Principal Component Analysis with external information on both subjects and variables

Analisi Statistica dei Dati Multidimensionali[1]

[1]Corso di Laurea in Scienze Statistiche e Attuariali

Facoltà di Scienze Economiche e Aziendali
Università degli Studi del Sannio

- $Im(\mathbf{X})$ denotes the subspace spanned by the column vectors of $\mathbf{X}$.
- Let $Ker(\mathbf{X}^T)$ be the kernel (null space) of $\mathbf{X}^T$.
- The data matrix $\mathbf{X}$ is often accompanied by external information concerning the rows and columns of data matrix.
- In order to consider several hypotheses and evaluate different effects, we consider several external information $\mathbf{H}$ and $\mathbf{Z}$ for the units and variables of the matrix $\mathbf{X}$.
- Let $\mathbf{H_X}$ be the external $n \times c$ row (units) information matrices on $\mathbf{X}$, respectively, with $c < n$.
- We denote a $p \times l$ column information matrix on $\mathbf{X}$ by $\mathbf{Z_X}$ with $l < p$.
- We assume that $\mathbf{Z_X}$ is also columnwise centered.

External information can assume several forms:

- if the rows of $\mathbf{X}$ represents subjects then we can collect the subject's demographic information in $\mathbf{H_X}$ and will try to explore how they are related to the set of variables of the main data;
- if we set $\mathbf{H_X}$ to $\mathbf{1}_n$ then we highlight the mean tendency across the subjects;
- we can consider for $\mathbf{H_X}$ matrix of dummy variables providing the subject's membership in some prespecified groups in order to analyze the differences among the groups;
- the external information $\mathbf{Z_X}$ for the variables collected in the matrix $\mathbf{X}$ can be unitary vectors with p component in order to capture the relationships among the columns of the matrix;

- if variables represent stimuli in a preference judgment study then $\mathbf{Z_X}$ can be a matrix of the descriptor variables of the stimuli;
- if we have different within-subjects experimental conditions then $\mathbf{Z_X}$ could be a matrix of contrasts; in this sense, orthogonal contrasts used in Analysis of Variance, by which different linear principal mean effects can be highlighted, are a particular case of external information;
- finally, if variables are repeated observations then $\mathbf{Z_X}$ could be a matrix of coefficients of orthogonal polynomials.

- Let $\mathbf{P}_{\mathbf{Z_X}/\mathbf{Q_X}} = \mathbf{Z_X}(\mathbf{Z_X}^T\mathbf{Q_X}\mathbf{Z_X})^{-}\mathbf{Z_X}^T\mathbf{Q_X}$ be the not symmetric oblique projection operator onto $Im(\mathbf{Z_X})$ along $Ker(\mathbf{Z_X}^T\mathbf{Q_X})$ (Takane and Shibayama, 1991).

- "External Analysis" of the Principal Component Analysis with external information on both subjects and variables (Takane and Shibayama, 1991; Takane and Hunter, 2001) considers the additive data decomposition of a single quantitative matrix $\mathbf{X}$ according to the row and column external information matrices $\mathbf{H_X}$ and $\mathbf{Z_X}$ under row and column metrics $\mathbf{L}$ and $\mathbf{Q_X}$, respectively:

$$\begin{aligned} \mathbf{X} = \quad & \mathbf{P}_{\mathbf{H_X}/\mathbf{L}}\mathbf{X}\mathbf{P}^{T}_{\mathbf{Z_X}/\mathbf{Q_X}} + \mathbf{P}^{\perp}_{\mathbf{H_X}/\mathbf{L}}\mathbf{X}\mathbf{P}^{T}_{\mathbf{Z_X}/\mathbf{Q_X}} + \\ & + \mathbf{P}_{\mathbf{H_X}/\mathbf{L}}\mathbf{X}\mathbf{P}^{\perp T}_{\mathbf{Z_X}/\mathbf{Q_X}} + \mathbf{P}^{\perp}_{\mathbf{H_X}/\mathbf{L}}\mathbf{X}\mathbf{P}^{\perp T}_{\mathbf{Z_X}/\mathbf{Q_X}} \end{aligned} \tag{1}$$

