An Approach to Nonlinear Principal Components Analysis Using Radially–symmetric Kernel Functions

Andrew R. Webb

Defence Research Agency, St Andrews Road, Malvern

Worcestershire, WR14 3PS, UK

May 1993
Revised, following referees comments, May, 1994.
Further revision, November, 1995: 651/2/JP95/8/2.0
Minor amendment, January, 1996: 651/2/JP95/8/3.0

Abstract

An approach to nonlinear principal components using radially–symmetric kernel basis functions is described. The procedure consists of two steps: a projection of the data set to a reduced dimension using a nonlinear transformation whose parameters are determined by the solution of a generalised symmetric eigenvector equation. This is achieved by demanding a maximum variance transformation subject to a normalisation condition (Hotelling's approach) and can be related to the homogeneity analysis approach of Gifi through the minimisation of a loss function. The transformed variables are the principal components whose values define contours, or more generally hypersurfaces, in the data space. The second stage of the procedure defines the fitting surface, the *principal* surface, in the data space (again as a weighted sum of kernel basis functions) using the definition of self–consistency of Hastie and Stuetzle. The parameters of this principal surface are determined by a singular value decomposition and cross validation is used to obtain the kernel bandwidths. The approach is assessed on four data sets.

Keywords: principal components analysis; principal curves; radial basis functions; homogeneity analysis; functional approximation; self–consistency; cross–validation; least–squares approximation; generalised eigenvalue problem.

1 Introduction

In this paper we propose an approach to nonlinear principal components analysis based on radially–symmetric kernel functions. We seek a nonlinear transformation from a space $\mathbb{R}^n$ to $\mathbb{R}^m$ ($m < n$) such that the variance of the data is maximised (subject to certain normalisation constraints) and we seek to model the transformation as a linear weighted sum of kernels. This may be regarded as a generalisation of Hotelling's derivation of principal components (Hotelling, 1933) to the nonlinear case.

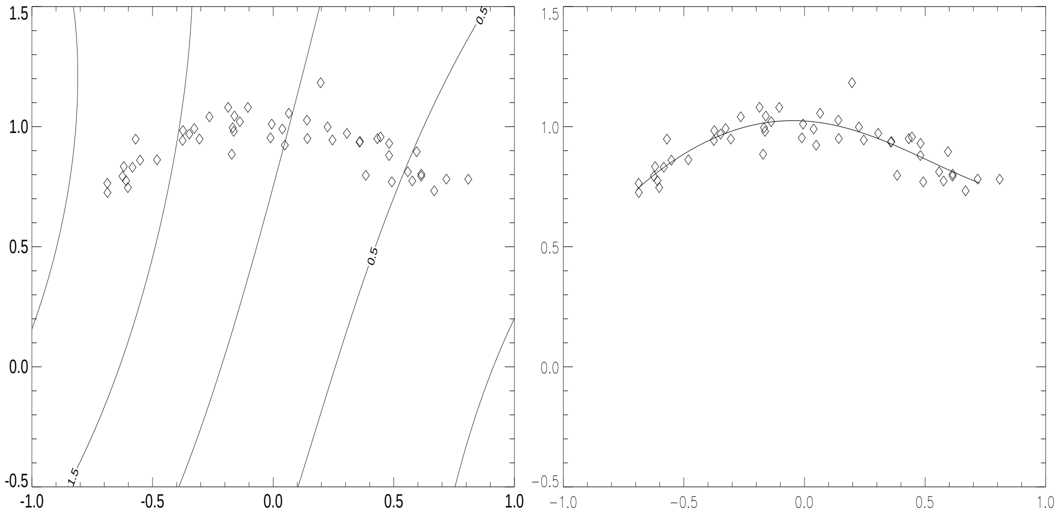

Figure 1: (a) Contour plot (left), and (b) fitting curve (right) for a set of 50 data points in $\mathbb{R}^2$.

In order to clarify the problem that we are addressing, consider the data illustrated in Figure 1. The data lie approximately on an arc of a circle. Performing a linear principal components analysis to produce a mapping to one dimension defines a transformation determined by the eigenvector of the covariance or correlation matrix corresponding to the largest eigenvalue. Each point in the reduced (one–dimensional space) corresponds to a line in the original data space of all points that project to it. Thus, a transformation obtained by linear principal components analysis can be thought of as defining straight lines (more generally hyperplanes) such that all points along the line project to the same value in the transformed space. A nonlinear principal components analysis defines contours (more generally hypersurfaces) in the original space such that all points along a contour project to the same value in the transformed space. These contours are shown in Figure 1a. Another view of linear principal components analysis is of defining lines and planes of closest fit. Similarly, nonlinear principal components defines curves and surfaces of best fit. Figure 1b plot the curve of best fit for a nonlinear principal components analysis of the illustrated data.

The study on kernel methods for nonlinear principal components presented in this paper is related to existing research in several different areas.

1. Principal curves (Hastie and Stuetzle, 1989) are a non–parametric fit to the data. Defined in terms of the concept of *self–consistency*, principal curves provide a nonlinear summary of the data. Further developments of the principal curves approach based on a mixture model have been made by Tibshirani (1992) and LeBlanc and Tibshirani (1994). The work in this paper presents an alternative derivation based on kernel models.

2. Two other different forms of nonlinear principal components analysis for discretized variables are multiple correspondence analysis and non–metric principal components analysis. These have both been integrated using the concept of homogeneity (Bekker and de Leeuw, 1988; de

Leeuw, 1982; Gifi, 1990), where a loss function, measuring the departure from homogeneity, is minimised. Developments of homogeneity analysis to situations in which the number of categories is large, or even to continuous variables, have been considered (Martin, 1988; van Rijckevorsel 1988). The work described in this paper is concerned with continuous variables, and the relationship to homogeneity analysis is considered in Appendix B.

3. Within the neural network literature, various approaches have been used to capture nonlinear relationships between variables for dimension reduction. Several authors have proposed approaches that generalise Pearson's definition of finding lines and planes of closest fit (Pearson, 1901) by finding surfaces of closest fit in the least–squares sense (for example, Kramer, 1991; Nakagawa, Ono and Hirata, 1991; Usui, Nakauchi and Nakano, 1990). The work presented in this paper differs from that in the neural networks literature in two main respects, namely, it is based on a kernel approach and secondly, that the nonlinear principal components are determined by solving a simple generalised symmetric eigenvector equation, rather than a nonlinear optimisation of a cost function.

4. Kernel estimators for nonparametric curve estimation have been used on a broad range of problems including regression, density estimation and discrimination. The use of radial basis functions for multivariable functional interpolation is comparatively recent (Powell, 1990) and least squares estimation of the weights in the context of functional approximation, rather than strict interpolation, was proposed by Broomhead and Lowe (1988). The work described in this paper uses such a kernel–based approach.

