=1}^{N}\left(\|\mathbf{L}_i - \mathbf{X}_i\|^2 - \|\mathbf{H}_i - \mathbf{X}_i\|^2\right) \approx \frac{1}{N\gamma}\sum_{i=1}^{N}\|\mathbf{L}_i - \mathbf{X}_i\|^2 \quad (6)$$

where N is the number of data points, $\mathbf{X}_i$ is a data point, $\mathbf{L}_i$ and $\mathbf{H}_i$ are the fitting/support points of the linear and nonlinear regressions respectively. $\gamma = 1/N\sum_{i=1}^{N}\|\mathbf{X}_i - \mathbf{m}\|^2$ is the variance of the data. The value of NLR is the amount of residual of linear fitting and indicates the nonlinearity of the data. Larger NLR means higher degree of nonlinearity.

PCA is the best linear fitting (to a reduced dimension). The fitting surface is a linear d-dimensional embedding in the direction of $\boldsymbol{v}$ passing through the mean $\mathbf{m}$ of the data.

$$\mathbf{L}_i = \sum_{k=1}^{d} D_{ik}\mathbf{v}_k + \mathbf{m} \qquad i = 1,2,...,N \quad (7)$$

where $\mathbf{D}_i$ is the projected point of $\mathbf{L}_i$ in the d-dimensional subspace.

PCA finds the best optimal set of fitting points, $\mathbf{L}_i$, by minimizing the squared-error criterion function [1]

$$\begin{aligned} E_d &= \sum_{i=1}^{N}\left\|(\sum_{k=1}^{d} D_{ik}\mathbf{v}_k + \mathbf{m}) - \mathbf{X}_i\right\|^2 = \sum_{i=1}^{N}\left\|\sum_{k=1}^{d} D_{ik}\mathbf{v}_k - (\mathbf{X}_i - \mathbf{m})\right\|^2 \\ &= \sum_{i=1}^{N}(\sum_{k=1}^{d} D_{ik}\mathbf{v}_k)^2 - 2\sum_{i=1}^{N}\sum_{k=1}^{d} D_{ik}\mathbf{v}_k^T(\mathbf{X}_i - \mathbf{m}) + \sum_{i=1}^{N}\|\mathbf{X}_i - \mathbf{m}\|^2 \end{aligned} \quad (8)$$

Partially differentiating with respect to D_{ik}, setting the derivative to zero while considering $\|\mathbf{v}_k\|$=1, we get

$$D_{ik} = \mathbf{v}_k^T(\mathbf{X}_i - \mathbf{m}) \quad (9)$$

Geometrically, this result merely means that we get a least-square solution by projecting $\mathbf{X}_i$ onto the d-dimensional subspace in the direction of $\mathbf{v}$. Then substitute D_{ik} in Eq. (9) into Eq. (8), we obtain

$$E_d = -\sum_{i=1}^{N}\sum_{k=1}^{d}[\mathbf{v}_k^T(\mathbf{X}_i - \mathbf{m})]^2 + \sum_{i=1}^{N}\|\mathbf{X}_i - \mathbf{m}\|^2 = -N\sum_{k=1}^{d}\mathbf{v}_k^T\mathbf{S}\mathbf{v}_k + \sum_{i=1}^{N}\|\mathbf{X}_i - \mathbf{m}\|^2 \quad (10)$$

where $\boldsymbol{S}$ is the covariance matrix. The vector $\mathbf{v}_k$ that minimizes E_d also maximizes the first term of Eq. (10). The Lagrange multipliers [1] are used to maximize the first term subject to constraint, $\|\mathbf{v}_k\|$=1. Let λ_k be the multiplier, we differentiate

$$U = \sum_{k=1}^{d}[\mathbf{v}_k^T\mathbf{S}\mathbf{v}_k - \lambda_k(\mathbf{v}_k^T\mathbf{v}_k - 1)] \quad (11)$$

with respect to $\mathbf{v}_k$, resulting

$$\frac{\partial U}{\partial \mathbf{v}_k} = 2\sum_{k=1}^{d}(\mathbf{S}\mathbf{v}_k - \lambda_k\mathbf{v}_k) \quad (12)$$

