SAVEFACE AND SIRFACE: APPEARANCE-BASED RECOGNITION OF FACES AND FACIAL EXPRESSIONS

Yangrong Ling[*], *Suchendra M. Bhandarkar*[*], *Xiangrong Yin*[+] *and QiQi Lu*[+]

[*]Department of Computer Science and [+]Department of Statistics
The University of Georgia, Athens, GA 30602–7404, USA

ABSTRACT

The problem of appearance-based recognition of faces and facial expressions is addressed. Previous work on sliced inverse regression (SIR) resulted in the formulation of an appearance-based face recognition technique termed as *Sirface* that is insensitive to large variation in lighting direction and facial expression. *Sirface* was shown to be superior to the well known *Fisherface* technique, that is based on Fisher's linear discriminant analysis (LDA), in terms of both, dimensionality reduction and classification accuracy. However, *Sirface*, which relies only on first-order statistics, is shown to be poor at discriminating between facial expressions. A novel statistical data dimensionality reduction technique based on sliced average variance estimation (SAVE) is shown to be effective in distinguishing between different facial expressions of the same face. SAVE, which exploits the difference in second-order statistics between the pattern classes, is shown to result in an optimal reduced dimensional subspace for quadratic discriminant analysis (QDA). The resulting appearance-based technique for recognition of faces and facial expressions, termed as *Saveface*, is experimentally compared to *Sirface* in terms of classification accuracy and data dimensionality reduction.

1. INTRODUCTION

Numerous algorithms have been proposed for appearance-based face recognition during the past few years. While much progress has been made toward recognizing faces under small variations in lighting, facial expression and pose, reliable techniques for distinguishing between facial expressions of the same face under variations of lighting and pose have proven elusive. In this paper, we propose a new technique for appearance-based recognition of faces and facial expressions; one that is insensitive to large variations in ambient illumination and pose. Note that variation in ambient illumination includes not only variation in light intensity, but also variation(s) in the direction(s) and number of light sources.

Conventional approaches to appearance-based recognition of faces exploit two key observations:
(1) All the images of a Lambertian surface, taken from a fixed viewpoint, but under varying illumination, lie in a 3-D linear subspace of the high-dimensional image space [1].
(2) Due to shadowing, specularities and variations in facial expression, the above observation does not hold exactly under all situations. In practice, certain regions of the face may exhibit considerable deviation from the aforementioned 3-D linear subspace. Consequently, these regions of the face are less reliable for the purpose of recognition [1].
The above observations are used to determine a linear projection of the faces from the high-dimensional image space to a significantly lower dimensional feature space such that the projection is insensitive to variations in ambient illumination, pose and facial expression. Thus, the use of appropriate data dimensionality reduction techniques is crucial in appearance-based recognition methods. Note that appearance-based approaches to face recognition preclude the use of an *a priori* model. In contrast to model-based approaches to face recognition where an explicit geometric and/or photogrammetric representation is needed, appearance-based approaches rely on the learning of an implicit model via selection of sample images of the face under varying conditions of illumination, pose, viewpoint and facial expression.

Techniques for appearance-based face recognition that are well described in the research literature include ones based on correlation, principal component analysis (termed as *Eigenface*), linear subspace projection and Fisher's linear discriminant analysis (termed as *Fisherface*). A comparison amongst these aforementioned techniques can be found in [1] where *Fisherface* is shown to be superior to the other techniques in terms of classification accuracy and data dimensionality reduction. An improved version of *Fisherface* based on sliced inverse regression (SIR) and termed as *Sirface* was presented in our previous work [7].

A common shortcoming of the *Fisherface* and *Sirface* techniques is that they exploit only the difference in the first-order statistics of the underlying pattern classes. If the distinguishing features between the pattern classes cannot be adequately represented by first-order statistics, then performance of both, *Fisherface* and *Sirface* can be expected to suffer. In this paper it is shown that differences in first-order statistics (i.e., class means) cannot adequately distinguish between different facial expressions of the same face. This provides the motivation for appearance-based recognition methods that are capable of exploiting the differences in second- and higher-order statistics amongst the underlying pattern classes.

