Principal Components Analysis with Spline Optimal Transformations for Continuous Data

Nuno Lavado, *Member, IAENG,* and Teresa Calapez

Abstract—**A new approach to generalize Principal Components Analysis in order to handle nonlinear structures has been recently proposed by the authors: quasi-linear PCA (qlPCA). It includes spline transformation of the original variables and the qualifier *quasi* was chosen to emphasize the exclusive use of linear splines. Alternating least squares fitting of a suitable objective loss function is the mechanism for achieving spline optimal transformation and nonlinear principal components. Optimal transformations are explicitly known after convergence and allow a straightforward projection of new observations onto the nonlinear principal components space as well as reconstruction the original variables. QlPCA reports model summary in a linear PCA fashion and allows the introduction of the *piecewise loadings* concept. This paper provides further details on qlPCA and its properties. Results of a simulation study are also presented.**

Index Terms—**nonlinear principal components analysis, qlPCA, linear PCA, CATPCA.**

I. Introduction

PRINCIPAL Components Analysis (PCA) being probably the most common descriptive multivariate technique for seeking linear structure in data, it is unsurprising to find a number of attempts at generalizing it in order to handle nonlinear structures. The concept behind PCA is to project the original data, which includes noise and redundant variables, into a latent space with the objective of capturing its true dimensionality.

All descriptive methods for dimension reduction share the same basic premise and general objectives: the original data can be viewed as a collection of n points in some high m-dimensional space, the points corresponding to sample individuals and the dimensions to measured variables, and we seek for a suitable low p-dimensional approximation in which the points are positioned such that as much information as possible is retained from the original space. By reducing the dimensionality, one can interpret few components rather than a large number of variables. Different interpretations of the phrase "as much information as possible" lead to the different multivariate techniques for dimension reduction.

One technique can be described as *linear* when the high-dimensional set of coordinates is replaced by another in a one-to-one *linear* relation with it. All attempts to generalize PCA in order to handle nonlinear structures, the generally denominated Nonlinear Principal Components Analysis (NLPCA), share the basic premise and general objectives mentioned, but they address the nonlinearity problem by relaxing the linear restrictions between spaces.

An early attempt to generalize PCA was made by Gnanadesikan and Wilk in the '60's. The idea was to extend the m-dimensional space by adding nonlinear functions of the original variables (quadratic and higher order terms) and then perform PCA on the expanded set of variables [1]. The key to this approach was to decide on the appropriate dimensionality of the extended space as well as the nonlinear relationships between the original variables needed to describe the system. This drawback was removed in the '90's by Schölkopf [2] using a function from the original m-dimensional space onto an arbitrarily high-dimensional space (known as feature space in the machine learning community) which "automatically" carries out the nonlinear mapping. It turns out that this mapping can be performed implicitly by using kernel functions and therefore does not need to be specified. This approach, known as kernel PCA, applies linear PCA on the feature space. Recently, Kruger [3] reviews existing work on NLPCA and points out that it can be divided into the utilization of autoassociative neural networks, principal curves and manifolds, kernel approaches or the combination of these approaches.

NLPCA's most known approaches among researchers dealing with continuous variables do not include the state-of-the-art to perform NLPCA for ordinal and nominal data, CATegorical PCA (CATPCA). We refer to a continuous variable as one with small marginal frequencies on every value, typically one or two. CATPCA performs quite well when applied to categorical variables and is more appropriate when [4]: (1) the data at hand contains categorical variables, and/or (2) the variables in the data are (or may be) nonlinearly related to each other. If the variables are nonlinearly related, CATPCA will be able to account for more of the variance in the data, and may enhance the interpretation of the solution compared to linear PCA. CATPCA's algorithm was developed in the '90's as an algorithm for categorical data analysis, thus for dealing with integer valued variables. Continuous data need to undergo a discretization process before the algorithm starts. Various discretization options are available for recoding continuous data within the procedure and one can always recode data beforehand. Our proposal is to adjust the algorithm to allow continuous values directly so that researchers dealing with continuous variables avoid thinking that some information is being neglect.

A new approach on generalizing PCA in order to handle nonlinear structures have been recently proposed by the authors [5], quasi-linear PCA (qlPCA). The proposed approach was inspired by the Gifi system [6], also called Homogeneity Analysis, in particular by its natural successor CATPCA. QlPCA recovers the spline based algorithm in CATPCA, and introduces continuous variables into the framework directly

Submitted October 27, 2011.

N. Lavado is with the Department of Physics and Mathematics, Coimbra Institute of Engineering (ISEC), Coimbra, Portugal and with the research unit Instituto Universitário de Lisboa (ISCTE-IUL), Unidade de Investigação em Desenvolvimento Empresarial (Unide-IUL), Lisboa, Portugal, e-mail: nlavado@isec.pt.

T. Calapez is with the research unit Instituto Universitário de Lisboa (ISCTE-IUL), Unidade de Investigação em Desenvolvimento Empresarial (Unide-IUL), Lisboa, Portugal, e-mail: teresa.calapez@iscte.pt.

without the need of any discretization process. Thus, this approach is more precise with regard to continuous variables and provides a better approximation of a strictly nonlinear analysis, becoming a valid option to perform NLPCA for those variables.

This paper provides further details on qlPCA and its properties. A brief review on splines is provided in the next section followed by an overview of the Gifi system and CATPCA in section III. A suitable objective (loss) function is defined in section IV as well as details on qlPCA's algorithm. The main properties of qlPCA are reported in section V: model summary, choosing the appropriate number of components, projecting new observations onto the nonlinear principal components' space, reconstruction and *piecewise loadings*' definition. Results of a simulation study are provided on section VI.

II. Spline's brief review

Low order polynomial spline functions play an important roll in our quasi-linear PCA (qlPCA) proposal. In this section it will be provided the main results on spline functions that will be useful for our purposes. For a comprehensive overview see [7], [8].

A function f over $[a, b]$, is a *polynomial spline* of degree v if within any subinterval is defined by a polynomial of degree v that join smoothly with adjacent ones. Although it is possible to define several degrees of smoothness at the boundaries of each subinterval [9], in the most common situations adjacent polynomials have matching derivatives up to order $v-1$ at each boundary point in the interior of $[a, b]$. Examples:

1) The first order (or zero degree) spline is a piecewise constant function, discontinuous at the interior knots;
2) The second order spline is a continuous piecewise linear function;
3) The third order spline is a piecewise quadratic function with matching first derivatives at the interior knots.

It can be shown [7], [8] that the set of splines of degree v with r interior knots is a linear space of functions, with dimension $w = v + 1 + r$, equal to the spline's order plus the number of interior knots. In 1966, Curry and Schoenberg [7] have built a basis, what they called *B-splines*, which revealed to be especially convenient for computation.