The outline of this paper is as follows. Section 2 presents a radial basis function approach to nonlinear principal components analysis, describing the constraints and the numerical solution. Section 3 describes the construction of principal curves using radial basis functions based on the self–consistency definition of Hastie and Stuetzle (1989). Section 4 considers the choice of kernel and Section 5 describes results of nonlinear principal components analysis applied to oxygen uptake data. Finally, we conclude with a discussion and summary.

2 Maximum Variance Approach

Consider a mapping $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$. For the purpose of illustration, we shall take $m = 1$ in the following analysis. The more general case is treated in Appendix A. We shall assume that f is a weighted sum of l basis functions of the form

$$f(\boldsymbol{x}; \boldsymbol{w}) \;=\; \sum_{i=1}^{l} w_i \phi_i(\boldsymbol{x}) \tag{1}$$

where $\phi_i, i = 1, \ldots, l$ is a nonlinear function of $\boldsymbol{x} \in \mathbb{R}^n$ and $\{w_i, i = 1, \ldots, l\}$ is a set of parameters to be determined. The functions ϕ_i themselves could be functions of several unknown parameters, but let us assume for the time being that the ϕ_i are fixed (known) functions of $\boldsymbol{x}$. We seek a solution for the w_i that maximises the variance

$$V \;=\; \langle (f - \overline{f})^2 \rangle \;=\; \boldsymbol{w}^* \boldsymbol{A} \boldsymbol{w}$$

The matrix $\boldsymbol{A}$ is the covariance matrix of the ϕ variables

$$\boldsymbol{A} \;\triangleq\; \langle (\boldsymbol{\phi} - \overline{\boldsymbol{\phi}})(\boldsymbol{\phi} - \overline{\boldsymbol{\phi}})^* \rangle$$

where $\boldsymbol{\phi} = (\phi_1(\boldsymbol{x}), \ldots, \phi_l(\boldsymbol{x}))^*$ and $\overline{\boldsymbol{\phi}}$ is the mean of $\boldsymbol{\phi}$. Imposing a normalisation constraint on $\boldsymbol{w}$ to prevent unconstrained maximisation, namely

$$\boldsymbol{w}^* \boldsymbol{B} \boldsymbol{w} \;=\; 1 \tag{2}$$

where $\boldsymbol{B}$ is a symmetric matrix leads to the maximisation of

$$\boldsymbol{w}^*\boldsymbol{A}\boldsymbol{w} - \lambda\boldsymbol{w}^*\boldsymbol{B}\boldsymbol{w} \tag{3}$$

where λ is a Lagrange multiplier. Maximising (3) yields for $\boldsymbol{w}$ a solution of the general symmetric eigenvector equation

$$\boldsymbol{A}\boldsymbol{w} = \lambda\boldsymbol{B}\boldsymbol{w}$$

If $\boldsymbol{B}$ is of full rank (of rank l), then taking for $\boldsymbol{w}$ the eigenvector (normalised to give $\boldsymbol{w}^*\boldsymbol{B}\boldsymbol{w}$ equal to unity) corresponding to eigenvalue λ gives the variance as

$$V = \boldsymbol{w}^*\boldsymbol{A}\boldsymbol{w} = \lambda$$

Thus, for maximum variance, we take the eigenvector corresponding to the largest eigenvalue and clearly, taking $\boldsymbol{B}$ to be the identity matrix gives the solution for $\boldsymbol{w}$ as the first principal component of the matrix $\boldsymbol{A}$.

However, the constraint matrix we choose to adopt in this work is

$$\boldsymbol{B}_{ij} = \left\langle \frac{\partial\phi_i^*}{\partial\boldsymbol{x}}\frac{\partial\phi_j}{\partial\boldsymbol{x}} \right\rangle \tag{4}$$

where

$$\frac{\partial\phi_i}{\partial\boldsymbol{x}} = \left(\frac{\partial\phi_i}{\partial x_1}, \ldots, \frac{\partial\phi_i}{\partial x_n}\right)^*$$

This constraint arises out of a condition on the gradient of the transformation f. The condition imposed is that the average magnitude squared of the gradient of f is unity; that is, for a transformation of the form (1)

$$\left\langle \frac{\partial f}{\partial\boldsymbol{x}}^*\frac{\partial f}{\partial\boldsymbol{x}} \right\rangle \equiv \boldsymbol{w}^*\boldsymbol{B}\boldsymbol{w} = 1 \tag{5}$$

In the case of a linear transformation (so that $l = n$ and $\phi_i(\boldsymbol{x}) = x_i, i = 1, \ldots, n$), this reduces to the unity normalisation constraint on the transformation vector since the matrix $\boldsymbol{B}$ is the identity matrix. In the nonlinear case, it allows the condition that the fitting surface is allowed to 'stretch' in some regions, provided that in some regions there is a compensating contraction so that the average magnitude of the gradient is unity.

The condition (5) differs from the 'unit–speed' parameterised curves in (Hastie and Stuetzle, 1989), where the gradient of the transformation is constrained to be unity along the direction of the principal curve. In the above, the average of the magnitude squared of the gradient of the transformation is constrained to be unity. In fitting a set of points by a line, the unit–speed condition may be thought of as fitting an inextensible string through the points and the condition above is equivalent to fitting a worm. Although the curves are identical, parameterisation along the curves is different owing to expansion and contraction along the worm's length. However, it is in fitting higher dimensional surfaces where the condition (5) is more appropriate. For example, it allows a hypersphere to be fitted by a rubber sheet, whereas the unit speed condition attempts to fit it by a sheet of paper, giving rise to folds in the surface.

This constraint (4) may give a constraint matrix $\boldsymbol{B}$ that is not positive definite. This depends on the form of the nonlinear functions ϕ_i. In such circumstances, there are three classes of solution to the general symmetric eigenvector equation. The first class is the set of eigenvectors corresponding to eigenvalues with the value zero. The second class is the set of eigenvectors with non–zero, but finite, eigenvalues. Finally, there are is the set of eigenvectors corresponding to those eigenvalues termed 'infinite eigenvalues' (Stewart, 1973, Chapter 7). We are not interested in this latter case since the eigenvectors lie in the null space of $\boldsymbol{B}$ and so it is not possible to satisfy the constraint $\boldsymbol{w}^*\boldsymbol{B}\boldsymbol{w} = 1$. A further complication arises if $\boldsymbol{A}$ and $\boldsymbol{B}$ have common null spaces in which case, the characteristic equation vanishes identically and any value of λ is an eigenvalue.

The numerical procedure for solving the general symmetric eigenvector equation in the case of singular $\boldsymbol{B}$ is the QZ algorithm (Stewart, 1973). This delivers two arrays, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, with the eigenvalues of the system being given by the ratio of the components of these vectors, *i.e.* $\lambda_i = \alpha_i/\beta_i$. We are interested in the eigenvector corresponding to the largest, but finite, eigenvalue. Clearly then, if we permit $\boldsymbol{B}$ to be singular, some threshold must be applied to differentiate between 'large and finite' and 'infinite'.

