Orthogonal rotation in PCAMIX*

Marie Chavent[1,2]†, Vanessa Kuentz-Simonet[3] and Jérôme Saracco[2,4]

[1] Université de Bordeaux, IMB, CNRS, UMR 5251, France
[2] INRIA Bordeaux Sud-Ouest, CQFD team, France
[3] Irstea, UR ADBX, France
[4] Institut Polytechnique de Bordeaux, France

Abstract

Kiers (1991) considered the orthogonal rotation in PCAMIX, a principal component method for a mixture of qualitative and quantitative variables. PCAMIX includes the ordinary Principal Component Analysis (PCA) and Multiple Correspondence Analysis (MCA) as special cases. In this paper, we give a new presentation of PCAMIX where the principal components and the squared loadings are obtained from a Singular Value Decomposition. The loadings of the quantitative variables and the principal coordinates of the categories of the qualitative variables are also obtained directly. In this context, we propose a computationaly efficient procedure for varimax rotation in PCAMIX and a direct solution for the optimal angle of rotation. A simulation study shows the good computational behavior of the proposed algorithm. An application on a real data set illustrates the interest of using rotation in MCA. All source codes are available in the R package "PCAmixdata".
Keywords: mixture of qualitative and quantitative data, principal component analysis, multiple correspondence analysis, rotation.

1 Introduction

Kaiser (1958) introduced the varimax criterion for the attainment of simple structure by orthogonal rotation in Principal Component Analysis (PCA). This criterion aims at maximizing the sum over the columns of the squared elements of the loading matrix. This matrix plays a significant part in the interpretation of the results since it contains the correlations between the variables and the principal components. The idea is to rotate the loading matrix and the standardized principal components so that groups of variables appear: having high loadings on the same component, moderate ones on a few components and negligible ones on the remaining components. Thus the variance gets redistributed along the newly rotated axes.

*Adv Data Anal Classif (2012), 6:131146, DOI 10.1007/s11634-012-0105-3
†marie.chavent@u-bordeaux2.fr

Despite the close relationship between PCA and Multiple Correspondence Analysis (MCA), rotation in MCA has not received much attention, while rotation in Correspondence Analysis (CA) has been recently studied and implemented in Matlab package, see for instance van de Velden and Kiers (2003, 2005), Greenacre (2006) or Urbano et al. (2009a, 2009b). However, Kiers (1991) did handle orthogonal rotation in MCA in the general context of rotation in PCAMIX, a principal component method for the mixture of qualitative and quantitative variables. More recently, Adachi (2004) considered oblique rotation in MCA. Oblique or orthogonal rotations involve the same problem of maximizing a simplicity criterion, only the imposed constraints differ. Since fewer constraints are imposed in oblique rotation, it is generally possible to obtain a solution more easily than in orthogonal rotation (Browne, 2001). Nevertheless, orthogonal rotations are commonly used in practice, as the orthogonally rotated loadings can be directly interpreted as correlations between the variables and the rotated components, and graphical representations remain possible.

MCA and PCA are particular cases of PCAMIX and Kiers (1991) extended the varimax criterion for the attainment of simple structure in this general context. For qualitative variables, the coefficient used to express the link between a variable and a component is the correlation ratio. The varimax criterion is then expressed with squared loadings defined as correlation ratios for qualitative variables and squared correlations for quantitative variables. Algorithms designed for the determination of an optimal orthogonal rotation in the context of PCA, as proposed for example by Kaiser (1958), Neudecker (1981) or Jennrich (2001) do not apply to this extended varimax criterion. Kiers (1991) proposed a matrix reformulation of the new varimax criterion in order to replace the rotation problem with a problem of simultaneous diagonalization of a set of symmetric matrices (ten Berge, 1984). To solve the latter one, he suggested to use the algorithm of de Leeuw and Pruzansky (1978). To the best of our knowledge, the resulting algorithm has never been presented in a single paper, so we have recalled for comparison purpose the main steps of the matrix reformulation and the simultaneous diagonalization. We shall refer to this algorithm as the rotation procedure based on Kiers' matrix reformulation.

In this paper we first present a new formulation of PCAMIX similar to that of Escofier (1979) and Pagès (2004) in the way quantitative and qualitative variables are transformed. However PCAMIX is reformulated as a Singular Value Decompostition (SVD) of the transformed data. This provides a direct way to determine both the component scores and the squared loadings and also the principal coordinates of the categories of the qualitative variables as well as the loadings of the qualitative variables. Next to find an optimal rotation for the PCAMIX varimax criterion, we use the iterative procedure suggested by Kaiser (1958) for PCA: we rotate pairs of dimensions according to an optimal angle θ, iteratively until the process converges. A new direct determination of this optimal angle is proposed. We shall refer to the resulting algorithm as the rotation procedure based on the SVD approach of PCAMIX. This algorithm leads to the same final rotation as Kiers' (1991) original approach, however a simulation study shows that it is computationally more efficient. When all the variables are quantitative, this new algorithm reduces to the Kaiser's (1958) procedure for orthogonal rotation in PCA with a new direct expression of the optimal planar angle θ.

This paper is organized as follows. Section 2 recalls Kiers' original PCAMIX method and proposes an alternative formulation using SVD. Section 3 deals with varimax rotation in PCAMIX. The optimization problem is given in Section 3.1. The determination of the optimal angle of rotation with Kiers' matrix reformulation approach is described in Section 3.2.1 for comparison purpose with the direct solution proposed in Section 3.2.2. A complete procedure for orthogonal rotation in more than two dimensions is given in Section 3.3. In section 4.1, a simulation study compares the computational time of the proposed rotation procedure with the rotation procedure based on Kiers (1991). In Section 4.2 a real data application illustrates the interest of rotation in MCA and shows some of the outputs and graphical representations available in the R package "PCAmixdata" we have developed. Concluding remarks are given in Section 5.

2 The PCAMIX method

Let us first introduce some notations used in the presentation of the PCAMIX method.