Set this gradient vector equal to zero, one can see that $\mathbf{v}_k$ and λ_k must be the eigenvector and eigenvaluse of the scatter matrix $\mathbf{S}$. Because $\mathbf{v}_k^T\mathbf{S}\mathbf{v}_k = \lambda_k\mathbf{v}_k^T\mathbf{v}_k = \lambda_k$, it follows that to minimize the criterion function, Eq. (10), we need to select the eigenvectors corresponding to the d largest eigenvaluses of the covariance matrix. Thus Eq. (10) can be written in the form of eigenvalue as

$$E_d = -N\sum_{k=1}^{d}\lambda_k + N\gamma \quad (13)$$

It shows that PCA finds the best fitting (with minimum squared-error, E_d) in the eigenvector directions (corresponding to the d largest eigenvaluses). If the sum of these d largest eigenvalues is equal to the variance γ, the PCA finds the perfect linear fitting (with E_d=0) of the data, which means the input data is linear in a d-dimensional embedding. As an n-dimensional data set is linear in an n-dimensional embedding, thus E_d=0 when d=n. Combining Eq. (6) and Eq. (13), we have the final form of the nonlinearity measure.

$$NLR = \frac{1}{N\gamma}E_d = 1 - \sum_{k=1}^{d}\lambda_k / \sum_{i=1}^{n}\lambda_i \quad (14)$$

4.2. Nonlinearity analysis of various data sets

The nonlinearity measure was tested on several artificial data sets and from the results we find that a data set has high degree of nonlinearity when the value of NRL is higher than 0.3. The nonlinearities of 200 training faces of ORL database and 150 training faces of Yale database are plotted in Figure 1. The NLR values of various dimensional manifolds indicate that ORL face data has higher degree of nonlinearity than Yale data in the same dimensional manifold. The NLR values of ORL and Yale in 60-dimensional manifold are about 0.11 and 0.03 respectively, indicating that the ORL data has low degree of nonlinearity in the reduced dimensional subspace used in the experiments and Yale data is almost linear in that space. They also explain why PCA has similar performance to nonlinear methods for dimension reduction on both ORL and Yale face databases in the experiments.

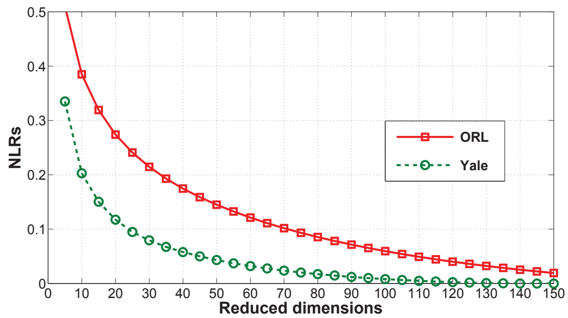

Fig. 1 *NLR* values of ORL face database and Yale face database.

For a further analysis, we compared the performance of PCA with LLE and CCA (followed by the NN classifier) again in reduced dimensions varying from 5 to 70 dimensions. The results are shown in Figure 2. We can draw four points from this result. First, the performance of three methods increases with dimensions, or lower nonlinearities lead to better performances. Second, PCA reaches reasonable performances with reduced dimensions of 40 in ORL (with *NLR* of 0.17) and 30 in Yale (with *NLR* of 0.08). With the increase in the number of reduced dimensions (decrease in nonlinearity), three methods have stable performances while nonlinear methods have similar or slightly better performances than PCA (less than 1% improvement in classification rate). Third, the classification rate of LLE is lower than that of PCA significantly at the reduced dimension of 15 in ORL and 5 in Yale. Figure 1 shows that *NLR* values of these two situations are higher than 0.3, which means both data sets become highly nonlinear in these low dimensional subspaces. LLE fails to preserve the structure of data sets. Fourth, in Yale database, nonlinear methods may outperform PCA in the reduced dimensions below 30 because the data set has started to show strong nonlinearity in these dimensions. However, all classification rates start to deteriorate.