In this paper, an appearance-based algorithm for recognition of faces and facial expressions, based on sliced average variance estimation (SAVE) [2] and termed as *Saveface*, is proposed. The reduced dimensional subspace determined by *Saveface* is equivalent to the one obtained using quadratic discriminant analysis (QDA). SAVE is a fairly recent technique in the area of statistical regression [2] and its connection to QDA has been established only recently [4]. On account of its ability to exploit the differences in second-order statistics between the pattern classes, *Saveface* is potentially more effective in distinguishing between different facial expressions of the same face compared to *Sirface* or *Fisherface*. Since *Sirface* has been shown to be superior to *Fisherface* [7], the primary aim of this paper is to compare *Saveface* to *Sirface* in the context of appearance-based recognition of faces and facial expressions.

2. APPEARANCE-BASED FACE RECOGNITION

The appearance-based face recognition problem can be simply stated as follows: Given a set of face images, each labeled with the person's identity (the *learning set*) and an unlabeled set of face images from the same group of people (the *test set*), identify the face of the person in each of the test images. The basic idea in appearance-based face recognition is to first, use the learning set to determine the classification rules and second, apply these classification rules to label the face in each test set image with the identity of the person. Formally, consider a set of n sample face images $\{\mathbf{X}_1,\ldots,\mathbf{X}_n\}$ where each image can be looked upon as a point in p-dimensional image space. Assume that each face image belongs to one of c classes $1,\ldots,c$ where the class label denotes the face identity. Thus, we need to determine the classification rules that map each sample image point $\mathbf{X}_i$ in the p-dimensional image space onto the right class label; a classical pattern classification problem. Note that p is typically large; $p = l \times k$ for an image of size $l \times k$ pixels. Hence it is desirable to reduce the p-dimensional image space to the smallest d-dimensional image subspace that retains all the necessary classification information. Formally, we need to determine a $p \times d$ matrix $\mathbf{B}$ where $\mathbf{B}^T\mathbf{X}$ is the desired d-dimensional image subspace such that each input pattern $\mathbf{X}_i$ is assigned to the same class regardless of whether the original p-dimensional image space or the reduced d-dimensional image subspace is used. The subspace spanned by the d columns of $\mathbf{B}$ is termed the central discriminant subspace [4].

Using the reduced d-dimensional image subspace as described above has several potential advantages. A reduced dimensional subspace serves to filter out the noisy or irrelevant portions of the input image space thus reducing both the classification error rate and the computational complexity. Also, if $d \leq 3$, one can visualize the sample data more easily. Although the input data in the p-dimensional space are expressed in their original scale, termed as the $\mathbf{X}$-scale, it is more convenient to transform the input data to an equivalent $\mathbf{Z}$-scale where $\mathbf{Z} = \mathbf{\Sigma}_{\mathbf{X}}^{-\frac{1}{2}}(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})$ and $\mathbf{\Sigma}_{\mathbf{X}}$ and $\boldsymbol{\mu}_{\mathbf{X}}$ are the covariance matrix and mean vector of $\mathbf{X}$, respectively. Here it is assumed that the matrix $\mathbf{\Sigma}_{\mathbf{X}}$ is nonsingular. If $\mathbf{\Sigma}_{\mathbf{X}}$ is singular then the dimensionality of the original $\mathbf{X}$ is reduced using principal components analysis (PCA) to the point where the resulting $\mathbf{\Sigma}_{\mathbf{X}}$ becomes nonsingular. The use of the $\mathbf{Z}$-scale as described above allows for easy comparison of various dimensionality reduction techniques.

3. SIRFACE AND SAVEFACE

Sliced inverse regression (SIR) [6] was originally developed for data dimensionality reduction in the context of statistical regression. Let $(\mathbf{Y}_i, \mathbf{X}_i)$ $i = 1,\ldots,n$ be an input sample, where $\mathbf{Y}$ is a response variable and $\mathbf{X}$ is a predictor vector. Li [6] considered the inverse mean of $\mathrm{E}(\mathbf{X}|\mathbf{Y})$ in the $\mathbf{Z}$-scale described above by constructing the SIR matrix $\mathbf{M}_{\mathrm{SIR}} = \mathrm{Var}(\mathrm{E}(\mathbf{Z}|\mathbf{Y}))$ where E denotes statistical expectation and Var the variance. In the context of pattern classification, the response variable $\mathbf{Y}$ is a categorical variable given by $\mathbf{Y} = 1,\ldots,c$. The SIR matrix is given by