- Let n denote the number of observation units, p_1 the number of quantitative variables, p_2 the number of qualitative variables and $p = p_1 + p_2$ the total number of variables.
- Let $\mathbf{z}_j$ be the column vector which contains the standardized scores of the n objects on variable j if the j-th variable is quantitative.
- Let $\mathbf{G}_j$ be the $n \times m_j$ indicator matrix for the variable j if it is qualitative with m_j categories and let $\mathbf{D}_j$ be the $m_j \times m_j$ diagonal matrix of observed frequencies.
- Let us denote by $m = m_1 + \cdots + m_{p_2}$ the total number of categories of the p_2 qualitative variables.
- Let $\mathbf{G} = (\mathbf{G}_1 | \cdots | \mathbf{G}_j | \cdots | \mathbf{G}_{p_2})$ be the $n \times m$ matrix of the indicator variables of the m categories of the p_2 qualitative variables and let $\mathbf{D} = \text{diag}(\mathbf{D}_1, \ldots, \mathbf{D}_j, \ldots, \mathbf{D}_{p_2})$ be the $m \times m$ diagonal matrix of frequencies of the m categories.

- Let $\mathbf{J} = \mathbf{I}_n - \mathbf{11}'/n$ be the centering operator where $\mathbf{I}_n$ denotes the $n \times n$ identity matrix and $\mathbf{1}$ the vector of order n with unit entries.

In the two following subsections, we give two formulations of the PCAMIX method and highlight their main differences.

2.1 The original PCAMIX procedure

Suppose k is the number of components required in PCAMIX. In Kiers (1991), the procedure computes the $n \times k$ matrix $\mathbf{X}$ of the standardized component scores, the variance of each component and the $p \times k$ matrix $\mathbf{C}$ of the squared loadings. The squared loadings are defined as squared correlation for quantitative variables and as correlation ratio for qualitative variables. This procedure is carried out according to the following steps:

1. For $j = 1, \ldots, p$: calculate the so-called $n \times n$ quantification matrix $\mathbf{S}_j$ with:

$$\begin{cases} \mathbf{S}_j = \frac{1}{n}\mathbf{z}_j\mathbf{z}_j' & \text{if variable } j \text{ is quantitative,} \\ \mathbf{S}_j = \mathbf{J}\mathbf{G}_j\mathbf{D}_j^{-1}\mathbf{G}_j'\mathbf{J} & \text{if variable } j \text{ is qualitative.} \end{cases}$$

2. Calculate the $n \times n$ matrix $\mathbf{S} = \sum_{j=1}^{p} \mathbf{S}_j$.

3. Perform an Eigenvalue Decomposition of $\mathbf{S}$. The desired matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_k)$ of the standardized component scores is given by the first k eigenvectors $\mathbf{x}_k$ of $\mathbf{S}$ normalized to n (such that $\mathbf{X}'\mathbf{X} = n\mathbf{I}_k$).

4. For $l = 1, \ldots, k$: calculate the variance of the l-th component given by $\mathbf{x}_l'\mathbf{S}\mathbf{x}_l$ where $\mathbf{x}_l$ denotes the l-th column of $\mathbf{X}$.

5. Calculate the $p \times k$ matrix $\mathbf{C} = (c_{jl})$ of the squared loadings of the p variables on the k components with $c_{jl} = \frac{1}{n}\mathbf{x}_l'\mathbf{S}_j\mathbf{x}_l$. For quantitative (resp. qualitative) variables, c_{jl} is the squared correlation (resp. correlation ratio) between the variable j and the component l.

When all the variables are quantitative (resp. qualitative), this procedure is equivalent to PCA (resp. MCA). But the loadings (correlations between the variables and the components) and the principal coordinates of the categories (the barycenters of the component scores) are not directly provided and must be calculated afterwards if desired. From a practical point of view this procedure requires the construction and the storage of p matrices of dimension $n \times n$ which can lead to memory size problems when n and p increase.

2.2 The SVD based PCAMIX procedure

This procedure is carried out according to the following steps:

1. Determine the $n \times (p_1 + m)$ matrix of interest $\mathbf{Z} = \frac{1}{\sqrt{n}}(\mathbf{Z}_1|\mathbf{Z}_2)$ where:

   - $\mathbf{Z}_1 = (\mathbf{z}_1|\cdots|\mathbf{z}_j|\cdots|\mathbf{z}_{p_1})$ is the $n \times p_1$ matrix of the standardized scores of the n observation units (objects) on the p_1 quantitative variables.
   - $\mathbf{Z}_2$ is the $n \times m$ matrix obtained by recoding $\mathbf{G}$ in the following way: $\mathbf{Z}_2 = \mathbf{JGD}^{-1/2}$.

2. Perform the SVD of $\mathbf{Z}$:
$$\mathbf{Z} = \mathbf{U}\Lambda^{1/2}\mathbf{V}', \tag{1}$$
where $\mathbf{U}'\mathbf{U} = \mathbf{V}'\mathbf{V} = \mathbf{I}_r$, Λ is the diagonal matrix of eigenvalues (in weakly descending order) and r is the rank of $\mathbf{Z}$.

3. Calculate the $n \times k$ matrix of the standardized component scores:
$$\mathbf{X} = \sqrt{n}\mathbf{U}_k \tag{2}$$
where $\mathbf{U}_k$ denotes the matrix of the first k columns of $\mathbf{U}$.

4. For $\ell = 1, \ldots, k$, the variance of the ℓ-th component is given by the ℓ-th eigenvalue in Λ.

5. Calculate the matrix:
$$\mathbf{A} = \mathbf{V}_k\Lambda_k^{1/2}, \tag{3}$$
where $\mathbf{V}_k$ denotes the matrix of the first k columns of $\mathbf{V}$ and Λ_k is the diagonal matrix of the k largest eigenvalues.