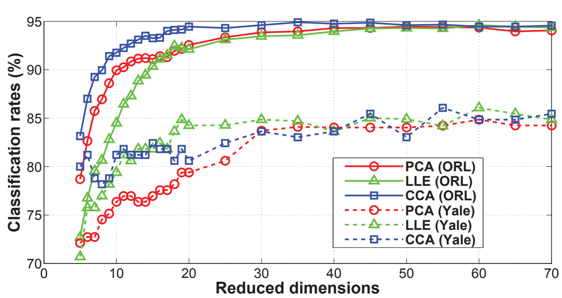

Fig. 2 Comparisons of PCA, LLE and CCA in various reduced dimensions.

5. CONCLUSIONS

In this paper, we compare the performances of PCA and nonlinear methods for dimensionality reduction in face recognition. Though nonlinear methods have capabilities for capturing nonlinear data structure, they do not often lead to significant improvement in performance in face recognition because real face data may distribute fairly linearly and the nonlinear capabilities of those nonlinear methods may not be effective in projecting these high-dimensional face data. The nonlinearity analysis also shows that by considering dimension reduction and overall information loss, good solutions can be achieved in subspaces of low nonlinearity, which leads similar or comparable performance of all projection methods, while PCA is much easier to implement and less computationally demanding.

6. REFERENCES

[1] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification* (2nd Ed.), Wiley, New York, 2001.

[2] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, pp. 71–86, 1991.

[3] B. Scholköpf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, pp. 1299-1319, 1998.

[4] S.T. Roweis and L.K. Saul., "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, pp. 2323-2326, 2000.

[5] J.B. Tenenbaum, V.D. Silva, and J.C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, pp. 2319-2323, 2000.

[6] P. Demartines and J. Hérault, "Curvilinear component analysis: a self-organizing neural network for nonlinear mapping of data sets," *IEEE trans. Neural Networks*, vol. 8, pp. 148-154, 1997.

[7] H. Yin, "Nonlinear dimensionality reduction and data visualization: a review," *Int. Journal of Automation and Computing*, vol. 4, pp. 294-303, 2007.

[8] Y.H. Pang, A.B.J. Teoh, E.K. Wong, and F.S. Abas, "Supervised locally linear embedding in face recognition," *Int. Symp. on Biometrics and Security Technologies*, pp. 1-6, 2008.

[9] M.H. Yang, "Kernel eigenfaces vs. kernel fisherfaces: face recognition using kernel methods," *Proc. 5th IEEE Int. Conf. on Automatic Face and Gesture Recognition*, Washington DC, pp. 215–220, 2002.

[10] M.H. Yang, N. Ahuja, and D. Kriegman, "Face recognition using kernel eigenfaces," *IEEE Int. Conf. on Image Processing*, pp. 37-40, 2000.

[11] M.H. Yang, "Extended isomap for pattern classification," *Proc. National Conference on Artificial Intelligence*, pp. 224-229, 2002.

[12] X. Tan, S. Chen, Z. Zhou, and F. Zhang, "Recognizing partially occluded, expression variant faces from single training image per person with SOM and soft *k*-NN ensemble," *IEEE trans. on Neural Networks*, vol. 16, pp. 875-886, 2005.

[13] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman, "Eigenfaces vs. fisherfaces: recognition using class specific linear projection," *IEEE trans. on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 711-720, 1997.

[14] C.J.C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, pp. 121-167, 1998.

[15] K. Emancipator and M.H. Kroll, "A quantitative measure of nonLinearity," *Clin. Chem*, vol. 39, pp. 766-772, 1993.

[16] K.C. Li, "Nonlinear confounding in high-dimensional regression," *Ann. Statist.*, vol. 25, pp. 577-612, 1997.

[17] H. Moisl, "Data nonlinearity in exploratory multivariate analysis of language corpora," *Proc. of 19th Meeting of the ACL Special Interest Group in Computational Morphology and Phonology*, Prague, Czech Republic, pp. 93–100, 2007.