Applied Mathematical Sciences, Vol. 2, 2008, no. 14, 659 - 666

The Study of Multivariate Temporal Data using PCA under Linear Constraints (PCA-LC)

Mohamed E. Baouche

Laboratoire de Mathématique
BP89 Sidi Bel-Abbès 22000, Algeria
elouard@univ-sba.dz

Abstract. PCA under Linear Constraints (PCA-LC) is a PCA in which we impose to the principal axis and components to belong to some sub-spaces. The Idea is to look for the principal axis and components by the optimization of a function defined on a special set of orthogonal basis.
The aim of this paper is to use some specifications and characterestics of PCA-LC in order to study multivariate temporal data as particular case of thre-way data.

Mathematics Subject Classification: 62H

Keywords: PCA, Rayleigh quotient, linear constraints, R-criteria, three-way data

1. Introduction

Linear Constrained PCA(PCA-LC)[8], is a generalization of the classic PCA to the case where linear constraints are introduced. Contrary to the traditional PCA, the PCA-LC gives different solutions all depends on the space on which we introduce the constraints firstly.

A three-way data is a data array with three indices, the first for the individuals, the second for the variables and the third for the occasions[1]. In the last 40 years, a large numbers of methods was developed to study them. We distinguish two large types of methods[5]: those which regard them as a cube of data, and those which pile up or juxtapose the tables. In this work we will study multivariate temporal data (where the third index corresponds to the time) by using the method of the PCA-LC.

In the first section we will give the notations used in this paper, after we will give a fast overview of the method of PCA-LC where some definitions are explained in the appendix. In the last section we will apply the method of PCA-LC in the study of multivariate temporal data.

2. Notation

Data are presented as a matrix $X_{n\times p}$, $p \leq n$, where each row corresponds to an object (a total of n), and each column to a variable. We consider individuals as elements of a vectorial space $E = \mathbb{R}^p$ and the variables as elements of another vectorial space $F = \mathbb{R}^n$. In this paper we assume that all variables have zero means. We can summarize our data in the next duality diagram:

$$\begin{array}{ccc} E & \xleftarrow{X^t} & F^* \\ {\scriptstyle V}\uparrow\downarrow{\scriptstyle M} & & {\scriptstyle W}\downarrow\uparrow{\scriptstyle N} \\ E^* & \xrightarrow{X} & Im(X) \subset F \end{array}$$

where:

- M and N are respectively the metrics defined on E and F.
- $V = X^tNX$ and $W = XMX^t$ are semi-metrics.
- We often consider N to be equal to D_p the diagonal weights matrix, in this case V is the variance-covariance matrix of the columns of X, and when the variables are standardized V become the correlation matrix.

We assume X to be a full rank matrix, this condition is equivalent to assume that X is injective, its implies that V is inversible.

The PCA of (X, M, N) consists to look for:

- principal axis u_i in the individuals space E, by the extraction of the eigen vectors of the M-symmetric operator VM, as result of the optimization of the inertia. To these axis we can associate:
  - principal factors $u_i^* = Mu_i$ in E^*.
  - principal components $U^i = XMu_i$ in F.
- principal components U^j in the variables space F, by the spectral analysis of the N-symmetric operator WN,as result of the optimization of the variance. To these components we can associate:
  - principal axis $u_i = \dfrac{X^tNU^j}{||U^j||_N}$ in E.
  - principal factors $u_i^* = Mu_i$ in E^*.

The classical approach use the inertia as a criterion to find the principal axis or the variance one to get the principal components, in both cases we get the same solutions.

Since RAO[6], many other criteria were proposed for the same goal. Croquette[2] have done a synthesis of these criteria where the major part, called R-criteria, was a function of a Rayleigh quotient depending on a symmetric operator.

3. The Linear Constrained PCA (PCALC)

PCA under Linear Constraints (PCA-LC) is a PCA in which we impose to the principal axis and components to belong to some sub-spaces.

Let $\bigoplus_{1}^{r} E_k$ be a decomposition in direct sum of the individuals space E, and $\bigoplus_{1}^{r} F_k \bigoplus F_0$ a decomposition in direct sum of the variables space F where $\bigoplus_{1}^{r} F_k = Im(X)$ and F_0 an additional space of $Im(X)$ in F. We denote by:

- P_k (resp. P_k^M) the cartesian projector on E_k (resp. The M-orthogonal projector on E_k).
- Q_k (resp. Q_k^N) the cartesian projector on F_k (resp. The N-orthogonal projector on F_k).
- $\forall k = 1, \ldots, r$: $p_k = dim(E_k) = dim(F_k)$, $\sum p_k = p$.

