

Multi-class Linear Feature Extraction by Nonlinear PCA

Robert P.W. Duin¹, Marco Loog^{2,3}, R. Haeb-Umbach³

¹Pattern Recognition Group, Department of Applied Physics
Delft University of Technology, The Netherlands
email: duin@ph.tn.tudelft.nl

²SSOR, Faculty of Information Technology and Systems
Delft University of Technology, The Netherlands

³Philips Research Laboratories Aachen, Germany

Abstract

The traditional way to find a linear solution to the feature extraction problem is based on the maximization of the class-between scatter over the class-within scatter (Fisher mapping). For the multi-class problem this is, however, sub-optimal due to class conjunctions, even for the simple situation of normal distributed classes with identical covariance matrices. We propose a novel, equally fast method, based on nonlinear PCA. Although still sub-optimal, it may avoid the class conjunction. The proposed method is experimentally compared with Fisher mapping and with a neural network based approach to nonlinear PCA. It appears to outperform both methods, the first one even in a dramatic way.

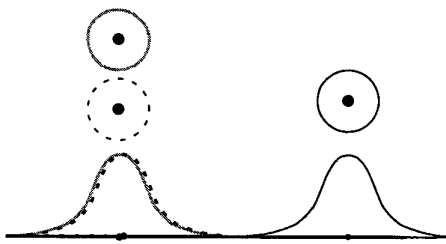
1. Introduction

In this paper we focus on the problem of linear feature extraction for the case of many pattern classes. In some applications this number of classes can be very large, e.g. in character recognition as well as in speech recognition the numbers of characters and phonemes that may be distinguished can vary from tens to hundreds. Also in person identification problems this number may be very large.

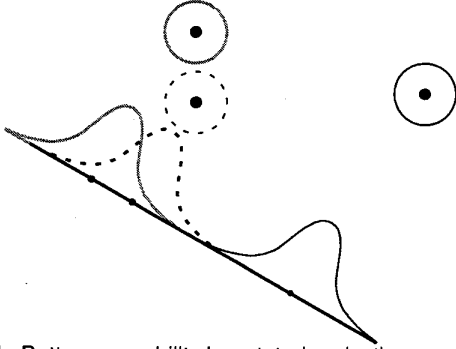
Even if the class distributions can be assumed to be very simple, e.g. Gaussian with identical covariance matrices, the number of classes c competes with the dimensionality of the feature space k in the following

way. After a pre-whitening step, i.e. rotation and scaling of the feature space R_k such that the common class covariance matrix becomes the identity matrix (causing a class variance of one in all directions), the problem is completely defined by the set of c class means. If c is smaller than k this set of class means constitutes a $c-1$ -dimensional linear subspace in which the between class distances are completely preserved. So, without loss of information R_k can be reduced to R_{c-1} . If the dimensionality of this space is still too large for performing accurate learning and computations like following the path of phonemes, a further reduction is desirable. However, any reduction of dimensionality below $c-1$ will disturb the class distances. So now the question arises: how do we find a subspace in which a projection of the class means preserves these distances such that the class separability is maintained as good as possible?

A simple and popular approach to linear dimension reduction is the Fisher mapping that optimises the class between scatter S_B with respect to the within scatter S_W . This results into an eigenvector decomposition of $S_W^{-1}S_B$ [4]. As a result of the pre-whitening step S_W^{-1} disappears (unity matrix). The resulting eigenvector decomposition of S_B is equivalent to a Principal Component Analysis (PCA) of the class means. This analysis emphasizes large class distances, by which preserving the distances of already well separated classes may result into an occlusion of neighbouring classes, see fig. 1. The more classes in the problem, the more likely that this happens.



a. Fisher Mapping



b. Better separability by rotated projection

Fig. 1. Example with three classes in R_2 for which a reduction to R_1 by Fisher mapping (a) results into an occlusion of two classes. Better reductions are possible (b).

There are several ways to deal with the above problem, e.g. by the Patrick-Fisher rule [4] and by nonlinear PCA [3] (NL-PCA) by neural networks. Note that in the last method “nonlinear” stands for the criterion and not for the resulting map. This map may still be linear. These NL-PCA methods change the optimization criterion in one way or the other and measure in a more effective way the total class separability. The price that has to be paid has been always, as far as the authors are aware, that the eigenvector procedure has to be replaced by an iterative optimization.

The usual drawbacks of iterative procedures are that they tend to be slow, need some stopping rule and may end up in some sub-optimal situation. How these arise can be understood from fig. 1. If the direction of projection is not rotated to the left (from a to b) but to the right, then the top class is projected in between the two other ones and is ‘locked’: the optimal situation (even more rotated to the left than shown) can only be reached after ‘passing’ one of the other classes, which makes this situation a local sub-optimum. In this paper

we will present an Eigenvector based procedure for NonLinear PCA (ENL-PCA) that may outperform Fisher mapping for large numbers of classes. It will be compared with a Neural Network based procedure for NonLinear PCA (NNL-PCA) as well.