In order to provide a spline's representation convenient from the points of view of application and computation, smoothness conditions are incorporated into a knot sequence $\{t\} = \{t_1, \ldots, t_{2v+r+2}\}$ where:

1) $t_1 \leq \ldots \leq t_{2v+r+2}$;
2) $t_1 = \ldots = t_{v+1} = a$;
3) $t_{v+r+2} = \ldots = t_{2v+r+2} = b$;
4) $t_{v+2}, \ldots, t_{v+r+1}$ are the r interior knots.

Given the knot sequence $\{t\}$, spline's basis of order $v+1$ is defined, for all $q = 1, 2, \ldots, w$, by the recursive relation

$$B_q^{[1]}(x) = \begin{cases} 1\,, & t_q \leq x < t_{q+1} \\ 0\,, & \text{otherwise} \end{cases} ,$$

$$B_q^{[v+1]}(x) = \frac{x - t_q}{t_{q+v} - t_q} B_q^{[v]}(x) + \frac{t_{q+v+1} - x}{t_{q+v+1} - t_{q+1}} B_{q+1}^{[v]}(x),$$

where

$$\frac{x - t_q}{t_{q+v} - t_q} B_q^{[v]}(x) \text{ and } \frac{t_{q+v+1} - x}{t_{q+v+1} - t_{q+1}} B_{q+1}^{[v]}(x)$$

are equal to zero when the denominators are zero.

Another set of basis splines particularly appealing to statisticians is the *M-spline* basis,

$$M_q^{[v+1]} = \frac{v+1}{t_{q+v+1} - t_q} B_q^{[v+1]} \quad , \; q = 1, \ldots, w. \tag{1}$$

It can be shown [8] that $M_q^{[v+1]}$ is positive and less than 1 over $]t_q, t_{q+v+1}[$, zero elsewhere and also that $\int_{-\infty}^{\infty} M_q^{[v+1]}(x)\, dx = 1$. Thus $M_q^{[v+1]}$ is a probability density function. Monotone transformations can be obtained using a monotone splines' basis together with nonnegative coefficients. Since each *M-spline* has the properties of a probability density function, another basis can be obtained using the corresponding distribution function: integrated splines or *I-splines* [10].

Given the knot sequence $\{t\}$ the *I-spline of order* $v+2$ is defined, for all $q = 1, 2, \ldots, w$ by

$$I_q^{[v+2]}(x) = \int_{-\infty}^{x} M_q^{[v+1]}(u)\, du. \tag{2}$$

Since each *M-spline* is a piecewise polynomial of degree v, the associated *I-spline* is a piecewise polynomial of degree $v+1$. Thus, the related space has dimension $w+1$, being w the associated *M-spline*'s dimensionality. However, by construction, there are only w independent *I-splines*, thus we can only get the subspace spanned by those. From now on w refers to this subspace's dimension being $w = v + r$ for a degree v spline with r interior knots.

From definition it follows that $I_q^{[v+2]}$ is non-constant over $]t_q, t_{q+v+1}[$, zero bellow t_q and one above t_{q+v+1}. *I-splines*' definition by a recurrence relation provides a convenient computational approach. As an illustrative example on how to write a *I-spline* on a given basis let's consider linear splines with one interior knot at the median and an input continuous variable x with minimum m_1, median m_2 and maximum m_3. The basis elements are the following:

$$I_1(x) = \begin{cases} 0, & x < m_1 \\ \frac{x - m_1}{m_2 - m_1}, & m_1 \leq x < m_2 \\ 1 & x \geq m_2 \end{cases}$$

$$I_2(x) = \begin{cases} 0, & x < m_2 \\ \frac{x - m_2}{m_3 - m_2}, & m_2 \leq x < m_3 \\ 1 & x \geq m_3 \end{cases}$$

To obtain a spline function of degree one with one interior knot by recurrence it will be necessary in the first place to compute the set of two *M-splines* basis functions of degree zero with one interior knot. As these basis functions are piecewise polynomial of degree zero, each I_i will be a piecewise polynomial of degree one as needed. Therefore this set of *I-splines* basis functions will have two elements. As the entire space of degree one spline functions with one interior knot is three-dimensional, the referred set can only generate one of its subspaces.

Figure 1 displays the family of *I-splines* of degree one defined on $[0, 1]$ with one interior knot at the median. Each *I-spline* is piecewise linear and non-constant over one interval. It also displays an example of the images obtained by each of the three functions (two basis functions and one spline) for a value of x above the median ($x = 0.58$, $I_1(0.58) = 1$, $I_2(0.58) = 0.29$ and $f(0.58) = 1.37$).

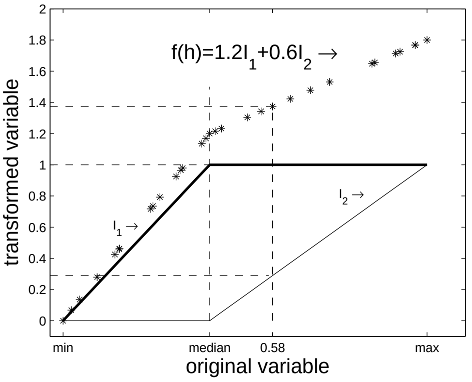

Fig. 1. *Spline* of degree one with one interior knot at the median. I_1 and I_2 are the basis spline functions. Stars represent the images of x through the *spline* obtained as a linear combination of I_1 and I_2 with coefficients 1.2 and 0.6, respectively.

A degree n polynomial is determined by $n + 1$ points. Therefore, each spline's segment in Figure 1 can be defined using the points associated with the minimum/median and median/maximum. However, qlPCA algorithm will search for the optimal (as defined in section IV) linear combination of *I-splines* by means of a multivariate linear regression with I_1 and I_2 as predictor variables and thus involving whole data and not only those two points.

III. Gifi system and CATPCA

For a comprehensive overview on the Gifi system see [6], a recent review is given by [11] and [4].

The central idea of the Gifi system is the notion of *optimal scaling* and its implementation through an alternating least squares (ALS) algorithm. The optimal scaling process as defined by the Gifi system is a transformation of variables by assigning quantitative values to qualitative variables in order to optimize a fixed criterion. Optimality is a relative notion, however, because it is always obtained with respect to the particular data set being analyzed and also depends on the class of admissible transformations. This process (optimal quantification, optimal scaling, optimal scoring) allows nonlinear transformations of variables. Variable transformation has become an important tool in data analysis over the last decades. For an historical overview see [4].

One of the optimal scaling procedures for dimension reduction and its SPSS implementation - CATPCA - was developed by the Data Theory Scaling System Group (DTSS), consisting of members of the departments of Education and Psychology of the Faculty of Social and Behavioral Sciences at Leiden University.