Definition 1. Introduce r linear constraints into E amounts forcing $dim(E_1)$ principal axes to belong to the subspace E_1, and $dim(E_2)$ principal axes to belong to E_2 and so on.
With the same manner we will define linear constraints in F. Introducing r linear constraints into $Im(X)$ amounts forcing $dim(F_1)$ principal components to belong to the subspace F_1, and $dim(F_2)$ principal components to belong to F_2 and so on.

In order to obtain the solutions that take into account the linear constraints, in E or F, we will define a class of functions, called weak R-criteria, when optimized gives us the required solutions.

Let (H, T) be an euclidian space of dimension m, and $\bigoplus_{1}^{r} H_k$ be a decomposition to direct sum of H. We denote by S_k the cartesian projector in H_k and S_k^T the T-orthogonal projector in H_k.

Definition 2.R-weak criteria
Let be A a T-symmetric operator. We denote by $\Delta(H, T)$ the set of T orthogonal basis in H.

- A T-orthogonal basis of H, u, is said **compatible** with the decomposition in direct sum $\bigoplus_{1}^{r} H_k$ if and only if:

$$\forall k \in \{1, \ldots, r\},\ \exists u(k)\ \subset u :\ u(k) \text{ is a basis of } H_k.$$

The set of such basis is denoted: $\Delta(\bigoplus_{1}^{r} H_k, T)$.

- Let $u \in \Delta(\bigoplus_1^r H_k, T)$. The metric $\widetilde{T}$ in H is defined by:

$$\forall\ u_i,\ u_j\ \in\ u:\ \widetilde{T}(u_i,u_j) = \begin{cases} T(u_i,u_j) & if\ \exists k:\ u_i,\ u_j \in H_k \\ 0 & else \end{cases}$$

  By construction $\widetilde{T}$ is an euclidean metric over H, and it coincide with T in each sub-space H_k. $\widetilde{T}$ is a diagonal bloc matrix, its diagonal blocs are exactly the restriction T_{ii} of T on H_i.
- We call R-weak criterion associated to the T-symmetric operator A, every numeric function F_A defined on $\Delta(\bigoplus_1^r H_k, T)$ by:

$$\begin{array}{ccc} \Delta(\bigoplus_1^r H_k, T) & \xrightarrow{F_A} & \mathbb{R} \\ u & \longmapsto & F_A(u) = f(q_A(u)) \end{array}$$

  with:
  $f \in \mathcal{F}(C_m)$ and q_A the rayleigh quotient associated to A. More details about $f \in \mathcal{F}(C_m)$ and q_A can be found in the Appendix.

Property 3.1. *The maximum of an R-weak criterion F_A is reached only and only for a T-orthogonal basis compound of the eigen vectors corresponding to the largest eigen values of the operator $\overline{A} = \sum_1^r S_k^T A S_k$.*

We notice that the maximum of an R-weak criterion is independent of the form of the R-weak criterion, it depends only on the T-orthogonal operator A.

3.1. **M-PCA-LC and N-PCA-LC. Definition 3.M-PCA-LC**
An M-PCA-LC of (X, M, N) specified by the decomposition in direct sum $\bigoplus_1^r E_k$ and in the sens of the M-symmetric operator A, consists on looking for an M-orthonormal basis u of E maximising the R-weak criterion F_A. The basis u allow us to extract the eigen values and eigen vectors of the operator $\overline{A} = \sum_1^r P_k^M A P_k$.
The basis U of $Im(X)$ corresponding to u is obtained by the spectral analysis of the operator $\overline{C} = X\widetilde{M}\overline{A}X^t\widetilde{W}^{-1}$, where: $\widetilde{M} = (M_{ii})$, $\widetilde{V} = (V_{ii})$, $\widetilde{W} = X^t\widetilde{M}X$. This M-PCA-Cl is denoted: M-PCA$(X, M, N, ; F_A/E_1, \dots E_r)$.