- The four submatrices in (1) are mutually trace-orthogonal in their respective metric matrices $\mathbf{L}$ and $\mathbf{Q_X}$ (Takane and Shibayama, 1991). For example, the first and the second terms in (1) are trace-orthogonal since we have simultaneously

$$trace[\underbrace{(\mathbf{P}_{\mathbf{Z_X}/\mathbf{Q_X}}\mathbf{X}^T\mathbf{P}^T_{\mathbf{H_X}/\mathbf{L}})\mathbf{L}}_{=\mathbf{A}^T}\underbrace{(\mathbf{P}^{\perp}_{\mathbf{H_X}/\mathbf{L}}\mathbf{X}\mathbf{P}^T_{\mathbf{Z_X}/\mathbf{Q_X}})\mathbf{Q_X}}_{=\mathbf{B}}] = 0$$

$$trace[\underbrace{\mathbf{L}(\mathbf{P}_{\mathbf{H_X}/\mathbf{L}}\mathbf{X}\mathbf{P}^T_{\mathbf{Z_X}/\mathbf{Q_X}})}_{=\mathbf{A}}\underbrace{\mathbf{Q_X}(\mathbf{P}_{\mathbf{Z_X}/\mathbf{Q_X}}\mathbf{X}^T\mathbf{P}^{\perp T}_{\mathbf{H_X}/\mathbf{L}})}_{=\mathbf{B}^T}] = 0$$

where two matrices of a same size, $\mathbf{A}$ and $\mathbf{B}$, are said to be trace-orthogonal when $tr(\mathbf{A}^T\mathbf{B}) = tr(\mathbf{A}\mathbf{B}^T) = 0$.

This property implies that the sum of squares of **X** can be exactly decomposed according to the sum of sums of squares corresponding to each submatrix in (1):

$$\begin{aligned} SS(\mathbf{X})_{\mathbf{L},\mathbf{Q_X}} = \; & SS(\mathbf{P}_{\mathbf{H_X}/\mathbf{L}}\mathbf{X}\mathbf{P}^{T}_{\mathbf{Z_X}/\mathbf{Q_X}})_{\mathbf{L},\mathbf{Q_X}} + SS(\mathbf{P}^{\perp}_{\mathbf{H_X}/\mathbf{L}}\mathbf{X}\mathbf{P}^{T}_{\mathbf{Z_X}/\mathbf{Q_X}})_{\mathbf{L},\mathbf{Q_X}} + \\ & + SS(\mathbf{P}_{\mathbf{H_X}/\mathbf{L}}\mathbf{X}\mathbf{P}^{\perp T}_{\mathbf{Z_X}/\mathbf{Q_X}})_{\mathbf{L},\mathbf{Q_X}} + SS(\mathbf{P}^{\perp}_{\mathbf{H_X}/\mathbf{L}}\mathbf{X}\mathbf{P}^{\perp T}_{\mathbf{Z_X}/\mathbf{Q_X}})_{\mathbf{L},\mathbf{Q_X}} \end{aligned}$$

where $SS(\mathbf{X})_{\mathbf{L},\mathbf{Q_X}} = tr(\mathbf{X}^T\mathbf{L}\mathbf{X}\mathbf{Q_X})$.

Each component in (1) has a specific statistical meaning:

- $\boxed{\mathbf{P}_{\mathbf{H_X}/\mathbf{L}}\mathbf{X}\mathbf{P}^{T}_{\mathbf{Z_X}/\mathbf{Q_X}}}$ represents the row and column constraints effects;
- $\boxed{\mathbf{P}^{\perp}_{\mathbf{H_X}/\mathbf{L}}\mathbf{X}\mathbf{P}^{T}_{\mathbf{Z_X}/\mathbf{Q_X}}}$ reflects the column constraints effect;
- $\boxed{\mathbf{P}_{\mathbf{H_X}/\mathbf{L}}\mathbf{X}\mathbf{P}^{\perp T}_{\mathbf{Z_X}/\mathbf{Q_X}}}$ reflects the row constraints effect
- the last term $\boxed{\mathbf{P}^{\perp}_{\mathbf{H_X}/\mathbf{L}}\mathbf{X}\mathbf{P}^{\perp T}_{\mathbf{Z_X}/\mathbf{Q_X}}}$ pertains to what can be explained by neither row nor column external information.