6. Write $\mathbf{A} = \begin{pmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \end{pmatrix}$ the concatenation of a $p_1 \times k$ matrix $\mathbf{A}_1$ and a $m \times k$ matrix $\mathbf{A}_2$.

   - The matrix $\mathbf{A}_1$ contains the loadings of the quantitative variables (the correlations between the quantitative variables and the components).
   - The matrix $\mathbf{D}^{-1/2}\mathbf{A}_2$ contains the principal coordinates of the categories of the qualitative variables.
   - Calculate the matrix $\mathbf{C}$ of the squared loadings of the p variables on the k components. This matrix is obtained from the matrix $\mathbf{A}$ as follows:
$$\begin{cases} c_{jl} = a_{jl}^2 & \text{if variable } j \text{ is quantitative,} \\ c_{jl} = \sum_{s \in I_j} a_{sl}^2 & \text{if variable } j \text{ is qualitative,} \end{cases}$$

where I_j is the set of row indices of $\mathbf{A}$ associated with the categories of the qualitative variable j. To simplify the notations, we note hereafter $c_{jl} = \sum_{s \in I_j} a_{sl}^2$ for both quantitative and qualitative variables with $I_j = \{j\}$ in the quantitative case.

Note that the matrix $\mathbf{X}$ of the standardized component scores is obtained from the SVD of the recoded data matrix $\mathbf{Z}$ whereas it was obtained from the Eigenvalue Decomposition of the matrix $\mathbf{S}$ (the sum of the quantification matrices $\mathbf{S}_j$) in Kiers' original approach. Also, the matrix $\mathbf{C}$ of the squared loadings (squared correlations or correlation ratios between the variables and the components) is calculated here from only one matrix $\mathbf{A}$ obtained with the SVD of $\mathbf{Z}$ whereas it was calculated from two matrices $\mathbf{X}$ and $\mathbf{S}_j$ in Kiers' original approach.

Contrary to the original PCAMIX approach, this procedure simultaneously provides the loadings of the quantitative variables and the principal coordinates of the categories of the qualitative variables. Moreover, when the data are mixed (quantitative and qualitative), the well known barycentric property in MCA remains true: the coordinates of the categories are the averages of the standardized component scores of the objects in those categories. The matrices $\mathbf{X}$, $\mathbf{A}_1$ and $\mathbf{D}^{-1/2}\mathbf{A}_2$ are then used to plot the observation units, the quantitative variables and the categories with the same interpretation rules as in PCA and MCA. Matrix $\mathbf{C}$ is used to plot the quantitative and qualitative variables on a same graphic.

3 Varimax rotation in PCAMIX

3.1 The optimization problem

Orthogonal rotation. As shown by Eckart and Young (1936), from the SVD in (1) and definitions of matrices $\mathbf{X}$ and $\mathbf{A}$ given in (2) and (3), the matrix $\mathbf{XA}'$ is a rank k least squares approximation of $\mathbf{Z}$. Let us introduce $\mathbf{T}$ an orthonormal rotation matrix: $\mathbf{TT}' = \mathbf{T}'\mathbf{T} = \mathbf{I}_k$. Let $\widetilde{\mathbf{X}} = \mathbf{XT}$ and $\widetilde{\mathbf{A}} = \mathbf{AT}$. As $\mathbf{XA}' = \widetilde{\mathbf{X}}\widetilde{\mathbf{A}}'$, this approximation is not unique over orthogonal rotations.

This non-uniqueness can be exploited to improve the interpretability of the original solutions. To simplify the interpretations, the matrices $\mathbf{X}$ and $\mathbf{A}$ are then rotated in such a way that when considering one variable, few squared loadings are large (close to 1) and as many as possible are close to zero.

The varimax problem. In PCA, since $\widetilde{\mathbf{A}}$ contains the loadings of the variables after rotation, the varimax rotation problem is formulated as

$$\max_{\mathbf{T}} \quad \{f(\mathbf{T}) | \mathbf{TT}' = \mathbf{T}'\mathbf{T} = \mathbf{I}_k\}, \tag{4}$$

where

$$f(\mathbf{T}) = \sum_{l=1}^{k}\sum_{j=1}^{p}(\tilde{a}_{jl}^2)^2 - \frac{1}{p}\sum_{l=1}^{k}\left(\sum_{j=1}^{p}\tilde{a}_{jl}^2\right)^2 \tag{5}$$

is the varimax function measuring the simplicity of the components after rotation.

In the SVD approach of PCAMIX, the varimax function f is defined by replacing in (5) the terms $\tilde{a}_{jl}^2$ by $\tilde{c}_{jl}$, where $\tilde{c}_{jl} = \sum_{s\in I_j}\tilde{a}_{sl}^2$ are the squared loadings after rotation:

$$f(\mathbf{T}) = \sum_{l=1}^{k}\sum_{j=1}^{p}(\tilde{c}_{jl})^2 - \frac{1}{p}\sum_{l=1}^{k}\left(\sum_{j=1}^{p}\tilde{c}_{jl}\right)^2. \tag{6}$$

Note that the squared loadings after rotation $\tilde{c}_{jl}$ are squared correlations (resp. correlation ratios) between the quantitative (resp. qualitative) variables and the rotated components.