The CATPCA algorithm is the state-of-the-art to perform nonlinear PCA for ordinal and nominal data [4] and is available since 1999 within SPSS Categories 10.0 onwards [12]. The traditional crisp coding of the categorical variables was maintained and the least squares estimation of the *spline* coefficients is performed by a multivariate regression on each iteration of the ALS procedure. This approach performs quite well when applied to categorical variables, but it needs an a priori discretization process for quantitative variables or categorical not coded in the traditional way. And by so it is no longer precise with regard to quantitative variables.

CATPCA procedure simultaneously quantifies *m* categorical variables while reducing the dimensionality of the data. The technique consists of finding object scores $\mathbf{X}$ of order $n \times p$ (i.e. n = number of cases-objects, p = number of dimensions) and sets of multiple category quantifications $\mathbf{Y}_j$ of order $k_j \times p$ (i.e. k_j = number of categories of each variable and $j = 1, \ldots, m$) so that the loss function:

$$\sigma(\mathbf{X}, \mathbf{Y}) = n^{-1} \sum_j tr\left[(\mathbf{X} - \mathbf{G}_j\mathbf{Y}_j)'(\mathbf{X} - \mathbf{G}_j\mathbf{Y}_j)\right] \quad (3)$$

is minimal, under the normalization restriction $\mathbf{X}'\mathbf{X} = n\mathbf{I}$, where:

- $\mathbf{G}_j$ is an indicator matrix for variable j, of order $n \times k_j$, whose elements are 0 when the *i-th* object is not in the *r-th* category of variable j and 1 when the *i-th* object is in the *r-th* category of variable j;
- $\mathbf{I}$ is the $p \times p$ identity matrix.

The algorithm uses ALS to minimize the loss function. It consists of two phases, a model estimation phase and an optimal scaling phase, iteratively alternated until convergence is reached. Both object scores and category quantifications are alternately updated until the optimum is found.

IV. Quasi-linear PCA

CATPCA's algorithm was developed in the '90's as an algorithm for categorical data analysis, thus for dealing with integer valued variables. Continuous data need to undergo a discretization process before the alternating least squares (ALS) algorithm starts. Various discretization options are available for recoding continuous data within the procedure and one can always recode data beforehand. Our proposal is to adjust the algorithm to allow continuous values directly so that researchers dealing with continuous variables avoid thinking that some information is being neglect.

The obvious advantages of incorporating continuous variables directly are:

- the user does not need to care about any discretization process;
- the relative distances within each variables' values are respected from the start without discretization losses of information.

Although defined and implemented for any given spline's order and number of interior knots, quasi-linear PCA (qlPCA) refers to linear splines. This approach allows modeling several degrees of nonlinear relationships between variables by increasing the number of knots while maintaining the transformations stepwise linear. Using this kind of splines, relative distances are not going to be lost during the optimization process being proportional after convergence within any two consecutive interior knots. The main advantages of this approach are related to explicitly defining the nonlinear optimal transformation and to identify *piecewise loadings* (section V).

A. Optimal scaling revisited

Let $\mathbf{H}$ be $n \times m$ standardized data matrix, p the number of retained principal components and $\mathbf{f}_j = f_j(\mathbf{h}_j)$ be the image vector of $\mathbf{h}_j$ under the function f_j, $j = 1, \ldots, m$. The loss funtion (3) can be re-written in a more general format into the loss function $\sigma : \mathrm{M}_{n \times p} \times \mathrm{M}_{pm \times n} \rightarrow \mathbb{R}$ so that

$$\sigma\left(\mathbf{X}, \mathbf{F}\right) = n^{-1} \sum_j tr \left[\left(\mathbf{X} - \mathbf{F}_j\left(\mathbf{h}_j\right)\right)^{'} \left(\mathbf{X} - \mathbf{F}_j\left(\mathbf{h}_j\right)\right) \right], \tag{4}$$

with normalization restriction $\mathbf{X}^{'}\mathbf{X} = n\mathbf{I}$ where:

- $\mathbf{F}_j = \left[\begin{array}{ccc} \mathbf{f}_{j1} & \ldots & \mathbf{f}_{jp} \end{array}\right]$ is the $n \times p$ matrix collecting the p (different) images of the (same) vector $\mathbf{h}_j$;
- f_{jt} is the transformed variable j associated with dimension t, $t = 1, \ldots, p$;
- $\mathbf{F} = \left[\begin{array}{ccc} \mathbf{F}_1 & \ldots & \mathbf{F}_m \end{array}\right]^{'}$ is an $nm \times p$ matrix.

Notice that the matrix $\mathbf{F}_j$ contains the images of the same vector $\mathbf{h}_j$, subject to p different transformations. If no restrictions are imposed upon the matrix $\mathbf{F}_j$, then each value of the j^{th} variable receives p different quantifications, one for each retained dimension, in what is commonly called *Multiple quantification* [6], [12], [13].

Usually, for continuous variables, order and distance restrictions are required, which can be imposed by the splines' parameters (degree, number of interior knots and its placement). A familiar way to implement those restrictions starts by imposing rank one restrictions on the matrix $\mathbf{F}_j$, on what is usually called *Single quantification* [6], [12], [13]. Details on *Single quantification* implementation in the proposed algorithm are given in the next section.

Having fixed the class of admissible transformations for each variable, the purpose is to find the object scores and the transformations that minimize the loss function (4). The main differences between the existing algorithms to solve the loss minimization problem are within the class of admissible transformations. In what splines are concerned, each class of transformations depends on the number of knots, spline's degree and knots placement. However, while the fitting problem is linear in the basis coefficients, it is highly nonlinear in the knots, and therefore it is desirable to avoid much optimization with respect to them [10]. The choice of a particular spline could be targeted according to the percentage of explained variance (section V.A), by trying different sets of parameters. However, like in all statistical models, nonlinear PCA via splines is susceptible to overfitting when there are too many parameters in the model. In order to prevent overfitting a reasonable amount of data should be in the vicinity of any interior knot [10].

Let $\mathbf{f}_j$ be the image vector of $\mathbf{h}_j$ under a spline of degree v with r interior knots, spanned by $w = v + r$ *I-splines*

$$\mathbf{f}_j = \sum_{i=1}^{w} \alpha_{ji} I_{ji}^{[v]}\left(\mathbf{h}_j\right) = \sum_{i=1}^{w} \alpha_{ji} I_{ji}^{[v]} = \mathbf{G}_j^{\triangle}\mathbf{y}_j, \tag{5}$$

where: $\mathbf{G}_j^{\triangle}$ is the pseudo-indicator matrix for variable j, of order $n \times w$ whose columns are the image vectors of the variable j by each of the *I-splines* basis functions; $\mathbf{y}_j$ is a vector of length w whose elements are the linear combination coefficients $\mathbf{y}_j = \left[\alpha_1 \alpha_2 \ldots \alpha_w\right]^{'}$.