$$\mathbf{M}_{\mathrm{SIR}} = \frac{1}{n}\Sigma_{i=1}^{c} n_i(\boldsymbol{\mu}_i\boldsymbol{\mu}_i^T) = \frac{1}{n}\mathbf{S}_B \tag{1}$$

where n_i is the number of images in the training set that belong to class i and $n = \Sigma_{i=1}^{c} n_i$ is the total number of images in the training set. The eigenvectors of $\mathbf{M}_{\mathrm{SIR}}$ that correspond to its non-zero eigenvalues span a reduced dimensional subspace of the original p-dimensional space of the input data. It can be shown that the reduced dimensionality $d_{sir} \leq c - 1$ for the pattern classification problem [7]. The input data can be projected onto this reduced dimensional subspace and classified therein. In the case of appearance-based face recognition, this data dimensionality reduction technique is termed as *Sirface*. Kent [5] and Cook and Yin [4] have shown the equivalence of the reduced dimensional subspaces determined by SIR and Fisher's LDA.

Sliced average variance estimation (SAVE) [2] was also originally developed for data dimensionality reduction in statistical regression. Cook and Weisberg [2] considered the following matrix expressed in the previously described $\mathbf{Z}$-scale: $\mathbf{M}_{\mathrm{SAVE}} = \mathrm{E}(\mathbf{I} - \mathbf{\Sigma}_{\mathbf{Z}|Y})^2$ where $\mathbf{I}$ is the identity matrix. The rank or dimension of $\mathbf{M}_{\mathrm{SAVE}}$ is determined using singular value decomposition (SVD). The eigenvectors of $\mathbf{M}_{\mathrm{SAVE}}$ that correspond to its non-zero eigenvalues also span a reduced dimensional subspace of the original p-dimensional space of the input data. Likewise, the input data can be projected onto this reduced dimensional subspace and classified therein. In the case of appearance-based face recognition, this data dimensionality reduction technique is termed as *Saveface*. In the context of pattern classification, let $\boldsymbol{\mu}_i$ and $\mathbf{\Sigma}_i$ denote the mean vector and covariance matrix for class i where $i = 1,\ldots,c$. The response variable is given by $\mathbf{Y} = 1,\ldots,c$ and the corresponding SAVE matrix $\mathbf{M}_{\mathrm{SAVE}}$ (in $\mathbf{Z}$-scale) is given by:

$$\mathbf{M}_{\mathrm{SAVE}} = \sum_{i=1}^{c} \frac{n_i}{n}(\mathbf{I} - \mathbf{\Sigma}_i)^2 \tag{2}$$

Cook and Yin [4] have formally developed the connection between SAVE and the classical QDA in the subspace sense. The optimal situation for classification using QDA is when the random variables $\mathbf{Z}|Y = i$ are normally distributed for each class i with different covariance matrices. Note that SAVE by itself does not require the assumption of normality of the random variables $\mathbf{Z}|Y = i$ for optimal classification. The subspace spanned by $\mathbf{M}_{\mathrm{SAVE}}$ can be shown to be $\mathcal{S}(\mathbf{I} - \mathbf{\Sigma}_i, i = 1,\ldots,c)$ [3] where $\mathcal{S}(\mathbf{A})$ denotes the subspace spanned by the eigenvectors of matrix $\mathbf{A}$ with non-zero eigenvalues. Under the assumptions of normality of the random variables $\mathbf{Z}|Y = i$, Odell [8], Tubbs *et al.* [9] and Young *et al.* [10] have shown the equivalence of the subspaces spanned by $\mathbf{M}_{\mathrm{SAVE}}$ and the QDA matrix. Thus SAVE is equivalent to QDA under assumptions of normality but is more general than QDA since it does not require the assumption of normality of the random variables $\mathbf{Z}|Y = i$ for optimal classification.