For comparison purpose, we recall Kiers' original expression of the varimax function in PCAMIX: the squared loadings after rotation $\tilde{c}_{jl}$ are given by $\frac{1}{n}\tilde{\mathbf{x}}_l'\mathbf{S}_j\tilde{\mathbf{x}}_l$, where $\tilde{\mathbf{x}}_l$ denotes the l-th column of $\widetilde{\mathbf{X}}$. Hence the varimax function (6) becomes:

$$f(\mathbf{T}) = \sum_{l=1}^{k}\sum_{j=1}^{p}\left(\frac{1}{n}\tilde{\mathbf{x}}_l'\mathbf{S}_j\tilde{\mathbf{x}}_l\right)^2 - \frac{1}{p}\sum_{l=1}^{k}\left(\sum_{j=1}^{p}\frac{1}{n}\tilde{\mathbf{x}}_l'\mathbf{S}_j\tilde{\mathbf{x}}_l\right)^2. \tag{7}$$

The iterative optimization procedure. Since a direct solution of (4) for the optimal $\mathbf{T}$ is not available, an iterative optimization procedure suggested by Kaiser (1958) for PCA can be used for PCAMIX. The idea is to consider at each iteration a planar rotation for which the rotation matrix $\mathbf{T}$ only depends on an angle θ (see below for details). This procedure rotates pairs of dimensions in the following way: the single-plane rotations are applied to dimensions 1 and 2, 1 and 3, …, 1 and k, 2 and 3,…, $(k-1)$ and k, iteratively until the process converges, i.e. until $k(k-1)/2$ successive rotations providing an angle of rotation equal to zero are obtained.

The key point of this rotation procedure is the definition of the single-plane rotation step. We give next details on the calculation of the optimal angle for planar rotation. Then we give the complete iterative procedure for rotation in more than two dimensions.

3.2 Planar rotation

Single planar rotations in $\mathbb{R}^2$ are obtained with a rotation matrix $\mathbf{T}$ defined by

$$\mathbf{T} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \tag{8}$$

where θ is the angle of rotation. The varimax rotation problem (4) is then rewritten as:

$$\max_{\theta\in\mathbb{R}} f(\theta).$$

For comparison purpose we recall first the solution based on Kiers' matrix reformulation before we give our direct solution.

3.2.1 Planar rotation using Kiers' matrix reformulation

Kiers (1991) proposes to use a procedure of simultaneous diagonalization of a set of symmetric matrices (ten Berge, 1984; de Leeuw and Pruzansky, 1978) to solve the global varimax optimization problem (4). For this purpose he gives the following matrix reformulation of the formula (7):

$$f(\mathbf{T}) = p^{-2} \sum_{j=1}^{p} \text{Trace}\,(\mathbf{T}'\mathbf{E}_j\mathbf{T}(\text{Diag}\ \ \mathbf{T}'\mathbf{E}_j\mathbf{T})) \tag{9}$$

where

$$\mathbf{E}_j = p\ \mathbf{X}'\mathbf{S}_j\mathbf{X} - n\Gamma \tag{10}$$

and Γ is the diagonal matrix with the k first eigenvalues of $\mathbf{S}$ on its diagonal.

From ten Berge (1984) and de Leeuw and Pruzansky (1978) it follows that the procedure for simultaneous diagonalization of the matrices $\mathbf{E}_j$ is equivalent to Kaiser's iterative optimization procedure with the optimal angle θ of single plane rotations defined by the equation:

$$\tan(4\theta) = \frac{a}{b}, \tag{11}$$

with

$$a = 4\sum_{j=1}^{p} e_{12}^j(e_{11}^j - e_{22}^j) \quad \text{and} \quad b = \sum_{j=1}^{p}(e_{11}^j - e_{22}^j)^2 - 4\sum_{j=1}^{p}(e_{12}^j)^2 \tag{12}$$

where $\mathbf{E}_j = \begin{pmatrix} e_{11}^j & e_{12}^j \\ e_{21}^j & e_{22}^j \end{pmatrix}$ is defined in (10).

As mentioned by several authors (see for instance Nevels, 1986; ten Berge, 1984; de Leeuw and Pruzansky, 1978 and Kaiser, 1958) equation (11) is only a necessary condition obtained upon setting the first order derivative of the objective function to zero. Both Kaiser (1958) and de Leeuw and Pruzansky (1978) developed a procedure for determining the optimal θ from the sign of the second order derivative of the objective function. They expressed the appropriate solution for every possible combination of signs of a and b in tabular form.

3.2.2 Planar rotation using the SVD approach of PCAMIX

The varimax function $f(\mathbf{T})$ defined with the SVD approach in (6) is written:

$$f(\theta) = \sum_{j=1}^{p}\left(\sum_{s\in I_j}\tilde{a}_{s1}^2\right)^2 + \sum_{j=1}^{p}\left(\sum_{s\in I_j}\tilde{a}_{s2}^2\right)^2 - \frac{1}{p}\left(\sum_{j=1}^{p}\sum_{s\in I_j}\tilde{a}_{s1}^2\right)^2 - \frac{1}{p}\left(\sum_{j=1}^{p}\sum_{s\in I_j}\tilde{a}_{s2}^2\right)^2 \tag{13}$$

with

$$\tilde{a}_{s1} = a_{s1}\cos(\theta) + a_{s2}\sin(\theta) \quad \text{and} \quad \tilde{a}_{s2} = -a_{s1}\sin(\theta) + a_{s2}\cos(\theta). \tag{14}$$

This function is equal to (see Appendix):

$$f(\theta) = f(0) + \frac{\rho}{4p}\big(\cos(4\theta - \psi) - \cos\psi\big) \tag{15}$$

where ρ and ψ are defined by:

$$\rho = (a^2 + b^2)^{1/2} \quad , \quad \cos\psi = b/\rho \quad , \quad \sin\psi = a/\rho \tag{16}$$

with a and b given by:

$$a = 2p\sum_{j=1}^{p} u_j v_j - 2\sum_{j=1}^{p} u_j \sum_{j=1}^{p} v_j \ , \quad b = p\sum_{j=1}^{p}({u_j}^2 - {v_j}^2) - \left(\sum_{j=1}^{p} u_j\right)^2 + \left(\sum_{j=1}^{p} v_j\right)^2 , \tag{17}$$

where u_j and v_j are defined by:

$$u_j = \sum_{s\in I_j}(a_{s1}^2 - a_{s2}^2) \quad \text{and} \quad v_j = 2\sum_{s\in I_j} a_{s1}a_{s2} \ . \tag{18}$$

The function f obtained in (15) is maximum for $\cos(4\theta - \psi) = 1 \Leftrightarrow 4\theta - \psi = 2k\pi$, thus the optimal angles are:

$$\theta = \frac{\psi}{4} + k\frac{\pi}{2}, \quad k \in \mathbb{Z}. \tag{19}$$

Note that the above expressions of u_j and v_j contain as special cases (take $I_j = \{j\}$) those defined by Kaiser (1958) for the PCA varimax solution. Note also that the classical necessary condition (11) immediately follows by setting the expression (23) of $pf'(\theta)$ given in the Appendix to zero (the coefficients a and b given by (12) on one side, and (17) and (18) on the other side are proportional).