In the previous section the optimal scaling or optimal quantification process was defined within categorical analysis as the transformation of variables by assigning quantitative values to qualitative variables in order to optimize equation (3). Using equation (5) it is possible to re-define the optimal quantification process as a ALS phase that, given the object scores, optimize the vectors $\mathbf{y}_j$ in order to minimize the loss function, or analogously, as the ALS phase that seeks to find the optimal linear combination for each basis spline, given the object scores from the model estimation phase, therefore obtaining the optimal spline transformation to each variable.

Notice that if the chosen class of admissible transformations are splines of degree one without interior knots the ALS optimization of equation (4) yields the traditional (linear) PCA solution.

B. QlPCA Algorithm Outline

Let $\mathbf{H}$ be $n \times m$ standardized data matrix, p the number of retained principal components, r the number of interior knots and v the spline's degree.

The qlPCA algorithm uses ALS to minimize (4) using rank one matrices $\mathbf{F}_j$. It consists of two phases iteratively alternated until convergence is reached, an optimal quantification phase and a estimation of object scores phase.

Let $\mathbf{H}$, p, r and v be the algorithm's inputs. The initial configuration is set as follows.

I. Initialization

```
1: $\mathbf{Z}_{n \times p} \leftarrow linearPCA(\mathbf{H})$
2: $[K, \Lambda^{1/2}, W] \leftarrow svd(\mathbf{Z})$
3: $\mathbf{X} \leftarrow \sqrt{n}\mathbf{K}\mathbf{W}^{'}$
4: $\mathbf{a}_j \leftarrow \frac{1}{n}\mathbf{X}^{'}\mathbf{h}_j, j = 1, \ldots, m$
5: $[\mathbf{G}_1^{\triangle} \ldots \mathbf{G}_m^{\triangle}] \leftarrow createIspline(H, v+1, r)$
```

Initialization starts by performing a linear PCA on $\mathbf{H}$ retaining p principal components on the object scores matrix $\mathbf{Z}$. This matrix is then orthonormalized (steps 2 and 3, being svd a singular value decomposition) such that $\mathbf{X}^{'}\mathbf{X} = n\mathbf{I}$. On step 4, $\mathbf{a}_j = \left[a_{1j} \ldots a_{sj} \ldots a_{pj}\right]^{'}$ for each j, being a_{sj} the Pearson correlation between variable j and the s^{th} principal component, also known as component loading. Step 5 computes the $n \times mw$ matrix $\mathbf{G}^{\triangle} = [\mathbf{G}_1^{\triangle} \ldots \mathbf{G}_m^{\triangle}]$ by the iterative procedure described in section II.

II. Alternating Least Squares: optimal quantification phase (loop across variables j):

```
1: $\tilde{\mathbf{y}}_j \leftarrow \mathbf{X}\mathbf{a}_j$
2: $\mathbf{y}_j \leftarrow \frac{1}{n}((\mathbf{G}_j^{\triangle})^{'}\mathbf{G}_j^{\triangle})^{-1}(\mathbf{G}_j^{\triangle})^{'}\tilde{\mathbf{y}}_j$
3: $\mathbf{f}_j = f_j\left(\mathbf{h}_j\right) \leftarrow \mathbf{G}_j^{\triangle}\mathbf{y}_j$
4: $\mathbf{f}_j \leftarrow \mathbf{f}_j - \overline{f}_j\mathbf{u}$
```

```
5: $\mathbf{f}_j \leftarrow \sqrt{n}\dfrac{\mathbf{f}_j}{\|\mathbf{f}_j\|}$
6: $\mathbf{a}_j \leftarrow \dfrac{1}{n}\mathbf{X}'\mathbf{f}_j$
```

Optimal quantification phase updates each variables' quantification $\mathbf{f}_j$ (steps 1 to 5) and the vectors $\mathbf{a}_j$ (step 6). Least squares estimation is being used in order to minimize the sum of the squares of the residuals between $\mathbf{X}\mathbf{a}_j$ and the transformed variable j (step 2). Notice that step 2 performs a multivariate linear regression with $\mathbf{X}\mathbf{a}_j$ as the dependent variable and the columns of $\mathbf{G}_j^{\triangle}$ as predictors. This updates $\mathbf{y}_j$, a vector whose elements are the linear combination coefficients for the *I-splines*' basis, which in turn will update each variables' quantification $\mathbf{f}_j$. Each transformed variable is mean centered (step 4) and normalized to $\sqrt{n}$ (step 5) so that its variance will be one. On step 6, $\mathbf{a}_j$ is updated, being now a_{sj}, the Pearson correlation between the transformed variable j and the s^{th} nonlinear principal component, from now on called nonlinear component loading.

III. Alternating Least Squares: estimation of object scores phase

```
1: $\mathbf{Z} \leftarrow \sum_j \mathbf{F}_j$ where $\mathbf{F}_j = \mathbf{f}_j\mathbf{a}_j'$
2: $[\mathbf{K}, \mathbf{\Lambda}^{1/2}, \mathbf{W}] \leftarrow svd(\mathbf{Z})$
3: $\mathbf{X} \leftarrow \sqrt{n}\mathbf{K}\mathbf{W}'$
```

The estimation of object scores phase updates $\mathbf{X}$. Rank one restrictions (*Single quantification*) are imposed on the $n \times p$ matrix $\mathbf{F}_j$ by defining $\mathbf{F}_j = \mathbf{f}_j\mathbf{a}_j' = [\mathbf{f}_j a_{1j} \ldots \mathbf{f}_j a_{pj}]$, therefore each column of $\mathbf{F}_j$ being collinear to $\mathbf{f}_j$. In its general format the loss function (4) used the $nm \times p$ matrix $\mathbf{F}$ to store multiple quantifications for each variable. This algorithm, using *single quantification* on every variable allows a re-definition of $\mathbf{F}$ to a $n \times m$ matrix where $\mathbf{F} = [\mathbf{f}_1 \ldots \mathbf{f}_m]$. Thus, by the previous definition, $\mathbf{F}$ is the transformed data matrix. Step 1 assigns to $\mathbf{Z}$ the sum of $\mathbf{F}_j$ over j (thus $\mathbf{Z}$ is a centered matrix). Matrix $\mathbf{Z}$ is then orthonormalized (steps 2 and 3) such that $\mathbf{X}'\mathbf{X} = n\mathbf{I}$; thus object scores will have unitary variance.

IV. Convergence test

The convergence test is performed by testing the difference between two successive values of the loss function (4) against 0.1×10^{-5}. The algorithm goes back to II if convergence is not reached or stops if the maximum number of iterations is reached.