When all the class covariance matrices are identical, i.e., $\mathbf{\Sigma}_i = \mathbf{\Sigma}$ for $i = 1,\ldots,c$, SAVE can be shown to be equivalent to SIR [6]; in which case the *Saveface* technique reduces to the *Sirface* technique. When not all $\mathbf{\Sigma}_i = \mathbf{\Sigma}$, i.e. when the covariance matrices contain class discriminatory information, it is possible that the SIR matrix $\mathbf{M}_{\mathrm{SIR}}$ may not capture all the relevant discriminatory information that the SAVE matrix $\mathbf{M}_{\mathrm{SAVE}}$ does. Broadly speaking, SIR captures the discriminatory information in the first (inverse) moment whereas SAVE captures the discriminatory information in the first two (inverse) moments. Since $\mathcal{S}(\mathbf{M}_{\mathrm{SIR}}) \subseteq \mathcal{S}(\mathbf{M}_{\mathrm{SAVE}})$, *Saveface* can be considered to be more comprehensive than *Sirface*. SAVE could also be looked upon as a generalized version of QDA since it removes all the redundant information (along the eigenvectors that correspond to the zero eigenvectors of $\mathbf{M}_{\mathrm{SAVE}}$) without requiring assumptions of normality. In a manner similar to *Sirface*, *Saveface* determines the smallest number of

predictors needed for classification while allowing one to use different classifiers such as the nearest-neighbor classifier, maximum likelihood classifier or the Bayesian classifier.

3.1. Test for determining the optimal reduced dimensionality d_{save}

In order to determine the reduced dimensionality d_{save}, we apply singular value decomposition (SVD) to the matrix $\mathbf{M}_{\mathrm{SAVE}}$ as follows:

$$\mathbf{M}_{\mathrm{SAVE}} = \mathbf{\Gamma}^T \begin{pmatrix} \mathbf{D} & 0 \\ 0 & 0 \end{pmatrix} \mathbf{\Gamma} \tag{3}$$

$\mathbf{\Gamma}$ is a $p \times p$ orthogonal matrix such that $\mathbf{\Gamma}^T = (\mathbf{\Gamma}_1, \mathbf{\Gamma}_0)$ where $\mathbf{\Gamma}_0$ is $p \times (p-d)$ matrix. $\mathbf{D}$ is a $d \times d$ diagonal matrix whose elements $\lambda_1 \geq \cdots \geq \lambda_d$ are the eigenvalues of $\mathbf{M}_{\mathrm{SAVE}}$. Since the optimum value of $d = d_{save}$ is typically unknown, it needs to be estimated from the underlying data. Given all the ordered eigenvalues of the sample matrix $\hat{\mathbf{M}}_{\mathrm{SAVE}}$, the value of $d = d_{save}$ is estimated to be the number of all the positive eigenvalues of the sample matrix $\hat{\mathbf{M}}_{\mathrm{SAVE}}$, such that $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_d > 0$ and $\hat{\lambda}_{d+1} = 0$. The corresponding eigenvectors $\hat{l}_1, \cdots \hat{l}_d$ are the basis vectors that span the reduced dimensional subspace. The best estimated value of $d = d_{save}$ is the one such that the cumulative proportion of the ordered eigenvalues $\Sigma_{i=1}^{d} \lambda_i^2 / \Sigma_{i=1}^{p} \lambda_i^2$ and the corresponding eigenvectors yield the best classification accuracy. Typically the search for d_{save} is started at a cumulative proportion value of 75% and the value of d incremented until the classification accuracy reaches a predefined acceptable rate.

4. EXPERIMENTAL RESULTS

In the first experiment, grayscale images of three different faces, each face with three different facial expressions, are considered. Figure 1 shows the faces of a subject with three different facial expressions and two images for each facial expression. Each facial expression is treated as a single class. The goal is to identify the different expressions of the same face using *Saveface* and *Sirface*. The results of the experiment are tabulated in Tables 1 and 2 respectively. The notations used in Tables 1 and 2 are as follows. N_c: number of classes in each dataset; N_{tc}: number of training samples per class; N_{ts}: number of samples in the test dataset; d: number of positive eigenvalues in the sample SAVE matrix; d_{save}: the best estimated reduced dimensionality using SAVE; d_{sir}: reduced dimensionality obtained using SIR, which is $\leq N_c - 1$; CP: cumulative proportion of eigenvalues from the sample SAVE matrix; and CA: classification accuracy.