3.3 The iterative rotation procedure

We consider now the case where the number k of dimensions in the rotation is greater than two. The iterative rotation procedure gives the matrix $\widetilde{\mathbf{X}}$ of the rotated standardized component scores and the matrix $\widetilde{\mathbf{A}}$ which is used to obtain the rotated squared loadings, the rotated loadings (correlations) of the quantitative variables and the rotated principal coordinates of the categories. This procedure is carried out according to the following steps:

1. Initialization: $\widetilde{\mathbf{X}} = \mathbf{X}$ and $\widetilde{\mathbf{A}} = \mathbf{A}$ where the $n \times k$ matrix $\mathbf{X}$ and the $(p_1 + m) \times k$ matrix $\mathbf{A}$ are given by the SVD based PCAMIX procedure given in Section 2.2.

2. For $l = 1, \ldots, k-1$ and $t = (l+1), \ldots, k$, calculate for the pair of dimensions (l, t):

- the angle of rotation $\theta = \psi/4$ with ψ defined in (16) . We choose:

$$\psi = \begin{cases} \arccos(\frac{b}{\sqrt{a^2+b^2}}) & \text{if} \quad a \geq 0, \\ -\arccos(\frac{b}{\sqrt{a^2+b^2}}) & \text{if} \quad a \leq 0. \end{cases} \tag{20}$$

  where a and b are defined in (17).
- the matrix of rotation $\mathbf{T} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$,
- the matrices $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{A}}$ updated by rotation of their l-th and t-th columns.

3. Repeat the previous step until the $k(k-1)/2$ angles θ are equal to zero.

4. Calculate:

- the matrix $\widetilde{\mathbf{C}}$ with $\tilde{c}_{jl} = \sum_{s \in I_j} \tilde{a}^2_{sl}$.
- the matrix $\widetilde{\mathbf{A}}_1$ of the first p_1 rows of $\widetilde{\mathbf{A}}$ which contains the rotated loadings of the quantitative variables.
- the matrix $\widetilde{\mathbf{A}}_2$ of the last m rows of $\widetilde{\mathbf{A}}$ and the matrix $\mathbf{D}^{-1/2}\widetilde{\mathbf{A}}_2$ which contains the rotated principal coordinates of the categories of the qualitative variables.

The main differences between this procedure and that constructed with Kiers' matrix reformulation are the following:

- The expressions of a and b in step (2): in this procedure they are expressed according to the matrix $\mathbf{A}$ of dimension $(p_1 + m) \times n$. With Kiers' matrix reformulation, a and b are expressed according to the p matrices $\mathbf{S}_j$ of dimension $n \times n$.

- The direct determination of the optimal angle in step (2): having an explicit expression for the solution is of theoretical interest and is more straightforward from a computational point of view.

- The outputs: this procedure directly provides the rotated loadings of the quantitative variables and the rotated principal coordinates of the categories which are used for graphical representations after rotation.

4 Numerical studies

The procedure proposed in this paper for varimax orthogonal rotation in PCAMIX has been implemented in R. A package called "PCAmixdata" is already available on the CRAN website. In this section, this algorithm is compared on simulated data with Kiers' rotation procedure. Then an application on a real data example illustrates the possible benefits of using rotation in MCA as particular case of PCAMIX.

4.1 A simulation study: comparison of computational times

An iterative rotation procedure based on Kiers' matrix reformulation has also been implemented in R. This procedure is the one proposed in Section 3.3 with the following modifications:

- Kiers' original PCAMIX procedure is used in the initialization step in place of the SVD based PCAMIX procedure.
- All the calculations and outputs based on the matrix $\mathbf{A}$ are removed because this matrix is not part of the original PCAMIX procedure.
- The coefficients a and b in step 2 are calculated according to their expressions (12) associated to Kiers' matrix reformulation. Note that the ratio $\frac{a}{b}$ is the same with the two approaches (SVD and matrix reformulation) so the optimal angle θ is the same.
- In step 4 the squared loadings are calculated with their expression in the original PCAMIX approach.

The computation time of the two rotation procedures (the one based on Kiers' matrix reformulation and the one based on the SVD approach of PCAMIX) is compared from simulated datasets with varying parameters: the number p of variables and the number n of observations. Each dataset is composed of $p/2$ quantitative variables and $p/2$ qualitative variables. The procedure used to build each dataset consists in simulating p quantitative variables and then recoding the last $p/2$ variables to get qualitative variables. More precisely the following two steps are used to simulate one dataset:

- A dataset with n observations and p variables is drawn from a multivariate normal distribution with a covariance matrix $\Sigma = \mathbf{Q}'\mathbf{Q}$ where $\mathbf{Q}$ is a $p \times p$ matrix drawn with p^2 independent elements from a uniform distribution on the interval $[0.2; 0.4]$.
- The last $p/2$ variables of this dataset are distributed in three equal-count categories. Each final dataset is then constituted of $p_1 = p/2$ quantitative variables, $p_2 = p/2$ qualitative variables and the total number of categories is $m = 3 \times p/2$.