Row sides metric matrix $\mathbf{L}$ can assume several forms:

1. if there are no differences in importance among the statistical units then $\mathbf{L}$ can be set to the identity matrix;
2. if there are differences in importance among the statistical units then special diagonal matrices are used as $\mathbf{L}$ in order to differentially weight the rows of $\mathbf{X}$;
3. finally, when rows of the data matrix are time points in single-subject multivariate time series data, Escoufier (1987) suggests to use the inverse of the matrix of serial correlations as $\mathbf{L}$.

- By using this approach, external information $\mathbf{H_X}$ and $\mathbf{Z_X}$ are incorporated within a single data set.
- It is evident that the Takane and Shibayama's basic decomposition (1991) is obtained from (1) by setting $\mathbf{L} = \mathbf{I}_n$ and $\mathbf{Q_X} = \mathbf{I}_p$:

$$\boxed{\mathbf{X} = \mathbf{P_{H_X}} \mathbf{X} \mathbf{P_{Z_X}} + \mathbf{P^{\perp}_{H_X}} \mathbf{X} \mathbf{P_{Z_X}} + \mathbf{P_{H_X}} \mathbf{X} \mathbf{P^{\perp}_{Z_X}} + \mathbf{P^{\perp}_{H_X}} \mathbf{X} \mathbf{P^{\perp}_{Z_X}}}$$

- An "Internal analysis" (PCA) is then performed on each component of decomposition.

- Consider the following model

$$\mathbf{X} = \mathbf{H_X} \mathbf{M} \mathbf{Z_X}^T + \mathbf{B} \mathbf{Z_X}^T + \mathbf{H_X} \mathbf{C} + \mathbf{E} \qquad (2)$$

where **M** $(c \times l)$, **B** $(n \times l)$, and **C** $(c \times p)$ are matrices of coefficients to be estimated, and **E** $(n \times p)$ a matrix of error components.

- The four terms in (2) explain portions of the original data matrix, **X**.
- The first term pertains to what can be explained by both $\mathbf{H_X}$ and $\mathbf{Z_X}$, the second term by $\mathbf{Z_X}$, the third term by $\mathbf{H_X}$, and the fourth term by neither $\mathbf{H_X}$ nor $\mathbf{Z_X}$.

- Let $\mathbf{X} = \mathbf{H_X}\mathbf{M}\mathbf{Z_X}^T + \mathbf{E}_1$, and consider the problem of estimating $\mathbf{M}$ so as to minimize $SS(\mathbf{E}_1) = tr(\mathbf{E}_1^T\mathbf{E}_1)$.
- We obtain

$$\hat{\mathbf{M}} = (\mathbf{H_X}^T\mathbf{H_X})^-\mathbf{H_X}^T\mathbf{X}\mathbf{Z_X}(\mathbf{Z_X}^T\mathbf{Z_X})^-$$

  where $(\mathbf{H_X}^T\mathbf{H_X})^-$ and $(\mathbf{Z_X}^T\mathbf{Z_X})^-$ are g-inverses of $(\mathbf{H_X}^T\mathbf{H_X})$ and $(\mathbf{Z_X}^T\mathbf{Z_X})$, respectively.
- The residual from the first term is now equal to

$$\hat{\mathbf{E}}_1 = \mathbf{X} - \mathbf{H_X}\hat{\mathbf{M}}\mathbf{Z_X}^T = \mathbf{X} - \mathbf{P}_{\mathbf{H_X}}\mathbf{X}\mathbf{P}_{\mathbf{Z_X}}$$

  where $\mathbf{P}_{\mathbf{H_X}}$ and $\mathbf{P}_{\mathbf{Z_X}}$ are orthogonal projection operators onto spaces spanned by the column vectors of $\mathbf{H_X}$ and $\mathbf{Z_X}$, respectively.

