Title	Matrix decomposition approaches to nonmetric multivariate analysis for dimension reduction
Author(s)	牧野, 直道
Citation	大阪大学, 2020, 博士論文
Version Type	VoR
URL	https://doi.org/10.18910/77470
rights	
Note	

Matrix decomposition approaches to nonmetric multivariate analysis for dimension reduction
行列分解法による次元縮約のための非計量多変量解析

Behavioral Sciences, Behavioral Statistics
Naomichi MAKINO
行動生態学講座　行動統計科学研究室
牧野直道
21A16822

2020, May, 14

要旨

主成分分析と探索的因子分析は, 量的変数間の関連性の可視化や観測された個体の特徴を得点として解釈をするための統計手法として, 一般的に用いられる統計手法である. 主成分分析と同様に, 探索的因子分析の解を特異値分解を用いて推定できることが近年の研究によって明らかになっている. 本論文では, このように特異値分解を用いて解を推定する方法を行列分解法と呼ぶ. 行列分解法により, 探索的因子分析と主成分分析が統一的に表現でき, 数理的観点から, 探索的因子分析と主成分分析を比較することができるという利点がある.

本論文の主題は 2 つから構成される. 第一の主題は, 数量化法を用いて, 行列分解法による主成分分析と探索的因子分析を非計量データにも適用できるように拡張することである. 非計量主成分分析と非計量探索的因子分析の推定法の開発や数理的観点からの両手法の比較を行った.

第二の主題は, 行列分解法を用いた非計量主成分分析と非計量探索的因子分析において, 解釈容易な解を得るための手法の開発である. 本論文では解釈容易な解を得る手法を 2 種類に整理した. 1 つ目は変数と因子/主成分の関係性を解釈容易にする手法, 2 つ目は個体の特徴を得点を用いた解釈を容易にする手法である. 前者の手法として回転法が一般的に用いられるが, 近年発展した前者の手法として, スパース法がある. 本研究を通じて, 回転法・スパース法が非計量主成分分析と非計量探索的因子分析にも拡張できることを明らかにした. 後者の目的を達成する手法として, 非計量主成分分析と非計量探索的因子分析による次元縮約と個体得点のクラスタリングを同時に行う手法があり, 本研究では, 非計量データにおける次元縮約と個体得点のクラスタリングを同時に行う手法の開発とともに, この手法に回転法・スパース法を適用できることを示した.

実データによる適用例を通じて, 開発手法の類似点・相違点の検討や提案した解釈容易な解を得る手法の有効性について, 検討を行った. 最後に, 数理的性質の検討や実例を用いた比較を通じて明らかにした内容について検討し, 今後の研究課題について議論を行った.

Summary

Principal component analysis (PCA) and exploratory factor analysis (EFA) are widely-used statistical techniques to depict the relationships between quantitative variables and interpret objects charactalistics with their scores. It is well-known that a PCA solution can be given by the singular value decomposition (SVD). Recently, it has been shown that an EFA solution can also be obtained via the SVD. It implies that both EFA and PCA can be treated as the unified framework via the SVD, which allows EFA to be easily comparable with PCA from mathematical viewpoint. In this thesis, the estimation approaches using the SVD are namad as matrix decomposition approaches.

The main theme of the thesis consists of two parts. First aim of the thesis is the extension of PCA and EFA based on the matrix decomposition approaches to handle nonmetric multivariate data analysis using the quantification technique. Nonmetric verision of PCA and EFA are compared with each other from mathematical viewpoint. The estimation algorithms are also developed.

Second aim of the thesis is to develop the methods to enhance the interpretability of nonmetric PCA and EFA solutions. These methods are classified into two types: the method for obtaining interpretable structures between variables and components/factors, and the method to facilitate an interpretation of objects charactalistics with their scores. Rotation is a popular technique to find interpretable structures between variables and components/factors. There is another recently-developted technique to obtain an interpretable solution, which is called as sparse technique. It is shown that both rotation and sparse techniques can be extended for nonmetric multivariate data analysis. To attain the goal for facilitating the interpretation of object scores, clustering is combined with nonmetric version of PCA and EFA estimations. Joint PCA/EFA estimations and clustering methods are developed to deal with nonmetric data.

The benefits of the proposed methods in this thesis are illustrated using two real data examples. We conclude the thesis with a discussion of practical implications of our findings and with some suggestions for future work.

Contents

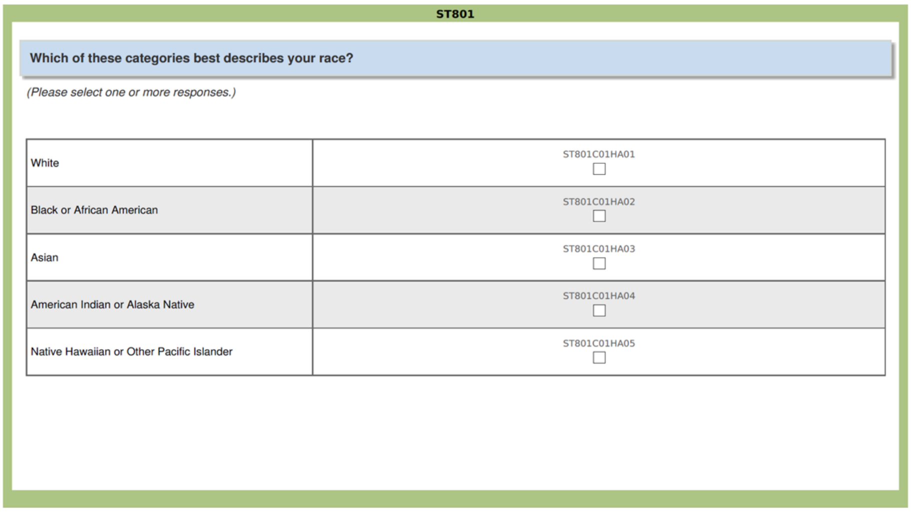
1

Introduction

Human science aims to expand our understanding of human being through diverse academic disciplines such as psychology, education, sociology, philosophy, anthropology, biology, physiology and/or behavioral science. Research objects in human science are often latent, and thus reseachers have developed various methods such as questionnaire, social survey, interview, psychological experiment and educational assessment to make them obvious.

Principal component analysis (PCA) and exploratory factor analysis (EFA) are widely-used statistical techniques for depicting hidden structures of quantitative observed data. From statistical viewpoint, the formulations of PCA and EFA can be mainly classified into two types. One is a probablisitic formulation in which components and factor scores are treated as latent random variables (e.g. Tipping and Bishop, 1999; Mulaik, 2010). The other is a descriptive formulation in which all parameters are regarded as fixed parameters. Recently, estimation techniques have been deveploed based on singular value decomposition (SVD), which is kinds of matrix decomposition technique. The SVD-based estimation techniques are named as matrix decomposition approaches in this thesis. The approaches have some theoritical advantages: PCA and EFA can be treated as the unified framework and their differences are easily discussed.

One of the limitations of PCA and EFA via the SVD-based matrix decomposition approaches is that only quantitative variables can be handled in PCA and EFA although qualitative variables are often encountered in human science researches. Here are examples such cases: in a sociological approach, researchers sometimes obtain categorical information about the religious affiliation, social stratification, nationality or race of their respondents via social survey or interview (Figure 1.1). When a psychological approach is adopted, categorical information about participants's mental state are observed. Another expamle in psychological approach is provided in psychometrics and educational measurements, which are the assigning of numerals to traits such as achievement, interest, attitudes, aptitudes, intelligence and performance. In this case, observed variables are often binary: 0 is obtained when incorrect response is observed and 1 is assigned to correct response (Table 1.1). The aim of the thesis is extension of matrix decomposition approaches for nonmetric multivariate analysis using the quantification technique to improve usability of study in human sciences. The quantification is the technique that qualitative variables are transformed so that quantitative data analysis methods can be applied. In the quantification, the parameters can be obtained by iterating the model parameter estimation and optimal scaling of qualitative variables. The quantification technique has the advantage: if a least squares procedure is known for analyzing quantitative data, it can be extended to qualitative data and the resulting algorithm will be convergent (Young, 1981). The matrix decomposition approaches are kinds of least square estimation procedures and thus the quantification is suitable method for extending to nonmetric data.



ST801

Which of these categories best describes your race?

(Please select one or more responses.)

White	ST801C01HA01 ☐
Black or African American	ST801C01HA02 ☐
Asian	ST801C01HA03 ☐
American Indian or Alaska Native	ST801C01HA04 ☐
Native Hawaiian or Other Pacific Islander	ST801C01HA05 ☐

Figure 1.1: An example of social survey (PISA 2018 national questionnaires in United states (OECD , 2019))

The techniques to facillitate the obtained solutions are also developed bacause it is desirable for practitioners to interpret the solutions easily. In this thesis, two types of interpretability of the solutions are discussed: one is the interpretability of relationships between variables and components/factors, and the other is the interpretability of objects scores.

The thesis is organized as follows. In chapter 2, some mathematical theories, PCA and EFA models are briefy introduced. The results introduced in chapter 2 are basis of the following

Table 1.1: An example of educational assessment for mathematics (Artificial data)

	Math item 1	Math item 2	Math item 3	Math item 4	$\cdots$	Math item p
Participant 1	1	0	1	1	$\cdots$	0
Participant 2	1	0	1	0	$\cdots$	1
Participant 3	1	1	0	0	$\cdots$	0
Participant 4	0	1	0	1	$\cdots$	0
Participant 5	1	0	0	1	$\cdots$	1
Participant 6	0	1	0	0	$\cdots$	0
Participant 7	0	0	1	0	$\cdots$	1
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
Participant n	0	1	1	0	$\cdots$	1

chapters. Interpretability between variables and components/factors is discussed in chapter 3. In this chapter, the methods for obtaining interpretable structures are also introduced. In chapter 4, the method for enhancing interpretability of components/factor scores is studied. Then, nonmetric versions of PCA and EFA are discussed in chapter 5 and 6, respectively. In these chapters, PCA and EFA based on matrix decomposition approaches and related methods to enhance the interpretability of the solutions are extended to handle nonmetric datasets. In chapter 7, real data examples are presented in order to demonstrate the benefits of the propopsed methods. In the last chapter, we conclude the thesis with a discussion of our findings and with some suggestions for future work.

It should be noted that this thesis is based on papers by Makino (2015) and Mori, Kuroda and Makino (2016) and further develop them.

2

Matrix decomposition approaches to multivariate analysis

Matrix decomposition apporaches are based on singular value decomposition (SVD) and ten Berge's theorem. In this chapter, these are briefly introduced before discussing nonmetric multivariate analysis.

2.1 Singular value decomposition

Let $\mathbf{X}$ be an arbitrary $n \times p$ matrix with $rank(\mathbf{X}) = r$. Then, any matrix $\mathbf{X}$ can be decomposed as

$$\mathbf{X} = \mathbf{K}\mathbf{\Delta}\mathbf{L}'. \tag{2.1}$$

Here, $\mathbf{K}$ $(n \times r)$ and $\mathbf{L}$ $(p \times r)$ are matrices that satisfy

$$\mathbf{K}'\mathbf{K} = \mathbf{L}'\mathbf{L} = \mathbf{I}_r \tag{2.2}$$

and $\mathbf{\Delta}$ is a $r \times r$ diagonal matrix whose elements are positive and arranged in decreasing order:

$$\delta_1 \geq \delta_2 \geq \cdots \geq \delta_r > 0. \tag{2.3}$$

This decomposition is called singular value decomposition (SVD). In the SVD, the columns of $\mathbf{K}$ and $\mathbf{L}$ are named as left singular vectors and right singular vectors respectively, and the diagonal elements of $\mathbf{\Delta}$ as singular values.

Assume $n \leq p \leq r = rank(\mathbf{X})$. Then, the SVD of $\mathbf{X}$ can be expressed as

$$\mathbf{X} = \tilde{\mathbf{K}}\tilde{\mathbf{\Delta}}\tilde{\mathbf{L}}'. \tag{2.4}$$

Here, $\tilde{\mathbf{K}}$ $(n \times p)$, $\tilde{\mathbf{\Delta}}$ $(p \times p)$, and $\tilde{\mathbf{L}}$ $(p \times p)$ are defined as

$$\tilde{\mathbf{K}} = [\mathbf{K}, \mathbf{K}_+], \tag{2.5}$$

$$\tilde{\mathbf{L}} = [\mathbf{L}, \mathbf{L}_+], \tag{2.6}$$

$$\tilde{\mathbf{\Delta}} = \begin{bmatrix} \Delta & {}_r\mathbf{O}_{p-r} \\ {}_{p-r}\mathbf{O}_r & {}_{p-r}\mathbf{O}_{p-r} \end{bmatrix}, \tag{2.7}$$

where $\mathbf{K}_+$ and $\mathbf{L}_+$ are arbitrary matrices that satisfy $\mathbf{K}'_+\mathbf{K}_+ = \mathbf{L}'_+\mathbf{L}_+ = \mathbf{I}_{p-r}$ and $\mathbf{K}'\mathbf{K}_+ = \mathbf{L}'\mathbf{L}_+ = {}_r\mathbf{O}_{p-r}$. Consequently, $\tilde{\mathbf{K}}$ and $\tilde{\mathbf{L}}$ satisfy $\tilde{\mathbf{K}}'\tilde{\mathbf{K}} = \tilde{\mathbf{L}}'\tilde{\mathbf{L}} = \mathbf{I}_p$. The existance of the SVD is proven in ten Berge (1993, pp.5-6)

2.2 ten Berge's theorem

The definition and properties of a suborthonormal matrix are firstly introduced, which is found in ten Berge (1983, 1993). Then, ten Berge's theorem (ten Berge, 1993) is discussed.

Denote that $\mathbf{\Gamma}$ is an $n \times n$ suborthonormal matrix of rank $r \leq n$, and $\mathbf{C}$ is a diagonal and nonnegative matrix whose diagonal elements are arranged in decreasing order. The matrix is suborthonormal if it can be completed to an orthonormal matrix, by appending rows or columns, or both. For example, $\mathbf{A}_{sub}$ is a suborthonormal matrix when $\mathbf{T}$ is an orthonormal matrix by appending matrices $\mathbf{A}_1$, $\mathbf{A}_2$ and $\mathbf{A}_3$:

$$\mathbf{T} = \begin{bmatrix} \mathbf{A}_{sub} & \mathbf{A}_1 \\ \mathbf{A}_2 & \mathbf{A}_3 \end{bmatrix}.$$

In other word, the suborthonormal matrix can be regarded as a submatrix of some orthonormal matrix. The suborthonormal matrix has the following properties:

- Every columwise or rowwise orthonormal matrix is suborthonornal.
- The product of any two suborthonormal matrices is also suborthonormal.

Then, ten Berge (1993) showed that the trace function $f(\mathbf{\Gamma}) = tr(\mathbf{\Gamma C})$ has the upper bound

$$f(\mathbf{\Gamma}) = tr(\mathbf{\Gamma C}) \leq c_1 + c_2 + \cdots + c_r, \tag{2.8}$$

where $c_1 + c_2 + \cdots + c_r$ is the sum of the r largest elements in $\mathbf{C}$. This upper bound is attained when

$$\mathbf{\Gamma} = \begin{bmatrix} \mathbf{I}_r & {}_r\mathbf{O}_{n-r} \\ {}_{n-r}\mathbf{O}_r & {}_{n-r}\mathbf{O}_{n-r} \end{bmatrix}. \tag{2.9}$$

This is named as ten Berge's theorem.

2.3 Maximization of bilinear form

Define the trace fucntion

$$b(\mathbf{\Gamma}_1, \mathbf{\Gamma}_2) = tr(\mathbf{\Gamma}_1' \mathbf{B} \mathbf{\Gamma}_2), \tag{2.10}$$

where $\mathbf{B}$ is a $n \times p$ matrix, $\mathbf{\Gamma}_1$ is $n \times r$ matrix and $\mathbf{\Gamma}_2$ is a $p \times r$ matrix. This kind of the function is called bilinear form.

Assume a problem that the maximization of the function (2.10) over $\mathbf{\Gamma}_1$ and $\mathbf{\Gamma}_2$ subject to $\mathbf{\Gamma}_1'\mathbf{\Gamma}_1 = \mathbf{\Gamma}_2'\mathbf{\Gamma}_2 = \mathbf{I}_r$. Denote the SVD of the matrix $\mathbf{B} = \mathbf{K\Delta L}'$. Then, the matrices $\mathbf{K}$, $\mathbf{L}$, $\mathbf{\Gamma}_1'$ and $\mathbf{\Gamma}_2'$ are the suborthogonal matrices. It is shown that the product of matrices $\mathbf{L}'\mathbf{\Gamma}_2\mathbf{\Gamma}_1'\mathbf{K}$ is also suborthogonal (ten Berge, 1993) and the rank of $\mathbf{L}'\mathbf{\Gamma}_2\mathbf{\Gamma}_1'\mathbf{K}$ is equal to or less than r. Using ten Berge's theorem, the following inequality holds:

$$\begin{aligned} b(\mathbf{\Gamma}_1, \mathbf{\Gamma}_2) &= tr(\mathbf{\Gamma}_1' \mathbf{B} \mathbf{\Gamma}_2) \\ &= tr(\mathbf{L}'\mathbf{\Gamma}_2\mathbf{\Gamma}_1'\mathbf{K\Delta}) \\ &\leq tr(\mathbf{\Delta}_r). \end{aligned} \tag{2.11}$$

The upper bound is attained when

$$\mathbf{\Gamma}_1 = \mathbf{K}_r, \tag{2.12}$$

$$\mathbf{\Gamma}_2 = \mathbf{L}_r. \tag{2.13}$$

Therefore, the optimal soltions $\mathbf{\Gamma}_1$ and $\mathbf{\Gamma}_2$ can be explicitly obtained using ten Berge's theorem. Some optimizations in matrix decomposition approaches reduce to maximization of bilinear form with column orthogonality constraints as will be shown in the following chapters.

2.4 Principal component analysis

Principal component analoysis (PCA) is one of the widely used method for analyzing numerical data and assigning the vectors of scores to categories and objects in reduced space. ten Berge and Kiers (1996) pointed out that the formulations of PCA can be classified into three types: the formulations based on Hotelling (1933), Pearson (1901) and Rao (1973). In this thesis, the formulation proposed by Pearson (1901) is introduced. In Pearson (1901), PCA can be formulated as

$$\mathbf{X} = \mathbf{F}\mathbf{A}' + \mathbf{E}, \tag{2.14}$$

where $\mathbf{X}$ is an n-individuals $\times$ p-variables standardized data matrix, $\mathbf{F}$ is an n-individuals $\times$ c-components principal component score matrix, $\mathbf{A}$ is a p-variables $\times$ c-components loading matrix, and $\mathbf{E}$ is an n-individuals $\times$ p-variables error matrix. For obtaining the principal component score matrix $\mathbf{F}$ and the loading matrix $\mathbf{A}$, PCA minimizes the the sum of the squares of the error $\mathbf{E}$,

$$f(\mathbf{F}, \mathbf{A}) = ||\mathbf{X} - \mathbf{F}\mathbf{A}'||^2, \tag{2.15}$$

subject to the following constraints,

$$n^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_c, \tag{2.16}$$

$$\mathbf{F} = \mathbf{J}\mathbf{F}. \tag{2.17}$$

Here, $\mathbf{J} = \mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}'_n$ denotes the centering matrix.