After convergence, qlPCA will produce the following outputs:

- $\mathbf{X}$ - $n \times p$ matrix containing the nonlinear objects scores;
- $\mathbf{F}$ - $n \times m$ matrix containing the optimally transformed variables;
- $\mathbf{A} = [\mathbf{a}_1 \ldots \mathbf{a}_m]$ - $p \times m$ matrix with nonlinear loadings;
- $\mathbf{y}$ - mw-vector with optimal coefficients associated with m *I-splines* basis.

C. Some notes on the computer program

QlPCA algorithm has been implemented through (freely downloadable) MATLAB m-files. MATLAB version 2009b was used and a MATLAB GUI (Graphical User Interface) and can be obtained from the correspondent author.

The computer program is in a preliminary form since at the moment it only allows the same type of splines for all variables.

For the computation of the *I-splines* basis we have designed our own program based on the theoretical foundations in [7], [8] and [10]. At the moment nothing in our algorithm ensures that a nondecreasing spline is obtained. However the basis coefficients are computed so that they minimize the loss function. Therefore, if a nondecreasing spline is the optimal transformation for the existing structure, it will emerge naturally. This will be exemplified in section VI.

Applying qlPCA involves trying out different options regarding spline's order and the number of interior knots. Linear splines have several advantages (see next section) but the user can always try higher order splines. The actual implementation does not allow users to choose specific locations for knots: they are always placed at equally spaced percentiles and thus each interval contains approximately the same number of data points. It was already mentioned that in order to prevent overfitting a reasonable amount of data should be in the vicinity of any interior knot [10]. The cause of overfitting is related to step II.2 of qlPCA's algorithm, where a multivariate regression with $w = v + r$ (spline's degree plus the number of interior knots) predictor variables is performed. With our knots placement option, preventing overfitting is a matter of the number of available individuals, n, versus w. There is no consensus in the literature regarding the minimum number of available observations per predictor on linear multivariate regression, but a value of 10 to 15 observations per predictor is commonly referred. Whatever the rule of thumb chosen, its adaption to the qlPCA framework implies multiplying that number by w.

V. Properties of quasi linear PCA

The qlPCA algorithm will take advantage of low order splines, without limitation concerning the number of interior knots, in order to achieve nonlinear PCA as a straightforward generalization of the traditional PCA including its measures of performance and interpretation. In this section, results on model summary, projecting new observations onto the nonlinear principal components space and reconstruction of the original variables are introduced. A new concept is also defined - *piecewise loadings* - as (piecewise) correlations between nonlinear principal components and the original variables.

Given a training $n \times m$ data matrix $\mathbf{H}$ the representation in the nonlinear principal components space can be found by minimizing the loss function (4) using the algorithm described in the previous section. Given the qlPCA parameters (p the number of retained principal components, r the number of interior knots and v the spline's degree and by consequence $w = v + r$ the linear space of spline functions' dimension), once the optimization is completed the following matrices and vectors, as defined in the previous section, can be considered known: $\mathbf{X}$,$\mathbf{F}$,$\mathbf{A}$ and $\mathbf{y}$.

A. Model summary

As in linear PCA, one is usually interested about the model's ability to account for the total variation in the original data matrix. However, PCA nonlinear varieties only report the nonlinear model's ability to account for the total variation in the optimally transformed data matrix.

Let's define *Variance Accounted For* per dimension ($s = 1, \ldots, p$) as

$$VAF_s = \sum_{j=1}^{m} a_{sj}^2, (\% \; of \; variance \; is \; VAF_s \times 100/m)$$

where a_{sj} is the Pearson correlation between variable j and the s^{th} principal component, also known as component loading.

Let $\mathbf{R}$ be the correlation matrix of $\mathbf{F}$. Each transformed variable is mean centered and normalized to $\sqrt{n}$, thus $\mathbf{R} = n^{-1}\mathbf{F}^{'}\mathbf{F}$. It can be shown that the first p eigenvalues of $\mathbf{R}$ equal $VAF_s, \; s = 1, \ldots, p$ and that the s^{th} diagonal element of $\mathbf{\Lambda}^{1/2}$ (step III.2) equals $\sqrt{n}VAF_s$.

Therefore VAF's definition introduces the concept of eigenvalues within qlPCA and is a straightforward generalization from linear PCA on the nonlinear model's ability to account for the total variation in the optimally transformed data matrix.

B. *Choosing the appropriate number of components*

As in linear PCA, the user must decide the adequate number of components to be retained in the solution. One of the most well-known criteria for this decision is to retain all principal components with eigenvalues greater than 1.0. However, there is broad consensus in the literature that this is among the least accurate methods [12], [14]. Alternative criterions include the scree plot, parallel analysis, Velicers partial correlation technique, cross-validation among others [14]. Being available in the most frequently used statistical software as well as more accurate than the "eigenvalues greater than 1.0" criterion, the scree plot criterion may well be the user's best available choice.

The scree test involves examining the plot of components' identification (on the x-axis) versus eigenvalues (on the y-axis) and looking for the break point, or "elbow", where the curve flattens out. The number of points above the "break" (i.e. not including the point at which the break occurs) is usually the appropriate number of components to retain, although it can be unclear if there are data points clustered together near the bend.

Since qlPCA algorithm minimizes the loss function for a given number p of retained principal components its solutions are usually not nested for different values of p. The first p eigenvalues obtained from the correlation matrix among the optimal transformed variables by qlPCA with p as input are usually not equal to the first p eigenvalues obtained from the correlation matrix among the optimally transformed variables by qlPCA with $p+1$ as input. Therefore, scree plots differ for different dimensionalities and the scree plots of the p, the $p-1$ and $p+1$ dimensional solutions should be compared [12]. The scree plots' "break" associated with qlPCA with p as input should be consistent with the scree plots' "break" associated with qlPCA with $p+1$ as input, so that the appropriate number of components to retain is p [12].

C. *Projecting new observations onto the nonlinear pc space*

Let $\mathbf{h}_{new}^{'}$ be a m-vector with new observations on the m variables from the same process as the original data matrix. The problem of projecting the new observations onto the nonlinear principal components space is to find the p-vector $\mathbf{x}_{new}^{'}$, the corresponding nonlinear object scores. If the original data matrix is not standardized, mean and variances vectors should be recorded and qlPCA algorithm applied on the standardized data matrix. A correction on $\mathbf{h}_{new}^{'}$ is applied using those vectors.

Notice that step III.3 can be re-written as $\mathbf{X} \leftarrow \sqrt{n}\mathbf{Z}\mathbf{W}\mathbf{\Lambda}^{-1/2}\mathbf{W}^{'}$ since by the singular value decomposition $\mathbf{Z} = \mathbf{K}\mathbf{\Lambda}^{1/2}\mathbf{W}^{'}$. Let $\mathbf{A}$ be the nonlinear loadings matrix, notice that (step III.1),

$$\mathbf{Z} = \mathbf{F}\mathbf{A}^{'}, \tag{6}$$

thus

$$\mathbf{X} = \sqrt{n}\mathbf{F}\mathbf{A}^{'}\mathbf{W}\mathbf{\Lambda}^{-1/2}\mathbf{W}^{'}, \tag{7}$$

where $\mathbf{W}$ and $\mathbf{\Lambda}^{-1/2}$ derive from the singular value decomposition on step III.2.