Definition 4.N-PCA-LC
An N-PCA-LC of (X, M, N) specified by the decomposition in direct sum

[1] $\widetilde{W}^- = NX\widetilde{V}^{-1}\widetilde{M}^{-1}\widetilde{V}^{-1}X^t\widetilde{N}$ is a g-inverse of $\widetilde{W}$ with $\widetilde{W}^-\widetilde{W}$ is N^{-1}-symmetric and $\widetilde{W}\widetilde{W}^-$ is N-symmetric

$\bigoplus_{1}^{r} F_k$ and in the sens of the N-symmetric operator A, consists on looking for an N-orthogonal basis U of $Im(X)$ maximising the R-weak criterion G_C by the extraction of eigen values and eigen vectors of the operator $\overline{C} = \sum_{1}^{r} Q_k^N C Q_k$.
The basis u of E corresponding to U is obtained by the spectral analysis of the operator $\overline{A} = X^t \widetilde{N} \overline{C} X \widetilde{V}^{-1}$, where: $\widetilde{N} = (N_{ii})$.
This N-PCA-CL is denoted: N-PCA$(X, M, N, ; G_C/F_1, \dots F_r)$

Remark 3.1.
To have the ability to make a PCA-LC, we have to determine:

- *The metrics M and N.*
- *The M-symmetric operator A if we start from the individuals space E, or the N-symmetric operator C in the other case. The choice of A and C depends on our goal behind the practice of the PCA-LC and require some practice.*

4. Multidimensional temporal data with PCA-LC

A multidimensional temporal data is a three-way data where the third index is time. Its obtained when measuring the same group of variables on the same group of individuals on many instants of time. The number of instants is less than that can allow us to use the process theory or the time series one[5].
Let be $X_1, X_2, \dots, X_r$ the data matrix such that: $\forall k$, X_k the matrix of the data collected at the instant k.
To each matrix X_k corresponds a duality diagram:

$$\begin{array}{ccc} E_k & \xleftarrow{X_k^t} & F^* \\ {\scriptstyle V_{kk}}\uparrow \;\downarrow{\scriptstyle M_{kk}} & & \uparrow{\scriptstyle N} \\ E_k^* & \xrightarrow{X_k} & F \end{array}$$

For two different matrix X_i et X_j, wa can associate the next duality diagram:

$$\begin{array}{ccccc} E_i & \xleftarrow{X_i^t} & F^* & \xrightarrow{X_j^t} & E_j \\ {\scriptstyle V_{ii}}\uparrow \;\downarrow{\scriptstyle M_{ii}} & & \uparrow{\scriptstyle N} & & {\scriptstyle V_{jj}}\uparrow \;\downarrow{\scriptstyle M_{jj}} \\ E_i^* & \xrightarrow{X_i} & F & \xleftarrow{X_j} & E_j^* \end{array}$$

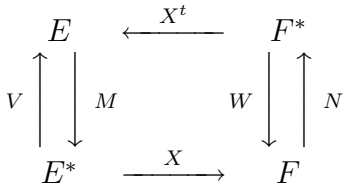
from this diagram wa can define a new application:

$$V_{ij} : E_j^* \longrightarrow E_i$$

such that: $V_{ij} = X_i^t N X_j$.
We suppose $\forall i, j : E_i \cap E_j = \phi$.

We pose: $E = \bigoplus_{1}^{r} E_k$.

and let $X = (X_1|X_2|\cdots|X_r)$ matrix obtained by the juxtaposition of $X_1 \ldots X_r$. A new duality digram can summarize the new data matrix:



with:

- $M = (M_{ij})$.
- $V = (V_{ij})$.

We have: $\widetilde{M} = (M_{ii})$, $\widetilde{V} = (V_{ii})$, $\widetilde{W} = X^t\widetilde{M}X$.

Once we determined the necessary elements to carry out a PCA-CL, it is to be noticed that the goal is to obtain axis and principal components which summarize the information of all the tables, from where the idea to impose linear constraints for which we reserve for each sub-space his part of the principal elements.

The choice of a PCA-LC, specified by $\bigoplus_{1}^{r} E_k$ and in the sens of an M-symmetric operator, is adopted because the decomposition in direct sum is introduced initially on the sub-space of the individuals E. The principal axes are obtained by the extraction of the eigen values and eigen vectors of the operator $\overline{A} = \sum P_k^M A P_k$, and the principal components are obtained from the operator $\overline{C} = X\widetilde{M}\overline{A}X^t\widetilde{W}^-$.

Appendix

Let (H, T) be an Euclidean space of dimension m, and A a T-symmetric operator on H. We denote by $\Delta(H, T)$ the set of the T-orthogonal basis of H. Each element of $\Delta(H, T))$ can be considered as a vector containing the elements of the basis ex: $u = (u_1, u_2, \ldots, u_m)$.