Let us consider minimizing the loss function (2.15) over the loading matrix $\mathbf{A}$. The loss function (2.15) can be expressed as

$$\begin{aligned}
||\mathbf{X} - \mathbf{F}\mathbf{A}'||^2 &= tr(\mathbf{X} - \mathbf{F}\mathbf{A}')'(\mathbf{X} - \mathbf{F}\mathbf{A}') \\
&= tr(\mathbf{X}'\mathbf{X} - 2\mathbf{X}'\mathbf{F}\mathbf{A}' + \mathbf{A}\mathbf{F}'\mathbf{F}\mathbf{A}') \\
&= tr(\mathbf{X}'\mathbf{X} - 2\mathbf{X}'\mathbf{F}(\mathbf{F}'\mathbf{F})^{-1/2}(\mathbf{F}'\mathbf{F})^{1/2}\mathbf{A}' \\
&\quad + \mathbf{A}(\mathbf{F}'\mathbf{F})^{1/2}(\mathbf{F}'\mathbf{F})^{1/2}\mathbf{A}' \\
&\quad + \mathbf{X}'\mathbf{F}(\mathbf{F}'\mathbf{F})^{-1/2}(\mathbf{F}'\mathbf{F})^{-1/2}\mathbf{F}'\mathbf{X} \\
&\quad - \mathbf{X}'\mathbf{F}(\mathbf{F}'\mathbf{F})^{-1/2}(\mathbf{F}'\mathbf{F})^{-1/2}\mathbf{F}'\mathbf{X}) \\
&= ||(\mathbf{F}'\mathbf{F})^{1/2}\mathbf{A}' - (\mathbf{F}'\mathbf{F})^{-1/2}\mathbf{F}'\mathbf{X}||^2 + c_{PCA_A}.
\end{aligned} \tag{2.18}$$

Here, $c_{PCA_A} = tr(\mathbf{X}'\mathbf{X} - \mathbf{X}'\mathbf{F}(\mathbf{F}'\mathbf{F})^{-1/2}(\mathbf{F}'\mathbf{F})^{-1/2}\mathbf{F}'\mathbf{X})$ is a constant, which is irrelevant to $\mathbf{A}$. As can be seen in the loss function (2.18), the minimization over $\mathbf{A}$ is attained when $(\mathbf{F}'\mathbf{F})^{1/2}\mathbf{A}' =$

$(\mathbf{F}'\mathbf{F})^{-1/2}\mathbf{F}'\mathbf{X}$. Thus, optimal $\mathbf{A}$ is described as

$$\begin{aligned}\mathbf{A} &= \mathbf{X}'\mathbf{F}(\mathbf{F}'\mathbf{F})^{-1} \\ &= n^{-1}\mathbf{X}'\mathbf{F}. \end{aligned} \tag{2.19}$$

This optimization can be viewed as regression (ten Berge, 1993). When the data matrix $\mathbf{X}$ is standardized, the loading matrix (2.19) expresses the correlations between the component scores and the observed data.

Next, the minimization over $\mathbf{F}$ is considered. Substituting (2.19) into the loss function (2.15), then

$$\begin{aligned} &tr(\mathbf{X}'\mathbf{X} - 2n^{-1}\mathbf{X}'\mathbf{F}\mathbf{F}'\mathbf{X} + n^{-1}\mathbf{X}'\mathbf{F}\mathbf{F}'\mathbf{X}) \\ &= -n^{-1}tr(\mathbf{X}'\mathbf{F}\mathbf{F}'\mathbf{X}) + c_{PCA_F} \end{aligned} \tag{2.20}$$

is obtained. Here, $c_{PCA_F} = tr(\mathbf{X}'\mathbf{X})$ is a constant that is irrelevant to $\mathbf{F}$. Therefore, the minimization of the loss fucntion over $\mathbf{F}$ reduces to the maximization of the function $tr(\mathbf{X}'\mathbf{F}\mathbf{F}'\mathbf{X})$ over $\mathbf{F}$ subject to the constraints (2.16) and (2.17). As shown in previous section, the maximization can be achieved using ten Berge's theorem. Denote the SVD of the data matrix $\mathbf{X} = \mathbf{K}\mathbf{\Delta}\mathbf{L}'$. Then, optimal $\mathbf{F}$ is expressed as

$$\mathbf{F} = \sqrt{n}\mathbf{K}_c, \tag{2.21}$$

where $\mathbf{K}_c$ is the matrix that contains first c left singular vectors. The optimal $\mathbf{F}$ is satisfied with (2.17) because $\mathbf{K}_c = \mathbf{X}\mathbf{L}_c\mathbf{\Delta}_c^{-1} = \mathbf{J}\mathbf{X}\mathbf{L}_c\mathbf{\Delta}_c^{-1} = \mathbf{J}\mathbf{K}_c$ holds when $\mathbf{X}$ is standardized.

Using (2.21), $\mathbf{A}$ can be described as

$$\begin{aligned}\mathbf{A} &= n^{-1}(\mathbf{K}\mathbf{\Delta}\mathbf{L}')'(\sqrt{n}\mathbf{K}_c) \\ &= \sqrt{n}^{-1}\mathbf{L}_c\mathbf{\Delta}_c. \end{aligned} \tag{2.22}$$

The result above implies that PCA solutions can be obtained by the SVD of the data matrix $\mathbf{X}$.

In some case, PCA parameters are estimated under the following constraint:

$$n^{-1}\mathbf{F}'\mathbf{F} = \mathbf{\Theta}. \tag{2.23}$$

Here, $\mathbf{\Theta}$ is diagonal matrix whose elements are arranged in descending order. Optimal $\mathbf{F}$ and $\mathbf{A}$ can be easily obtained by the SVD of $\mathbf{X}$:

$$\mathbf{F} = \sqrt{n}\mathbf{K}_c\mathbf{\Delta}_c, \tag{2.24}$$

$$\mathbf{A} = \sqrt{n}^{-1}\mathbf{L}_c. \tag{2.25}$$

Equation (2.24) is satisfied with the constraint (2.23) and squered singular values $\mathbf{\Delta}^2$ (eigen values of $\mathbf{X}'\mathbf{X}$) correspond to variances of components.

2.5 Exploratory factor analysis

2.5.1 The Model in exploratory factor analysis

Exploratory factor analysis (EFA) is a popular statistical method that aims to explain the interrelationships among observed variables by two types of latent scores called common and unique factors. The relationships between observed variables and latent factors are described as a linear model. The model can be expressed in the following manners:

$$\mathbf{x}_{(i)} = \mathbf{A}\mathbf{f}_{(i)} + \mathbf{\Psi}\mathbf{u}_{(i)}, \tag{2.26}$$

$$\mathbf{x}_j = \mathbf{F}\mathbf{a}_{(j)} + \psi_j\mathbf{u}_j, \tag{2.27}$$

$$\mathbf{X} = \mathbf{F}\mathbf{A}' + \mathbf{U}\mathbf{\Psi}, \tag{2.28}$$

where $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_p] = [\mathbf{x}'_{(1)}, \ldots, \mathbf{x}'_{(n)}]'$ is an n-observations $\times$ p-variables standardized data matrix, $\mathbf{F} = [\mathbf{f}_1, \ldots, \mathbf{f}_c] = [\mathbf{f}'_{(1)}, \ldots, \mathbf{f}'_{(n)}]'$ is an n-observations $\times$ c-factors matrix whose contains common factor scores, $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_c] = [\mathbf{a}'_{(1)}, \ldots, \mathbf{a}'_{(p)}]'$ is a p-variables $\times$ c-factors matrix of factor loadings, $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_p] = [\mathbf{u}'_{(1)}, \ldots, \mathbf{u}'_{(n)}]'$ is an n-observations $\times$ p-variables matrix whose contains unique factor scores, and $\mathbf{\Psi} = diag(\psi_1, \ldots, \psi_p)$ is a p-variables $\times$ p-variables diagonal matrix of uniqueness. In EFA model, variances of observed data are decomposed into common and unique parts; common factors explain the part of variances that share with each other, but unique factors explain the variances that do not share with any variables. That is, common and unique factors are assumed to be mutually uncorrelated. Unique factors are also assumed to be uncorrelated with each other. From these assumption, the model of correlation matrix $\mathbf{\Sigma}$ can be written as

$$\mathbf{\Sigma} = \mathbf{A\Phi A}' + \mathbf{\Psi}^2, \tag{2.29}$$

where $\mathbf{\Phi}$ is a $c \times c$ matrix of common factors correlations. If the common factors are mutually uncorrelated, (2.29) reduces to

$$\mathbf{\Sigma} = \mathbf{AA}' + \mathbf{\Psi}^2. \tag{2.30}$$

The model parameters of EFA are traditionally estimated by minimizing the differences between a model correlation structure expressed in (2.30) and a sample correlation matrix using maximum likelihood (ML) or least square (LS) discrepancy measures (e.g., Harman, 1976; Mulaik, 2010).

2.5.2 Maximun likelihood approach

In ML approach, the common and unique factors are assumed to be distributed according to the multivariate normal distributions:

$$\mathbf{f}_{(i)} \backsim N_c(\mathbf{0}_c, \mathbf{I}_c), \tag{2.31}$$

$$\mathbf{u}_{(i)} \backsim N_p(\mathbf{0}_p, \mathbf{\Psi}^2). \tag{2.32}$$

The assumptions (2.31) and (2.32) imply that $\mathbf{x}_{(i)}$ is distributed according the following multivariate normal distribution:

$$\mathbf{x}_{(i)} \backsim N_p(\mathbf{0}_p, \mathbf{\Sigma}). \tag{2.33}$$

Then, the log likelihood function for a sample correlation matrix $\mathbf{R}$ can be formulated as

$$ml(\mathbf{A}, \mathbf{\Psi}) = \frac{n}{2} log|(\mathbf{AA}' + \mathbf{\Psi}^2)^{-1}\mathbf{R}| - \frac{n}{2} tr\{(\mathbf{AA}' + \mathbf{\Psi}^2)^{-1}\mathbf{R}\}. \tag{2.34}$$

The function (2.34) can be translated in an equivalent function:

$$ML(\mathbf{A}, \mathbf{\Psi}) = -log|(\mathbf{AA}' + \mathbf{\Psi}^2)^{-1}\mathbf{R}| + tr\{(\mathbf{AA}' + \mathbf{\Psi}^2)^{-1}\mathbf{R}\}. \tag{2.35}$$

The maximization of (2.34) is equivalent to the minimization of (2.35). The optimization can be achieved by various procedures such as gradient algorithm (Joreskog, 1967) or EM algorithm (Adachi, 2013; Rubin and Thayer, 1982).

2.5.3 Least square approach

In LS approach, the discrepancy between the model and the sample correlations can be described as

$$LS(\mathbf{A}, \mathbf{\Psi}) = ||\mathbf{R} - \mathbf{A}\mathbf{A}' - \mathbf{\Psi}||^2. \tag{2.36}$$

Differenciating the function (2.36) with respect to $\mathbf{A}$ and $\mathbf{\Psi}$, then we have

$$(\mathbf{R} - \mathbf{A}\mathbf{A}' - \mathbf{\Psi}^2)\mathbf{A} = \mathbf{O}, \tag{2.37}$$

$$\mathbf{\Psi}^2 = diag(\mathbf{R} - \mathbf{A}\mathbf{A}'). \tag{2.38}$$

Here, it is assumed that $\mathbf{A}\mathbf{A}' = \mathbf{\Omega}$ ($\mathbf{\Omega}$ is a diagonal) for eliminating the indeterminacy. Then, equation (2.37) can be re-expressed as $(\mathbf{R} - \mathbf{\Psi}^2) = \mathbf{A}\mathbf{\Omega}$. Consequently, the minimization of the function (2.36) over $\mathbf{A}$ with a fixed $\mathbf{\Psi}$ is attained by

$$\mathbf{A} = \mathbf{E}_c\mathbf{\Delta}_c^{1/2}, \tag{2.39}$$

where $\mathbf{E}_c$ is a matrix of first c eigen vectors with eigen value decomposition (EVD) $(\mathbf{S} - \mathbf{\Psi}^2) = \mathbf{E}\mathbf{\Delta}\mathbf{E}'$. Minimizing the function (2.36) over $\mathbf{\Psi}$ with a fixed $\mathbf{A}$ is attained by equation (2.38). The parameters can be estimated by alternating these minimization when the convergence criterion is reached.

2.5.4 Matrix decomposition factor analysis (MDFA)

Recently, an alternative method for estimating model parameters has been developed, in which all parameters are treated as fixed and unknown. The approach is named as matrix decomposition factor analysis (MDFA) (Adachi, 2012; de Leeuw, 2004, 2008; Unkel and Trendafilov, 2010, 2013). The MDFA loss function is defined as the minimization of the difference between the observed data matrix and the EFA model parameters:

$$g(\mathbf{F}, \mathbf{U}, \mathbf{A}, \mathbf{\Psi}) = ||\mathbf{X} - \mathbf{F}\mathbf{A}' - \mathbf{U}\mathbf{\Psi}||^2. \tag{2.40}$$

The model parameters are then obtained by minimizing (2.40) over $\mathbf{F}$, $\mathbf{A}$, $\mathbf{U}$, and $\mathbf{\Psi}$, subject to the constraints of common and unique factors being mutually uncorrelated:

$$\mathbf{J}\mathbf{F} = \mathbf{F}, n^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_c, \tag{2.41}$$

$$\mathbf{JU} = \mathbf{U}, n^{-1}\mathbf{U}'\mathbf{U} = \mathbf{I}_p, \tag{2.42}$$

$$n^{-1}\mathbf{F}'\mathbf{U} = \mathbf{O}, \tag{2.43}$$

$$\mathbf{\Psi} \text{ is diagonal.} \tag{2.44}$$

The constraints described above are mathematical expression of the assumption of common and unique factors.

In contrast to other EFA estimation techniques, the data matrix is directly fitted to EFA model parts. Consequently, MDFA allows simultaneous estimation of all model parameters, although factor scores cannot be uniquely identified. In other estimation methods, the factor score matrices **F** and **U** (if required) can be obtained only via a two-stages procedure; **A** and **Ψ** are firstly estimated and then factor scores are estimated or predicted as a function of **A**, **Ψ**, and **X** (e.g. Mulaik, 2010). This form of indeterminacy that factor scores cannot be uniquely determined is well-known as factor indeterminacy. In spite of the condition of factor indeterminacy, non-uniqueness of the common and unique factor scores is not a problem from an algorithmic viewpoint in MDFA. In fact, the MDFA estimation procedures avoid the conceptual problem of factor indeterminacy and facilitate the estimation of both common and unique factor scores (de Leeuw, 2004; Unkel and Trendafilov, 2010).

In the MDFA estimation algorithm, the factor score estimation and the factor loading estimation steps are iterately (de Leeuw, 2004, 2008). The minimization of the loss function (2.40) over loadings and uniqueness is considered. The loss function (2.40) can be expanded as

$$||\mathbf{X} - \mathbf{FA}' - \mathbf{U\Psi}||^2 = n||n^{-1}\mathbf{X}'\mathbf{F} - \mathbf{A}||^2 + n||n^{-1}diag(\mathbf{X}'\mathbf{U}) - \mathbf{\Psi}||^2 + c_{A\Psi}, \tag{2.45}$$

where $c_{A\Psi}$ is a constant which is irrelevant to **A** and **Ψ**. Optimal **A** and **Ψ** are obtained by

$$\mathbf{A} = n^{-1}\mathbf{X}'\mathbf{F}, \tag{2.46}$$

$$\mathbf{\Psi} = n^{-1}\text{diag}(\mathbf{X}'\mathbf{U}). \tag{2.47}$$

It should be noted that the loading matrix **A** can be viewed as the correlation between factors and variables because both of **X** and **F** are assumed to be standardized.

Define $\mathbf{S} = [\mathbf{F}, \mathbf{U}]$ and $\mathbf{\Lambda} = [\mathbf{A}, \mathbf{\Psi}]$ with dimensions $n \times (p + c)$ and $p \times (p + c)$. Then, $\mathbf{S}$ is satisfied with

$$n^{-1}\mathbf{S}'\mathbf{S} = \mathbf{I}_{p+c}, \tag{2.48}$$

$$\mathbf{S} = \mathbf{J}\mathbf{S}. \tag{2.49}$$

Using $\mathbf{S}$ and $\mathbf{\Lambda}$, the loss function (2.40) can be rewritten as

$$\|\mathbf{X} - \mathbf{F}\mathbf{A}' - \mathbf{U}\mathbf{\Psi}\|^2 = -tr(\mathbf{S}'\mathbf{X}\mathbf{\Lambda}) + c_S, \tag{2.50}$$

where c_S is a constant which is irrelevant to $\mathbf{S}$. Minimizing the loss fucntion over $\mathbf{S}$ is attained by maximizing $tr(\mathbf{S}'\mathbf{X}\mathbf{\Lambda})$ over $\mathbf{S}$. The maximization of the trace function $tr(\mathbf{S}'\mathbf{X}\mathbf{\Lambda})$ over $\mathbf{S}$ is achieved by

$$\mathbf{S} = \sqrt{n}\mathbf{K}\mathbf{L}' = \sqrt{n}\mathbf{K}_1\mathbf{L}'_1 + \sqrt{n}\mathbf{K}_2\mathbf{L}'_2. \tag{2.51}$$

Here, $\mathbf{K} = [\mathbf{K}_1, \mathbf{K}_2]$ and $\mathbf{L} = [\mathbf{L}_1, \mathbf{L}_2]$ are block matrices of left and right singular vectors with the SVD of $\mathbf{X}\mathbf{\Lambda}$. $\mathbf{S}$ is satisfied with $\mathbf{S} = \mathbf{J}\mathbf{S}$ because $\mathbf{X}$ is assumed to be standardized. Note that the rank of $\mathbf{X}\mathbf{\Lambda}$ is at most p. It implies that $\mathbf{K}_2$ and $\mathbf{L}_2$ cannot be uniquely determined.

The whole algorithm in MDFA can be described as follows:

Step 1: Set the number of factors c.

Step 2: Initialize the parameters.

Step 3: Minimize $\mathbf{A}$ and $\mathbf{\Psi}$ given $\mathbf{F}$ and $\mathbf{U}$, which is achieved by (2.46) and (2.47).

Step 4: Minimize $\mathbf{F}$ and $\mathbf{U}$ given $\mathbf{A}$ and $\mathbf{\Psi}$, which is achieved by (2.51).

Step 5: Finish if convergence is reached; otherwise, back to Step 3.

2.6 Matrix decomposition approach

As shown in the previous sections, the PCA and MDFA solustions can be obtained via the SVD. In this thesis, this type of estimation algorithms are named as matrix decomposition approaches to multivariate analysis for dimension reduction.