2. Eigenvector based NonLinear PCA

As argued above, Fisher mapping becomes after a pre-whitening step a standard PCA procedure on the class means. In [1] and [2] it is shown that the between scatter matrix of the class means represented by row vectors m_i ($i=1,c$) can be written in terms of the distances between these means in the following way:

$$S_B = \sum_{i=1}^{c-1} \sum_{j=i+1}^c p_i p_j (m_i - m_j)(m_i - m_j)^T, \quad (1)$$

in which p_i and p_j are the prior probabilities of the classes i and j . The problem with using standard PCA is that large class distances unnecessarily dominate the first eigenvectors of this matrix. In order to lower their influence a weights based on the error-function (erf) may be used. This relates the class distances to their error contribution. The analysis in [1] yields the following modified scatter matrix

$$S_B = \sum_{i=1}^{c-1} \sum_{j=i+1}^c p_i p_j w(d_{ij})(m_i - m_j)(m_i - m_j)^T \quad (2)$$

Now each term is weighted by a function $w(d_{ij})$ of the square root of the Mahalanobis distance d_{ij} between the classes i and j . As a result of the pre-whitening, this is equal to the Euclidean distance. For the weighting function the error function $\text{erf}(\cdot)$ is used:

$$w(d) = \frac{1}{2d^2} \text{erf}\left(\frac{d}{2\sqrt{2}}\right) \quad (3)$$

in which $\text{erf}(\cdot)$ is defined as:

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt \quad (4)$$

The first term in (3) normalises the size of the distances, while the second term weighs them according to their error contribution as it relates the distance d between two class means to the overlap of the two classes if they have a standard normal distribution. In this way, small class distances (see fig. (1)) are empha-

sized according to their error contribution. The eigenvectors corresponding to the largest eigenvalues of S_B are used for obtaining the linear map. Note that this is still not optimal as only two-class relations are considered. Moreover, this modifies the scatter matrix *before* mapping. A truly optimal result should be based on the class overlap *after* mapping. An attempt to this is made in the next section, losing the possibility of a straight forward eigenvector approach.

3. Neural Network based NonLinear PCA

The PCA subspace is a linear subspace with maximum explained variance. The unexplained variance is thereby minimum. An $[n,k,n]$ auto-associative neural network with n inputs, a single hidden layer, here also called bottleneck layer of k neurons and n output neurons finds such a k -dimensional subspace if all neurons are linear, in the hidden layer as well as in the output layer. The targets of such a network are set equal to the input vectors. Consequently, the mean square error on the output is equivalent to the error of a linear reconstruction and thereby to the unexplained variance in the hidden layer.

There are various ways to transform such a linear PCA network into a nonlinear network. One is the addition of a nonlinear layer between the bottleneck

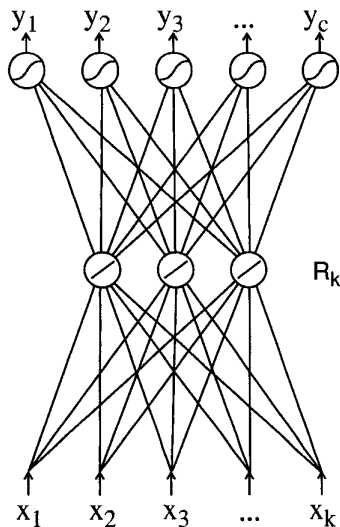


Fig. 2. The network used for nonlinear PCA. In the hidden layer of k neurons the linear subspace is realized for which the linear classification is optimized for the mse on the sigmoidal output units.

and the output. Another possibility is to use a nonlinear transfer function on the outputs. In these ways a nonlinear reconstruction or a nonlinear error weight is realized. The mapping itself, from input to bottleneck layer, is still linear. Such a network may be used for nonlinear PCA [3], in which the adjective ‘nonlinear’ describes the criterion and not the mapping.

Our final goal is to find a subspace in which the classes are optimal separable. Therefore we replace the output reconstruction layer of n neurons by a classification layer of c neurons (i.e. equal to the number of classes) with a sigmoidal output function. This function, like the error function in the eigenvector approach, prevents the domination of well separable classes. This network is sketched in fig. 2. It can be trained by standard procedures like backpropagation or Levenberg-Marquardt. In the below experiment we used the latter.