3 Principal Curves and Self–consistency

The analysis of the previous section provides us with a method of projecting data to a reduced dimension. There is no explicit sense of curve fitting, rather we are defining hypersurfaces in the data space, $\mathbb{R}^{\mathrm{n}}$, of points which map to a given point in $\mathbb{R}^{\mathrm{m}}$. That is, given a point $\boldsymbol{y} \in \mathbb{R}^{\mathrm{m}}$, the transformation defines a family of surfaces given by $\{\boldsymbol{x} \in \mathbb{R}^{\mathrm{n}} | \boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{y}\}$. This was illustrated in Figure 1a for some (broadly) one–dimensional data lying in a plane. The lines shown are contours in $\mathbb{R}^2$ for which $\boldsymbol{f}(\boldsymbol{x})$ is constant. The surfaces are indexed by values of $\boldsymbol{y}$ at constant separation and they cross the data approximately orthogonal to the direction of the line of the data. In linear principal components, $\boldsymbol{f}$ is a linear transformation and we have

$$\boldsymbol{y} \;=\; \boldsymbol{U}^*(\boldsymbol{x} - \boldsymbol{\mu}) \tag{6}$$

where $\boldsymbol{U} = \{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m\}$ is the matrix whose columns are the first m eigenvectors of the covariance matrix and $\boldsymbol{\mu}$ is the mean vector. Equation (6) defines hyperplanes of dimension $\mathbb{R}^{\mathrm{n-m}}$ in $\mathbb{R}^{\mathrm{n}}$ of points that project to the same point in $\mathbb{R}^{\mathrm{m}}$.

In some cases, we may want the best fitting surface to the data points in the original space, $\mathbb{R}^{\mathrm{n}}$, rather than simply the projected data. This defines a transformation back from the reduced dimension space to the original space. This allows a measure of fit analogous to proportion of variance explained by a linear principal components analysis to be defined as follows.

In linear principal components, a point $\boldsymbol{x}$ projects to $\hat{\boldsymbol{x}} = \boldsymbol{U}\boldsymbol{U}^*(\boldsymbol{x} - \boldsymbol{\mu}) + \boldsymbol{\mu}$. In the nonlinear case, we may use the definition of 'self–consistency' (Hastie and Stuetzle, 1989) (see also Flury (1993) for the use of the concept of self–consistency to define the principal points of a random vector) to define the fitting surface, that is the surface $\boldsymbol{g}(\boldsymbol{y}) \in \mathbb{R}^{\mathrm{n}}$ that satisfies

$$E\{\boldsymbol{x} | \boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{y}\} \;=\; \boldsymbol{g}(\boldsymbol{y}) \quad \text{for all } \boldsymbol{y}$$

or

$$\int_{\boldsymbol{f}(\boldsymbol{x})=\boldsymbol{y}} \boldsymbol{x} p(\boldsymbol{x}) d\boldsymbol{x} \;=\; \boldsymbol{g}(\boldsymbol{y}) \tag{7}$$

Thus a point on the fitting surface is the mean over the hypersurface that passes through that point. This is the definition of a self–consistent surface. It may be regarded as regression over a curved trajectory.

The fitting surface for the data in Figure 1a is shown in Figure 1b. Note from the figures that each point $\boldsymbol{x}$ does not project to the nearest point on the surface, but along the contour on which it lies. A point $\boldsymbol{x}^*$ projects to

$$\hat{\boldsymbol{x}} \;=\; \int_{\boldsymbol{f}(\boldsymbol{x})=\boldsymbol{f}(\boldsymbol{x}^*)} \boldsymbol{x} p(\boldsymbol{x}) d\boldsymbol{x}$$

Although in principle Equation (7) can be evaluated using numerical integration, it does require an estimate of the distribution $p(\boldsymbol{x})$ and is impractical for high dimensional problems. In order to obtain a fitting surface, we again adopt a kernel–based approach. We define a transformation from $\mathbb{R}^{\mathrm{m}}$ to $\mathbb{R}^{\mathrm{n}}$ by

$$\boldsymbol{h}(\boldsymbol{y}) \;=\; \sum_{j=1}^{k} \boldsymbol{v}_j \theta_j(\boldsymbol{y}) + \boldsymbol{v}_0 \tag{8}$$

for some kernels $\theta_j, (j = 1, \ldots, k)$, weights $\boldsymbol{V} = (\boldsymbol{v}_1|\ldots|\boldsymbol{v}_k)$, and bias term $\boldsymbol{v}_0$, where k is the number of kernels and $\boldsymbol{v}_i \in \mathrm{I\!R}^{\mathrm{n}}, (\mathrm{i} = 1, \ldots, \mathrm{k})$. We seek a solution for the weights that minimises

$$E \triangleq \frac{1}{N}\sum_{i=1}^{N}|\boldsymbol{x}_i - \boldsymbol{h}(\boldsymbol{y}_i)|^2 \tag{9}$$

where $\boldsymbol{y}_i = \boldsymbol{f}(\boldsymbol{x}_i)$ and $\{\boldsymbol{x}_i, i = 1, \ldots, N\}$ is the given data set. Thus, $\boldsymbol{h}(\boldsymbol{y}_i) = \boldsymbol{h}(\boldsymbol{f}(\boldsymbol{x}_i))$ is the projection of $\boldsymbol{x}_i$ on the fitting surface and we seek the surface of best fit in the least squares sense. Equation (9) may be written as

$$E = \frac{1}{N}\sum_{i=1}^{N}|\boldsymbol{x}_i - \sum_{j=0}^{k}\boldsymbol{v}_j\theta_j(\boldsymbol{y}_i)|^2 \tag{10}$$

where $\theta_o() \equiv 1$. In the limit of a large number of samples this tends to

$$\int |\boldsymbol{x} - \sum_{j=0}^{k}\boldsymbol{v}_j\theta_j(\boldsymbol{y}(\boldsymbol{x}))|^2 p(\boldsymbol{x})d\boldsymbol{x} \tag{11}$$

The solution for the augmented $n \times (k+1)$ matrix $\boldsymbol{V}_0 = [\boldsymbol{v}_0|\boldsymbol{V}]$ that minimises (11) also minimises

$$\int |\boldsymbol{z} - \sum_{j=1}^{k}\boldsymbol{v}_j\theta_j(\boldsymbol{y}_i)|^2 p(\boldsymbol{x})d\boldsymbol{x}$$

where $\boldsymbol{z} = E\{\boldsymbol{x}|\boldsymbol{y}\} = \boldsymbol{g}(\boldsymbol{y})$. Thus the function given by (8) is a least squares approximation to the self–consistent function in the limit of a large number of samples.