The models of PCA and MDFA seem to be similar to each other, but a crucial difference is found in the error term (Adachi, 2016; Adachi and Trendafilov, 2019). No assumption is made for the error term **E** in PCA. Therefore, the errors in PCA are correlated to all variables. In contrast, the error term in EFA is assumed to be uncorrelated each other, and unique to each variables, which corresponds to the constraints (2.42) and (2.44). Here are examples that show the difference between PCA and EFA (figure 2.1 and 2.2).

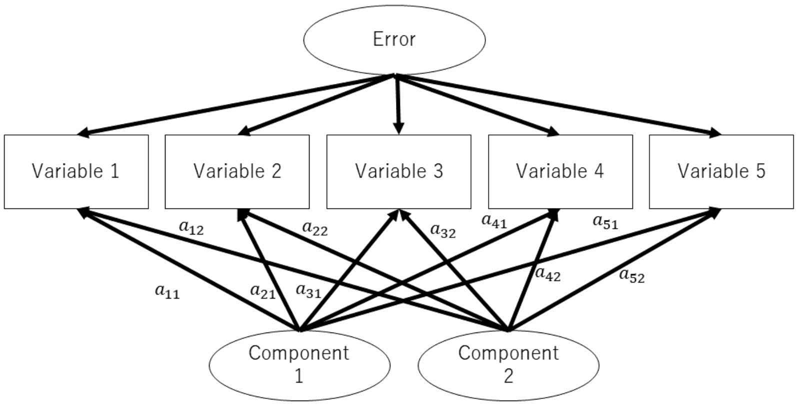


Figure 2.1: An example of PCA model with two components

From another point of view, PCA and MDFA differ in the rank of model parts. In PCA, the number of components c is assumed to be less than or equal to p. Thus, the loss function $||\mathbf{X} - \mathbf{FA}'||^2$ can be considered as the low rank approximation problem because $rank(\mathbf{FA}') = c$. In MDFA, the loss function can be rewritten as

$$||\mathbf{X} - \mathbf{S}\mathbf{\Lambda}'||^2 = ||\mathbf{S} - \mathbf{X}\mathbf{\Lambda}||^2 + c_S. \tag{2.52}$$

In this reformulation, MDFA can be viewed as the higher rank approximation because $\mathbf{X}\mathbf{\Lambda}$ with $rank(\mathbf{X}\mathbf{\Lambda}) = p$ is fitted to $\mathbf{S}$ with $rank(\mathbf{S}) = c + p > p$ (Adachi, 2015). Recall that $\mathbf{S}$ in MDFA is uniquely undetermined but $\mathbf{F}$ in PCA is uniquely determined.

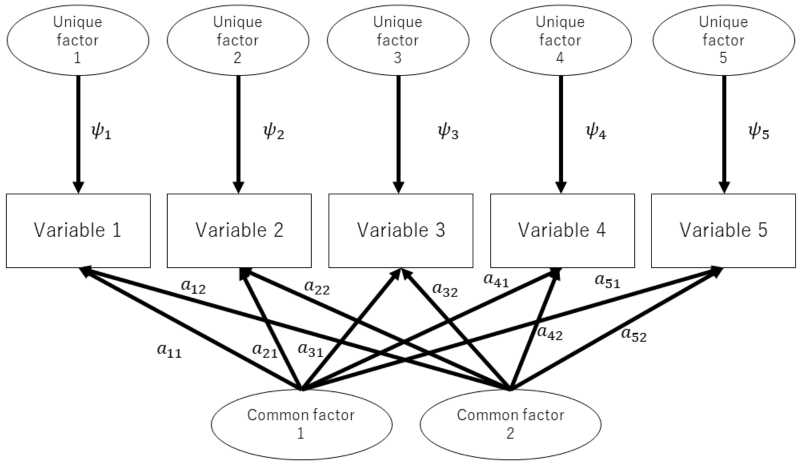


Figure 2.2: An example of EFA model with two factors

2.7 Extension matrix decomposition approaches to nonmetric multivariate analysis by quantification

There are various methods to extend multivariate analysis to handle nonmetric data. Among them, we introduce the technique called as quantification or optimal scaling. The quantification is a widely used statistical technique for analyzing categorical variables. In the qunatification, the observed categorical variables are transformed into quantitative scores such that the data analysis model most closely matches the observations (Gifi, 1990; Young, 1981). Young (1981) pointed out that data analysis techniques for quantitative data can be extended for nonmetric data if least square estimation algorithms are known. That is, using the quantification technique, a variety of quantitative data analysis model can be extended easily. In nonmetric data analysis using the quantification technique, the parameters are divided into two groups:

1. The parameters of the model part (which is referred to as model parameters)
2. The parameters of data part (which is refered to as optimal scaling parameters or quantification parameters)

The loss function is optimized by alternately optimizing over the model parameters with fixed the quantification parameters and then over the quantification parameters with fixed model param-

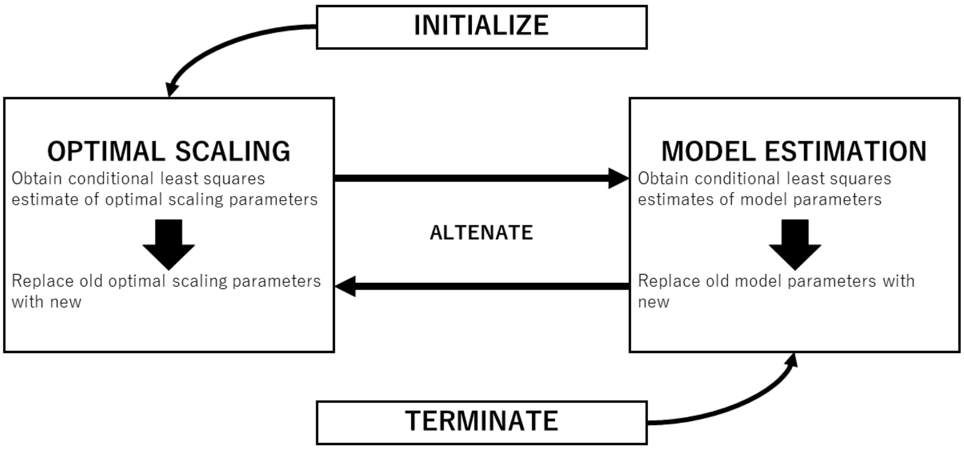


Figure 2.3: Flow of estimation in quantification (Young, 1981)

eters. Figure 2.3 shows the optimization process of the quantification. In the optimal scaling part, qualitative data is transformed into quantified data so that existing data analysis for quantitative data can be adopted. Then, the model parmeters are estimated given the quantified data. It should be noted that components scores and loadings in PCA, and factors scores, loadings and uniqueness in EFA correspond to the model parameters in this thesis. Thus, using the quantification technique, nonmetric version of PCA and MDFA can be defined as

$$\mathbf{X}^* \cong \mathbf{F}\mathbf{A}', \tag{2.53}$$

$$\mathbf{X}^* \cong \mathbf{F}\mathbf{A}' + \mathbf{U}\boldsymbol{\Psi}, \tag{2.54}$$

where $\mathbf{X}^*$ is an $n \times R$ quantified data matrix that is also estimated in the quantification and $\cong$ denotes a least square approximation. That is, the loss function of nonmetric version of PCA is described as

$$f_{npca}(\mathbf{X}^*, \mathbf{F}, \mathbf{A}) = ||\mathbf{X}^* - \mathbf{F}\mathbf{A}'||^2, \tag{2.55}$$

and that of nonmetric version of MDFA does

$$f_{nmdfa}(\mathbf{X}^*, \mathbf{F}, \mathbf{A}, \mathbf{U}, \boldsymbol{\Psi}) = ||\mathbf{X}^* - \mathbf{F}\mathbf{A}' - \mathbf{U}\boldsymbol{\Psi}||^2. \tag{2.56}$$

We will develop optimization algorithms for the loss functions described above and discuss some mathematical properties of the proposed methods in the following chapters.

3

Interpretability in relationship between variables and components/factors

Interpretability is one of the important themes in data analysis. In this thesis, we classify two cases that users intepret obtained solutions. One is the case that users focus on relationships between components/factors and variables, which is discussed in this chapter. As shown in the previous chapter, the loading matrix expresses relationships between variables and components/factors. Therefore, we will discuss techniques that make the loading matrix interpretable in the following sections.

3.1 Interpretability of relationships between components/factors and variables

To consider the interpretability of relationships between components/factors and variables, we must firstly define what the interpretable structure is. Thurstone (1947) provided five rules for the interpretable structure, which is called as ideal simple structure:

1. Each row of the factor matrix should have at least one zero.
2. If there are c common factors each column of the factor matrix should have at least c zeros.
3. For every pair of columns of the factor matrix there should be several variables whose entries vanish in one column but not in the other.
4. For every pair of columns of the factor matrix, a large proportion of the variables should have vanishing entries in both columns when there are four or more factors.
5. For every pair of columns of the factor matrix there should be only a small number of variables with non-vanishing entries in both columns.

An example of the ideal simple structure that reflects Thurstone's five rules is shown in table 3.1, where **x** indicates a nonzero value. Ideal simple structure has two features (Adachi, 2016):

1. Sparse, i.e., a number of elements are zero
2. Well classified, i.e., different variables load different components or factors

First feature allows users to focus on the nonzero elements to capture the relationships between variables and components/factors. Second feature clarifies the differences between the components/factors. If the ideal simple structure is obtained, users can easily interpret the solutions. However, there are some cases when the ideal simple structure cannot be acquired. To handle these cases, the approximate simple structure is considered as an alternative to the ideal simple structure. An example of approximate simple structure is expressed in table 3.2. In table 3.2, ” small ” stands for values which are close to zero, and ” **Large** ” expresses values with large absolute values. When users neglect small values below certain magnitudes, say 0.25 and focus on large values in approximate simple structure, then they interpret the solution in the same way as the ideal simple structure.

In PCA and EFA literature, the ideal simple structure and the approximate simple structure can be expressed as in figure 3.1 - 3.4. The connections from components/factors to variables are clear in the ideal simple strucutre (figure 3.1 and 3.2); each variables are explained by one components/factors in the ideal simple strucutre and the users can easily interpret the PCA and EFA solutions. In figure 3.3 and 3.4, the bold arrows show that components/factors are connected strongly to variables and the thin arrows represent that weak connections exist between variables and components/factors. The relationships between variables and components/factors in figure 3.3 and 3.4 are a little obscure comparing with the ideal simple structure, but the users can sufficiently interpret the PCA and EFA solutions in the approximate simple structure.

3.2 Rotation technique

In PCA and EFA, the loss function can be transformed by an arbitrary nonsingular matrix $\mathbf{T}$ of order $c \times c$ without affecting the fit of solutions:

$$||\mathbf{X} - \mathbf{F}\mathbf{A}'||^2 = ||\mathbf{X} - \mathbf{F}\mathbf{T}'^{-1}\mathbf{T}'\mathbf{A}'||^2, \tag{3.1}$$

$$||\mathbf{X} - \mathbf{F}\mathbf{A}' - \mathbf{U}\mathbf{\Psi}||^2 = ||\mathbf{X} - \mathbf{F}\mathbf{T}'^{-1}\mathbf{T}'\mathbf{A}' - \mathbf{U}\mathbf{\Psi}||^2. \tag{3.2}$$

Table 3.1: Ideal simple structure

Ideally simple			An example of Ideally simple		
Comp/Factor 1	Comp/Factor 2	Comp/Factor 3	Comp/Factor 1	Comp/Factor 2	Comp/Factor 3
x	0	0	**0.8**	0	0
x	0	0	**0.8**	0	0
x	0	0	**0.9**	0	0
x	0	0	**0.7**	0	0
0	**x**	0	0	**0.9**	0
0	**x**	0	0	**0.7**	0
0	**x**	0	0	**0.8**	0
0	0	**x**	0	0	**-0.8**
0	0	**x**	0	0	**-0.8**
0	0	**x**	0	0	**-0.8**

Table 3.2: Approximate simple structure

Approximately simple			An example of Approximately simple		
Comp/Factor 1	Comp/Factor 2	Comp/Factor 3	Comp/Factor 1	Comp/Factor 2	Comp/Factor 3
Large	small	small	**0.8**	0.2	0.1
Large	small	small	**0.8**	0.2	0.1
Large	small	small	**0.9**	0.1	0.1
Large	small	small	**0.7**	0.1	0.1
small	**Large**	small	0.2	**0.9**	0.2
small	**Large**	small	0.25	**0.7**	0.2
small	**Large**	small	0.1	**0.8**	0.2
small	small	**Large**	0.15	0.1	**-0.8**
small	small	**Large**	0.2	0.1	**-0.8**
small	small	**Large**	0.1	0.1	**-0.8**

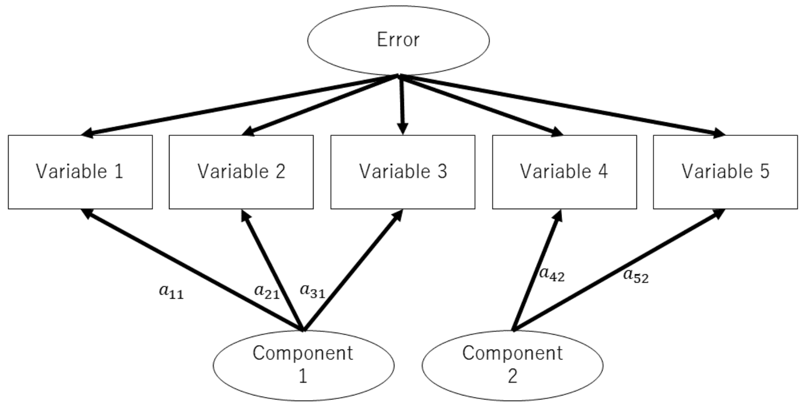


Figure 3.1: An example of the ideal simple strucutre in PCA model with two components

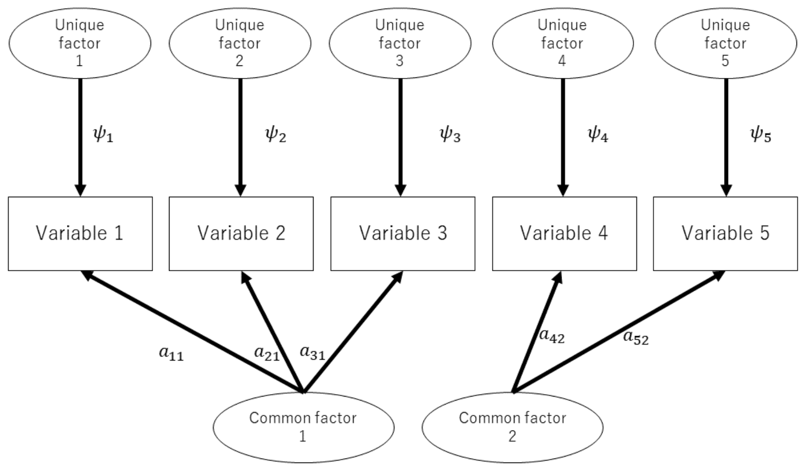


Figure 3.2: An example of the ideal simple strucutre in EFA model with two factors

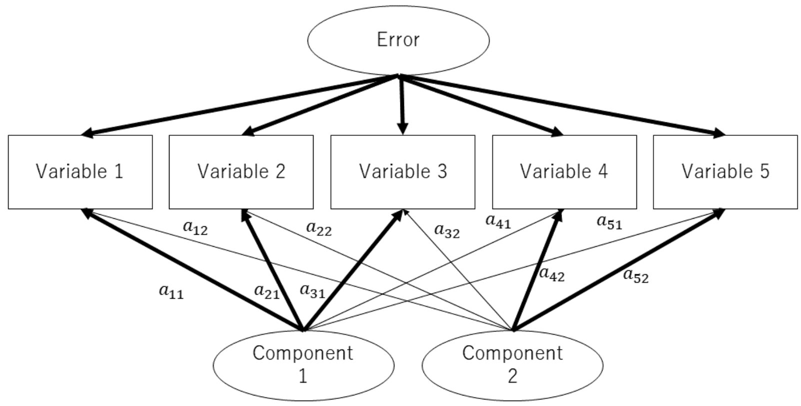


Figure 3.3: An example of the approximate simple strucutre in PCA model with two components

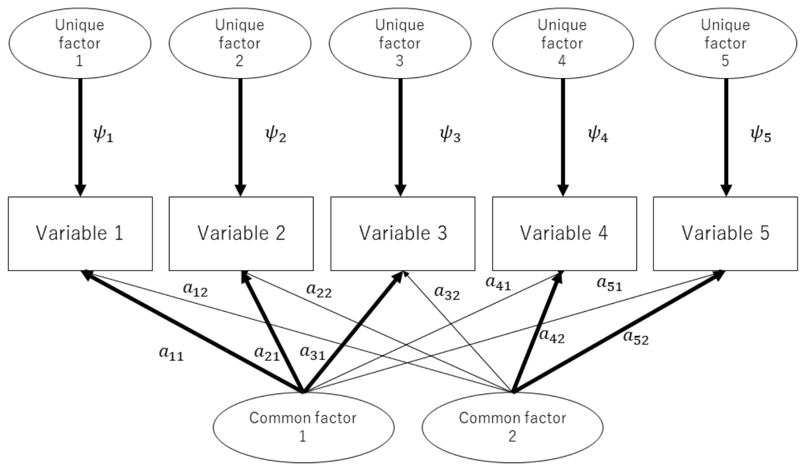


Figure 3.4: An example of the appoximate simple strucutre in EFA model with two factors

This property is known as rotational freedom or rotational indeterminacy (Browne, 2001). Using this property, estimated solutions are transformed into simple structure that is defined as Thurstone's five rules. This transformation is named as rotation. The rotation is the tandem technique; one of the PCA and/or EFA solutions (initial solution) are firstly estimated, and then these solutions are transformed (rotated) into the simple structure. If no identification condition is imposed on $\mathbf{T}$, arbitrary nonsingular transformation can be applied to the PCA and/or EFA solutions. However, the normalization conditions such as

$$\mathbf{T}'\mathbf{T} = \mathbf{I}_c, \tag{3.3}$$

$$diag(\mathbf{T}^{-1}\mathbf{T}'^{-1}) = \mathbf{I}_c, \tag{3.4}$$

are imposed in the rotation. The correlations of rotated components/factors under the condition $\mathbf{T}'\mathbf{T} = \mathbf{I}_c$ are considered:

$$n^{-1}\mathbf{T}'\mathbf{F}'\mathbf{F}\mathbf{T} = \mathbf{T}'\mathbf{T} = \mathbf{\Phi} = \mathbf{I}_c. \tag{3.5}$$

This result shows that component/factor correlations are unchanged by the transformation under the normalization condition $\mathbf{T}'\mathbf{T} = \mathbf{I}_c$ because components/factors are estimated under the assumption $\mathbf{\Phi} = \mathbf{I}_c$. This transformation is called orthogonal rotation. Next, let us consider the components/factor correlations rotated under the the condition $diag(\mathbf{T}^{-1}\mathbf{T}'^{-1}) = \mathbf{I}_c$:

$$\begin{aligned} n^{-1}\mathbf{T}^{-1}\mathbf{F}'\mathbf{F}\mathbf{T}'^{-1} &= \mathbf{T}^{-1}\mathbf{T}'^{-1} = \mathbf{\Phi} \\ &= \begin{pmatrix} 1 & \cdots & \phi_{1i} & \cdots & \phi_{1c} \\ \vdots & \ddots & & & \vdots \\ \phi_{i1} & & 1 & & \phi_{ic} \\ \vdots & & & \ddots & \vdots \\ \phi_{c1} & \cdots & \phi_{ci} & \cdots & 1 \end{pmatrix}. \end{aligned} \tag{3.6}$$

When the initial solutions are rotated under $diag(\mathbf{T}^{-1}\mathbf{T}'^{-1}) = \mathbf{I}_c$, rotated components/factors are no longer mutually uncorrelated. This rotation is called as oblique rotation. Mathematically, orthogonal rotation is the special case of oblique rotation. In practice, oblique rotation is thought useful for finding meaningful dimensions in data, since many concepts in human sciences are not independent of each other and thus the dimensions associated with the concepts need not be orthogonal.