For each set of parameters (n, p), 20 simulations are drawn. Because the two rotation procedures iterate planar rotations until convergence, we compare their computation time for $k = 2$. For each procedure and couple of parameters (n, p), the median of computation times (over the 20 replications) is given in Table 1 and the ratio between these medians is given in Table 2.

		p=10	p=50	p=100	p=200
n=50	Matrix reformulation	0.05	0.12	0.22	0.44
n=50	SVD	0.02	0.06	0.12	0.27
n=100	Matrix reformulation	0.14	0.33	0.56	1.04
n=100	SVD	0.02	0.09	0.17	0.34
n=200	Matrix reformulation	0.55	1.12	1.86	3.38
n=200	SVD	0.02	0.11	0.26	0.53
n=400	Matrix reformulation	2.15	4.32	7.1	12.65
n=400	SVD	0.03	0.16	0.37	0.89
n=800	Matrix reformulation	10.06	19.27	30.54	error
n=800	SVD	0.05	0.25	0.58	1.79

Table 1: Medians of computation times (in seconds) of two PCAMIX rotation procedures: the one based on Kiers' matrix reformulation and the one based on the SVD approach of PCAMIX.

	p=10	p=50	p=100	p=200
n=50	2.9	2.0	1.8	1.6
n=100	8.7	3.8	3.3	3.0
n=200	23.2	10.3	7.0	6.4
n=400	69.4	27.7	19.0	14.2
n=800	214.1	77.4	52.9	error

Table 2: Ratios between medians of computation times of the two rotation procedures (matrix reformulation/SVD).

Table 1 shows that the proposed rotation procedure (based on the SVD approach of PCAMIX) is faster than the one based on Kiers' matrix reformulation for all configurations. For configurations where $p = 10$, Table 2 shows that our proposed rotation procedure is from 3 times faster for $n = 50$ to 214 times faster for $n = 800$ than the rotation procedure based on Kiers' matrix reformulation. For configurations with large values of p, this ratio is less important but still increases with n. For the configuration when n and p are large ($n = 800$ and $p = 200$), an error occurs with the rotation procedure based on Kiers' matrix reformulation. The maximum capacity of memory size of the computer was reached in that case. This error occurs during the calculation of the p matrices $\mathbf{S}_j$ of size $n \times n$. This exhibits the computational efficiency of the proposed rotation procedure.

4.2 A real data application

This real data application illustrates the interest of rotation in MCA. A food habits survey[1] was carried out in 1999 on students living in the region "Aquitaine" in South West of France. We focus on the answers of 2885 students to 12 binary questions concerning their consumption at breakfast (coffee, cereals, eggs...). The PCAMIX method (equivalent here to MCA) has been applied to this dataset and the first 4 components have been rotated.

In Figure 1 the association of the variables with the first two components is obviously easier after rotation. This rotation of the first four components leads in Table 3 to clear associations between the binary variables: coffee is associated with milk, eggs with cheese and deli, bread with jam and cereals with pure milk. The effect of the rotation on the objects scores and on the categories coordinates can also be visualized in Figures 2 and 3. The interpretation rule associated with the barycentric property remains true after rotation.

	Before rotation				*After rotation*			
	1	2	3	4	1	2	3	4
coffee	0.23	0.22	0.06	0.05	**0.49**	0.00	0.00	0.07
tea	0.05	0.01	0.06	0.18	0.05	0.02	0.06	0.17
milk	0.15	0.16	0.08	0.00	**0.37**	0.00	0.01	0.01
milk chocolate	0.43	0.18	0.01	0.06	**0.62**	0.00	0.01	0.05
pure milk	0.02	0.00	0.05	0.40	0.01	0.00	0.02	**0.44**
cheese	0.18	0.23	0.01	0.00	0.00	**0.42**	0.00	0.00
deli	0.20	0.27	0.00	0.05	0.00	**0.51**	0.00	0.01
eggs	0.20	0.37	0.00	0.01	0.00	**0.58**	0.00	0.00
jam	0.06	0.02	0.49	0.02	0.00	0.00	**0.59**	0.00
honey	0.00	0.05	0.14	0.20	0.00	0.02	0.20	0.16
bread	0.11	0.01	0.45	0.00	0.01	0.01	**0.53**	0.03
cereals	0.01	0.01	0.12	0.22	0.05	0.01	0.04	**0.27**

Table 3: Correlation ratio (squared loadings) between the variables and the first 4 components before and after rotation.

Note that for binary variables MCA and PCA lead to equivalent object scores and squared loadings (correlations are equal to correlation ratios). Then considering the data as quantitative in PCAMIX (equivalent to PCA in that case) gives the same results except for the plots of the categories which are not defined in that case.

5 Conclusion

We have given in this paper a SVD based formulation of the PCAMIX method. This new formulation leads to an efficient procedure for varimax rotation in PCAMIX where a direct solution for the optimal angle of planar rotation has been obtained. The numerical results have shown on simulations that this procedure is computationally more efficient than the procedure based on Kiers' matrix reformulation. Also the numerical results on a real data

[1]This survey was realized by the Bordeaux School of Public Health (Institut de Santé Publique, d'Epidémiologie et de Développement - ISPED)