Definition 5. A Rayleigh quotient q_A, associated to the T-symmetric operator A, is a function defined from $H - \{0\}$ to $\mathbb{R}$ by:

$$q_A(h) = \frac{T(Ah, h)}{T(h, h)} \quad \forall h \in H - \{0\}$$

Property 4.1. [2] *The Rayleigh quotient, associated to the T-symmetric operator A, defined over $\Delta(H,T)$ by:*

$$\begin{array}{ccc} \Delta(H,T) & \xrightarrow{q_A} & \mathbb{R}^m \\ u & \longmapsto & q_A(u) = (q_A(u_1), q_A(u_2), \ldots, q_A(u_m)) \end{array}$$

satisfies:
1) $q_A(\Delta(H,T)) = C_m$ *is a convex, compact and symmetric polyhedra of* $\mathbb{R}^m$
2) $C_m = \left\{ (z_1, \ldots, z_m) \in \mathbb{R}^m, \forall J \subset \{1, \ldots, m\} : \sum_{j=1}^{card(J)} \lambda_j^{<} \leq \sum_{j \in J} z_j \leq \sum_{j=1}^{card(J)} \lambda_j^{>} \right\}$
where $\{\lambda_j^{>}, j = 1, \ldots, m\}$ *(respectively* $\{\lambda_j^{<}, j = 1, \ldots, m\}$*) are the eigen values of A set in decreasing order (respectively increasing order).*

Definition 6. Let H_1 be a convex and symmetric subset of $\mathbb{R}^m$. A function f from H_1 to $\mathbb{R}$ is called "strictly s-convex" function if:

- f is symmetric on H_1
- $\forall z = (z_1, \ldots, z_m) \in H_1$ such that $z_1 \neq z_2$, the function g_z defined by: $\begin{array}{ccc} [0, 1/2] & \xrightarrow{g_z} & \mathbb{R} \\ t & \longmapsto & g_z(t) = f(z_1 - t(z_1 - z_2), z_2 + t(z_1 - z_2), z_3, \ldots, z_m) \end{array}$ is a decreasing function.

The set of the strictly s-convex functions on H_1 is denoted by: $\mathcal{F}(H_1)$

Definition 7. An R-criteria is a numerical function F_A defined over $\Delta(H,T)$ by:

$$\begin{array}{ccc} \Delta(H,T) & \xrightarrow{F_A} & \mathbb{R} \\ u & \longmapsto & F_A(u) = f(q_A(u)) \end{array}$$

with: $f \in \mathcal{F}(C_m)$ and q_A the Rayleigh quotient associated to the T-symmetric operator A.

The next property, given by Alain Croquette[2], is very important because it gives the necessary and sufficient condition in order to get the maximum of an R-criterion.

Property 4.2. [2] *Let F_A be an R-criterion associated to the T-symmetric operator A defined by:*

$$\forall u \in \Delta(H,T) : \; F_A(u) = F_A(u_1, \ldots, u_m) = f(q_A(u_1), \ldots, q_A(u_m))$$

where: $f \in \mathcal{F}(C_m)$ and q_A the rayleigh quotient associated to the operator A. F_A admit a maximum over $\Delta(H,T)$ only and only for the T-orthogonal basis of H composed of the m eigen vectors of A associated to the m biggest eigen values of A.

Remark 4.1. *From the property 4.2 we notice that the maximum of an R-criterion F_A is independent of the form of the strictly s-convex function f, it depends only on the T-symmetric operator A.*

References

[1] R. Coppi. *An introduction to multiway data and their analysis*, Computational Statistics & Data Analysis, 1994. VOL. 18. 3-13.

[2] A. Croquette. *Quelques resultats synthétiques en analyse des données multidimentionnelles: Optimalité et metriques à effet relationnels*, Thése 3eme cycle, université PAUL SABATIER, TOULOUSE 1980.

[3] D. Grau, *Mesures des effets relationnels-Applications*, Thése 3eme cycle, université PAUL SABATIER, TOULOUSE 1983.

[4] I.T. Jolliffe, *Principal Component Analysis*, Springer 2002.

[5] C. Lavit, *Analyse conjointe de tableaux quantitatifs*, Edition MASSON 1988.

[6] Rao, cr, *The use and interpretation of principal component analysis in applied research*, SANKHAIA, Série A, Volume. 26.

[7] Yves Schektman, *Contribution à la mise en facteurs dans les sciences expérimentales et à la mise en oeuvre des calculs statistiques*, Thèse d'état, Université PAUL SABATIER, TOULOUSE 1978.

[8] F.O Tebboune, *Contribution aux analyses en composantes principales sous contraintes linéaires*, Thése 3eme cycle, université PAUL SABATIER, TOULOUSE 1981.

Received: October 20, 2007