Rotation is computationally defined as the maximization of simplicity or minimization of complexity of $\mathbf{AT}$ subject to $\mathbf{T}'\mathbf{T} = \mathbf{I}_c$ or $diag(\mathbf{T}^{-1}\mathbf{T}'^{-1}) = \mathbf{I}_c$. Denote that $rot(\mathbf{AT})$ is the function of $\mathbf{T}$ that stands for how well $\mathbf{A}_T = \mathbf{AT}$ approximates the ideal simple structure. A number of rotation techniques that are aimed for obtaining the simple structure have been proposed. These techniques differ in terms of how to define the function $rot(\mathbf{AT})$. In this thesis, several popular criteria that define $rot(\mathbf{AT})$ are introduced. It should be noted that various rotation techniques exist, but it is difficult to have the ideal simple structure using the rotation because $\mathbf{T}$ cannot be chosen so that some elements of $\mathbf{A}_T$ are exactly zero. It is possible to acquire the approximate simple structure by the rotation (Adachi, 2016).

3.2.1 Crawford-Ferguson criterion

Crawford and Ferguson (1970) proposed the criterion that is a weighted sum of a measure of complexity of the rows and columns. Denote the elements of $\mathbf{AT}$ as at_{ij}. Then, the criterion can be expressed as

$$(1-\kappa)\sum_{i=1}^{p}\sum_{j=1}^{c}\sum_{l\neq j}^{c} at_{ij}^2 at_{il}^2 + \kappa\sum_{j=1}^{c}\sum_{i=1}^{p}\sum_{k\neq i}^{p} at_{ij}^2 at_{kj}^2, \tag{3.7}$$

where κ is a parameter that the users define before the rotation ($0 \leq \kappa \leq 1$). The first term in (3.7) expresses the row complexity and the second term in (3.7) does the column complexity. The degree of row and column complexities is controlled by κ. Crawford-Ferguson criterion includes the other rotation criteria. In orthogonal rotation, Crawford-Ferguson criterion is equivalent to orthomax criterion. Various orthogonal rotation criteria can be expressed by changing the parameter κ. The table 3.3 shows the relationships between Crawford-Ferguson criterion and orthomax criterion. Among orthomax criteria, varimax rotation is widely used in PCA and/or EFA.

3.2.2 Simplimax rotation

In Crawford-Ferguson criterion, complexity is defined mathematically. We introduce another technique called Procrustes rotation. Procrustes rotation refers to a class of rotation technique to rotate loadings so that resulting $\mathbf{A}_T$ is matched with a target matrix. In Procrustes rotation, the target matrix is assumed to have the ideal or approximate simple structure and minimize the

Table 3.3: The relationships Crawford-Ferguson criterion and orthomax criterion

	κ
Quartimax	0
Biquartimax	1/2p
Varimax	1/p
Equamax	c/2p
Parsimax	c-1/p+c-2
Factor parsimony	1

differences between the target matrix and $\mathbf{A}_T$. Let the target matrix be $\mathbf{B}$ ($p \times c$). Procrustes rotation can be formulated as

$$||\mathbf{B} - \mathbf{AT}||^2 = ||\mathbf{B} - \mathbf{A}_T||^2. \tag{3.8}$$

The function (3.8) is minimized over $\mathbf{T}$ subject to (3.3) or (3.4). Various Procrustes rotation techniques for obtaining the simple structure have been proposed such as promax rotation (Hendrickson and White, 1964) or simplimax rotation (Kiers, 1994). A target matrix in promax rotation does not guarantee to have the simple structure. Thus, simplimax rotation is introduced among the various Procrustes rotation techniques in this thesis.

The loss function of simplimax rotation can be formulated as the minimization of the function (3.8) with the constraints that the target matrix has l zero elements and $(pc - l)$ arbitrary elements, but position of zeros are unknown. The problem to be solved by simplimax rotation is that of finding the truly simple target matrix with l zero elements that can be approximated best by $\mathbf{A}_T$. An example of the target matrix is shown in the table 3.4.

To handle the minimization problem in simplimax rotation, we introduce $p \times c$ binary indicator matrix $\mathbf{W}_{ind}$ which has l zero elements at the position of the l zeros, and unit elements elsewhere. That is,

$$w_{ind_{ik}} = \begin{cases} 0 & \text{(if } (i, j) \text{ th element in the target matrix is assigned to be zero)} \\ 1 & \text{(otherwise)} \end{cases} \tag{3.9}$$

Then, the loss function to be minimized in simplimax rotation can be rewritten as

$$SR(\mathbf{B}, \mathbf{T}, \mathbf{W}_{ind}) = ||\mathbf{W}_{ind} \bullet \mathbf{B} - \mathbf{AT}||^2. \tag{3.10}$$

Table 3.4: An example of target matrix in simplimax rotation

Comp/Factor 1	Comp/Factor 2	Comp/Factor 3
?	0	0
?	0	0
?	0	0
?	0	0
0	**?**	0
0	**?**	0
0	**?**	0
0	0	**?**
0	0	**?**
0	0	**?**

Here, $\bullet$ expresses operator of Hadamard product (element-wise product). The loss function (3.10) is expanded in element-wise and we have

$$\begin{aligned} SR(\mathbf{B}, \mathbf{T}, \mathbf{W}_{ind}) &= \sum_{i=1}^{p}\sum_{j=1}^{c}(w_{ind_{ij}}b_{ij} - at_{ij})^2 \\ &= \sum_{i=1}^{p}\sum_{j=1}^{c}(1 - w_{ind_{ij}})at_{ij}^2 + \sum_{i=1}^{p}\sum_{j=1}^{c} w_{ind_{ij}}(at_{ij} - b_{ij})^2. \end{aligned} \tag{3.11}$$

In the function (3.11), the first term pertains to the elements of the target matrix **B** that are contraints to be zero, and the second term petains to the unconstrained elements of the target matrix **B**. By taking $b_{ij} = at_{ij}$ if $w_{ind_{ij}} = 1$ and $b_{ij} = 0$ if $w_{ind_{ij}} = 0$ for $i = 1, \ldots, p,\ j = 1, \ldots, c$, then **B** can be eliminated from the loss function. Thus, the problem reduces to minimize

$$\sum_{i=1}^{p}\sum_{j=1}^{c}(1 - w_{ind_{ij}})at_{ij}^2, \tag{3.12}$$

subject to the constraints (3.3) or (3.4). The minimization can be achieved by alternately minimize the loss function (3.12) over **T** given $\mathbf{W}_{ind}$ and over $\mathbf{W}_{ind}$ given **T**. The minimization over **T** given $\mathbf{W}_{ind}$ can be attained by solving the orthogonal or oblique Procrustes analysis (Browne, 1972a,b; Jennrich, 2001, 2002; ten Berge and Nevel, 1977). Next, the minimization over $\mathbf{W}_{ind}$ given **T** is considered. Because $w_{ind_{ij}}$ is restricted to be binary, the minimum is found by

$$w_{ind_{ij}} = \begin{cases} 0 & (\text{if } at_{ij}^2 \leq at_{[l]}^2) \\ 1 & (\text{otherwise}), \end{cases} \tag{3.13}$$

where $at^2_{[l]}$ is the l smallest values of at^2_{ij}.

The estimation algorithm of simplimax can be summarized as follows:

Step 1: Set the number of zeros l.

Step 2: Initialize the parameters.

Step 3: Minimize over $\mathbf{T}$ given $\mathbf{W}_{ind}$, which is achieved by solving the orthogonal or oblique Procrustes problem.

Step 4: Minimize over $\mathbf{W}_{ind}$ given $\mathbf{T}$, which is attained by (3.13).

Step 5: Finish if convergence is reached; otherwise, back to Step 3.

Note that the algorithm is rather sensitive to hitting the local minima (Kiers, 1994). Therefore, it is advised to use many different ramdom starts for $\mathbf{T}$ in the initialization.

3.3 Sparse technique

Rotation is one of useful and widely-used technique for obtaining the simple structure. However, it is difficult to have the ideal simple structure in rotation. Recently, alternative methods for acquiring the ideal simple structure have been proposed. These methods are called as sparse PCA and sparse EFA, respectively. In the methods, loadings are estimated by imposing the constraints that they must have exactly zero elements. A major approach to obtain sparse PCA solutions is by using penalty function (Shen and Huang, 2008; Trendafilov, 2014; Zou *et al.*, 2006). This has also been employed in sparse EFA estimation (Choi *et al.*, 2010; Hirose and Yamamoto, 2014, 2015; Ning and Georgiou, 2011). In this type of sparse EFA, the penalty function is combined with maximum likelihood EFA objective function. In MDFA literature, another type of sparse EFA estimation method has been proposed, in which parameters can be obtained under the constraint that the loading matrix must have l zero elements: the loss function $||\mathbf{X} - \mathbf{F}\mathbf{A}' - \mathbf{U}\boldsymbol{\Psi}||^2$ is minimized over the model parameters, imposing the model parameters constraints and

$$Card(\mathbf{A}) = l, \tag{3.14}$$

where $Card(\mathbf{A})$ expresses that $\mathbf{A}$ has the exact l zero elements. The number of zeros l is the pre-specified integer by users. This type of sparse method is called as unpenalized sparse factor analysis (Adachi and Trendafilov, 2014). The unpenalized sparse estimation approach can be applied to sparse PCA estimation (Adachi and Trendafilov, 2016).

It is easy to control the degree of sparseness in the unpenalized sparse methods; the unpenalized sparse PCA and sparse EFA are convenient for users who wish to have a loading matrix with a specified number of zero elements. Furthermore, unpenalized sparse methods are compatible with matrix decomposition approaches. In the following subsections, unpenalized sparse PCA and EFA are introduced.

3.3.1 Unpenalized sparse principal component analysis

Unpenalized sparse loading PCA (USLPCA) is defined as the minimization of the loss function

$$||\mathbf{X} - \mathbf{F}\mathbf{A}'||^2, \tag{3.15}$$

subject to the constraints (2.16) and (3.14). Adachi and Trendafilov (2016) pointed out that the loss function (3.15) can be rewriten as

$$||\mathbf{X} - \mathbf{F}\tilde{\mathbf{A}}'||^2 + n||\tilde{\mathbf{A}} - \mathbf{A}||^2, \tag{3.16}$$

with $\tilde{\mathbf{A}}$ being the cross-product matrix of $p \times c$:

$$\tilde{\mathbf{A}} = n^{-1}\mathbf{X}'\mathbf{F}. \tag{3.17}$$

The decomposition above shows that the second term in (3.16) is only relevant to $\mathbf{A}$. Thus, the minimization with respect to $\mathbf{A}$ reduces to that of the second term $||\tilde{\mathbf{A}} - \mathbf{A}||^2$ for a fixed $\mathbf{F}$.

The solutions for USLPCA can be obtained by alternately iterating the updates of $\mathbf{F}$ and $\mathbf{A}$. Firstly, the minimization over $\mathbf{F}$ with a fixed $\mathbf{A}$ is discussed. The loss function of USLPCA can be expanded as

$$||\mathbf{X} - \mathbf{F}\mathbf{A}'||^2 = tr(\mathbf{X}'\mathbf{X}) - 2tr(\mathbf{X}'\mathbf{F}\mathbf{A}') + n\mathbf{A}\mathbf{A}'. \tag{3.18}$$

Thus, the minimization of $\mathbf{F}$ with a fixed $\mathbf{A}$ is equivalent to the maximization of $tr(\mathbf{X}'\mathbf{F}\mathbf{A}')$. This maximization can be attained for

$$\mathbf{F} = \sqrt{n}\mathbf{K}_c\mathbf{L}'_c, \tag{3.19}$$

where $\mathbf{K}_c$ is the first c left singular vectors and $\mathbf{L}_c$ is the first c right singular vectors which are given by the SVD of $\mathbf{XA}$.

Then, the minimization of $\mathbf{A}$ with a fixed $\mathbf{F}$ is considered. The loss fucntion (3.16) can be expressed in element-wise as

$$\sum_{(i,j)\in O} \tilde{a}_{ij}^2 + \sum_{(i,j)\in \bar{O}} (\tilde{a}_{ij} - a_{ij})^2, \tag{3.20}$$

where O denotes the set of the l indexes (i, j)'s indicating the locations of the loadings to be zero. The compliment set $\bar{O}$ is the set containing the $pc - l$ indexes (i, j)'s of the nonzero loadings. It is obvious that the following inequality holds:

$$\sum_{(i,j)\in O} \tilde{a}_{ij}^2 + \sum_{(i,j)\in \bar{O}} (\tilde{a}_{ij} - a_{ij})^2 \geq \sum_{(i,j)\in O} \tilde{a}_{ij}^2. \tag{3.21}$$

When a_{ij} with $(i, j) \in \bar{O}$ are taken equal to corresponding $\tilde{a}_{ij}$, then (3.21) attains its lower limit $\sum_{(i,j)\in O} \tilde{a}_{ij}^2$. That is, we have optimal $\mathbf{A}$ as

$$a_{ij} = \begin{cases} 0 & (\text{iff } \tilde{a}_{ij}^2 \leq \tilde{a}_l^2) \\ \tilde{a}_{ij} & (\text{otherwise}) \end{cases} \tag{3.22}$$

Here, $\tilde{a}_l^2$ expresses the l smallest values of $\tilde{a}_{ij}^2$.

The whole algorithm of USLPCA can be summarized as follow:

Step 1: Pre-specify the number of zero elements $card(A) = l$ and the number of components.

Step 2: Initialize the parameters.

Step 3: Update $\mathbf{F} = \sqrt{n}\mathbf{K}_c\mathbf{L}'_c$ and $\tilde{\mathbf{A}} = n^{-1}\mathbf{X}'\mathbf{F}$.

Step 4: Update $\mathbf{A}$ with (3.22).

Step 5: Finish if convergence is reached; otherwise, back to Step 3.

3.3.2 Unpenalized sparse exploratory factor analysis

Next, unpenalized sparse factor analysis is considered. Adachi and Trendafilov (2014) proposed sparse orthogonal factor analysis (SOFA) as unpenalized sparse factor analysis. In SOFA, the loss function

$$SOFA(\mathbf{F}, \mathbf{A}, \mathbf{U}, \mathbf{\Psi}) = ||\mathbf{X} - \mathbf{F}\mathbf{A}' - \mathbf{U}\mathbf{\Psi}||^2 \tag{3.23}$$

is minimized over $\mathbf{F}$, $\mathbf{A}$, $\mathbf{U}$ and $\mathbf{\Psi}$, subject to the constraints (2.41), (2.42), (2.43), (2.44) and (3.14). In the same manner as USLPCA, the loss function of the unpenalized sparse factor analysis can be rewritten as

$$||\mathbf{X} - \mathbf{F}\mathbf{A}' - \mathbf{U}\mathbf{\Psi}||^2 = ||\mathbf{X} - \mathbf{F}\tilde{\mathbf{A}}' - \mathbf{U}\mathbf{\Psi}||^2 + n||\mathbf{A} - \tilde{\mathbf{A}}||^2, \tag{3.24}$$

where $\tilde{\mathbf{A}}$ in (3.24) is defined as $\tilde{\mathbf{A}} = n^{-1}\mathbf{X}'\mathbf{F}$.

The reexpression (3.24) indicates that the minimizations over $\mathbf{F}$, $\mathbf{U}$ and $\mathbf{\Psi}$ are equlivalent to those of MDFA. Thus, optimal solutions can be obtained by

$$\mathbf{S} = [\mathbf{F}, \mathbf{U}] = \sqrt{n}\mathbf{K}\mathbf{L}' = \sqrt{n}\mathbf{K}_1\mathbf{L}'_1 + \sqrt{n}\mathbf{K}_2\mathbf{L}'_2, \tag{3.25}$$

$$\mathbf{\Psi} = n^{-1}diag(\mathbf{X}'\mathbf{U}). \tag{3.26}$$

Here, $\mathbf{K} = [\mathbf{K}_1, \mathbf{K}_2]$ and $\mathbf{L} = [\mathbf{L}_1, \mathbf{L}_2]$ are left and right singular vectors and given by the SVD of $\mathbf{X}[\mathbf{A}, \mathbf{\Psi}]$. Note that the loading matrix $\mathbf{A}$ is constrainted to have l zero elements and the SVD of $\mathbf{X}[\mathbf{A}, \mathbf{\Psi}]$ differs from ordinary MDFA estimation.

Lastly, the minimization over $\mathbf{A}$ is considered. The same estimation algorithm in USLPCA can be used because the second term in (3.24) is only relevant to $\mathbf{A}$: the optimal $\mathbf{A}$ can be given by (3.22).

The whole algorithm of unpenalized sparse factor analysis is described as follows:

Step 1: Pre-specify the number of zero elements $card(A) = l$ and the number of factors.

Step 2: Initialize the parameters.

Step 3: Update $\mathbf{F}$ and $\mathbf{U}$ with (3.25).

Step 4: Update $\mathbf{\Psi}$ with (3.26).

Step 5: Update $\mathbf{A}$ with (3.22).

Step 6: Finish if convergence is reached; otherwise, back to Step 3.

3.3.3 Fitness of sparse PCA and MDFA

Recall that the loss functions of sparse PCA and MDFA, and the following inequalities hold:

$$||\mathbf{X} - \mathbf{F}\tilde{\mathbf{A}}'||^2 + n||\tilde{\mathbf{A}} - \mathbf{A}||^2 \geq ||\mathbf{X} - \mathbf{F}\tilde{\mathbf{A}}'||^2, \tag{3.27}$$

$$||\mathbf{X} - \mathbf{F}\tilde{\mathbf{A}}' - \mathbf{U\Psi}||^2 + n||\mathbf{A} - \tilde{\mathbf{A}}||^2 \geq ||\mathbf{X} - \mathbf{F}\tilde{\mathbf{A}}' - \mathbf{U\Psi}||^2. \tag{3.28}$$

The lower bounds are attained by $\tilde{\mathbf{A}} = \mathbf{A} = n^{-1}\mathbf{X}'\mathbf{F}$. Here, right-handed side of inequalities (3.27) and (3.28) are equivalent to ordinary PCA and MDFA, respectively. It indicates that ordinary PCA and MDFA fit better than sparse ones. The solutions that have ideal simple structure can be obtained by sparse techniques, but the fitness of sparse techniques is not greater than that of the rotation.