The results tabulated in Tables 1 and 2 show that *Saveface* outperforms *Sirface* in distinguishing between different facial expressions of the same face. The classification accuracy of *Saveface* is seen to improve monotonically with an increasing number of training samples per class (Table 1) as per expectation. The classification accuracy of *Sirface* does not exhibit such a trend (Table 2). This shows that first-order statistics are not adequate to distinguish between different facial expressions of the same face; second- or higher-order statistics are needed. Although the optimal reduced dimensionality for *Saveface* is greater than that of *Sirface* (which is 2 since each dataset has only 3 classes), it is still very small compared to the dimensionality of the input image space which is of the order of 10^5. Table 3 depicts how the best reduced dimensionality d=d_{save} for *Saveface* is determined. For example, in the case of dataset 1, three values for d, namely, 10, 11, and 12, with corresponding CP values of 75.8%, 83.2% and 90.7% respectively are shown. It can be seen that when $d = 12$, the classification accuracy CA =100%. Hence d_{save} is chosen to be 12. The same procedure is followed for all the datasets from all subjects.

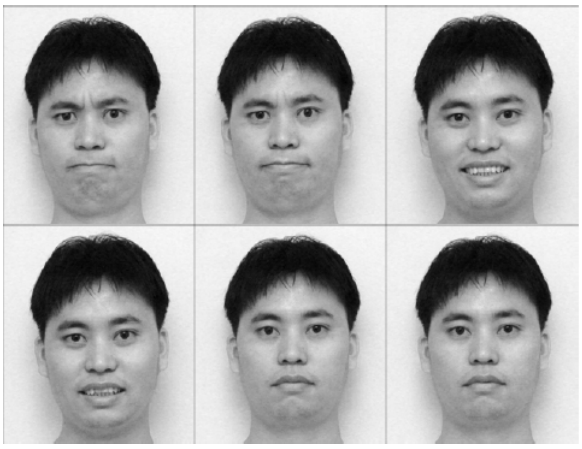

Fig. 1. Face images with varying facial expressions

Table 1. Performance of *Saveface* for 3 different faces

F	D	N_c	N_{tc}	N_{ts}	d	d_{save}	CP%	CA%
1	1	3	5	53	14	12	90.7	100
	2	3	7	47	20	16	88.4	100
	3	3	10	38	29	27	95.5	100
2	1	3	5	60	14	14	100	71.7
	2	3	10	45	29	29	100	97.8
3	1	3	5	54	14	14	100	94.4
	2	3	10	39	29	29	100	100

F: Face Number, D: Dataset Number

In the second experiment, the dataset consists of a combination of different faces with different facial expressions. *Saveface* and *Sirface* are used to determine the face identity and facial expression as shown in Tables 4 and 5. Dataset 1 consists of 6 subjects where each subject has 3 facial expressions and each facial expression has 5 training samples. This results in a total of 90 training images. The 6 different subjects are treated as 6 distinct pattern classes. Dataset 2 consists of 5 subjects, where each subject has 3 facial expressions and each facial expression has 10 training samples. This results in a total of 150 training images. Each facial expression for each subject is treated as a single class, resulting in 15 classes. For both experiments, the reduced dimension for *Saveface* is chosen to be $d - 1$ (the dimension corresponding to the smallest non-zero eigenvalue is ignored) and the reduced dimension for *Sirface* is chosen to be $N_c - 1$. From Tables 4 and 5 it can be seen that *Sirface* outperforms *Saveface*. This seems counterintuitive because the *Sirface* dimensions are contained within the *Saveface* dimensions. Hence *Saveface* should be expected to outperform *Sirface*. However, *Sirface* captures mainly the mean differences between the pattern classes whereas *Saveface* captures the mean differences and variance differences simultaneously. When the dataset contains different faces as classes, regardless of whether or not different facial expressions are used in creating the classes,

Table 2. Performance of *Sirface* for 3 different faces

F	D	N_c	N_{tc}	N_{ts}	d_{sir}	CA%
1	1	3	5	53	2	85.1
	2	3	7	47	2	83.0
	3	3	10	38	2	78.9
2	1	3	5	60	2	71.7
	2	3	10	45	2	66.7
3	1	3	5	54	2	74.1
	2	3	10	39	2	79.5

F: Face Number, D: Dataset Number

the differences in face identities usually dominate the differences in facial expressions. That is, the mean differences between the pattern classes clearly dominate the variance differences. This is the reason why *Sirface* usually performs better than *Saveface* in this case (Tables 4 and 5; dataset 1). On the other hand, when the dataset contains different faces with different expressions as classes, the eigenvalues of the Saveface matrix may be diluted. In this case the *Saveface* matrix produces many non-zero eigenvalues, many of which may correspond to the differences in facial expressions. The corresponding eigenvectors may not be relevant for distinguishing between different faces. It is also possible that the same facial expression on different faces may correspond to a different set of eigenvalues and hence a different set of eigenvectors. Therefore, in order to achieve robust recognition of faces and facial expressions in practice, the best technique is the combination of *Sirface* and *Saveface* where *Sirface* is used to classify different faces based on identity and *Saveface* is used to classify different expressions of the same face.