The solution for $\boldsymbol{V}_0$ that minimises (10) is

$$\boldsymbol{V}_0^* = (\boldsymbol{\Theta}_0^*\boldsymbol{\Theta}_0)^{-1}\boldsymbol{\Theta}_0^*\boldsymbol{X} \tag{12}$$

provided that the inverse exists, where the $N \times (k+1)$ matrix $\boldsymbol{\Theta}_0 = [\boldsymbol{1}|\boldsymbol{\Theta}]$; the ith column of $\boldsymbol{\Theta}^*$ is $\boldsymbol{\theta}(\boldsymbol{y}_i)$; $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^*$ and $\boldsymbol{X}^* = (\boldsymbol{x}_1|\ldots|\boldsymbol{x}_N)$. Thus, the fitting surface is obtained in two stages:

1. solution of the generalised eigenvector equation (Equation (1)) to obtain the weights for the dimension reducing transformation (3).
2. solution for $\boldsymbol{V}_0$ that minimises the least–squares expression (10), given by (12).

One of the difficulties with any of the methods of intrinsic dimensionality extraction is the determination of the value of the intrinsic dimensionality itself. In the case of a linear transformation, the ratio, η,

$$\eta = \frac{\sum_{i=1}^{m}\lambda_i}{\sum_{i=1}^{n}\lambda_i}$$

of the sum of the m largest eigenvalues of the data covariance matrix to the total sum (the trace of the covariance matrix) gives the fraction of the variance retained in the m–dimensional space defined by the eigenvectors associated with those eigenvalues. One method of estimating the intrinsic dimensionality is to take the value of m for which η is sufficiently large. For example, a value of 0.95 is taken in (Wyse, Dubes and Jain, 1980).

In the nonlinear principal components analysis situation, we may take a measure of performance to be the normalised error the square root of the ratio of the square error in the approximation to the square error between the data and its mean, given by

$$\epsilon = \left(\frac{||\boldsymbol{H} - \boldsymbol{X}||^2}{||\boldsymbol{X} - \hat{\boldsymbol{X}}||^2}\right)^{\frac{1}{2}} \tag{13}$$

where $\boldsymbol{H}^* = (\boldsymbol{h}_1|\ldots|\boldsymbol{h}_N)$ and $\boldsymbol{h}_i = \boldsymbol{h}(\boldsymbol{f}(\boldsymbol{x}_i))$ is the projection on to the fitting surface of the point $\boldsymbol{x}_i$; $\hat{\boldsymbol{X}}$ is a $N \times n$ matrix whose rows are the mean vector of the data matrix $\boldsymbol{X}$, *i.e.* $\hat{\boldsymbol{X}} = \frac{1}{N}\mathbf{1}(\mathbf{1}^*\boldsymbol{X})$. For a linear transformation using the first m principal components

$$\epsilon = \left(\frac{\sum_{i=m+1}^{n}\lambda_i}{\sum_{i=1}^{n}\lambda_i}\right)^{\frac{1}{2}}$$

which gives the relationship between η and the normalised error as

$$\epsilon^2 = 1 - \eta$$

Thus, in the linear case, the condition $\eta > 0.95$ corresponds to $\epsilon^2 < 0.05$ or $\epsilon < 0.224$. In the nonlinear case, we may increase the value of m incrementally and monitor the value of ϵ^2.

4 Choice of Basis Functions

Radial basis functions are related to many other statistical models including nonparametric density estimation based on kernel functions, mixture models of density estimation, kernel approaches to regression and the coding of variables and least–squares estimation of coefficients in homogeneity analysis. However, one major difference between the form of the basis functions used for functional approximation and kernel basis functions in density estimation is that the former are not restricted to be bounded and in fact certain unbounded functions give better performance on some interpolation problems (Powell, 1990; Lowe, 1995). Also, the number of kernels is usually taken to be less than the number of data samples.

In order to specify our model completely, we need to specify the number and positions of the centres, the functional form of the nonlinearities and the smoothing parameters that control the basis function widths.

Choice of functional form

There are many possible functions that can be used for basis functions. Some of the functions adopted are identical (apart from a normalising factor) to the kernels used in density estimation, while others such as thin plate splines ($\phi(r) = r^2\log(r)$) clearly are not (Powell, 1990; Silverman, 1986). It is conceivable that the functional form of the basis functions is different in different parts of the data space, but usually all basis functions are assumed to be identical. The ith basis function is of the form

$$\phi_i(\boldsymbol{x}) = \phi(|\boldsymbol{x} - \boldsymbol{c}_i|)$$

where $\boldsymbol{c}_i$ is a centre. In the experiments described in the following section, we use both the Gaussian form, $\phi(r) = \exp(-r^2/(2\sigma^2))$, and the thin plate spline, $\phi(r) = (r^2/\sigma^2)\log(r/\sigma)$, with a smoothing parameter σ.

Number of centres

In kernel density estimation and indeed in many of the early applications of kernel methods to regression, the number of centres is take to be equal to the number of data points; that is, a kernel is positioned at each sample ($\boldsymbol{c}_i = \boldsymbol{x}_i, i = 1,\ldots,N$) and a sum formed. In a radial basis function approach, with weights determined by a least–squares optimisation procedure, the number of centres is usually a parameter to specify. Thus, there is not necessarily a centre at every data point. There are two main reasons for this. One is computational. Having a centre at each data point leads to a least squares problem with matrices of order $N \times N$. For large N, not only does computation time become excessive, but the matrices involved may become ill–conditioned.

A second reason concerns performance. Having a few well–placed centres may give better performance, in terms of a reduction in the error estimate of the approximator, than having a centre at every data sample. However, having less centres than data samples does not mean that data are thrown away, since all data samples are used in the optimisation process.

Choice of centres

The centres can be chosen randomly from the data set for many applications. An alternative procedure is to cluster the data, perhaps using a k–means algorithm or a normal mixture model with parameters estimated using the EM algorithm for example, and then to choose the cluster means as centres. A k–means procedure is adopted in the experiments of Section 5.

Smoothing parameter

Many of the procedures for estimating the window width in kernel density estimation (Hand, 1981; Hand, 1982; Silverman, 1986) can be employed to determine the smoothing parameter. These include heuristic procedures based on distances to the kth nearest neighbour as well as cross validation approaches. The smoothing parameters may also be variable or adaptive.

5 Results and Comparison

5.1 Implementation Details

Although the solution of the generalised eigenvector equation $\boldsymbol{A}\boldsymbol{w} = \lambda \boldsymbol{B}\boldsymbol{w}$ can be found using the QZ algorithm, the procedure adopted in the following numerical experiments was to use the double precision Nag routine for solving the generalised symmetric eigenvalue problem.

