PCA with Missing Data

Yoshio Takane
University of Victoria/McGill University

Symposium in Osaka, September 2013

Summary

Missing data arise in virtually all data analysis situations, and how to deal with them is one of the most important concerns for every data analyst. There are at least two conventional approaches to missing data in PCA. One is based on homogeneity analysis (HA), and the other on weighted low rank approximations (WLRA). We review some properties of these two approaches, emphasizing their similarities and differences, and suggest some extensions. PCA with missing data is also important as a preprocessing step to ICA (whitening) when missing data exist.

Frequently Used Projectors

- Orthogonal projectors:

$$\mathbf{P}_Z = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-}\mathbf{Z}' \quad \text{and} \quad \mathbf{Q}_Z = \mathbf{I} - \mathbf{P}_Z.$$

- Oblique projectors: Let $\mathbf{W}$ be an *nnd* matrix such that $\text{rank}(\mathbf{WZ}) = \text{rank}(\mathbf{Z})$.

$$\mathbf{P}_{Z/W} = \mathbf{Z}(\mathbf{Z}'\mathbf{WZ})^{-}\mathbf{Z}'\mathbf{W}$$

and

$$\mathbf{Q}_{Z/W} = \mathbf{I} - \mathbf{P}_{Z/W}.$$

Some Notations

- $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_p]$ $(n \times p)$: The raw data matrix
- $\mathbf{1}_n$ $(n \times 1)$: The vector of ones
- $\mathbf{D}_{w_j}$ $(n \times n)$: A diagonal matrix such that its ith diagonal element is 1 if the ith element of $\mathbf{x}_j$ is observed, and 0 otherwise
- $\mathbf{F}$ $(n \times r)$: The matrix of component scores
- $\mathbf{u}_j'$ $(1 \times r)$ and $\mathbf{u}_{0j}'$ $(1 \times r)$: Vectors of weights applied to $\mathbf{x}_j$ and $\mathbf{1}_n$
- $\mathbf{A}' = [\mathbf{a}_1, \cdots, \mathbf{a}_p]$ $(r \times p)$: The matrix of component loadings
- $\mathbf{x}_j^*$: The jth data vector centered with respect to non-missing observations

Homogeneity Analysis

- The HA Approach: Minimize

$$\phi = \sum_{j=1}^{p} \mathrm{SS}(\mathbf{F} - \mathbf{x}_j\mathbf{u}_j' - \mathbf{1}_n\mathbf{u}_{0j}')_{D_{w_j}},$$

  wrt to $\mathbf{F}$, $\mathbf{u}_j'$, and $\mathbf{u}_{0j}'$ $(j = 1, \cdots, p)$, where $\mathrm{SS}(\mathbf{Z})_W = \mathrm{tr}(\mathbf{Z}'\mathbf{W}\mathbf{Z})$.
- Originally proposed to deal with missing data in multiple correspondence analysis (Meulman, 1982). (Meulman actually used the centered data $\mathbf{x}_j^*$ and set $\mathbf{u}_{0j}' = \mathbf{0}_r'$.)
- The ϕ is the criterion commonly used in multiple-set canonical correlation analysis (when each of the p data sets consists of more than one variable). The meet loss (Gifi, 1990).

Weighted Low Rank Approximations

- The WLRA Approach: Minimize

$$\tau = \sum_{j=1}^{p} \mathrm{SS}(\mathbf{x}_j^* - \mathbf{F}\mathbf{a}_j')_{D_{w_j}},$$

  with respect to $\mathbf{F}$ and $\mathbf{a}_j$.
- Originally proposed by Gabriel and Zamir (1979).
- We could have the raw data vector $\mathbf{x}_j$ and the constant term $-\mathbf{1}_n m_j$ in the minimization criterion. (This was never done before.)
- The τ is the criterion used in PCA. The join loss (Gifi, 1990).

The Two Criteria

- The two criteria were first(?) introduced by Meredith (1964).
- They are simply related (one can be turned into the other) when there are no missing data (i.e., $\mathbf{D}_{w_j} = \mathbf{I}_n$ for all j). In general, however, they are not simply related (Gifi, 1990).
- The minimization of ϕ leads to closed-form solutions, while that of τ iterative solutions.

Two Alternative Solutions for HA

- Two alternative solutions (Takane & Oshima-Takane, 2003):
  (1) The Missing-Data Passive (MDP) Method
  (2) The Test Equating (TE) Method.
- The MDP Method: Developed in the context of multiple correspondence analysis (Gifi, 1990) with missing data.
- The TE Method: Developed in the context of test equating (Shibayama, 1988). University entrance examinations create incomplete data because not all applicants take the same examinations.