Table 3. Face 1: Determining the value of d_{save}

F	D	d_p	CP%	CA%
1	1	10	75.8	64.2
		11	83.2	88.7
		12	90.7	100
	2	14	78.1	91.5
		15	83.2	91.5
		16	88.4	100
	3	25	88.5	65.8
		26	92.0	97.4
		27	95.5	100

F: Face Number, D: Dataset Number

Table 4. Performance of *Saveface* on grouped datasets

D	N_c	N_{tc}	N_{ts}	d	d_{save}	CA%
1	6	15	328	90	89	53.7
2	15	10	193	150	149	70.0

D: Dataset Number

5. CONCLUSIONS AND FUTURE WORK

A novel statistical data dimensionality reduction technique based on sliced average variance estimation (SIR) was proposed and implemented. The resulting appearance-based face recognition technique termed as *Saveface* was shown to be effective in distinguishing between different facial expressions of the same face. *Saveface* complements existing appearance-based face recognition methods; in particular, *Sirface*. While *Sirface* exploits the differences in the class means for pattern classification (i.e., first-order statistics), *Saveface* exploits, additionally, the differences in the class covariance matrices (i.e., second-order statistics). When the differences in the class means are the dominant factor, such as in the identification of different faces, *Sirface* is preferable. However, when the differences in the class covariance matrices is the dominant factor, such as when distinguishing between different facial expressions of the same face, *Saveface* is the preferred technique. Future research will consider appearance-based face recognition techniques that exploit higher-order statistics (greater than second-order) for more robust and comprehensive classification.

Table 5. Performance of *Sirface* on grouped datasets

D	N_c	N_{tc}	N_{ts}	d_{sir}	CA%
1	6	15	328	5	81.7
2	15	10	193	14	85.5

D: Dataset Number

6. REFERENCES

[1] P.N. Belhumeur, J.P. Hespanha and D.J. Kriegman, Eigenfaces vs. Fisherfaces: Recognition using class-specific linear projection, *IEEE Trans. PAMI*, 19(7), pp. 711-720, 1997.

[2] R.D. Cook, and S. Weisberg, Discussion of Li (1991). *Jour. Amer. Stat. Assoc.*, 86, pp. 328-332, 1991.

[3] R.D. Cook, and F. Critchley, Identifying outliers and regression mixtures graphically, *Jour. Amer. Stat. Assoc.*, 95, pp. 735-749, 2000.

[4] R. D. Cook and X. Yin, Dimension reduction and visualization in discriminant analysis(with discussion). *Aus.& New Zealand Jour. Statistics*, 43(2), 147–199, 2001.

[5] J.T. Kent, Discussion of Li (1991), *Jour. Amer. Stat. Assoc.*, 86, pp. 336-337, 1991.

[6] K.C. Li, Sliced inverse regression for dimension reduction (with discussion), *Jour. Amer. Stat. Assoc.*, 86, pp. 316-342, 1991.

[7] Y. Ling, X. Yin, and S.M. Bhandarkar, Sirface vs. Fisherface: Recognition using class specific linear projection, *Proc. IEEE ICIP*, Barcelona, Spain, pp. 885-888, 2003.

[8] P.L. Odell, A model for dimension reduction in pattern recognition using continuous data, *Pattern Recog.*, 11, pp. 51-54, 1979.

[9] J.D. Tubbs, W.A. Coberly and D.M. Young, Linear dimension reduction and Bayes classification with unknown population parameters, *Pattern Recog.*, 15, pp. 167-172, 1982.

[10] D.M. Young, V.R. Marco and P.L. Odell, Quadratic discrimination: some results on optimal low-dimensional representation. *Jour. Stat. Plan. Inf.*, 17, pp. 307-319, 1987.