- We now separately fit the second and the third terms to $\hat{\mathbf{E}}_1$:

$$\hat{\mathbf{E}}_1 = \mathbf{B}\mathbf{Z}_{\mathbf{X}}^T + \mathbf{E}_2$$
$$\hat{\mathbf{E}}_1 = \mathbf{H}_{\mathbf{X}}\mathbf{C} + \mathbf{E}_3$$

- We obtain a least squares estimate of $\mathbf{B}$ that minimizes $SS(\mathbf{E}_2)$ by

$$\hat{\mathbf{B}} = \mathbf{P}_{\mathbf{H}_{\mathbf{X}}}^{\perp}\mathbf{X}\mathbf{Z}_{\mathbf{X}}(\mathbf{Z}_{\mathbf{X}}^T\mathbf{Z}_{\mathbf{X}})^{-}$$

where $\mathbf{P}_{\mathbf{H}_{\mathbf{X}}}^{\perp} = (\mathbf{I} - \mathbf{P}_{\mathbf{H}_{\mathbf{X}}})$ orthogonal projection operator to $\mathbf{P}_{\mathbf{H}_{\mathbf{X}}}$ (that is $\mathbf{P}_{\mathbf{H}_{\mathbf{X}}}^{\perp}\mathbf{P}_{\mathbf{H}_{\mathbf{X}}} = \mathbf{0}$ and $\mathbf{P}_{\mathbf{H}_{\mathbf{X}}} + \mathbf{P}_{\mathbf{H}_{\mathbf{X}}}^{\perp} = \mathbf{I}$).
- Similarly, we obtain

$$\hat{\mathbf{C}} = (\mathbf{H}_{\mathbf{X}}^T\mathbf{H}_{\mathbf{X}})^{-}\mathbf{H}_{\mathbf{X}}^T\mathbf{X}\mathbf{P}_{\mathbf{Z}_{\mathbf{X}}}^{\perp}$$

that minimizes $SS(\mathbf{E}_3)$

- Now, the estimate of the fourth term is given by

$$\begin{aligned}\hat{\mathbf{E}} =& \mathbf{H_X}\hat{\mathbf{M}}\mathbf{Z}_{\mathbf{X}}^{T} + \hat{\mathbf{B}}\mathbf{Z}_{\mathbf{X}}^{T} + \mathbf{H_X}\hat{\mathbf{C}} \\ =& \mathbf{X} - \mathbf{P}_{\mathbf{H_X}}\mathbf{X}\mathbf{P}_{\mathbf{Z_X}} - \mathbf{P}_{\mathbf{H_X}}^{\perp}\mathbf{X}\mathbf{P}_{\mathbf{Z_X}} - \mathbf{P}_{\mathbf{H_X}}\mathbf{X}\mathbf{P}_{\mathbf{Z_X}}^{\perp} \\ =& \mathbf{P}_{\mathbf{H_X}}^{\perp}\mathbf{X}\mathbf{P}_{\mathbf{Z_X}}^{\perp}\end{aligned}$$

- By substituting the least squares estimates for the corresponding parameters, we obtain the following decomposition of the data matrix, **X**:

$$\begin{aligned}\mathbf{X} =& (\mathbf{P}_{\mathbf{H_X}} + \mathbf{P}_{\mathbf{H_X}}^{\perp})\mathbf{X}(\mathbf{P}_{\mathbf{Z_X}} + \mathbf{P}_{\mathbf{Z_X}}^{\perp}) \\ =& \mathbf{P}_{\mathbf{H_X}}\mathbf{X}\mathbf{P}_{\mathbf{Z_X}} + \mathbf{P}_{\mathbf{H_X}}^{\perp}\mathbf{X}\mathbf{P}_{\mathbf{Z_X}} + \mathbf{P}_{\mathbf{H_X}}\mathbf{X}\mathbf{P}_{\mathbf{Z_X}}^{\perp} + \mathbf{P}_{\mathbf{H_X}}^{\perp}\mathbf{X}\mathbf{P}_{\mathbf{Z_X}}^{\perp}\end{aligned}$$