According to (3.27) and (3.28), we have the following inequality:

$$ntr(\mathbf{A}'\mathbf{A}) \leq ntr(\tilde{\mathbf{A}}'\tilde{\mathbf{A}}) \Longleftrightarrow ||\mathbf{F}\mathbf{A}'||^2 \leq ||\mathbf{F}\tilde{\mathbf{A}}'||^2. \tag{3.29}$$

Here, $||\mathbf{F}\mathbf{A}'||^2$ and $||\mathbf{F}\tilde{\mathbf{A}}'||^2$ expresses variances in common parts. It should be noted that the inequality (3.29) holds in both of PCA and MDFA. Therefore, common variances in ordinary PCA and MDFA are always greater than or equal to those one for sparse PCA and MDFA.

4

Interpretability in component/factor scores of individuals

In chapter 3, interpretability in loading matrix was considered. We discuss interpretability in component and common factor scores of individuals in this chapter. Usually, a number of individuals are observed in human sciences, and it may be time-consuming to interpret all of individual scores. There are various possible ways to tackle its difficulty. Cluster analysis or simply called clustering is introduced among them in this thesis. Clustering is the procedure for classifying individuals into groups (cluster) so that similar individuals are classified into the same group and dissimilar ones are assigned to different groups. Using clustering, the users do not necessarily interpret all of the observed individuals and the problem reduces to the interpretation of clusters. Thus, clustering is used as the method for facilitating the interpretability of individuals in this thesis. There are various procedures for performing clustering. Among these procedures, one of the most popular method, called k-means clustering (MacQueen, 1967), is introduced.

4.1 Clustering the objects by k-means

The model of k-means can be described as

$$\mathbf{X} = \mathbf{PC} + \mathbf{E}, \tag{4.1}$$

where $\mathbf{X}$ is a data matrix (n-individuals $\times$ p-variables), $\mathbf{P}$ is a membership matrix (n-observations $\times$ γ-clusters), and a cluster centroid matrix which expresses how each cluster is characterized by variables (γ-clusters $\times$ p-variables). The membership matrix is assumed to be satisfied with

$$p_{ik} = \begin{cases} 1 & \text{(if individual } i \text{ belongs to cluster } k\text{)} \\ 0 & \text{(otherwise)} \end{cases} \tag{4.2}$$

and

$$\mathbf{G1}_{\gamma} = \mathbf{1}_n. \tag{4.3}$$

These constraints suggest that each individual belongs to only one cluster.

In k-means clustering, the number of cluster γ is pre-specified by users before the anlaysis. The loss fucntion of k-means is defined as

$$KMC(\mathbf{P}, \mathbf{C}) = ||\mathbf{X} - \mathbf{PC}||^2. \tag{4.4}$$

The function (4.4) is minimized over **P** and **C** subject to (4.2) and (4.3). The loss function (4.4) is expressed as

$$\begin{aligned} KMC(\mathbf{P}, \mathbf{C}) &= tr(\mathbf{X}'\mathbf{X} - 2\mathbf{X}'\mathbf{PC} - \mathbf{C}'\mathbf{P}'\mathbf{PC}) \\ &= tr(\mathbf{X}'\mathbf{X} - 2\mathbf{X}'\mathbf{PC} - \mathbf{C}'\mathbf{P}'\mathbf{PC} + \mathbf{X}'\mathbf{P}(\mathbf{P}'\mathbf{P})^{-1/2}(\mathbf{P}'\mathbf{P})^{-1/2}\mathbf{P}'\mathbf{X} \\ &- \mathbf{X}'\mathbf{P}(\mathbf{P}'\mathbf{P})^{-1/2}(\mathbf{P}'\mathbf{P})^{-1/2}\mathbf{P}'\mathbf{X}) \\ &= ||(\mathbf{P}'\mathbf{P})^{1/2}\mathbf{C} - (\mathbf{P}'\mathbf{P})^{-1/2}\mathbf{P}'\mathbf{X}||^2 + const_C. \end{aligned} \tag{4.5}$$

Here, $const_C$ is a constant, which is irrelevant to **C**. Thus, the optimal **C** is obtained by

$$\mathbf{C} = (\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'\mathbf{X}. \tag{4.6}$$

Let us consider (4.6) in more detail. $\mathbf{P}'\mathbf{P}$ becomes a diagonal matrix whose kth diagonal elements is the numner of individuals belonging to cluster k because of the constraints (4.2) and (4.3). This fact indicates that elements of **C** is the mean of the data within cluster for each variable:

$$c_{kj} = \frac{1}{n_k} \sum_{i \in \text{cluster } k} x_{ij}, \tag{4.7}$$

where n_k is the number of individuals in cluster k.

Next, the minimization over **P** is considered. Let $\mathbf{x}_{(i)}$ be the i-th row vector of **X** and $\mathbf{p}_{(i)}$ be the i-th row vector of **P**. Then, the loss function (4.4) can be rewriten as

$$||\mathbf{X} - \mathbf{PC}||^2 = \sum_{i=1}^{n} ||\mathbf{x}_{(i)} - \mathbf{p}_{(i)}\mathbf{C}||^2. \tag{4.8}$$

As can be seen in (4.8), the optimal $\mathbf{p}_{(i)}$ can be obtained by minimizing

$$||\mathbf{x}_{(i)} - \mathbf{p}_{(i)}\mathbf{C}||^2, \tag{4.9}$$

over $\mathbf{p}_{(i)}$, subject to (4.2) and (4.3) with a fixed **C**. It implies that optimal **P** can be given by minimizing (4.9) over $\mathbf{p}_{(i)}$ for each i ($i = 1, \ldots, p$).

Because of the constraints (4.2) and (4.3), $\mathbf{p}_{(i)}$ is filled with zeros except for one element taking 1. It indicates that (4.9) can be reexpressed as

$$||\mathbf{x}_{(i)} - \mathbf{c}_{(k)}||^2, \tag{4.10}$$

when individual i belongs to cluster k. Here, $\mathbf{c}_{(k)}$ denotes the k-th row of $\mathbf{C}$. That is, the function (4.9) takes one of γ distinct values:

$$||\mathbf{x}_{(i)} - \mathbf{p}_{(i)}\mathbf{C}||^2 = \begin{cases} ||\mathbf{x}_{(i)} - \mathbf{c}_{(1)}||^2 & \text{if } \mathbf{p}_{(i)} = [1, 0, \ldots, 0] \\ ||\mathbf{x}_{(i)} - \mathbf{c}_{(2)}||^2 & \text{if } \mathbf{p}_{(i)} = [0, 1, \ldots, 0] \\ \vdots & \\ ||\mathbf{x}_{(i)} - \mathbf{c}_{(\gamma)}||^2 & \text{if } \mathbf{p}_{(i)} = [0, 0, \ldots, 1] \end{cases} \tag{4.11}$$

Therefore, the miminization over $\mathbf{p}_{(i)}$, subject to (4.2) and (4.3) with a fixed $\mathbf{C}$ reduces to choose the vector $\mathbf{p}_{(i)}$ corresponding the minimal one among (4.11). This selection can be summarized as

$$p_{ik} = \begin{cases} 1 & \text{if } ||\mathbf{x}_{(i)} - \mathbf{c}_{(k)}||^2 \text{attains the minimum among (4.11)} \\ 0 & \text{otherwise.} \end{cases} \tag{4.12}$$

The whole algorithm of k-means clustering can be described as follow:

Step 1: Set the number of cluster γ.

Step 2: Initialize the parameters.

Step 3: Obtain the cluster centroid matrix $\mathbf{C}$ with (4.6).

Step 4: Update indicator matrix $\mathbf{P}$ with (4.12).

Step 5: Finish if convergence is reached; otherwise, back to Step 3.

4.2 Joint dimension reduction and clustering

k-means clustering is the useful method, but it is often difficult to be graphically represented when a number of variables are observed. To tackle such cases, dimension reduction methods such as PCA or EFA are sometimes applied to observed data; reducing the number of variables is expected to simplify the interpretation of the features and be graphically represented in a low-dimensional space (figure 4.1).

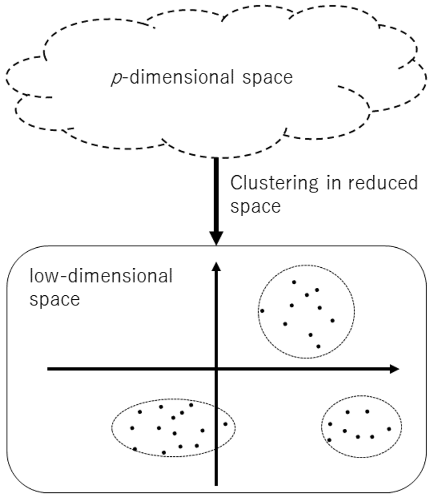


Figure 4.1: Clustering in reduced dimension

A commonly used approach is to apply dimension reduction techniques firstly to obtain a low dimensional solutions, and then to perform clustering participants. This is called as tandem analysis or tandem approach. Although it is easy for practitioners to use tandem approach, it is pointed out that the tandem approach may not provide the correct taxonomic information (Arabie and Hubert, 1994; De Soete and Caroll, 1994; Timmerman, Ceulmans, Kiers and Vichi, 2010; Vichi and Kiers, 2001). To address the problem in tandem analysis, dimension reduction and clustering are performed jointly. This is formulated as combining the loss function of dimension reduction with k-means clustering, in which a low-dimensional representation of variables is obtained in such a way that it also contains the useful information for clustering participants. The differences between tandem and joint analysis is summarized in figure 4.2.

There are various approaches to perform dimension reduction and clustering jointly such as reduced k-means (De Soete and Caroll, 1994), factorial k-means (Vichi and Kiers, 2001) and generalized reduced clustering (Yamamoto and Hwang, 2014). In these methods, PCA with the constraint (2.23) is used for dimension reduction and is combined with k-means. In EFA literature, Mitsuhiro and Yadohisa (2018) proposed simultaneous analysis method for reduction and

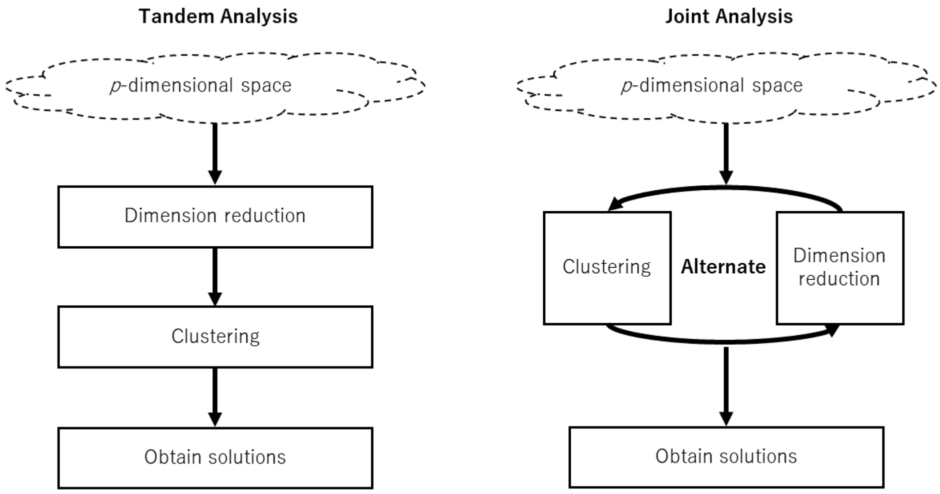


Figure 4.2: Flows of tandem and joint analysis

clustering (SARC), in which MDFA is combined with k-means and the parameters are simultaneously estimated. The loss function of SARC can be described as

$$\alpha_1 \|\mathbf{X} - \mathbf{H}\mathbf{\Omega}'\|^2 + \alpha_2 \|\mathbf{H} - \mathbf{P}\mathbf{C}\|^2. \tag{4.13}$$

When $\mathbf{H} = \mathbf{F}$ and $\mathbf{\Omega} = \mathbf{A}$, (4.13) can be considered as the loss function of joint dimension reduction via PCA and k-means clustering. Furthermore, set $\mathbf{H}$ as $\mathbf{S} = [\mathbf{F}, \mathbf{U}]$ and $\mathbf{\Omega}$ as $\mathbf{\Lambda} = [\mathbf{A}, \mathbf{\Psi}]$ in (4.13), then we have the loss function of joint MDFA and clustering. That is, SARC is a simultaneous framework that is composed of dimension reduction methods with matrix decomposition approaches integrated with k-means clustering.

4.2.1 Joint principal component analysis and k-means

Firstly, the loss function of joint PCA and k-means is defined. Combining the loss function of PCA with that of k-means, we have

$$\alpha_1 \|\mathbf{X} - \mathbf{F}\mathbf{A}'\|^2 + \alpha_2 \|\mathbf{F} - \mathbf{P}\mathbf{C}\|^2. \tag{4.14}$$

The loss function (4.14) is minimized over $\mathbf{F}$, $\mathbf{P}$ and $\mathbf{C}$ subject to

$$\mathbf{J}\mathbf{F} = \mathbf{F}, \;\; n^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_c, \tag{4.15}$$

$$p_{ik} = \begin{cases} 1 & \text{(if individual } i \text{ belongs to cluster } k\text{)} \\ 0 & \text{(otherwise)} \end{cases} \tag{4.16}$$

and

$$\mathbf{P1}_\gamma = \mathbf{1}_n. \tag{4.17}$$

Here, α_1 and α_2 are weights that are satisfied with

$$\alpha_1 + \alpha_2 = 1. \tag{4.18}$$

These weights are pre-specified by users. If they judges that classification is more important than dimension reduction, the weights are set to $\alpha_1 < \alpha_2$, and vice versa.

The minimizing (4.14) with respect to **A** can be attained by

$$\mathbf{A} = n^{-1}\mathbf{X}'\mathbf{F}, \tag{4.19}$$

because **A** is only relevant to the first term in (4.14) and it indicates that the optimal solution can be given in the same manner as the PCA estimation procedure. The minimization of the loss function over **P** and **C** is considered. The second term is only associated with these parameters. Thus, optimal **P** and **C** can be obtained using the k-means estimation procedure:

$$\mathbf{C} = (\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'\mathbf{F}, \tag{4.20}$$

and

$$p_{ik} = \begin{cases} 1 & \text{if } ||\mathbf{f}_{(i)} - \mathbf{c}_{(k)}||^2 \text{attains the minimum among (4.14)} \\ 0 & \text{otherwise.} \end{cases} \tag{4.21}$$

Lastly, the miminization over **F** is considered. The loss function (4.14) can be expanded as

$$\begin{aligned} &\alpha_1\{tr(\mathbf{X}'\mathbf{X} - 2\mathbf{X}'\mathbf{FA}' + \mathbf{AF}'\mathbf{FA}')\} + \alpha_2\{tr(\mathbf{F}'\mathbf{F} - 2\mathbf{F}'\mathbf{PC} + \mathbf{C}'\mathbf{P}'\mathbf{PC})\} \\ &= -\alpha_1 tr(\mathbf{F}'\mathbf{XA}) - \alpha_2 tr(\mathbf{F}'\mathbf{PC}) + c_{JDR_F} \\ &= -tr\{\mathbf{F}'(\alpha_1\mathbf{XA} + \alpha_2\mathbf{PC})\} + c_{JDR_F}, \end{aligned} \tag{4.22}$$

where c_{JDR_F} is a constant, which is irrelevant to **F**. Thus, the minimization of (4.14) over **F** reduces to the maximization of (4.22) over **F** subject to (4.15). This maximization can be achieved by

$$\mathbf{F} = \hat{\mathbf{K}}_c\hat{\mathbf{L}}'_c, \tag{4.23}$$

where the SVD of $\alpha_1\mathbf{XA} + \alpha_2\mathbf{PC}$ is defined as

$$\alpha_1\mathbf{XA} + \alpha_2\mathbf{PC} = \hat{\mathbf{K}}\hat{\mathbf{\Delta}}\hat{\mathbf{L}}'. \tag{4.24}$$

The whole algorithm in joint PCA and k-means is described as follows:

Step 1: Set the number of components c and that of cluster γ.

Step 2: Initialize the parameters .

Step 3: Update $\mathbf{F}$ with (4.23).

Step 4: Update $\mathbf{A}$ with (4.19).

Step 5: Update $\mathbf{P}$ with (4.21).

Step 6: Update $\mathbf{C}$ with (4.20).

Step 7: Finish if convergence is reached; otherwise, back to Step 3.

4.2.2 Joint matrix decomposition factor analysis and k-means

4.2.2.1 Mitsuhiro and Yadohisa (2018) method

Next, the loss function of joint MDFA and k-means is considered. In Mitsuhiro and Yadohisa (2018), the loss function of SARC is defined as

$$\alpha_1\|\mathbf{X} - \mathbf{FA}' - \mathbf{U\Psi}\|^2 + \alpha_2\|[\mathbf{F},\ \mathbf{U}] - \mathbf{PC}\|^2 = \alpha_1\|\mathbf{X} - \mathbf{S\Lambda}'\|^2 + \alpha_2\|\mathbf{S} - \mathbf{PC}\|^2, \tag{4.25}$$

which is minimized over $\mathbf{S}$, $\mathbf{\Lambda}$, $\mathbf{P}$ and $\mathbf{C}$, subject to the constraints (4.16), (4.17), (4.18),

$$n^{-1}\mathbf{S}'\mathbf{S} = \mathbf{I}_{(p+c)},\quad \mathbf{S} = \mathbf{JS}, \tag{4.26}$$

and

$$\mathbf{\Psi} \text{ is a diagonal matrix.} \tag{4.27}$$

The term $||\mathbf{X} - \mathbf{S}\mathbf{\Lambda}'||^2$ is only relevant to $\mathbf{\Lambda}$, and thus minimization of the loss function (4.25) over Λ with fixed other parameters can be attained by

$$\mathbf{A} = n^{-1}\mathbf{X}'\mathbf{F}, \tag{4.28}$$

$$\mathbf{\Psi} = n^{-1} diag(\mathbf{X}'\mathbf{U}). \tag{4.29}$$

Next, the minimizations over $\mathbf{P}$ and $\mathbf{C}$ are discussed. The term $||\mathbf{S} - \mathbf{PC}||^2$ is only relevant to $\mathbf{P}$ and $\mathbf{C}$. Thus, optimal $\mathbf{P}$ and $\mathbf{C}$ are obtained by

$$\mathbf{C} = (\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'\mathbf{S}, \tag{4.30}$$

and

$$p_{ik} = \begin{cases} 1 & \text{if } ||\mathbf{s}_{(i)} - \mathbf{c}_{(k)}||^2 \text{attains the minimum among (4.25)} \\ 0 & \text{otherwise.} \end{cases} \tag{4.31}$$

Finally, the minimization over $\mathbf{S}$ is considered. The loss fucntion (4.25) can be expanded as

$$\begin{aligned} &\alpha_1\{tr(\mathbf{X}'\mathbf{X} - 2\mathbf{X}'\mathbf{S}\mathbf{\Lambda}' + \mathbf{\Lambda}\mathbf{S}'\mathbf{S}\mathbf{\Lambda}')\} + \alpha_2\{tr(\mathbf{S}'\mathbf{S} - 2\mathbf{S}'\mathbf{PC} + \mathbf{C}'\mathbf{P}'\mathbf{PC})\} \\ &= -\alpha_1 tr(\mathbf{S}'\mathbf{X}\mathbf{\Lambda}) - \alpha_2 tr(\mathbf{S}'\mathbf{PC}) + c_{JDRMDFA_S} \\ &= -tr\{\mathbf{S}'(\alpha_1\mathbf{X}\mathbf{\Lambda} + \alpha_2\mathbf{PC})\} + c_{JDRMDFA_S}, \end{aligned} \tag{4.32}$$

where $c_{JDRMDFA_S}$ is a constant, which is irrelevant to $\mathbf{S}$. The optimal $\mathbf{S}$ can be given by

$$\mathbf{S} = \dot{\mathbf{K}}_{(c+p)}\dot{\mathbf{L}}'_{(c+p)}, \tag{4.33}$$

because the minimization over $\mathbf{S}$ reduces to the maximization over $\mathbf{S}$. Here, $\dot{\mathbf{K}}_{(c+p)}$ and $\dot{\mathbf{L}}_{(c+p)}$ are left and right singular vectors of $(\alpha_1\mathbf{X}\mathbf{\Lambda} + \alpha_2\mathbf{PC})$.