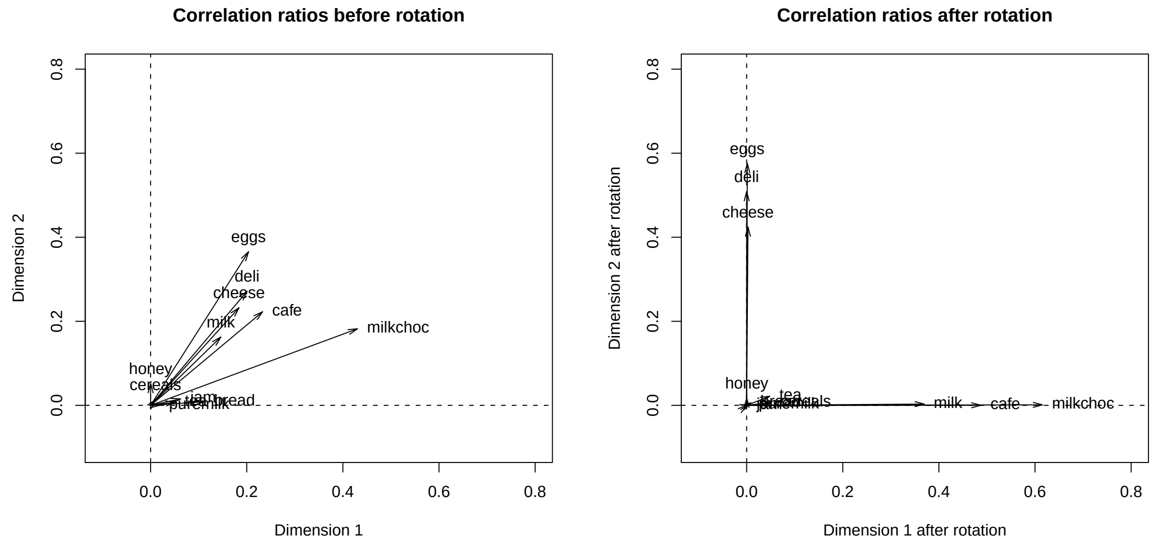

Figure 1: Plots of the correlation ratios between the variables and the two first components before rotation and after rotation.

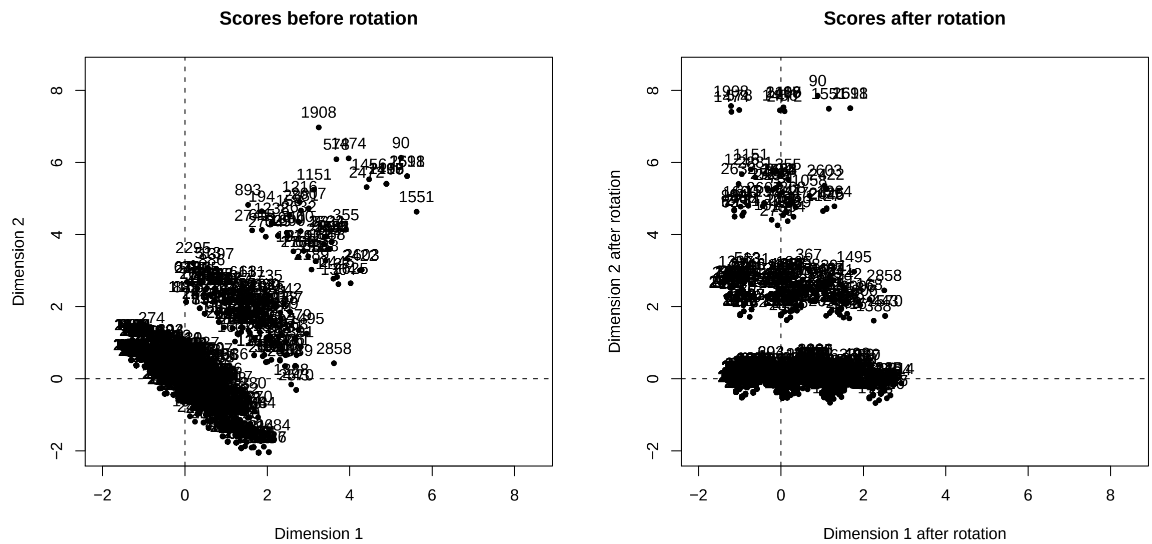

Figure 2: Plots of the (standardized) scores of the 2885 students on the first two components before and after rotation.

application have shown the interest of this algorithm in the context of MCA with graphical representations of both variables and categories after rotation. The PCAMIX procedure as well as the rotation procedure have been implemented in the R package “PCAmixdata”.

Appendix

Define the complex numbers:

$$
\begin{aligned}
a_s &\stackrel{\text{def}}{=} a_{s,1} + i a_{s,2} \ , & \tilde{a}_s &\stackrel{\text{def}}{=} e^{-i\theta} a_s = \tilde{a}_{s,1} + i\tilde{a}_{s,2} \ , \\
t_j &\stackrel{\text{def}}{=} \textstyle\sum_{s \in I_j} a_s^2 = u_j + i v_j \ , & \tilde{t}_j &\stackrel{\text{def}}{=} \textstyle\sum_{s \in I_j} \tilde{a}_s^2 = e^{-2i\theta} t_j = \tilde{u}_j + i\tilde{v}_j \ ,
\end{aligned}
$$

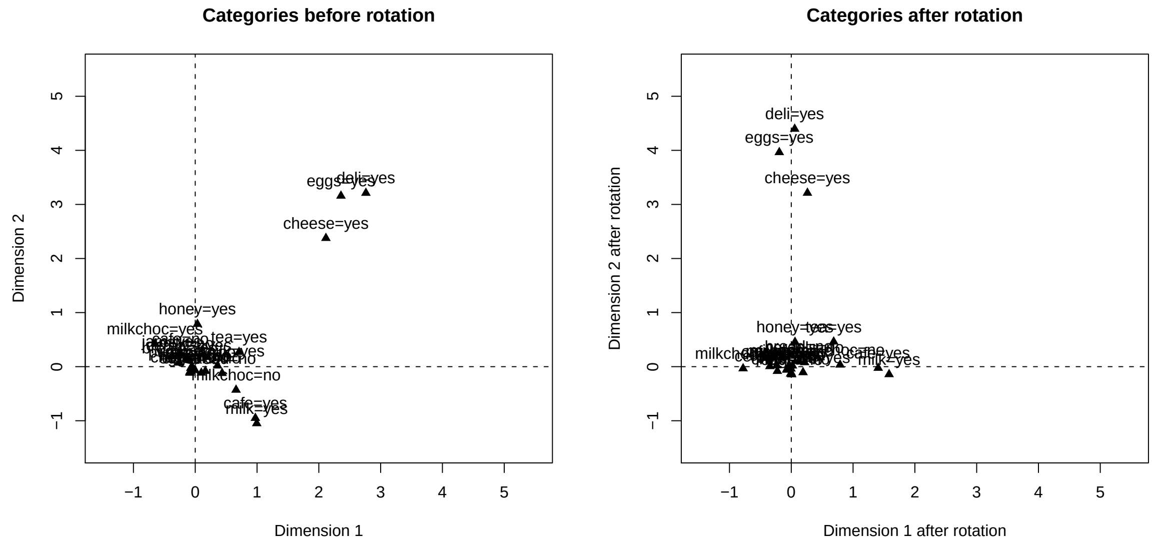

Figure 3: Plots of the category coordinates on the first two components before and after rotation.