The MDP Method (1)

- $\mathbf{J}_n = \mathbf{1}_p \otimes \mathbf{I}_n$ $(np \times n)$
- $\mathbf{D}_1 = \mathbf{I}_p \otimes \mathbf{1}_n$ $(np \times p)$
- $\mathbf{D}_X$ $(np \times p)$: The block diagonal matrix with $\mathbf{x}_j$ as the jth diagonal block
- $\mathbf{D}_w$ $(np \times np)$: The diagonal matrix with $\mathbf{D}_{w_j}$ as the jth diagonal block
- $\mathbf{U}' = [\mathbf{u}_1, \cdots, \mathbf{u}_p]$ $(r \times p)$, and $\mathbf{U}'_0 = [\mathbf{u}_{01}, \cdots, \mathbf{u}_{0p}]$ $(r \times p)$
- We rewrite the HA criterion as follows (without using a summation mark):

$$\phi = \mathrm{SS}(\mathbf{J}_n\mathbf{F} - \mathbf{D}_X\mathbf{U} - \mathbf{D}_1\mathbf{U}_0)_{D_w}.$$

The MDP Method (2)

- The MDP method minimizes ϕ in the order of

$$\min_{F} \min_{U|F} \min_{U_0|F,U} \phi$$

subject to the restrictions that $\mathbf{1}_n'\mathbf{F} = \mathbf{0}_r'$ and $\mathbf{F}'\mathbf{Q}\mathbf{F} = \mathbf{I}_r$ (Note 1), where $\mathbf{Q} = \sum_{j=1}^{p} \mathbf{D}_{w_j}\mathbf{Q}_{1_n/D_{w_j}}$.

- This leads to the generalized eigen-equation

$$(\mathbf{J}_n'\tilde{\mathbf{Q}}\mathbf{P}_{D_X/\tilde{Q}}\mathbf{J}_n)\mathbf{F} = (\mathbf{J}_n'\tilde{\mathbf{Q}}\mathbf{J}_n)\mathbf{F}\Delta^2$$

to be solved for $\mathbf{F}$, where $\tilde{\mathbf{Q}} = \mathbf{D}_w\mathbf{Q}_{D_1/D_w}$ (Note 0).

The TE Method (1)

- The TE method minimizes the same criterion, but in different order. This leads to different restrictions under which ϕ is minimized and different parameter estimates.
- We rewrite the HA criterion using different symbols for parameter matrices:

$$\psi = \mathrm{SS}(\mathbf{J}_n\mathbf{G} - \mathbf{D}_X\mathbf{V} - \mathbf{D}_1\mathbf{V}_0)_{D_w}.$$

The TE Method (2)

- The TE method minimizes ψ in the order of

$$\min_{V} \min_{V_0|V} \min_{G|V,V_0} \psi$$

  subject to $\mathbf{1}_p'\mathbf{V}_0 = \mathbf{0}_r'$ and $\mathbf{V}'\mathbf{S}\mathbf{V} = \mathbf{I}_r$ (Note 1), where $\mathbf{S} = \mathbf{D}_X'\tilde{\mathbf{Q}}\mathbf{D}_X$.
- This leads to the generalized eigen-equation

$$\mathbf{D}_X'(\tilde{\mathbf{Q}} - \mathbf{Q}^* + \mathbf{Q}^*\mathbf{P}_{D_1/Q^*})\mathbf{D}_X\mathbf{V} = \mathbf{S}\mathbf{V}\Delta^2$$

  to be solved for $\mathbf{V}$, where $\mathbf{Q}^* = \mathbf{D}_w\mathbf{Q}_{J_n/D_w}$ (Note 0).

The Relationship between the Two (1)

- $\mathbf{X}^* = [\mathbf{Q}_{1_n/D_{w_1}}\mathbf{x}_1, \cdots, \mathbf{Q}_{1_n/D_{w_p}}\mathbf{x}_p]$
- $\mathbf{Q} = \sum_{j=1}^{p} \mathbf{D}_{w_j}\mathbf{Q}_{1_n/D_{w_j}} = \mathbf{J}_n'\tilde{\mathbf{Q}}\mathbf{J}_n$.
- The generalized eigen-equations solved in the two methods can be rewritten as

$$\mathbf{X}^*\mathbf{S}^{-1}\mathbf{X}^{*'}\mathbf{F} = \mathbf{Q}\mathbf{F}\Delta^2,$$

and

$$\mathbf{X}^{*'}\mathbf{Q}^{+}\mathbf{X}^*\mathbf{V} = \mathbf{S}\mathbf{V}\Delta^2.$$

- Note for the latter we used $\mathbf{D}_X'(\tilde{\mathbf{Q}} - \mathbf{Q}^*\mathbf{Q}_{D_1/Q^*})\mathbf{D}_X = \mathbf{D}_X'(\tilde{\mathbf{Q}}\mathbf{P}_{J_n/\tilde{Q}})\mathbf{D}_X = \mathbf{X}^{*'}\mathbf{Q}^{+}\mathbf{X}^*$ (Note 2).