For the low–dimensional example (example 1), Gaussian kernels of the form

$$\phi_i(\boldsymbol{x}) = \exp\left\{-\frac{|\boldsymbol{x}-\boldsymbol{c}_i|^2}{\sigma^2}\right\}$$

were used both for the data reducing (variance maximising) transformation and for the surface approximation transformation that determines the principal curve. These had smoothing parameters σ_1 and σ_2 respectively. For a given number of basis functions, the matrix $\boldsymbol{B}$ was found to be nonsingular over a range of values for the parameter σ_1. Therefore, in all experiments, σ_1 was restricted to this range. There is some empirical evidence (Lowe, 1995) to suggest that in high dimensions, non–compact forms for the kernels offer better performance. Therefore, for the higher–dimensional examples (examples 2, 3 and 4), we used kernels of the form

$$\phi_i(\boldsymbol{x}) = \frac{|\boldsymbol{x}-\boldsymbol{c}_i|^2}{\sigma^2}\ln\left(\frac{|\boldsymbol{x}-\boldsymbol{c}_i|}{\sigma}\right)$$

for both the data reducing and the surface approximation transformations. A cross–validation procedure was used to obtain an estimate of the reconstruction error as a function of σ_1^2 and σ_2^2 (for several values of l, the number of centres) as these parameters were stepped over a coarse grid.

The centres were chosen as a result of a k–means procedure, and the numbers of centres for the data reducing transformation and the surface approximation transformation were taken to be equal. Of course, when the number of centres is a high proportion of the number of data samples, many of the positions produced by the k–means procedure will be at data points for the first transformation.

5.2 Results

Example 1

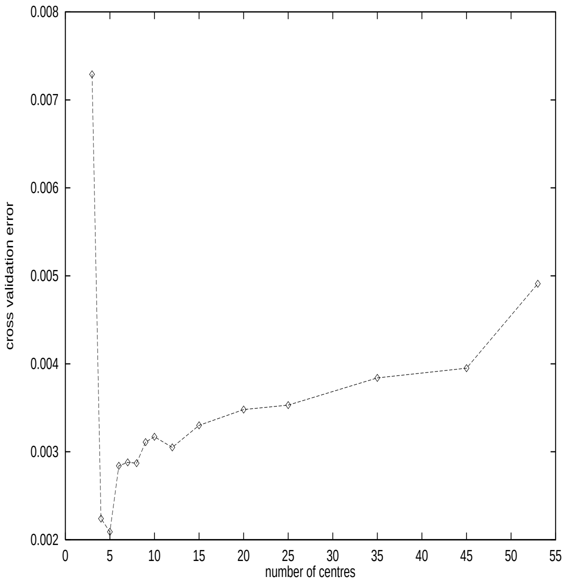

Figure 2: Minimum cross–validation error as a function of the number of centres (the numbers in each transformation were taken to be equal).

This data set used to illustrate a nonlinear principal components analysis for determining relationships between variables is taken from Hand *et al* (1994) and comprises 53 measurements of oxygen uptake and expired ventilation. An analysis of this data is given by Bennett (1988). Here, we seek a projection to one dimension in order to determine the relationship between the variables. Before performing a nonlinear principal components analysis, the data were normalised to zero mean, unit variance.

A cross–validation estimate of the error E was obtained as a function of σ_1^2 and σ_2^2. Figure 2 plots the minimum cross validation error as a function of the number of centres. There is a minimum at 5 centres. Thus, better performance, in terms of minimum cross–validation error, is achieved with having less centres than data points.

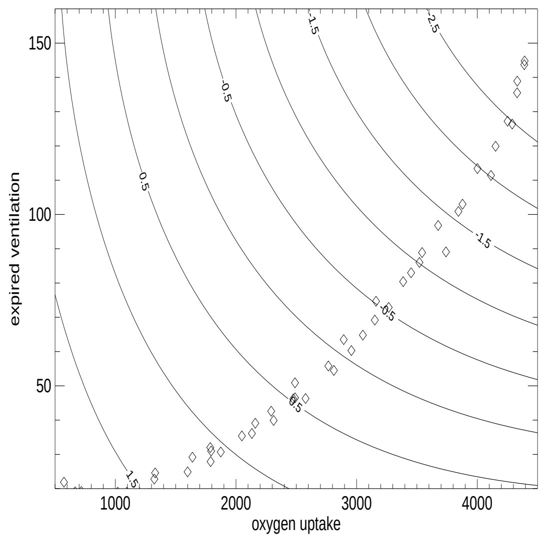

Figure 3: Contour plot for anaerobic threshold data for 5 centres ($\sigma_1^2 = 100$).

Figure 3 plots the contours of constant $\boldsymbol{f}(\boldsymbol{x})$ in the data space (after the reduced–dimension variable has been normalised to zero mean); that is, each point on the contour maps to the same value. The nonlinear relationship between the variables is obtained by plotting expired ventilation against oxygen uptake as the reduced dimension variable is scanned over the range -3.0 to 2.0. Figure 4 plots the fitting surface. A value of the normalised error of $\epsilon = 2.99 \times 10^{-2}$ was found. This corresponds to a value of η of 0.999.

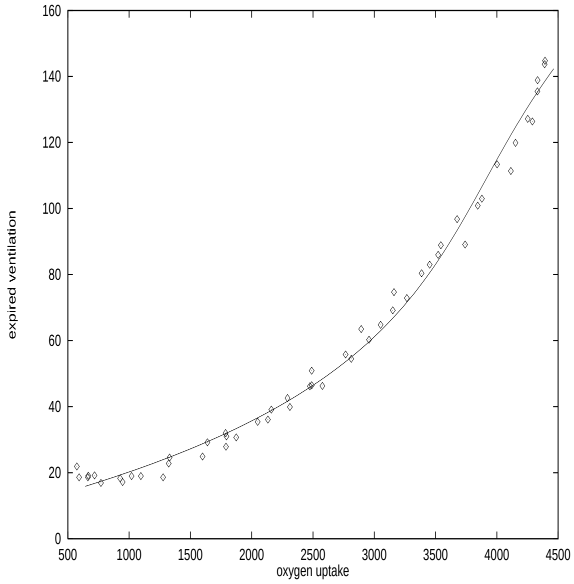

Figure 4: Kernel fit for 5 centres for the anaerobic threshold data ($\sigma_1^2 = 100$, $\sigma_2^2 = 70$).

Example 2

This data set comprises simulated in–situ spectroscopic ellipsometry measurements for complex $Si_{1-x}Ge_x/Si$ structures, where x is the proportion of Ge. Ellipsometry is a technique for determining the optical properties and morphology of a surface from measurements of the change in the polarisation state of light reflected by it. The data have been generated as part of an investigation into techniques for estimating film composition from measurements made as a function of the ellipsometric polarisation angles Ψ and Δ. In this example, the Ψ spectrum is used, sampled at 29 equi–spaced intervals and 51 measurements are made as a uniform layer is deposited (with x=0.15). Six examples of the spectra are shown in Figure 5 (measurements 1, 11, 21, 31, 41, and 51 as the layer is grown).