where $\tilde{a}_{s,1}, \tilde{a}_{s,2}$ have been defined in (14), u_j, v_j in (18), and where $\tilde{u}_j, \tilde{v}_j$ are given by the same formula as u_j, v_j, but with a tilde over $a_{s,1}, a_{s,2}$.

We introduce now a complex-valued varimax function $F(\theta)$ of the rotation angle θ by:

$$F(\theta) \stackrel{\text{def}}{=} p \sum_{j=1}^{p} \tilde{t}_j^2 - (\sum_{j=1}^{p} \tilde{t}_j)^2 = e^{-4i\theta} F(0) \ ,$$

where $F(0)$ is given by the formula for $F(\theta)$ after suppressing the tilde. $F(\theta)$ can be decomposed as follows:

$$F(\theta) = \underbrace{p \sum_{j=1}^{p} (\tilde{u}_j^2 - \tilde{v}_j^2) - (\sum_{j=1}^{p} \tilde{u}_j)^2 + (\sum_{j=1}^{p} \tilde{v}_j)^2}_{g(\theta)} + \underbrace{2i\{p \sum_{j=1}^{p} \tilde{u}_j \tilde{v}_j - \sum_{j=1}^{p} \tilde{u}_j \sum_{j=1}^{p} \tilde{v}_j\}}_{i\,h(\theta)} . \tag{21}$$

Comparison with the formula (16), (17), (18) defining b, a, ρ, ψ shows that:

$$F(0) = g(0) + ih(0) = b + ia = \rho\, e^{i\psi} \ .$$

Hence we get:

$$F(\theta) = \rho\, e^{i(\psi - 4\theta)} = \rho \left\{ \cos(4\theta - \psi) - i \sin(4\theta - \psi) \right\} \ .$$

But taking derivatives of the varimax function $f(\theta)$ defined in (13) gives, using the fact that $a'_{s,1}(\theta) = a_{s,2}(\theta)$ and $a'_{s,2}(\theta) = -a_{s,1}(\theta)$:

$$\begin{aligned} pf'(\theta) &= 2\{p \sum_{j=1}^{p} \tilde{u}_j \tilde{v}_j - \sum_{j=1}^{p} \tilde{u}_j \sum_{j=1}^{p} \tilde{v}_j\} \\ &= h(\theta) = -\rho \sin(4\theta - \psi) && (22) \\ &= a \cos 4\theta - b \sin 4\theta, && (23) \end{aligned}$$

and (22) proves (15) by integration.

References

Adachi, K., (2004), Oblique promax rotation applied to solutions in multiple correspondence analysis, *Behaviormetrika*, **31**, 1-12.

Browne, M. W., (2001), An overview of analytic rotation in exploratory factor analysis, *Multivariate behavioral research*, **36**, 111-150.

de Leeuw, J., and Pruzansky, S., (1978), A new computational method to fit the weighted Euclidean distance model, *Psychometrika*, **43**, 479-490.

Escofier, B., (1979), Traitement simultané de variables qualitatives et quantitatives en analyse factorielle [Simultaneous treatment of qualitative and quantitative variables in factor analysis], *Cahiers de l'Analyse des Données*, **4**, 137-146.

Greenacre, M.J., (2006), Tying up the loose ends in simple correspondence analysis, *http://www.econ.upf.es/docs/papers/downloads/940.pdf*

Jennrich, R.I., (2001), A simple general procedure for orthogonal rotation, *Psychometrika*, **66**(2), 289-306.

Kaiser, H.F., (1958), The varimax criterion for analytic rotation in factor analysis, *Psychometrika*, **23**(3), 187-200.

Kiers, H.A.L., (1991), Simple structure in Component Analysis Techniques for mixtures of qualitative and quantitative variables, *Psychometrika*, **56**, 197-212.

Neudecker, H., (1981), On the matrix formulation of Kaiser's varimax criterion, *Psychometrika*, **46**, 343-345.

Nevels, K., (1986), A direct solution for pairwise rotations in Kaiser's varimax method, *Psychometrika*, **51**(2), 327-329.

Pagès, J., (2004), Analyse Factorielle de données mixtes [Factor Analysis for Mixed Data], *Revue de Statistique Appliquée*, **52**(4), 93-11.

ten Berge, J.M.F., (1984), A joint treatment of varimax rotation and the problem of diagonalizing symmetric matrices simultaneously in the least-squares sense, *Psychometrika*, **49**, 347-358.

Urbano, L.-S., van de Velden, M., and Kiers, H. A. L., (2009a), CAR: a matlab package to compute correspondence analysis with rotation, *British journal of mathematical and statistical psychology*, **62**, 583-600.

Urbano, L.-S., van de Velden, M., and Kiers, H. A. L., (2009b), Oblique rotation in correspondence analysis: a step forward in the search of the simplest interpretation, *Journal of statistical software*, **31**(8), 1-14.

van de Velden, M., and Kiers, H. A. L., (2003), An application of rotation in correspondence analysis, *In H. Yanai, A. Okada, K. Shigemasu, Y. Kano, and J.J. Meulman, (Eds.), New Developments in Psychometrics*, Tokyo: Springer Verlag, 471-478.

van de Velden, M., and Kiers, H. A. L., (2005), Rotation in correspondence analysis, *Journal of Classification*, **22**, 251-271.