The Relationship between the Two (2)

- These two eigen-equations are simply related. They are both simply related to GSVD$(\mathbf{Q}^{+}\mathbf{X}^{*}\mathbf{S}^{-1})_{Q,S}$ denoted as $\mathbf{Q}^{+}\mathbf{X}^{*}\mathbf{S}^{-1} = \mathbf{F}\Delta\mathbf{V}'$.
- More specific relationships between parameters:
  $\mathbf{F} = \mathbf{Q}_{1_n}\mathbf{G}\Delta^{-1}$.
  $\mathbf{U} = \mathbf{V}\Delta$ $(\mathbf{V} = \mathbf{U}\Delta^{-1})$.
  The relationship between $\mathbf{U}_0$ and $\mathbf{V}_0$ is rather complicated in general. When there are no missing data, we have
  $\mathbf{U}_0 = (\mathbf{V}_0 - \mathbf{1}_p\mathbf{1}_n'\mathbf{G}/n)\Delta$, or $\mathbf{G} = \mathbf{F}\Delta - \mathbf{1}_n\mathbf{1}_p'\mathbf{U}_0\Delta^{-1}$.

WLRA

- The WLRA method allows very flexible weighting schemes in low-rank approximations to data matrices.
- Let $\mathbf{x}^* = \text{vec}(\mathbf{X}^*)$, and $\mathbf{x}_0 = \text{vec}(\mathbf{FA}')$. Consider minimizing

$$\tau = (\mathbf{x}^* - \mathbf{x}_0)'\mathbf{W}^*(\mathbf{x}^* - \mathbf{x}_0),$$

  where $\mathbf{W}^*$ is an *np* by *np nnd* weight matrix.
- This minimization cannot be solved in closed form except for special cases in which $\mathbf{W}^* = \mathbf{L} \otimes \mathbf{K}$, where $\mathbf{L}$ is a *p* by *p nnd* matrix and $\mathbf{K}$ is an *n* by *n nnd* matrix, in which case the problem reduces to GSVD$(\mathbf{X}^*)_{K,L}$.

General Solutions

- $\mathbf{x}_0$ can be rewritten in two alternative ways: $\mathbf{x}_0 = (\mathbf{A} \times \mathbf{I}_n)\text{vec}(\mathbf{F}) = (\mathbf{I}_p \otimes \mathbf{F})\text{vec}(\mathbf{A}')$, which form a basis for an iterative updating of $\mathbf{f} = \text{vec}(\mathbf{F})$ and $\mathbf{a} = \text{vec}(\mathbf{A}')$.
- Let $\mathbf{A}^* = \mathbf{A} \otimes \mathbf{I}_n$. Then $\mathbf{f}$ can be updated by $\mathbf{f} = (\mathbf{A}^{*'}\mathbf{W}^*\mathbf{A}^*)^{-1}\mathbf{A}^{*'}\mathbf{W}^*\mathbf{x}^*$ for fixed $\mathbf{A}$.
- Let $\mathbf{F}^* = \mathbf{I}_p \otimes \mathbf{F}$. Then $\mathbf{a}$ can be updated by $\mathbf{a} = (\mathbf{F}^{*'}\mathbf{W}^*\mathbf{F}^*)^{-1}\mathbf{F}^{*'}\mathbf{W}^*\mathbf{x}^*$ for fixed $\mathbf{F}$.
- An ALS (alternating least squares) algorithm, which is monotonically convergent.

Simplification (1)

- When $\mathbf{W}^*$ is diagonal, the algorithm can be simplified considerably.
- The τ can be rewritten as

$$\tau = \sum_{j=1}^{p} \mathrm{SS}(\mathbf{x}_j^* - \mathbf{F}\mathbf{a}_j)_{D_{w_j}}.$$

The $\mathbf{a}_j$ can be separately updated.