Table 1 shows the proportion of variance explained by the first 2 principal components for linear and nonlinear principal components. We see that 87% of the variance is accounted for by the first

r :	1	2
Linear PCA:	0.680	0.961
Nonlinear PCA:	0.874	0.999

Table 1: Proportion of variance explained by the first r principal components for ellipsometry data.

nonlinear principal component, compared to 68% for linear PCA.

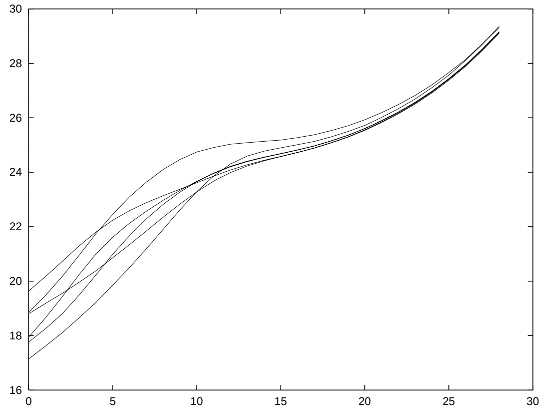

Figure 5: Sample ellipsometry spectra (polarisation angle Ψ against variable number (power)).

Example 3

The data comprise measurements made on a radar focal–plane array. Briefly, an array of detectors is placed on the back of a high dielectric lens. Radiation from a far–field source is focused on to the array and measurements are made on the array outputs as the source is scanned across the field of view in azimuth and elevation. There are 12 detectors in the array and measurements are made for a source on a 19×19 grid. We seek a low–dimensional representation of the data which we would expect to lie approximately on a 2–dimensional manifold in $\mathbb{R}^{12}$. However, there are some discontinuities in the gradient of the manifold at the null of the array element beamshapes.

Table 2 shows the proportion of variance explained by the first 5 principal components for linear and nonlinear principal components. We see that 92% of the variance is accounted for by the first

r :	1	2	3	4	5
Linear PCA:	0.310	0.527	0.687	0.787	0.854
Nonlinear PCA:	0.530	0.923	0.936	0.968	0.984

Table 2: Proportion of variance explained by the first r principal components for radar data.

two nonlinear principal components, compared to 53% for linear PCA.

Example 4

For this example, we take the states data (Becker and Chambers, 1984). These data have been analysed by LeBlanc and Tibshirani (1994) using their adaptive principal surfaces procedure and are used here to provide a comparison with that procedure. The data comprise measurements on 7 variables (population estimate as of July 1, 1975; per capita income, 1974; illiteracy percent of population, 1970; life expectancy in years, 1969–1970; murder and non–negligent manslaughter rate per 100,000 population, 1976; percent high–school graduates, 1970; mean number of days with minimum temperature below freezing, 1931–1960, in capital or large city) for 50 US states.

Following LeBlanc and Tibshirani (1994), we standardised the data so that each variable had median zero and the mean absolute deviation (over 0.67) was unity and we seek a two–dimensional projection for the data. The projection on to the first two linear principal components accounts for 74% of the variance. A nonlinear principal components analysis was performed. The values of the model parameters resulting in the minimum cross–validation estimate of the error were $l = 5, \sigma_1 = 0.01, \sigma_2 = 0.1$. This gives a projection on to two dimensions that captures 78% of the variance, a result identical to that of LeBlanc and Tibshirani.

6 Summary and Discussion

In this paper we have presented a nonlinear nonparametric approach to dimension reduction using spherically–symmetric kernel functions. It provides a generalisation to linear principal components analysis. The basic concept is simple and can be easily implemented using standard routines for solving the generalised eigenvector equation and singular value decomposition routines. The most computationally demanding part of the procedure as presented in this paper is the choice of smoothing parameters for the kernels used. However, close to optimal performance is achieved over a wide range of smoothing parameter values and it is to be expected that some of the simple heuristics used in density estimation (Hand, 1981) for determining the smoothing parameter will offer acceptable performance. This is an area for further study.

There are several other areas that require further investigation. The kernel functions used in the experiments reported were all identical, namely spherically–symmetric gaussian or thin–plate spline kernels, for both the dimension–reducing transformation and the principal surface transformation. Other forms may be more appropriate for functional approximation and indeed variable or adaptive kernel procedures may prove important for certain classes of problem. The effects of bias have not been considered. The self–consistent estimate produces a biased estimate of the surface for distributions of the form $\boldsymbol{f}(\theta)+\boldsymbol{n}$ (Hastie and Stuetzle, 1989; Tibshirani, 1992). Tibshirani proposes an attractive alternative definition of principal surfaces, based on a mixture model, that does not suffer from this defect. As presented in this paper, the procedure will not produce principal surfaces that are closed, unlike that of (Hastie and Stuetzle, 1989). This is because the principal surface is the mean of the data along a contour of the dimension reducing transformation and this transformation is continuous and differentiable. The incorporation of topological constraints might be important for certain classes of problem. Although the procedure is straightforward to apply using standard numerical routines, we have not explored the use of other numerical procedures, perhaps based on the alternating least squares principle (Gifi, 1990). In some signal processing applications, efficient algorithms are essential.

To summarise, a simple procedure for nonlinear principal components has been presented and illustrated on four data sets. A nonlinear transformation can be achieved by the solution of a single generalised eigenvector equation.

References

Becker, R.A. and Chambers, J.M. (1984). *S. An Interactive Environment for Data Analysis and Graphics.* Wadsworth Statistics/Probability Series, Belmont, California.

Bekker, P. and de Leeuw, J. (1988). Relations between variants of non–linear principal components analysis. In J.L.A. van Rijckevorsel and J. de Leeuw, editors, *Component and Correspondence Analysis. Dimension Reduction by Function Approximation*, Wiley Series in Probability and Mathematical Statistics, pages 1–31. John Wiley and Sons Ltd., New York.

Bennett, G.W. (1988). Determination of anaerobic threshold. *The Canadian Journal of Statistics*, 16(3), 307–316.

Broomhead, D.S. and Lowe, D. (1988). Multi–variable functional interpolation and adaptive networks. *Complex Systems*, 2(3), 269–303.

de Leeuw, J. (1982). Nonlinear principal components analysis. In H. Caussinus, editor, *COMPSTAT '82. Proceedings in Computational Statistics.* Physica–Verlag, Vienna.

Flury, B.D. (1993). Estimation of principal points. *Applied Statistics*, 42(1), 139–151.

Gifi, A. (1990). *Nonlinear Multivariate Analysis.* Wiley Series in Probability and Mathematical Statistics. John Wiley and Sons Ltd., New York.