New observations' representation on the nonlinear principal components space is achieved in two steps. Firstly transformed values of $\mathbf{h}_{new}^{'}$ are computed (loop across variables $j, \; j = 1, \ldots, m$):

```
if $min(\mathbf{h}_j) \leq h_{new,j} \leq max(\mathbf{h}_j)$ then
    $f_{new,j} \leftarrow interp(\mathbf{h}_j, \mathbf{f}_j, h_{new,j})$,
else
    $f_{new,j} \leftarrow extrap(\mathbf{h}_j, \mathbf{f}_j, h_{new,j})$,
end if
```

For within range values linear interpolation is performed to find $f_{new,j}$, the value of the underlying optimal spline function f_j at the value $h_{new,j}$, the new observation on variable j. Notice that after convergence the optimal linear spline with r interior knots is completely defined with $r+2$ points: r associated with the interior knots at two associated with the minimum and maximum of each variable. Therefore interpolation only needs those $r+2$ points to achieve the exact spline's value. For out of range values linear extrapolation is performed.

Secondly, nonlinear object scores for the new observations are obtained using equation (7) with the $1 \times m$ optimal transformed matrix $\mathbf{F}_{new} = [f_{new,1} \cdots f_{new,m}]$.

D. *Reconstruction*

The problem of reconstruction the original data is concerned with finding an estimate $\hat{\mathbf{H}}$ of $\mathbf{H}$ using its nonlinear object scores representation. This process includes two steps: firstly from nonlinear object scores $\mathbf{X}$ to an approximation $\hat{\mathbf{F}}$ of the transformed data; secondly from $\hat{\mathbf{F}}$ to an estimate $\hat{\mathbf{H}}$ of the original data.

When $p < m$ nonlinear principal components are used $\mathbf{X}\mathbf{a}_j$ is an approximation of the optimal transformed variable $\mathbf{f}_j$ (qlPCA's algorithm step II.2), thus $\mathbf{X}\mathbf{A}$ is an estimate $\hat{\mathbf{F}}$ of $\mathbf{F}$. Therefore

$$\mathbf{F} = \mathbf{X}\mathbf{A} + (\mathbf{F} - \hat{\mathbf{F}}) \tag{8}$$

where the first term on the right-hand side of the equation represents the contribution due to the retained nonlinear principal components and the second term represents the amount that is not explained by the qlPCA model - the residual.

Having used qlPCA with linear splines each column of $\mathbf{XA}$ is thus an approximation of the optimal linear spline's images. Inverse linear interpolation gives the exact inverse function of a stepwise linear function. Therefore, using inverse linear interpolation, with values of $\mathbf{Xa}_j$ within the range of the j^{th} spline and inverse linear extrapolation for values outside, $\hat{\mathbf{H}}$ is achieved.

E. Piecewise Loadings

Let $\mathbf{x} \leftarrow \mathbf{x}_s$ be the s^{th} principal component, $\mathbf{f} \leftarrow \mathbf{f}_j$ the j^{th} transformed variable and $\mathbf{h} \leftarrow \mathbf{h}_j$ the j^{th} (original) variable. Within qlPCA, as within any other nonlinear PCA approach, the term loading refers to the Pearson correlation between $\mathbf{x}$ and $\mathbf{f}$. In this section, $\mathbf{x}_i$, $\mathbf{f}_i$ and $\mathbf{h}_i$ will denote the truncated vectors obtained from $\mathbf{x}$, $\mathbf{f}$ and $\mathbf{h}$, respectively, with elements from the i^{th} piece of those vectors, $i = 1 \dots r+1$, defined by two consecutive knots (minimum, interior knots and maximum). The idea of *piecewise loadings* is to achieve (piecewise) correlations between truncated vectors $\mathbf{x}_i$ and $\mathbf{h}_i$ by means of $\mathbf{f}_i$. Therefore, *piecewise loadings* can reveal stepwise behavior between nonlinear principal components and the original variables.

The authors have shown in [5] an analytic expression of *piecewise loadings* for the particular case of linear splines with one interior knots. However, analytic expressions on *piecewise loadings* with more interior knots demands the use of the related analytic expression of those optimal spline transformation. Those expressions are not available as a direct algorithm output, thus results should refer not to the analytic expressions but to the truncated vectors $\mathbf{x}_i$, $\mathbf{f}_i$ and $\mathbf{h}_i$.

Consider the following auxiliar result: let $\mathbf{y}$ and $\mathbf{z}$ be two vectors (not necessarily mean centered neither standardized) of the same dimension and g a linear function, being $\mathbf{g} = g(\mathbf{z}) = m\mathbf{z} + b\mathbf{u}$ the image vector of $\mathbf{z}$ throughout g, where $\mathbf{u}$ is a vector with ones. It can be shown that:

$$corr(\mathbf{y}, g(\mathbf{z})) = \begin{cases} corr(\mathbf{y}, \mathbf{z}), & m > 0 \\ -corr(\mathbf{y}, \mathbf{z}), & m < 0 \end{cases} \tag{9}$$

After qlPCA convergence $\mathbf{x}$ and $\mathbf{f}$ are available, thus $corr(\mathbf{x}_i, \mathbf{f}_i)$ can be computed for every piece i. Since linear splines are being used, within each piece $\mathbf{f}_i$ and $\mathbf{h}_i$ are related by $\mathbf{f}_i = m_i\mathbf{h}_i + b_i\mathbf{u}$ for a known m_i and b_i. Therefore by the previous result:

$$corr(\mathbf{x}_i, \mathbf{h}_i) = \begin{cases} corr(\mathbf{x}_i, \mathbf{f}_i), & m_i > 0 \\ -corr(\mathbf{x}_i, \mathbf{f}_i), & m_i < 0 \end{cases} \tag{10}$$

VI. Simulation study

We now discuss an example in which known nonlinear functions are simulated. A version of this problem, called the cylinder problem, has been used by [10], [6] and several other authors to test their nonlinear approaches to principal components analysis.

Consider twelve variables, where ten of them are defined as nonlinear functions of the remaining two.

All variables with the exception of 8 and 9 are log-linear while 8 and 9 can be put into linear form by the transformations $\log \circ \tan$ and $\log \circ \cot$, respectively. Thus, there exists a set of twelve monotone transformations which will cause the transformed data matrix to have a two-dimensional structure with scores for each observation being $\log$ altitude and $\log$ base area.