Simplification (2)

- $\mathbf{x}_i^{*'}$: The ith row vector of $\mathbf{X}^*$
- $\mathbf{f}_i'$: The ith row vector of $\mathbf{F}$
- $\mathbf{D}_{w_i'}$: The diagonal weight matrix whose jth diagonal element is 1 if the jth element of $\mathbf{x}_i^{*'}$ is observed, and 0 otherwise.
- Then, τ can be rewritten in another way as

$$\tau = \sum_{i=1}^{n} \mathrm{SS}(\mathbf{x}_i^{*'} - \mathbf{f}_i'\mathbf{A}')_{D_{w_i'}},$$

and $\mathbf{f}_i'$ can be separately updated.

Including the Constant Term

- It is possible to include a constant term as in HA. However, the algorithm becomes a bit unwieldy.
- Suppose the data are not centered. This requires the constant term $-\mathbf{1}_n m_j$ to be included in the criterion. We minimize $\tau^* = \sum_{j=1}^{p} \mathrm{SS}(\mathbf{x}_j - \mathbf{F}\mathbf{a}_j - \mathbf{1}_n m_j)_{D_{w_j}}$ with respect to m_j and $\mathbf{a}_j$, which leads to $m_j = (\mathbf{1}_n \mathbf{D}_{w_j} \mathbf{1}_n)^{-1} \mathbf{1}_n' \mathbf{D}_{w_j} (\mathbf{x}_j - \mathbf{F}\mathbf{a}_j)$, and $\mathbf{a}_j = (\mathbf{F}'\mathbf{Q}_j\mathbf{F})^{-1}\mathbf{F}'\mathbf{Q}_j\mathbf{x}_j$, where $\mathbf{Q}_j = \mathbf{D}_{w_j}\mathbf{Q}_{1/D_{w_j}}$. This matrix is not diagonal. The τ^* is no longer separable with respect to i. A full version of the algorithm is necessary.

Concluding Remarks

- I tend to favor the HA approach:
  1. Closed-form solutions
  2. Nested solutions
  3. No need to prescribe the number of components
- Monte Carlo experiments are necessary to compare the two approaches systematically.

References

- Gabriel, K. R., and Zamir, S. (1979). *Technometrics*, **21**, 489–498.
- Gifi, A. (1990). *Nonlinear multivariate analysis*. Chichester: Wiley.
- Meredith, W. (1964). *Psychometrika*, **29**, 187–206 .
- Meulman, J. (1982). *Homogeneity analysis of incomplete data*. Leiden: DSWO Press.
- Shibayama, T. (1988). Unpublished Doctoral Dissertation. University of Tokyo, (in Japanese).
- Takane, Y., and Oshima-Takane, Y. (2003). *Behaviormetrika*, **30**, 145–154.
- Takane, Y., and Zhou, L. (2011). *Linear Algebra and Its Applications*, **436**, 2567–2577.

Note 0

- The $\mathbf{U}_0$ and $\mathbf{U}$ in the MDP method are given by:

$$\mathbf{U}_0 = (\mathbf{D}_1'\mathbf{D}_w\mathbf{D}_1)^{-1}\mathbf{D}_1'\mathbf{D}_w(\mathbf{J}_n\mathbf{F} - \mathbf{D}_X\mathbf{U}),$$

and

$$\mathbf{U} = (\mathbf{D}_X'\tilde{Q}\mathbf{D}_X)^{-1}\mathbf{D}_X'\tilde{Q}\mathbf{J}_n\mathbf{F}.$$

- The $\mathbf{G} and \mathbf{V}_0$ in the TE method are given by:

$$\mathbf{G} = (\mathbf{J}_n'\mathbf{D}_w\mathbf{J}_n)^{-1}\mathbf{J}_n'\mathbf{D}_w(\mathbf{D}_X\mathbf{V} + \mathbf{D}_1\mathbf{V}_0),$$

nd

$$\mathbf{V}_0 = -(\mathbf{D}_1'\mathbf{Q}^*\mathbf{D}_1)^+\mathbf{D}_1'\mathbf{Q}^*\mathbf{D}_X\mathbf{V},$$

where $^+$ indicates a Moore-Penrose inverse.

Notes 1 & 2

- Note 1. These restriction are necessary, although arbitrary in form, because $\mathbf{J}_n$ and $\mathbf{D}_1$ are not disjoint, and the homogeneity criterion can be trivially made zero by setting all parameter matrices to zero matrices.
- Note 2. Takane and Zhou (2012; Lemma 3). Let $\mathbf{Z} = [\mathbf{M}, \mathbf{N}]$. Then,

$$\mathbf{Q}_{[M,N]/K} = \mathbf{Q}_{M/K}\mathbf{Q}_{N/KQ_{M/K}} = \mathbf{Q}_{N/K}\mathbf{Q}_{M/KQ_{N/K}}.$$