The alternating least square algorithm of joint MDFA and k-means is described as

Step 1: Set the number of components c and that of cluster γ.

Step 2: Initialize the parameters.

Step 3: Update $\mathbf{S}$ with (4.33).

Step 4: Update $\mathbf{\Lambda}$ with (4.28) and (4.29).

Step 5: Update $\mathbf{P}$ with (4.31).

Step 6: Update $\mathbf{C}$ with (4.30).

Step 7: Finish if convergence is reached; otherwise, back to Step 3.

4.2.2.2 Modified joint matrix decomposition factor analysis and k-means

In (4.25), $\mathbf{P}$ and $\mathbf{C}$ are fitted to $n \times (c + p)$ factor scores matrix. It indicates that it may be difficult to obtain well-clustered $\mathbf{F}$ because the loss function (4.25) is defined to estimate well-clustered $\mathbf{S}$. There is another drawback that the common factor and cluster centroid cannot be jointly displayed in a low-dimensional space. To address the problem, we propose a modified joint MDFA and k-means clustering method, in which the loss function is described as

$$\alpha_1||\mathbf{X} - \mathbf{S\Lambda}'||^2 + \alpha_2||\mathbf{F} - \mathbf{PC}||^2. \tag{4.34}$$

That is, second term in (4.25) is replaced to $||\mathbf{F} - \mathbf{PC}||^2$ so that well-clustered $\mathbf{F}$ can be estimated.

The loss function (4.34) can be expanded as

$$\begin{aligned}
&\alpha_1\{tr(\mathbf{X}'\mathbf{X} - 2\mathbf{X}'\mathbf{S\Lambda}' + \mathbf{\Lambda S}'\mathbf{S\Lambda}')\} + \alpha_2\{tr(\mathbf{F}'\mathbf{F} - 2\mathbf{F}'\mathbf{PC} + \mathbf{C}'\mathbf{P}'\mathbf{PC})\} \\
&= -\alpha_1 tr(\mathbf{F}'\mathbf{XA}) - \alpha_2 tr(\mathbf{F}'\mathbf{PC}) + c_{JDRMDFA_F} \\
&= -tr\{\mathbf{F}'(\alpha_1\mathbf{XA} + \alpha_2\mathbf{PC})\} + c_{JDRMDFA_F}.
\end{aligned} \tag{4.35}$$

Here, $c_{JDRMDFA_F}$ is a constant, which is irrelevant to $\mathbf{F}$. Recall that $\mathbf{F}$ must satisfy with the constraint $\mathbf{F}'\mathbf{U} = \mathbf{O}$. Then, $\mathbf{F}$ can be re-expressed as

$$\mathbf{F} = (\mathbf{I} - \mathbf{P}_u)\mathbf{F}, \tag{4.36}$$

where $\mathbf{P}_u = \mathbf{U}(\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'$. In this re-expression, we have $\mathbf{F}'(\mathbf{I}-\mathbf{P}_u)\mathbf{U} = \mathbf{F}'\mathbf{U} - \mathbf{F}'\mathbf{U}(\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'\mathbf{U} = \mathbf{F}'\mathbf{U} - \mathbf{F}'\mathbf{U} = \mathbf{O}$. Substituting (4.36) into (4.35), then the loss function can be expressed as

$$-tr\{\mathbf{F}'(\alpha_1(\mathbf{I}-\mathbf{P}_u)\mathbf{X}\mathbf{A} + \alpha_2(\mathbf{I}-\mathbf{P}_u)\mathbf{P}\mathbf{C})\} + c_{JDRMDFA_F}. \tag{4.37}$$

The minimization (4.37) can be achieved by the SVD of $(\alpha_1(\mathbf{I}-\mathbf{P}_u)\mathbf{X}\mathbf{A} + \alpha_2(\mathbf{I}-\mathbf{P}_u)\mathbf{P}\mathbf{C})$.

Next, the minimization over $\mathbf{U}$ is considered. The loss function can be described as

$$\begin{aligned}&\alpha_1\{tr(\mathbf{X}'\mathbf{X} - 2\mathbf{X}'\mathbf{S}\mathbf{\Lambda}' + \mathbf{\Lambda}\mathbf{S}'\mathbf{S}\mathbf{\Lambda}')\} + \alpha_2\{tr(\mathbf{F}'\mathbf{F} - 2\mathbf{F}'\mathbf{P}\mathbf{C} + \mathbf{C}'\mathbf{P}'\mathbf{P}\mathbf{C})\}\\ &= -\alpha_1 tr(\mathbf{U}'\mathbf{X}\mathbf{\Psi}) + c_{JDRMDFA_U}.\end{aligned} \tag{4.38}$$

$\mathbf{U}$ can also be reparametrized as

$$\mathbf{U} = (\mathbf{I}-\mathbf{P}_f)\mathbf{U}, \tag{4.39}$$

where $\mathbf{P}_f = \mathbf{F}(\mathbf{F}'\mathbf{F})^{-1}\mathbf{F}'$. The loss function can be described as

$$-\alpha_1 tr(\mathbf{U}'(\mathbf{I}-\mathbf{P}_f)\mathbf{X}\mathbf{\Psi}) + c_{JDRMDFA_U}, \tag{4.40}$$

where $c_{JDRMDFA_U}$ is a constant, which is irrelevant to $\mathbf{U}$. In the same manner as $\mathbf{F}$, the optimal $\mathbf{U}$ can be obtained by the SVD of $(\mathbf{I}-\mathbf{P}_f)\mathbf{X}\mathbf{\Psi}$.

In the loss function (4.34), $\mathbf{\Lambda}$, $\mathbf{P}$ and $\mathbf{C}$ can be estimated by using (4.28), (4.29), (4.31) and (4.30).

The alternating least square algorithm of joint MDFA and k-means is described as follows:

Step 1: Set the number of components c and that of cluster γ.

Step 2: Initialize the parameters.

Step 3: Update $\mathbf{F}$ by the SVD of $(\alpha_1(\mathbf{I}-\mathbf{P}_u)\mathbf{X}\mathbf{A} + \alpha_2(\mathbf{I}-\mathbf{P}_u)\mathbf{P}\mathbf{C})$.

Step 4: Update $\mathbf{U}$ by the SVD of $(\mathbf{I}-\mathbf{P}_f)\mathbf{X}\mathbf{\Psi}$.

Step 5: Update $\mathbf{\Lambda}$ with (4.28) and (4.29).

Step 6: Update $\mathbf{P}$ with (4.31).

Step 7: Update $\mathbf{C}$ with (4.30).

Step 8: Finish if convergence is reached; otherwise, back to Step 3.

4.3 Interpretability between variables and components/factors in joint dimension reduction and clustering

4.3.1 Rotational freedom in joint dimension reduction and clustering

The rotational freedom exists in (4.14) and (4.34). That is, the value of the loss function cannot be affected by nonsingular matrix **T**:

$$
\begin{aligned}
&\alpha_1||\mathbf{X}-\mathbf{F}\mathbf{A}'||^2+\alpha_2||\mathbf{F}-\mathbf{P}\mathbf{C}||^2\\
&=-\alpha_1 tr(\mathbf{F}'\mathbf{X}\mathbf{A}')-\alpha_2 tr(\mathbf{F}'\mathbf{P}\mathbf{C})+c_{JDR_F}\\
&=-\alpha_1 tr(\mathbf{F}'\mathbf{X}\mathbf{X}'\mathbf{F}(\mathbf{F}'\mathbf{F})^{-1})-n\alpha_2 tr(\mathbf{F}'\mathbf{P}(\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'\mathbf{F}(\mathbf{F}'\mathbf{F})^{-1})+c_{JDR_F}\\
&=-\alpha_1 tr(\mathbf{T}\mathbf{F}'\mathbf{X}\mathbf{X}'\mathbf{F}\mathbf{T}(\mathbf{T}'\mathbf{F}'\mathbf{F}\mathbf{T})^{-1})-n\alpha_2 tr(\mathbf{T}'\mathbf{F}'\mathbf{P}(\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'\mathbf{F}\mathbf{T}(\mathbf{T}'\mathbf{F}'\mathbf{F}\mathbf{T})^{-1})+c_{JDR_F}\\
&=-n^{-1}\alpha_1 tr(\mathbf{T}\mathbf{F}'\mathbf{X}\mathbf{X}'\mathbf{F}\mathbf{T}\mathbf{T}^{-1}\mathbf{T}'^{-1})-\alpha_2 tr(\mathbf{T}'\mathbf{F}'\mathbf{P}(\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'\mathbf{F}\mathbf{T}\mathbf{T}^{-1}\mathbf{T}'^{-1})+c_{JDR_F}\\
&=-n^{-1}\alpha_1 tr(\mathbf{F}'\mathbf{X}\mathbf{X}'\mathbf{F})-\alpha_2 tr(\mathbf{F}'\mathbf{P}(\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'\mathbf{F})+c_{JDR_F},
\end{aligned}
\tag{4.41}
$$

and

$$
\begin{aligned}
&\alpha_1||\mathbf{X}-\mathbf{F}\mathbf{A}'-\mathbf{U}\boldsymbol{\Psi}||^2+\alpha_2||\mathbf{F}-\mathbf{P}\mathbf{C}||^2\\
&=-\alpha_1 tr(\mathbf{F}'\mathbf{X}\mathbf{A}')-\alpha_2 tr(\mathbf{F}'\mathbf{P}\mathbf{C})+c_{JDRMDFA_F}\\
&=-\alpha_1 tr(\mathbf{F}'\mathbf{X}\mathbf{X}'\mathbf{F}(\mathbf{F}'\mathbf{F})^{-1})-n\alpha_2 tr(\mathbf{T}'\mathbf{F}'\mathbf{P}(\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'\mathbf{F}(\mathbf{F}'\mathbf{F})^{-1})+c_{JDRMDFA_F}\\
&=-n^{-1}\alpha_1 tr(\mathbf{T}\mathbf{F}'\mathbf{X}\mathbf{X}'\mathbf{F}\mathbf{T}\mathbf{T}^{-1}\mathbf{T}'^{-1})-\alpha_2 tr(\mathbf{T}'\mathbf{F}'\mathbf{P}(\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'\mathbf{F}\mathbf{T}\mathbf{T}^{-1}\mathbf{T}'^{-1})+c_{JDRMDFA_F}\\
&=-n^{-1}\alpha_1 tr(\mathbf{F}'\mathbf{X}\mathbf{X}'\mathbf{F})-\alpha_2 tr(\mathbf{F}'\mathbf{P}(\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'\mathbf{F})+c_{JDRMDFA_F}.
\end{aligned}
\tag{4.42}
$$

Thus, the loading matrix can be rotated towards simple structure in the same manner as PCA or MDFA.

4.3.2 Sparse technique in joint dimension reduction and clustering

As discussed in the previous subsection, the interpretable loading matrix can be estimated by rotation in joint PCA/MDFA and clustering. The sparse constraint $card(A) = l$ can be imposed on joint dimension reduction and clustering methods:

$$
\alpha_1||\mathbf{X}-\mathbf{F}\mathbf{A}'||^2+\alpha_2||\mathbf{F}-\mathbf{P}\mathbf{C}||^2=\alpha_1||\mathbf{X}-\mathbf{F}\tilde{\mathbf{A}}'||^2+\alpha_2||\mathbf{F}-\mathbf{P}\mathbf{C}||^2+\alpha_1 n||\mathbf{A}-\tilde{\mathbf{A}}||^2 \tag{4.43}
$$

and

$$\alpha_1 ||\mathbf{X} - \mathbf{S}\boldsymbol{\Lambda}'||^2 + \alpha_2 ||\mathbf{S} - \mathbf{PC}||^2 = \alpha_1 ||\mathbf{X} - \mathbf{F}\tilde{\mathbf{A}}' - \mathbf{U}\boldsymbol{\Psi}||^2 + \alpha_2 ||\mathbf{F} - \mathbf{PC}||^2 + \alpha_1 n ||\mathbf{A} - \tilde{\mathbf{A}}||^2. \quad (4.44)$$

These reformulations indicate that the parameters in (4.43) and (4.44) can be estimated by combining the algorithm of sparse method with that of joint dimension reduction and k-means method. Therefore, estimation algorithms joint sparse PCA/MDFA and k-means can be descreibed as follows:

Joint sparse PCA and k-means

Step 1: Set c, γ and $card(\mathbf{A}) = l$.

Step 2: Initialize the parameters.

Step 3: Update $\mathbf{F}$ with (4.23).

Step 4: Update $\mathbf{A}$ with (3.22).

Step 5: Update $\mathbf{P}$ with (4.21).

Step 6: Update $\mathbf{C}$ with (4.20).

Step 7: Finish if convergence is reached; otherwise, back to Step 3.

Joint sparse MDFA and k-means

Step 1: Set c, γ and $card(\mathbf{A}) = l$.

Step 2: Initialize the parameters.

Step 3: Update $\mathbf{S}$ with (4.33).

Step 4: Update $\mathbf{A}$ with (3.22).

Step 5: Update $\boldsymbol{\Psi}$ with (3.26).

Step 6: Update $\mathbf{P}$ with (4.21).

Step 7: Update $\mathbf{C}$ with (4.20).

Step 8: Finish if convergence is reached; otherwise, back to Step 3.

5

Extension of principal component analysis to nonmetric data

In chapter 5, we extend PCA and its related methods to handle nonmetric data via the quantification technique. It is also shown that nonmetric version of PCA is equivalent to multiple correspondence analysis (MCA).

5.1 Multiple correspondence analysis/Homogeneity analysis

MCA or sometime called as homogeneity analysis (HA) is a widely-used technique to analyze categorical data and aims to reduce large sets of variables into smaller sets of components that summarize the information contained in the data. The purpose of MCA is the same to that of PCA. There are various approaches to formulate an MCA (e.g. Beh and Lombardo, 2014; Gifi, 1990; Greenacre, 1988; Nishisato, 2006), but they have proved to give essentially equivalent solutions originating in different theoretical foundations (Tenenhaus and Young, 1985). Among these formulations, two fomulations are introduced. Denote $\mathbf{G} = [\mathbf{G}_1, \ldots, \mathbf{G}_p]$ as an n-individuals and K-categories indicator matrix of the observed categorical data, where K_i is the number of categories in the i-th variable and K is the sum of all categories in observed variables. The k-th element g_{ijk} in the i-th row $\mathbf{g}_{ij}$ is defined as

$$g_{ijk} = \begin{cases} 1 & \text{(if } i\text{-th individual choiced category } k \text{ in } j\text{-th variable)} \\ 0 & \text{(otherwise).} \end{cases} \tag{5.1}$$

It implies that the indicator matrices $\mathbf{G}_1, \ldots, \mathbf{G}_p$ are satisfied with

$$\mathbf{G}_j \mathbf{1}_{K_j} = \mathbf{1}_n, \tag{5.2}$$

$$\mathbf{G'}_j \mathbf{G}_j = \mathbf{D}_j, \tag{5.3}$$

where $\mathbf{D}_j$ is a diagonal matrix whose elements are category frequencies in j-th variable. We also denote $\mathbf{D} = b - diag(\mathbf{D}_1, \ldots, \mathbf{D}_p)$, where $b - diag$ is the operator of block diagonal matrix.

In the first formulation, the loss function can be defined as

$$||\mathbf{JGD}^{-1/2} - n^{-1}\mathbf{FW}'\mathbf{D}^{1/2}||^2. \tag{5.4}$$

Here, $\mathbf{F}$ is an $n \times c$ component score matrix and $\mathbf{W}$ is a $K \times c$ category score matrix. It is also assumed that $\mathbf{F}$ is satisfied with $n^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_c$ and $\mathbf{JF} = \mathbf{F}$. This apporach is originally based on simple correspondence analysis of contingencey table, which is formulated as weighted least squares approximation of the column-centered contingency table by a model paramters, with weights in inverse proportion to the marginal totals of the contingency table. When the contingency table in simple correspondence analysis is replaced to the indicator matrix $\mathbf{G}$, then we have the loss function (5.4) and minimize it over $\mathbf{F}$ and $\mathbf{W}$ subject to $n^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_c$ and $\mathbf{JF} = \mathbf{F}$ (Greenacre, 1988).

Homogeneity analysis is based on the homogeneity assumption: the scores for individuals should be homogeneous to the scores for the categories which individuals choose (Gifi, 1990). The loss function is expressed as

$$||\mathbf{F} - \mathbf{GW}||^2 = \sum_{i=1}^{p} ||\mathbf{F} - \mathbf{G}_i\mathbf{W}_i||^2, \tag{5.5}$$

which is minimized over $\mathbf{F}$ and $\mathbf{W}$ subject to $n^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_c$ and $\mathbf{JF} = \mathbf{F}$.