Hand, D.J. (1981). *Discrimination and Classification.* John Wiley and Sons, New York.

Hand, D.J. (1982). *Kernel Discriminant Analysis*, volume 2 of *Pattern Recognition and Image Processing Research Studies Series.* Research Studies Press, Letchworth, Herts.

Hand, D.J., Daley, F., Lunn, A.D., McConway, K.J. and Ostrowski, E. (1994). *A Handbook of Small Data Sets.* Chapman and Hall, London.

Hastie, T. and Stuetzle, W. (1989). Principal curves. *Journal of the American Statistical Association*, 84(406), 502–516.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24, 417–441, 498–520.

Kramer, M.A. (1991). Nonlinear principal component analysis using autoassociative neural networks. *American Institute of Chemical Engineers Journal*, 37(2), 233–243.

LeBlanc, M. and Tibshirani, R. (1994). Adaptive principal surfaces. *Journal of the American Statistical Association*, 89(425), 53–664.

Lowe, D. (1995). On the use of nonlocal and non positive definite basis functions in radial basis function networks. Fourth IEE International Conference on Artificial Neural Networks, Cambridge, 206–211. IEE Conference Publication 409.

Martin, J.-F. (1988). On probability coding. In J.L.A. van Rijckevorsel and J. de Leeuw, editors, *Component and Correspondence Analysis. Dimension Reduction by Function Approximation*, Wiley Series in Probability and Mathematical Statistics, pages 103–114. John Wiley and Sons Ltd., New York.

Nakagawa, S., Ono, Y. and Hirata, Y. (1991). Dimensionality reduction of dynamical patterns using a neural network. In B.H. Juang, S.Y. Kung, and C.A. Kamm, editors, *Neural Networks for Signal Processing, Proceedings of the 1991 IEEE Workshop*, pages 256–265, Princeton, New Jersey.

Pearson, K. (1901). On lines and planes of closest fit. *Philosophical Magazine*, 6, 559–572.

Powell, M.J.D. (1990). The theory of radial basis function approximation in 1990. DAMPT Numerical Analysis Report 1990/NA11, University of Cambridge, Department of Applied Mathematics and Theoretical Physics, Silver Street, Cambridge CB3 9EW.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*, volume 26 of *Monographs on Applied Probability and Statistics.* Chapman and Hall, London.

Stewart, G.W. (1973). *Introduction to Matrix Computations.* Computer Science and Applied Mathematics. Academic Press, London.

Tibshirani, R. (1992). Principal curves revisited. *Statistics and Computing,* 2(4), 183–190.

Usui, S., Nakauchi, S. and Nakano, M. (1990). Reconstruction of Munsell color space by a five–layered neural network. In *IJCNN Int. Joint Conf. Neural Networks*, volume II, pages 515–520, San Diego.

van Rijckevorsel, J.L.A. (1988). Fuzzy coding and B–splines. In J.L.A. van Rijckevorsel and J. de Leeuw, editors, *Component and Correspondence Analysis. Dimension Reduction by Function Approximation*, Wiley Series in Probability and Mathematical Statistics, pages 33–54. John Wiley and Sons Ltd., New York.

Wyse, N., Dubes, R. and Jain, A.K. (1980). A critical evaluation of intrinsic dimensionality algorithms. In E.S. Gelsema and L.N. Kanal, editors, *Pattern Recognition in Practice*, pages 415–425. North–Holland Publishing Company, Amsterdam.

Appendix A Maximum Variance Solution for Arbitrary Dimensions

From Section 2, the solution for $\boldsymbol{w}$ that minimises the variance $\boldsymbol{w}^*\boldsymbol{A}\boldsymbol{w}$ subject to $\boldsymbol{w}^*\boldsymbol{B}\boldsymbol{w} = 1$ is given by the solution of the generalised eigenvector equation

$$\boldsymbol{A}\boldsymbol{w} = \lambda\boldsymbol{B}\boldsymbol{w}$$

with the largest (finite) eigenvalue. For a mapping to a dimension $m > 1$, we impose the constraint that

$$\left\langle \frac{\partial f_i}{\partial \boldsymbol{x}}^* \frac{\partial f_j}{\partial \boldsymbol{x}} \right\rangle \equiv [\boldsymbol{W}^*\boldsymbol{B}\boldsymbol{W}]_{ij} = [\boldsymbol{I}_m]_{ij} \quad 1 \leq i,j \leq m$$

where $\boldsymbol{W} = [\boldsymbol{w}_1|\ldots|\boldsymbol{w}_m]$ and $\boldsymbol{I}_m$ is the m–dimensional identity matrix. Thus, the $\boldsymbol{w}_i$ satisfy $\boldsymbol{w}_i^*\boldsymbol{B}\boldsymbol{w}_j = \delta_{ij}$. We seek stationary points of the variance $\boldsymbol{w}^*\boldsymbol{A}\boldsymbol{w}$ subject to this constraint. The solutions of the generalised eigenvector equation above satisfy this condition and may be ordered in terms of decreasing eigenvalues.

Appendix B Links with Homogeneity Analysis

Another approach to nonlinear principal components is that provided by homogeneity analysis (Gifi, 1990; van Rijckevorsel and de Leeuw, 1988). Let us assume that we have a set of N observations and measurements are made on n variables. Let the $N \times n$ matrix of (zero–mean) observations be $\boldsymbol{H}$, with jth column the N–dimensional vector $\boldsymbol{h}_j, j = 1, \ldots, n$. In the homogeneity analysis approach, a sum–of–squares loss function of the form

$$s(\boldsymbol{t}; \boldsymbol{w}) \;=\; \frac{1}{n} \sum_{j=1}^{n} |\boldsymbol{t} - a_j \boldsymbol{h}_j|^2 \tag{14}$$

is minimised with respect to the *weights* $\boldsymbol{a}$ and the *scores* $\boldsymbol{t}$ (an N–dimensional vector). Unconstrained minimisation will obviously lead to $s = 0$. Therefore, a constraint is applied either to the scores or the weights. For example, minimising (14) subject to the constraint on the weights $\boldsymbol{a}$

$$\boldsymbol{a}^* \boldsymbol{D} \boldsymbol{a} \;=\; 1$$

where $\boldsymbol{D}$ is the $n \times n$ diagonal matrix, $\mathrm{diag}\{\boldsymbol{H}^* \boldsymbol{H}\}$, leads to the eigenvector equation

$$\frac{1}{n} \boldsymbol{H}^* \boldsymbol{H} \boldsymbol{a} \;=\; (1 - \lambda) \boldsymbol{D} \boldsymbol{a}$$

where λ is a Lagrange multiplier. This may be rewritten as

$$\left(\boldsymbol{H} \boldsymbol{D}^{-\frac{1}{2}} \right)^* \left(\boldsymbol{H} \boldsymbol{D}^{-\frac{1}{2}} \right) \boldsymbol{D}^{\frac{1}{2}} \boldsymbol{a} \;=\; n(1 - \lambda) \boldsymbol{D}^{\frac{1}{2}} \boldsymbol{a}$$

showing that $\boldsymbol{D}^{\frac{1}{2}} \boldsymbol{a}$ is an eigenvector of $(\boldsymbol{H} \boldsymbol{D}^{-\frac{1}{2}})^* (\boldsymbol{H} \boldsymbol{D}^{-\frac{1}{2}})$, or a singular vector of $\boldsymbol{H} \boldsymbol{D}^{-\frac{1}{2}}$. The loss function s takes the value $s = \lambda$ and therefore the solution for minimum loss is that for which $1 - \lambda$ is largest; the solution for $\boldsymbol{D}^{\frac{1}{2}} \boldsymbol{a}$ is the eigenvector corresponding to the largest eigenvalue.