TABLE I
CYLINDER PROBLEM.

Variable	Formula
1. Altitude	a
2. Base Area	b
3. Base Perimeter	$2\sqrt{b\pi}$
4. Side Area	$2a\sqrt{b\pi}$
5. Volume	ab
6. Moment of Inertia	$ab^2/2\pi$
7. Slenderness Ratio	$a/\sqrt{2b\pi}$
8. Diagonal-Base Angle	$\arctan[a\sqrt{\pi}/(2\sqrt{b})]$
9. Diagonal-Side Angle	$\operatorname{arccot}[a\sqrt{\pi}/(2\sqrt{b})]$
10. Electrical Resistance	a/b
11. Conductance	b/a
12. Torsional Deformability	$2a\pi/b^2$

As a consequence, it is expected that a nonlinear PCA over the original data matrix (as well as an ordinary PCA over the transformed data matrix) reveals an almost perfect fit with only two dimensions and that the optimal spline transformations behave approximately like logarithmic functions.

A. Design

There were made three simulations, having as a starting point 51 different cylinders defined by using uniformly distributed pseudorandom numbers for the altitudes and base areas. For each of these 51 cylinders values of the remain ten variables listed in table I were computed.

We wanted to illustrate our procedure with data having some level of noise and consequently we added normal, zero-mean, pseudorandom disturbances. Each variable was first transformed by the appropriate linearizing transformation, then a disturbance was added to this value, and finally the disturbed value inversely transformed. The dispersion of the noise term varies from variable to variable, as it is a proportion of the standard deviation of each variable. Two sets of disturbed data were generated, one with a 10% noise level, and the other with a 25% noise level.

Having in mind the suggestion on prevent overfitting (section IV.C) and since there are 51 available individuals results of qlPCA are based on linear splines with two interior knots (chosen to be approximately at the 33rd and 67th percentiles) and three interior knots (chosen to be approximately at the quartiles) for each variable.

B. Results

In the following table are given the results obtained after applying the proposed algorithm (qlPCA) using linear splines with two interior knots (qlPCA2), linear splines with three interior knots (qlPCA2) and linear Principal Components Analysis (PCA) on the three data sets (no noise, 10% noise and 25% noise). The fit is expressed in terms of percentage of variance accounted for two dimensions.

In all cases the proposed algorithm (qlPCA) performed much better than linear PCA. The theoretical structure imposed on data makes reasonable to assume that a two-dimensional structure should be considered. In order to test

TABLE II
VARIANCE ACCOUNTED FOR TWO DIMENSIONS (%).

Algorithm	No noise	10% noise	25% noise
qlPCA2	97.77	96.90	92.59
qlPCA3	98.43	97.68	94.41
PCA	82.16	80.18	72.62

it on the worst situation, data with 25% noise level, scree plots analysis have been conducted to check this assumption on both PCA and qlPCA. Regarding linear PCA's scree plot, presented in Figure 2, eigenvalues of the correlation matrix of the original variables were used. This plot does not show a clear "elbow" like the ones on Figure 3 for qlPCA, as the slope changes abruptly not once but twice on the third and on the sixth component. Therefore based on the referred plot, either two or five components solutions are defendable.

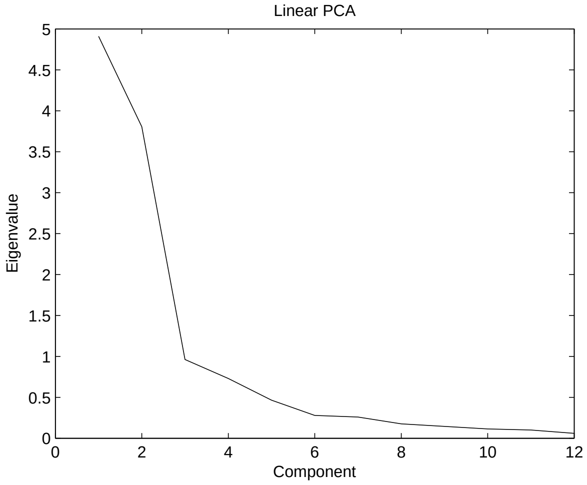

Fig. 2. Scree plot from linear PCA on cilinder data with 25% noise level. On the y-axis are the eigenvalues of the correlation matrix of the original variables.

Regarding qlPCA's scree plot, it used eigenvalues of the correlation matrix of the transformed variables from a two-dimensional solution to check this assumption that two components were the appropriate number of components to retain. From this plot, presented in Figure 3 (upper plot), we concluded that the elbow is located at the third component. Remember that qlPCA solutions are not nested (section V.B), so a scree plot for a three-dimensional solution can be different from a scree plot for a two-dimensional solution, with the position of the elbow moving from the third to the second component. In the present analysis, different dimensionalities consistently place the elbow at the third component, as shown in Figure 3. Therefore, the information from both scree plots suggests that two is the appropriate number of components to retain as expected.

Figure 4 displays the optimal spline transformation of variable 6 from a two dimensional qlPCA on cilinder data without noise. This plot is the typical one among the twelve transformations plots. This type of logarithmic shaped plots was, as expected, the most commonly obtained among the twelve transformation plots. However, an exception - transformation of variable twelve - was observed, as it can be seen in Figure 5.

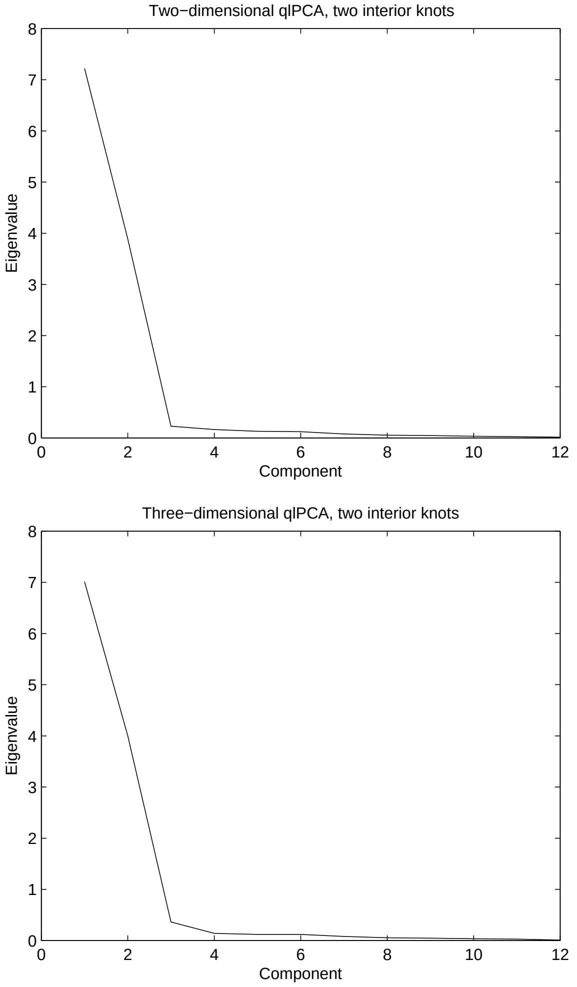

Fig. 3. Scree plot from a two and three dimensional qlPCA on cilinder data with 25% noise level. On the y-axis are the eigenvalues of the correlation matrix of the transformed variables.