The loss functions (5.4) and (5.5) can be expanded as

$$\begin{aligned}
||\mathbf{JGD}^{-1/2} - n^{-1}\mathbf{FW}'\mathbf{D}^{1/2}||^2 &= tr(\mathbf{D}^{-1/2}\mathbf{G}'\mathbf{JGD}^{-1/2} - 2n^{-1}\mathbf{D}^{-1/2}\mathbf{G}'\mathbf{JFW}'\mathbf{D}^{1/2} \\
&\quad + n^{-1}\mathbf{D}^{1/2}\mathbf{WW}'\mathbf{D}^{1/2}) \\
&= K - p + n^{-1}tr(-2\mathbf{D}^{-1/2}\mathbf{G}'\mathbf{JFW}'\mathbf{D}^{1/2} + \mathbf{D}^{1/2}\mathbf{WW}'\mathbf{D}^{1/2}) \\
&= K - p + n^{-1}tr(-2\mathbf{F}'\mathbf{GW} + \mathbf{W}'\mathbf{DW})
\end{aligned} \tag{5.6}$$

and

$$\begin{aligned}
||\mathbf{F} - \mathbf{GW}||^2 &= tr(n\mathbf{I}_c - 2\mathbf{F}'\mathbf{GW} + \mathbf{W}'\mathbf{DW}) \\
&= nc + tr(-2\mathbf{F}'\mathbf{GW} + \mathbf{W}'\mathbf{DW}).
\end{aligned} \tag{5.7}$$

Therefore, the loss functions (5.6) and (5.7) are equivalent without constants and scale that are irrelevant to the parameters. The optimization based on (5.7) is considered below.

The optimal parameters can be obtained by solving the regression problem and maximizing the bilinear-form function. The loss function can be rewritten as

$$\begin{aligned}\sum_{i=1}^{p} ||\mathbf{F} - \mathbf{G}_i\mathbf{W}_i||^2 = \sum_{i=1}^{p} tr(&\mathbf{F}'\mathbf{F} - 2\mathbf{F}'\mathbf{G}_i\mathbf{W}_i - \mathbf{W}'_i\mathbf{G}'_i\mathbf{G}_i\mathbf{W}_i \\ &+ \mathbf{F}'\mathbf{G}_i(\mathbf{G}'_i\mathbf{G}_i)^{-1/2}(\mathbf{G}'_i\mathbf{G}_i)^{-1/2}\mathbf{G}'_i\mathbf{F} - \mathbf{F}'\mathbf{G}_i(\mathbf{G}'_i\mathbf{G}_i)^{-1/2}(\mathbf{G}'_i\mathbf{G}_i)^{-1/2}\mathbf{G}'_i\mathbf{F}) \\ &= \sum_{i=1}^{p} ||(\mathbf{G}'_i\mathbf{G}_i)^{1/2}\mathbf{W}_i - (\mathbf{G}'_i\mathbf{G}_i)^{-1/2}\mathbf{G}'_i\mathbf{F}||^2 + const_{HA_W}, \end{aligned} \tag{5.8}$$

where $const_{HA_W}$ is the constant which is irrelevant to the $\mathbf{W}$. Thus, optimal $\mathbf{W}$ can be obtained by

$$\begin{aligned}\mathbf{W} &= [(\mathbf{G}'_1\mathbf{G}_1)^{-1}\mathbf{G}'_1\mathbf{F}, \ldots, (\mathbf{G}'_p\mathbf{G}_p)^{-1}\mathbf{G}'_p\mathbf{F},] \\ &= \mathbf{D}^{-1}\mathbf{G}'\mathbf{F}. \end{aligned} \tag{5.9}$$

Then, the minimization over $\mathbf{F}$ is considered. Substituting (5.9) to the loss function (5.5), we have

$$\begin{aligned}&\sum_{i=1}^{p} tr(\mathbf{F}'\mathbf{F} - 2\mathbf{F}'\mathbf{G}_i(\mathbf{G}'_i\mathbf{G}_i)^{-1}\mathbf{G}'_1\mathbf{F} - \mathbf{F}'\mathbf{G}_i(\mathbf{G}'_i\mathbf{G}_i)^{-1}\mathbf{G}'_i\mathbf{F}) \\ &= -tr(\mathbf{F}' \sum_{i=1}^{p}\{\mathbf{G}_i(\mathbf{G}'_i\mathbf{G}_i)^{-1}\mathbf{G}'_i\}\mathbf{F}) + c_{HA_F} \\ &= -tr(\mathbf{F}'\mathbf{J}\mathbf{G}\mathbf{D}^{-1}\mathbf{G}'\mathbf{J}\mathbf{F}) + c_{HA_F}, \end{aligned} \tag{5.10}$$

where $const_{HA_F}$ is the constant which is irrelevant to the $\mathbf{F}$. Thus the minimization of the loss function (5.5) over $\mathbf{F}$ is equivalent to the maximization of $tr(\mathbf{F}'\mathbf{J}\mathbf{G}\mathbf{D}^{-1}\mathbf{G}'\mathbf{J}\mathbf{F})$ over $\mathbf{F}$. This maximization can be achieved by the SVD of $\mathbf{J}\mathbf{G}\mathbf{D}^{-1/2} = \mathbf{K}\mathbf{\Delta}\mathbf{L}'$. That is, the optimal $\mathbf{F}$ can be attained by

$$\mathbf{F} = \sqrt{n}\mathbf{K}_c. \tag{5.11}$$

The equation (5.11) is substituted into the equation (5.9), the optimal $\mathbf{W}$ can be also expressed as

$$
\begin{aligned}
\mathbf{W} &= \mathbf{D}^{-1}\mathbf{G}'\mathbf{F} \\
&= \mathbf{D}^{-1/2}\mathbf{D}^{-1/2}\mathbf{G}'\mathbf{J}\mathbf{F} \\
&= \mathbf{D}^{-1/2}\mathbf{L}\boldsymbol{\Delta}\mathbf{K}'(\sqrt{n}\mathbf{K}_c) \\
&= \sqrt{n}\mathbf{D}^{-1/2}\mathbf{L}_c\boldsymbol{\Delta}_c. \qquad (5.12)
\end{aligned}
$$

This result implies that optimal $\mathbf{F}$ and $\mathbf{W}$ can be explicitly obtained by the SVD of $\mathbf{JGD}^{-1/2} = \mathbf{K}\boldsymbol{\Delta}\mathbf{L}'$.

5.2 Extension of principal component analysis via quantification

When variables are categorical, they are transformed into quantitative scores $\mathbf{X}^*$ by a quantification procedure. Then $\mathbf{X}_i^* = \mathbf{G}_i\mathbf{Q}_i$ represents multiple quantification of i-th variable where $\mathbf{Q}_i$ (K_i-categories $\times$ R_i-dimensions) is the quantification parameter matrix of the i-th variable. Replacing i-th column vector of data matrix $\mathbf{x}_{(i)}$ to $\mathbf{G}_i\mathbf{Q}_i$ in the function (2.15), then we have

$$NPCA(\mathbf{F}, \mathbf{A}, \mathbf{Q}) = \sum_{i=1}^{p} \|\mathbf{G}_i\mathbf{Q}_i - \mathbf{F}\mathbf{A}'_i\|^2. \qquad (5.13)$$

Thus, nonmetric version of PCA can be defined as the problem that the function (5.13) is minimized over $\mathbf{F}$, $\mathbf{A}_i$, and $\mathbf{Q}_i$ ($i = 1, \ldots, p$), subject to the following constraints

$$n^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_c, \ \mathbf{JF} = \mathbf{F}, \qquad (5.14)$$

$$\mathbf{1}'_n\mathbf{G}_i\mathbf{Q}_i = \mathbf{0}'_{K_i}, \qquad (5.15)$$

$$n^{-1}\mathbf{Q}'_i\mathbf{G}'_i\mathbf{G}_i\mathbf{Q}_i = \mathbf{I}_{R_i}. \qquad (5.16)$$

The constraints (5.15) and (5.16) imply that the quantified data are assumed to be standardized.

The loss function (5.13) can be rewritten as

$$\begin{aligned}
&\sum_{i=1}^{p} tr(\mathbf{G}_i\mathbf{Q}_i - \mathbf{F}\mathbf{A'}_i)'(\mathbf{G}_i\mathbf{Q}_i - \mathbf{F}\mathbf{A'}_i) \\
&= \sum_{i=1}^{p} tr(\mathbf{Q'}_i\mathbf{G'}_i\mathbf{G}_i\mathbf{Q}_i - 2\mathbf{Q'}_i\mathbf{G'}_i\mathbf{F}\mathbf{A'}_i + \mathbf{A}_i\mathbf{F'}\mathbf{F}\mathbf{A'}_i) \\
&= \sum_{i=1}^{p} tr(\mathbf{Q'}_i\mathbf{G'}_i\mathbf{G}_i\mathbf{Q}_i - 2\mathbf{Q'}_i\mathbf{G'}_i(\mathbf{F'F})^{-1/2}(\mathbf{F'F})^{1/2}\mathbf{A'}_i + \mathbf{A}_i(\mathbf{F'F})^{1/2}(\mathbf{F'F})^{1/2}\mathbf{A'}_i \\
&+ \mathbf{Q'}_i\mathbf{G'}_i\mathbf{F}(\mathbf{F'F})^{-1/2}(\mathbf{F'F})^{-1/2}\mathbf{F'}\mathbf{G}_i\mathbf{Q}_i - \mathbf{Q'}_i\mathbf{G'}_i\mathbf{F}(\mathbf{F'F})^{-1/2}(\mathbf{F'F})^{-1/2}\mathbf{F'}\mathbf{G}_i\mathbf{Q}_i) \\
&= \sum_{i=1}^{p} ||(\mathbf{F'F})^{1/2}\mathbf{A'}_i - (\mathbf{F'F})^{-1/2}\mathbf{F'}\mathbf{G}_i\mathbf{Q}_i||^2 + c_{NPCA_A}.
\end{aligned} \tag{5.17}$$

Here c_{NPCA_A} is a constant, which is irrelevant to $\mathbf{A}$. As can be seen in reexpressed the loss function (5.17), the minimization over $\mathbf{A}$ is achieved when $(\mathbf{F'F})^{1/2}\mathbf{A'}_i = (\mathbf{F'F})^{-1/2}\mathbf{F'}\mathbf{G}_i\mathbf{Q}_i$ for all variables. Define $\mathbf{B_Q} = b - diag(\mathbf{Q}_1, \dots, \mathbf{Q}_p)$, then optimal $\mathbf{A}$ can be described as

$$\begin{aligned}
\mathbf{A} &= \mathbf{B'}_Q\mathbf{G'}\mathbf{F}(\mathbf{F'F})^{-1} \\
&= n^{-1}\mathbf{B'}_Q\mathbf{G'}\mathbf{F}.
\end{aligned} \tag{5.18}$$

The equation (5.18) is equivalent to (2.19) when $\mathbf{X}$ is replaced to $\mathbf{GB}_Q$. Because of the constraints (5.15) and (5.16), the equation (5.18) can be regared as correlations between quantified variables and components.

The loss function (5.13) can also be reexpressed as

$$\begin{aligned}
&\sum_{i=1}^{p} ||\mathbf{G}_i\mathbf{Q}_i - \mathbf{F}\mathbf{A'}_i||^2 \\
&= \sum_{i=1}^{p} tr(\mathbf{Q'}_i\mathbf{G'}_i\mathbf{G}_i\mathbf{Q}_i - 2\mathbf{Q'}_i\mathbf{G'}_i\mathbf{F}\mathbf{A'}_i + \mathbf{A}_i\mathbf{F'}\mathbf{F}\mathbf{A'}_i + \mathbf{F'F} - \mathbf{F'F}) \\
&= \sum_{i=1}^{p} ||\mathbf{F} - \mathbf{G}_i\mathbf{Q}_i\mathbf{A}_i||^2 + c_{NPCA_Q} \\
&= \sum_{i=1}^{p} ||\mathbf{F} - \mathbf{G}_i(\mathbf{Q}_i\mathbf{A}_i + \mathbf{W}_i - \mathbf{W}_i)||^2 + c_{NPCA_Q} \\
&= \sum_{i=1}^{p} ||\mathbf{F} - \mathbf{G}_i(\mathbf{Q}_i\mathbf{A}_i + \mathbf{W}_i - \mathbf{W}_i)||^2 + c_{NPCA_Q} \\
&= \sum_{i=1}^{p} ||\mathbf{F} - \mathbf{G}_i\mathbf{W}_i||^2 + \sum_{i=1}^{p} ||\mathbf{D}_i^{1/2}\mathbf{W}_i - \mathbf{D}_i^{1/2}\mathbf{Q}_i\mathbf{A}_i||^2 + c_{NPCA_Q},
\end{aligned} \tag{5.19}$$

where c_{NPCA_Q} is the constant, which is irrelevant to $\mathbf{Q}_i$ ($i = 1, \ldots, p$) and $\mathbf{W}_i = \mathbf{D}_i^{-1}\mathbf{G}'_i\mathbf{F}$. This reexpression follows that the second term in the loss function (5.19) is only relevant to the quantification parameters. Thus, from now on, we consider the minimization of $\sum_{i=1}^{p} \|\mathbf{D}_i^{1/2}\mathbf{W}_i - \mathbf{D}_i^{1/2}\mathbf{Q}_i\mathbf{A}_i\|^2$ over $\mathbf{Q}_i$ ($i = 1, \ldots, p$). Solving the regression problem (ten Berge, 1993), the optimal loading $\mathbf{A}_i$ can be given by

$$\mathbf{A}_i = n^{-1}\mathbf{Q}'_i\mathbf{D}_i^{1/2}\mathbf{D}_i^{1/2}\mathbf{W}_i. \tag{5.20}$$

Substituting (5.20) into $\sum_{i=1}^{p} \|\mathbf{D}_i^{1/2}\mathbf{W}_i - \mathbf{D}_i^{1/2}\mathbf{Q}_i\mathbf{A}_i\|^2$, we have

$$\begin{aligned} &\sum_{i=1}^{p} tr(\mathbf{W}'_i\mathbf{D}_i\mathbf{W}_i - \mathbf{Q}'_i\mathbf{D}_i^{1/2}\mathbf{D}_i^{1/2}\mathbf{W}_i\mathbf{W}'_i\mathbf{D}_i^{1/2}\mathbf{D}_i^{1/2}\mathbf{Q}_i) \\ &= \sum_{i=1}^{p} -tr(\mathbf{Q}'_i\mathbf{D}_i^{1/2}\mathbf{D}_i^{1/2}\mathbf{W}_i\mathbf{W}'_i\mathbf{D}_i^{1/2}\mathbf{D}_i^{1/2}\mathbf{Q}_i) + c_{NPCA_{Q1}}. \end{aligned} \tag{5.21}$$

Here, $c_{NPCA_{Q1}}$ is the constant, which is irrelevant to the quantification parameters. We reparametrize $\mathbf{W}_i$ using the SVD as

$$\mathbf{W}_i = \tilde{\mathbf{K}}\mathbf{H}\mathbf{H}^{-1}\tilde{\mathbf{\Delta}}\tilde{\mathbf{L}}', \tag{5.22}$$

where $\tilde{\mathbf{K}}$ and $\tilde{\mathbf{L}}$ contain left and right singular vectors, respectively, and $\mathbf{H}$ is an arbitrary nonsingular matrix. $\mathbf{D}_i^{1/2}\mathbf{Q}_i$ is the column-orthogonal matrix because it is satisfied with $n^{-1}\mathbf{Q}'_i\mathbf{G}'_i\mathbf{G}_i\mathbf{Q}_i = \mathbf{I}_{R_i}$. Thus, the loss function (5.21) can be minimized by maximizing $tr(\mathbf{Q}'_i\mathbf{D}_i^{1/2}\mathbf{D}_i^{1/2}\mathbf{W}_i\mathbf{W}'_i\mathbf{D}_i^{1/2}\mathbf{D}_i^{1/2}\mathbf{Q}_i)$. This maximization can be achieved by

$$\mathbf{D}_i^{1/2}\mathbf{Q}_i = \mathbf{D}_i^{1/2}\tilde{\mathbf{K}}\mathbf{H} \Longleftrightarrow \mathbf{Q}_i = \tilde{\mathbf{K}}\mathbf{H} \tag{5.23}$$

when we have $\mathbf{H}$ that is satisfied with

$$n^{-1}\mathbf{H}'\tilde{\mathbf{K}}'\mathbf{D}_i\tilde{\mathbf{K}}\mathbf{H} = \mathbf{I}_{R_i}. \tag{5.24}$$

Define the eigen value decompsotion (EVD) of $n^{-1}\tilde{\mathbf{K}}'\mathbf{D}_i\tilde{\mathbf{K}}$ as

$$n^{-1}\tilde{\mathbf{K}}'\mathbf{D}_i\tilde{\mathbf{K}} = \tilde{\mathbf{R}}\mathbf{\Xi}^2\tilde{\mathbf{R}}'. \tag{5.25}$$

Here, $\tilde{\mathbf{R}}$ contains eigen vectors and $\mathbf{\Xi}^2$ is a diagonal matrix whose elements are eigenvalues. The rank of $\mathbf{W}_i$ is $min(c, K_i - 1) = R_i$. Thus, the ranks of $\tilde{\mathbf{R}}$ and $\mathbf{\Xi}^2$ are also R_i and $\tilde{\mathbf{R}}$ and $\mathbf{\Xi}^2$ are nonsingular matrices. Then, optimal $\mathbf{H}$ can be obtained if we set

$$\mathbf{H} = \tilde{\mathbf{R}}\mathbf{\Xi}^{-1}. \tag{5.26}$$

It follows that the optimal quantification parameters can be given by

$$\mathbf{Q}_i = \tilde{\mathbf{K}}\tilde{\mathbf{R}}\boldsymbol{\Xi}^{-1}. \tag{5.27}$$

It should be noted that the product of $\mathbf{Q}_i$ and $\mathbf{A}_i$ can be expressed as

$$\begin{aligned}\mathbf{Q}_i\mathbf{A}_i &= n^{-1}\mathbf{Q}_i\mathbf{Q}'_i\mathbf{D}_i^{1/2}\mathbf{D}_i^{1/2}\mathbf{W}_i \\ &= (\tilde{\mathbf{K}}\mathbf{H})(n^{-1}\mathbf{H}'\tilde{\mathbf{K}}'\mathbf{D}_i\tilde{\mathbf{K}}\mathbf{H}\mathbf{H}^{-1}\tilde{\boldsymbol{\Delta}}\tilde{\mathbf{L}}') \\ &= \mathbf{W}_i. \end{aligned} \tag{5.28}$$

It implies that category score matrix $\mathbf{W}_i$ can be decomposed into the quantification and loading matrices for each i.