If we take the matrix $\boldsymbol{H}$ to be the matrix of nonlinear functions of the observations $\boldsymbol{x}_i, i = 1, \ldots, N$ (that is, the coded observations) given by

$$\boldsymbol{H}_{ij} = \boldsymbol{\Phi}_{ij} \;\triangleq\; \phi_j(\boldsymbol{x}_i) - \overline{\phi}_j \quad i = 1, \ldots, N; j = 1, \ldots, n$$

so that the ith column of $\boldsymbol{\Phi}^*$ is $\boldsymbol{\phi}(\boldsymbol{x}_i) - \overline{\boldsymbol{\phi}}$, and $\overline{\phi}_j$ is the j component of $\overline{\boldsymbol{\phi}} = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{\phi}(\boldsymbol{x}_i)$, then $\boldsymbol{D}^{-\frac{1}{2}} \boldsymbol{H}^* \boldsymbol{H} \boldsymbol{D}^{-\frac{1}{2}}$ is the correlation matrix in the space spanned by the outputs of the nonlinear functions ϕ and the minimum loss solution for $\boldsymbol{D}^{\frac{1}{2}} \boldsymbol{a}$ is the first principal component.

Let us now develop this approach by introducing the symmetric positive semi–definite constraint matrix $\boldsymbol{B}$, which can be written $\boldsymbol{B} = \boldsymbol{U} \boldsymbol{\Sigma} \boldsymbol{U}^*$, where $\boldsymbol{U}$ is an orthogonal matrix and $\boldsymbol{\Sigma}$ is diagonal. Now, if we allow $\boldsymbol{H}$ to incorporate the constraint matrix by taking it to be the matrix $\boldsymbol{\Phi} \boldsymbol{U}$, then minimising the loss subject to $\boldsymbol{a}^* \boldsymbol{D} \boldsymbol{a} = 1$, where now

$$\boldsymbol{D} \;=\; \mathrm{diag}\{\boldsymbol{H}^* \boldsymbol{H}\} \;=\; \mathrm{diag}\{\boldsymbol{U}^* \boldsymbol{\Phi}^* \boldsymbol{\Phi} \boldsymbol{U}\},$$

yields the equation for $\boldsymbol{a}$

$$\boldsymbol{U}^* \boldsymbol{\Phi}^* \boldsymbol{\Phi} \boldsymbol{U} \boldsymbol{a} \;=\; n(1 - \lambda) \boldsymbol{D} \boldsymbol{a}, \tag{15}$$

with the minimum loss solution for $\boldsymbol{D}^{\frac{1}{2}} \boldsymbol{a}$ being the eigenvector of $(\boldsymbol{\Phi} \boldsymbol{U} \boldsymbol{D}^{-\frac{1}{2}})^* (\boldsymbol{\Phi} \boldsymbol{U} \boldsymbol{D}^{-\frac{1}{2}})$ corresponding to the largest eigenvalue.

How do we justify the particular choice of $\boldsymbol{\Phi} \boldsymbol{U}$ for $\boldsymbol{H}$? Let us rewrite (15) as

$$\boldsymbol{\Phi}^* \boldsymbol{\Phi} \boldsymbol{w} \;=\; n(1 - \lambda)(\boldsymbol{U} \boldsymbol{D} \boldsymbol{U}^*) \boldsymbol{w} \tag{16}$$

where $\boldsymbol{w} = \boldsymbol{U}\boldsymbol{a}$. Then since the matrix $\boldsymbol{\Phi}^*\boldsymbol{\Phi}$ is proportional to the sample covariance matrix $\boldsymbol{A}$ of the ϕ variables ($\boldsymbol{\Phi}\boldsymbol{\Phi}^* = (N-1)\boldsymbol{A}$), Equation (16) can also be derived by finding the linear transformation of the variables that maximises the variance $\boldsymbol{w}^*\boldsymbol{\Phi}^*\boldsymbol{\Phi}\boldsymbol{w}$ subject to the normalisation constraint

$$\boldsymbol{w}^*\boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^*\boldsymbol{w} = 1 \qquad (17)$$

Thus, the minimum loss solution can be derived in a similar manner to the nonlinear principal components analysis derivation of Section 2, namely by maximising a variance subject to a constraint.

But why should we choose this constraint? If we assume that $\boldsymbol{B}$ is nonsingular ($\boldsymbol{\Sigma}^{-1}$ exists) and denoting $\boldsymbol{C} = \boldsymbol{U}\boldsymbol{\Sigma}^{\frac{1}{2}}$, this constraint can be written

$$\boldsymbol{w}^*\boldsymbol{C}\boldsymbol{D}'\boldsymbol{C}^*\boldsymbol{w} = 1 \qquad (18)$$

where

$$\boldsymbol{D}' = \mathrm{diag}\{\boldsymbol{C}^{-1}\boldsymbol{A}(\boldsymbol{C}^{-1})^*\}$$

The condition (18) is a normalisation constraint on the matrix $\boldsymbol{A}$ when a constraint matrix $\boldsymbol{B}$ is present. Clearly, if $\boldsymbol{B}$ were the identity matrix, then $\boldsymbol{D}' = \mathrm{diag}(\boldsymbol{A})$ and the constraint on the weights $\boldsymbol{w}$ is equivalent to finding the eigenvectors of the correlation matrix, rather than the covariance matrix, $\boldsymbol{A}$. In general, the condition normalises the matrix $\boldsymbol{A}$ and incorporates a constraint matrix $\boldsymbol{B}$ into the adaptation to heterogeneous variances of the basic homogeneity analysis approach (Gifi, 1990, Chapter 3).

Thus, the approach described in Section 2 can be phrased as minimising a loss function subject to appropriate constraints on the weights. The link with homogeneity analysis also opens up the possibility of efficient algorithms based on the alternating least–squares approach.