4. Experiments

In order to test the two NL-PCA procedures described in the section 2 and 3 we performed two experiments, one on artificial data and one on real data. In the artificial problem we assume known class means and equal class variances in all directions, simulating the pre-whitened situation. A set of 30 class means is generated from a 30-dimensional normal distribution with a variance of 4 in all directions. Subspaces with dimensions 1 to 29 are computed. The linear separability of these subspaces is estimated by a

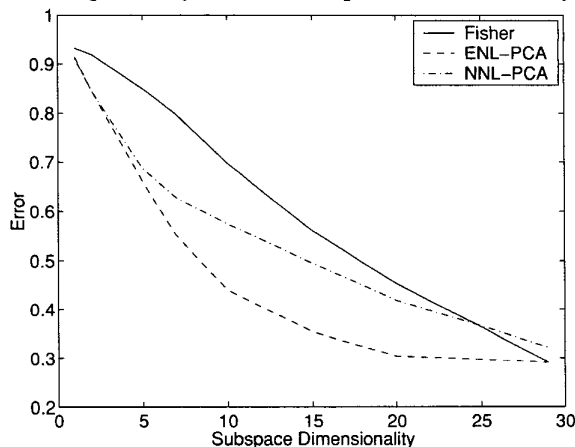


Fig. 3. Classification error as function of the subspace dimensionality for an artificial problem with 30 classes and 30 features.

Monte-Carlo procedure. This experiment is repeated 10 times (10 different sets of means) and the results are averaged.

Fig. 3 presents the classification errors for standard PCA (corresponding to Fisher Mapping before the pre-whitening), the Eigenvector based NL-PCA (ENL-PCA) and the Neural Network based NL-PCA (NNL-PCA). This experiment shows the possibility of a dramatic improvement of NL-PCA over Fisher Mapping. The neural network procedure is for low dimensions equivalent with the eigenvector approach. For larger values it performs significant worse. This, of course, is dependent on implementation of the optimization procedure used in this experiment. (We used Matlab's Neural Network Toolbox). For small dimensionalities the computing times of all three methods are comparable. For larger dimensionalities the computational effort of the neural network procedure is much larger.

For the real dataset we used the Landsat dataset [5], as used in the Statlog project [6]. This is a 6-class problem with 36 features. It has been chosen as its relative large sample size (6435 objects) may show accurate results. A fixed training set of 4435 objects is used for training, leaving 2000 for testing (the same sets as used in Statlog). The three methods are used again for finding subspaces of 1 to 5 dimensions, this time including the pre-whitening step as means and variances are unknown. For each subspace a linear classi-

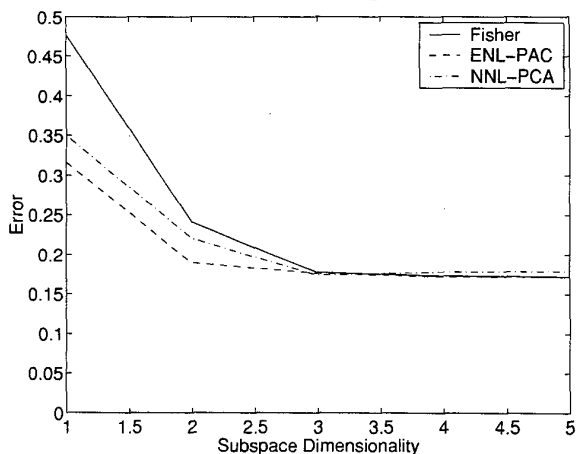


Fig. 4. Classification error as function of the subspace dimensionality for a real problem with 6 classes and 35 features.

fier assuming normal densities is computed on the training data used for obtaining the mapping. The test data is projected in the subspaces and classified. The resulting errors, shown in fig [4], confirm the results of the artificial problem: the NL-PCA methods improve the Fisher Mapping.

5. Conclusions

In this paper we show that the performance of linear dimension reduction in multi-class problems can be dramatically improved by using nonlinear PCA techniques instead of Fisher mapping. These techniques obtain linear maps by a nonlinear criterion that is more sensitive for small class differences. The results hold for normally distributed data with equal covariance matrices.

We studied two methods for nonlinear PCA. One, based on an eigenvector approach, another using neural networks. Both improve the Fisher mapping. The first is almost equally fast and performs in our experiments better than the much slower neural network procedure. This last method, however, may be improved further as it is based in our experiments on just a standard neural network implementation.

6. References

- [1] M. Loog, *Approximate Pairwise Accuracy Criteria for Multi-class Linear Dimension Reduction: Generalisations of the Fisher Criterion*, WBBM Report Series 44, Delft University Press, Delft, The Netherlands, 1999.
- [2] M. Loog, R.P.W. Duin, R. Haeb-Umbach, *Multi-class Linear Dimension Reduction through Weighted Pairwise Fisher Criteria*, 2000, submitted.
- [3] E. Oja, The nonlinear PCA learning rule in independent component analysis, *Neurocomputing*, vol. 17, no. 1, 1997 Sep 30, 25-45.
- [4] P.A. Devijver and J. Kittler, *Pattern recognition: A statistical approach*, Prentice/Hall, London, 1982.
- [5] A. Srinivasan, *Landsat satellite data*, ftp.ics.edu/pub/machine-learning-databases/statlog/satimage, Website maintained by C.J. Merz and P.M. Murphy.
- [6] D. Michie, D.J. Spiegelhalter, and C.C. Taylor, *Machine Learning, Neural and Statistical Classification*, Ellis Horwood, New York, 1994.