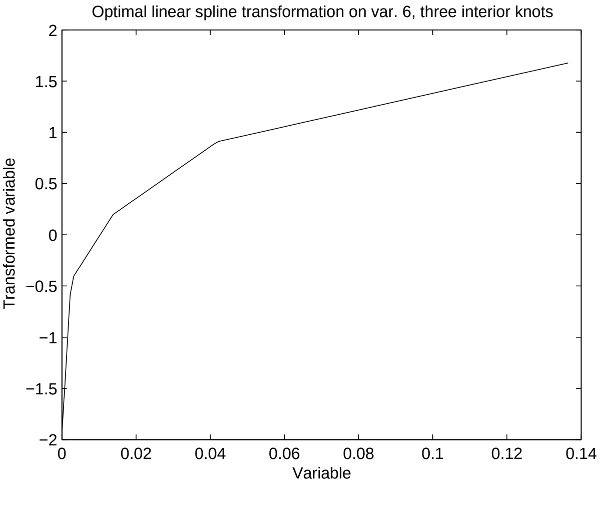

Fig. 4. Optimal spline transformations with three interior knots of variable 6 from a two dimensional qlPCA on cilinder data without noise.

Further inspection of this variable shows that the interior knots are placed at 3.85, 10.32 and 25.38 (the quartiles) and that its maximum is 6498.7. This is an example where an alternative knot placement could have improved the variable's transformation.

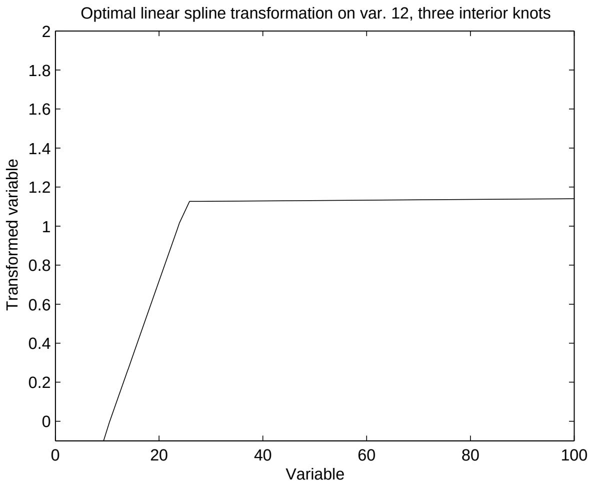

Fig. 5. Optimal spline transformations with three interior knots of variable 12 from a two dimensional qlPCA on cilinder data without noise.

VII. Conclusion

A new approach on Nonlinear Principal Components Analysis, so-called quasi-linear PCA (qlPCA) was presented in detail in this paper. The qualification *quasi* emphasize that it refers to the use of linear splines. QlPCA algorithm uses alternating least squares to minimize a suitable objective loss function. It consists of two phases iteratively alternated until convergence is reached, an optimal quantification phase and a estimation of object scores phase.

Nonlinear PCA techniques usually report its solution with relational measures between the nonlinear transformed variables obtained after convergence and the associated objects scores. Considering low order splines some relations between nonlinear principal components and the original variables can be revealed. Optimal transformations are explicitly known after convergence and it was shown that optimization over linear splines have several advantages. It allows projecting new observations onto the nonlinear principal components' space and reconstruction the original observations. A new concept was also defined - *piecewise loadings* - as (piecewise) correlations between nonlinear principal components and the original variables.

QlPCA recovers a spline based algorithm designed for categorical variables (CATPCA) and introduces continuous variables into the framework without the need for a discretization process. It should be emphasized that the proposed algorithm is not intended to be a competitor of CATPCA but rather a different approach.

Nonlinear PCA's most known approaches among researchers dealing with continuous variables are autoassociative neural networks, principal curves and manifolds, kernel approaches or the combination of these [3]. Therefore, comparisons studies with qlPCA will be taken.

Applying qlPCA involves trying out different options regarding spline's order and the number of interior knots. Researchers are invited to try it out using a MATLAB GUI (Graphical User Interface) that can be obtained from the correspondent author.

References

[1] W. Krazanowski and F. Marriott, *Multivariate Analysis Part I - Distributions, Ordination and Inference*, Edward Arnold, 1994.

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

[3] U. Kruger, J. Zhang and Lei Xie, "Developments and Applications of Nonlinear Principal Component Analysis - a Review," in: *Principal Manifolds for Data Visualization and Dimension Reduction 2007*, pp. 1-43.

[4] J. Meulman, A. Kooij and W. Heiser, "Principal Components Analysis with Nonlinear Optimal Scaling Transformations for Ordinal and Nominal Data," in *The Sage Handbook of Quantitative Methodology for the Social Sciences 2004*, pp. 49-70.

[5] N. Lavado and T. Calapez, "Quasi-Linear PCA: Low Order Spline's Approach to NonLinear Principal Components", in *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering 2011, WCE 2011, 6-8 July, 2011, London, U.K.*, pp. 360-364.

[6] A. Gifi, *Nonlinear Multivariate Analysis*, Wiley, 1991.

[7] C. de Boor, *A Practical Guide to Splines*, Springer, 1978.

[8] L. Schumaker, *Spline Functions: Basic Theory*, Wiley, 1981.

[9] J. Ramsay, "Monotone Regression Splines in Action," *Statistical Science*, vol. 3, pp. 425-461, 1988.

[10] S. Winsberg and J. Ramsay, "Monotone spline transformations for dimension reduction," *Psychometrika*, vol. 48, pp. 575-595, 1983.

[11] G. Michailidis and J. De Leeuw, "The GIFI system of descriptive multivariate analysis," *Statist. Sci.*, vol. 13, pp. 307-336, 1998.

[12] M. Linting, J. Meulman, P. Groenen and A. Van der Kooij, "Nonlinear principal components analysis: Introduction and application" *Psychological Methods*, vol.12, pp. 336-358, 2007.

[13] J. De Leeuw and J. Van Rijckevorsel, "Beyond Homogeneity Analysis," in: *Component and Correspondence Analysis: Dimension Reduction by Functional Approximation 1988*, pp. 55-80.

[14] J. Jackson, *A User's Guide to Principal Components*, Wiley, 1991.