Finally, the minimization over $\mathbf{F}$ is considered. Substituting (5.18) into the loss function, then

$$\begin{aligned}&\sum_{i=1}^{p} tr(\mathbf{Q}'_i\mathbf{G}'_i\mathbf{G}_i\mathbf{Q}_i - 2n^{-1}\mathbf{Q}'_i\mathbf{G}'_i\mathbf{F}\mathbf{F}'\mathbf{G}_i\mathbf{Q}_i + n^{-1}\mathbf{Q}'_i\mathbf{G}'_i\mathbf{F}\mathbf{F}'\mathbf{G}_i\mathbf{Q}_i) \\ &= -n^{-1}\sum_{i=1}^{p} tr(\mathbf{Q}'_i\mathbf{G}'_i\mathbf{F}\mathbf{F}'\mathbf{G}_i\mathbf{Q}_i) + c_{NPCA_F} \\ &= -n^{-1}tr(\mathbf{B}'_Q\mathbf{G}'\mathbf{F}\mathbf{F}'\mathbf{G}\mathbf{B}_Q) + c_{NPCA_F}\end{aligned}$$

is obtained. Here, $c_{NPCA_F} = tr(\mathbf{B}'_Q\mathbf{G}'\mathbf{G}\mathbf{B}_Q)$ is the constant that is irrelevant to $\mathbf{F}$. As will be shown in subsection 5.5.2, $tr(\mathbf{B}'_Q\mathbf{G}'\mathbf{F}\mathbf{F}'\mathbf{G}\mathbf{B}_Q)$ is reexpressed as

$$tr(\mathbf{B}'_Q\mathbf{G}'\mathbf{F}\mathbf{F}'\mathbf{G}\mathbf{B}_Q) = ntr(\mathbf{F}'\mathbf{J}\mathbf{G}\mathbf{D}^{1/2}\mathbf{D}^{-1/2}\mathbf{G}'\mathbf{J}'\mathbf{F}). \tag{5.29}$$

Therefore, the minimization of the loss fucntion over $\mathbf{F}$ is equivalent to maximize the function $tr(\mathbf{F}'\mathbf{J}\mathbf{G}\mathbf{D}^{1/2}\mathbf{D}^{-1/2}\mathbf{G}'\mathbf{J}'\mathbf{F})$ subject to the constraint (5.14). The optimal $\mathbf{F}$ is described as

$$\mathbf{F} = \sqrt{n}\mathbf{K}_c \tag{5.30}$$

in the same way as (5.11).

5.3 Equivalence to mutiple correspondence analysis and homogeneity analysis

As shown in the previous section, MCA can also be formulated as approximating an data matrix by a lower rank matrix using the quantification technique in the same way as PCA formulation

(Murakami, Kiers, and ten Berge, 1999; Murakami, 1999; Adachi and Murakami, 2011; Mori, Kuroda and Makino, 2016) . In ordinary MCA, the number of quantification dimensions R_i is fixed to $\min(K_i - 1, r)$, and the parameters can be uniquely determined. When the number of dimensions can be set in $1 \leq R_i < \min(K_i - 1, r)$, the solutions cannot be uniquely determined and thus the parameters can be estimated by alternate iterations of the operations that quantify the categorical variables and that obtain component score and loading matrices. This type of MCA is referred to as rank-restricted MCA (Murakami *et al*, 1999), which includes ordinary MCA as a special case. When all categorical variables are unidimensionally quantified, quantification parameters are assumed to be $\mathbf{1}'_n\mathbf{G}_i\mathbf{q}_i = \mathbf{0}'_{K_i}$, $n^{-1}\mathbf{q}'_i\mathbf{G}'_i\mathbf{G}_i\mathbf{q}_i = 1$. Then, (rank-restricted) MCA reduces to nonmetric PCA (NCA) (Adachi and Murakami, 2011). In the case of numerical variables, standardized numerical variables also satisfy the constraints above. In other words, standardization of numerical variable is a restricted version of single quantification ($R_i = 1$), in which quantification parameters are already known, and NCA is equivalent to PCA when all variables are numerical. The hierarchical relationships among MCA, NCA, and PCA are summarized in Table 5.1.

Table 5.1: Relationships among MCA, nonmetric PCA, and PCA

Quantification	PCA model
Non	PCA
Single	NCA
Multiple	MCA

5.4 Rotation in multiple correspondence analysis

MCA also has the rotational freedom and the rotation of MCA solutions has been dealt with in previous studies (Adachi, 2004; Chavent *et al*, 2012; Kiers, 1991). The loss function (5.6) cannot be affected by an nonsingular matrix $\mathbf{T}$:

$$||\mathbf{JGD}^{-1/2} - n^{-1}\mathbf{FT}'^{-1}\mathbf{T}'\mathbf{W}'\mathbf{D}^{1/2}||^2 = ||\mathbf{JGD}^{-1/2} - n^{-1}\mathbf{FT}'^{-1}\mathbf{T}'\mathbf{W}'\mathbf{D}^{1/2}||^2. \tag{5.31}$$

That is, the MCA solutions can be obliquely-rotated based on the formulation above. However, only orthogonal rotation is allowed in homogeneity analysis formulation:

$$||\mathbf{F} - \mathbf{GW}||^2 = ||\mathbf{FT}_{orth} - \mathbf{GWT}_{orth}||^2, \tag{5.32}$$

with an arbitrary orthogonal matrix $\mathbf{T}_{orth}$. There are various MCA formulations that give the same solutions, but they differ from viewpoint of rotational freedom.

When the solutions are obliquely rotated based on (5.31), they cannot be interpreted as the graphical representation bacause homogeneity assumption is violated (Adachi, 2004):

$$||\mathbf{F} - \mathbf{GW}||^2 \neq ||\mathbf{FT}'^{-1} - \mathbf{GWT}||^2. \tag{5.33}$$

The violation does not occur when orthogonal rotation is adopted. Oblique rotation is useful to obtain simple structure, but the graphical representation limitation exists in rotation of MCA solution base on Adachi (2004).

There are some differences between the rotation in PCA and MCA: the relation between the loadings and component scores are known for PCA, $\mathbf{A} = n^{-1}\mathbf{F}'\mathbf{X}$, but this does not generally hold in MCA (Kiers, 1991). In previous studies on MCA rotation, there are two mainly options for rotating solutions (Adachi, 2004). These options are refered as a category option and an item option, respectively. Another option can be considered based (5.19), which is called as loading option in this thesis. Three options are introduced in the following subsections.

5.4.1 Category option

In the category option, the component structure matrix is rotated towards simple structure, which is defined as standardized inner products between inticator matrix $\mathbf{G}$ and component score matrix $\mathbf{F}$: the component structure matrix is defined as

$$\begin{aligned} n^{-1}\mathbf{D}^{-1/2}\mathbf{G}'\mathbf{JF} &= n^{-1/2}(\mathbf{L}\boldsymbol{\Delta}\mathbf{K}')\mathbf{K}_c \\ &= n^{-1/2}(\mathbf{L}_c\boldsymbol{\Delta}_c) = n^{-1}\mathbf{D}^{1/2}\mathbf{W}. \end{aligned} \tag{5.34}$$

In this option, the function $f(\mathbf{T}) = n^{-1}\mathbf{D}^{1/2}\mathbf{WT}$ is optimized so that the component structure matrix has the approximate simple structure.

5.4.2 Item option

In the item option, the item structure matrix is rotated to simple structure. Elements of the item structre matrix contain discrimination measure (Gifi, 1990), which shows a contribution of the inertia of a variable that is accounted for. The discrimination measure can be defined as

$$DM = n^{-1}\mathbf{f}'_j\mathbf{G}_i\mathbf{D}_i^{-1}\mathbf{G}'_i\mathbf{f}_j. \tag{5.35}$$

Substituting $\mathbf{W}_i = \mathbf{D}_i^{-1}\mathbf{G}'_i\mathbf{F} = [\mathbf{D}_i^{-1}\mathbf{G}'_i\mathbf{f}_1, \ldots, \mathbf{D}_i^{-1}\mathbf{G}'_i\mathbf{f}_c] = [\mathbf{w}_{i1}, \ldots, \mathbf{w}_{ic}]$ into (5.35), then we have

$$DM = n^{-1}\mathbf{f}'_j\mathbf{G}_i\mathbf{D}_i^{-1}\mathbf{G}'_i\mathbf{f}_j = n^{-1}\mathbf{f}'_j\mathbf{G}_i\mathbf{w}_{ij}, \tag{5.36}$$

$$= n^{-1}\mathbf{w}'_{ij}\mathbf{D}_i\mathbf{w}_{ij}. \tag{5.37}$$

As shown in above, the discrimination measures can be expressed in three different manners. Adachi (2004) pointed out that oblique rotation makes the three expressions of the discrimination measures unequal, while othogonal rotation preserves their equality: after oblique rotation, (5.35), (5.36) and (5.37) can be expressed as

$$n^{-1}\tilde{\mathbf{t}}'_j\mathbf{f}'_j\mathbf{G}_i\mathbf{D}_i^{-1}\mathbf{G}'_i\mathbf{f}_j\tilde{\mathbf{t}}_j, \tag{5.38}$$

$$n^{-1}\tilde{\mathbf{t}}'_j\mathbf{f}'_j\mathbf{G}_i\mathbf{w}_{ij}\mathbf{t}_j, \tag{5.39}$$

$$n^{-1}\mathbf{t}'_j\mathbf{w}'_{ij}\mathbf{D}_i\mathbf{w}_{ij}\mathbf{t}_j. \tag{5.40}$$

Here, $\mathbf{t}_j$ is the jth column vector of $\mathbf{T}$ and $\tilde{\mathbf{t}}_j$ is the jth column vector of $\mathbf{T}'^{-1}$. In item option, Adachi (2004) chose (5.38) for oblique rotation among them.

5.4.3 Loading option

In (5.13), another rotational option can be considered:

$$\sum_{i=1}^{p} ||\mathbf{G}_i\mathbf{Q}_i - \mathbf{F}\mathbf{A}'_i||^2 = \sum_{i=1}^{p} ||\mathbf{G}_i\mathbf{Q}_i\mathbf{T}_{orth(i)} - \mathbf{F}\mathbf{T}'\mathbf{T}'^{-1}\mathbf{A}'_i\mathbf{T}_{orth(i)}||^2, \tag{5.41}$$

where $\mathbf{T}_{orth(i)}$ $(R_i \times R_i)$ is an orthonormal matrix ($i = 1, \ldots, p$). Here, $\mathbf{A}_i$ expresses the correlations between components $\mathbf{F}$ and the quantified data $\mathbf{G}_i\mathbf{Q}_i$ for each i. It indicates that the loading can be defined in the same manner as PCA.

However, there is a difference between the rotation of PCA loadings and that of MCA loadings: the loading matrix can be rotated by pre-and post-multiplied matrices. Murakami *et al* (1999) and Adachi *et al* (2011) modified an orthomax criterion in three-way PCA for MCA solutions in which the loading matrix is rotated towards the simple structure by pre- and post-multiplied rotation matrices.

5.5 Reformulation of MCA loss function

The result (5.41) suggests that correlations between quantified data and components can be rotated towards simple structure in the same manner as PCA. Thus, we discuss the method that the loading matrix is transformed towards simple structure. In Murakami *et al* (1999) and Adachi and Murakami (2011), only orthogonal rotation can be applied to the MCA solution because the quantification parameters have only orthogonal rotational freedom in (5.13). Due to this limitation, the loading matrix is restricted to be orthgonally rotated in Murakami *et al* (1999) and Adachi and Murakami (2011). However, as discussed in chapter 3, oblique rotation is preferred to obtain the simple structure. To tackle the limitation, we show that MCA can be rotated obliquely without affecting homogeneity assumption and that a joint oblique rotation method is proposed.

As shown in section 5.2, the loss function (5.13) can be described as

$$\begin{aligned}
\sum_{i=1}^{p} ||\mathbf{F} - \mathbf{G}_i\mathbf{Q}_i\mathbf{A}_i||^2 &= \sum_{i=1}^{p} tr(\mathbf{F}'\mathbf{F} - 2\mathbf{F}'\mathbf{G}_i\mathbf{Q}_i\mathbf{A}_i + \mathbf{A}'_i\mathbf{Q}'_i\mathbf{G}'_i\mathbf{G}_i\mathbf{Q}_i\mathbf{A}_i) \\
&= \sum_{i=1}^{p} tr(nc - \mathbf{F}'\mathbf{G}_i\mathbf{Q}_i(\mathbf{Q}'_i\mathbf{G}'_i\mathbf{G}_i\mathbf{Q}_i)^{-1}\mathbf{Q}'_i\mathbf{G}'_i\mathbf{F}) \\
&= \sum_{i=1}^{p} tr(nc - \mathbf{F}'\mathbf{P}_{gq_i}\mathbf{F}) \\
&= \sum_{i=1}^{p} tr(nc - n(\mathbf{F}'\mathbf{F})^{-1}\mathbf{F}'\mathbf{P}_{gq_i}\mathbf{F}) \\
&= \sum_{i=1}^{p} tr(nc - n\mathbf{P}_{gq_i}\mathbf{P}_f) \qquad (5.42)
\end{aligned}$$

with orthogonal projection matrices $\mathbf{P}_{gq_i} = \mathbf{G}_i\mathbf{Q}_i(\mathbf{Q}'_i\mathbf{G}'_i\mathbf{G}_i\mathbf{Q}_i)^{-1}\mathbf{Q}'_i\mathbf{G}'_i$ for each i and $\mathbf{P}_f = \mathbf{F}(\mathbf{F}'\mathbf{F})^{-1}\mathbf{F}'$.

The orthogonal projection matrices $\mathbf{P}_{gq_i}$ and $\mathbf{P}_f$ are unchanged in nonsingular transformation:

$$\begin{aligned}\mathbf{P}_{gq_i} &= \mathbf{G}_i\mathbf{Q}_i(\mathbf{Q'}_i\mathbf{G'}_i\mathbf{G}_i\mathbf{Q}_i)^{-1}\mathbf{Q'}_i\mathbf{G'}_i \\ &= \mathbf{G}_i\mathbf{Q}_i\mathbf{T}_i^*(\mathbf{T}_i^{*'}\mathbf{Q'}_i\mathbf{G'}_i\mathbf{G}_i\mathbf{Q}_i\mathbf{T}_i^*)^{-1}\mathbf{T}_i^{*'}\mathbf{Q'}_i\mathbf{G'}_i \\ &= \mathbf{G}_i\mathbf{Q}_i\mathbf{T}_i^*(n\mathbf{T}_i^{*'}\mathbf{T}_i^*)^{-1}\mathbf{T}_i^{*'}\mathbf{Q'}_i\mathbf{G'}_i \\ &= n^{-1}\mathbf{G}_i\mathbf{Q}_i\mathbf{T}_i^*\mathbf{T}_i^{*^{-1}}\mathbf{T}_i^{*'^{-1}}\mathbf{T}_i^{*'}\mathbf{Q'}_i\mathbf{G'}_i \\ &= \mathbf{G}_i\mathbf{Q}_i(\mathbf{Q'}_i\mathbf{G'}_i\mathbf{G}_i\mathbf{Q}_i)^{-1}\mathbf{Q'}_i\mathbf{G'}_i \end{aligned} \tag{5.43}$$

and

$$\begin{aligned}\mathbf{P}_f &= \mathbf{F}(\mathbf{F'F})^{-1}\mathbf{F'} \\ &= \mathbf{FT}(\mathbf{T'F'FT})^{-1}\mathbf{T'F'} \\ &= \mathbf{FT}(n^{-1}\mathbf{T'T})^{-1}\mathbf{T'F'} \\ &= n^{-1}\mathbf{FTT}^{-1}\mathbf{T'}^{-1}\mathbf{T'F'} \\ &= \mathbf{F}(\mathbf{F'F})^{-1}\mathbf{F'}, \end{aligned} \tag{5.44}$$

where $\mathbf{T}_i^*$ and $\mathbf{T}$ are arbitrary nonsingular matrices. It is well-known that the nonsingular matrix has the following property:

$$(\mathbf{T'T})^{-1} = \mathbf{T}^{-1}\mathbf{T'}^{-1}. \tag{5.45}$$

This property was used in (5.43) and (5.44).

Recall that the loading matrix can be expressed as

$$\begin{aligned}\mathbf{A}_i &= (\mathbf{Q'}_i\mathbf{G'}_i\mathbf{G}_i\mathbf{Q}_i)^{-1}\mathbf{Q'}_i\mathbf{G'}_i\mathbf{F}, \\ \mathbf{A} &= (\mathbf{B'}_Q\mathbf{G'GB}_Q)^{-1}\mathbf{B'}_Q\mathbf{G'F}.\end{aligned}$$

Therefore, the loading matrix can be obliquely rotated by pre-and post-multiplied matrices:

$$\begin{aligned}\mathbf{A}_i^* &= (\mathbf{T}_i^{*'}\mathbf{Q'}_i\mathbf{G'}_i\mathbf{G}_i\mathbf{Q}_i\mathbf{T}_i^*)^{-1}\mathbf{T}_i^{*'}\mathbf{Q'}_i\mathbf{G'}_i\mathbf{FT} \\ &= n^{-1}\mathbf{T}_i^{*^{-1}}\mathbf{Q'}_i\mathbf{G'}_i\mathbf{FT}, \end{aligned} \tag{5.46}$$

$$\begin{aligned}\mathbf{A}^* &= (\tilde{\mathbf{T}}'\mathbf{B'}_Q\mathbf{G'GB}_Q\tilde{\mathbf{T}})^{-1}\tilde{\mathbf{T}}'\mathbf{B'}_Q\mathbf{G'FT} \\ &= n^{-1}\tilde{\mathbf{T}}^{-1}\mathbf{B'}_Q\mathbf{G'FT},\end{aligned}$$

with $\tilde{\mathbf{T}}^{-1} = b - diag(\mathbf{T}_1^{*^{-1}}, \ldots, \mathbf{T}_p^{*^{-1}})$.

5.5.1 Joint simplimax rotation

An alternative joint oblique rotation technique is needed to be developed in the proposed reformulation. We propose a joint simplimax rotation algorithm based on three-way simplimax rotation to obtain simple core matrix (Kiers, 1998). The loss function of joint simplimax rotation is formulated as

$$\begin{aligned} JS(\mathbf{T}, \mathbf{T}_{oblq}, \mathbf{B}) &= \sum_{i=1}^{p} \|\mathbf{T}_{oblq_i}^{-1}\mathbf{A}_i\mathbf{T} - \mathbf{B}_i\|^2 \\ &= \|\mathbf{T}_{oblq}^{-1}\mathbf{A}\mathbf{T} - \mathbf{B}\|^2 \end{aligned} \tag{5.47}$$

with $\mathbf{T}$ and $\mathbf{T}_{oblq}^{-1} = b - diag(\mathbf{T}_{oblq_1}^{-1}, \ldots, \mathbf{T}_{oblq_p}^{-1})$ are rotation matrices, which satisfy with

$$diag(\mathbf{T}'\mathbf{T}) = \mathbf{I} \tag{5.48}$$

$$diag(\mathbf{T}'_{oblq}\mathbf{T}_{oblq}) = \mathbf{I}, \tag{5.49}$$

and $\mathbf{B}$ is a target matrix which satisfy with

$$\mathbf{B} \text{ has } l \text{ zero elements.} \tag{5.50}$$

The loss function is alternately minimized over $\mathbf{T}$, $\mathbf{T}_{oblq}$ and $\mathbf{B}$, subject to (5.48), (5.49), and (5.50). It should be noted that when the quantification dimensions are all set to be one, $\mathbf{T}_{oblq}$ reduces to the identity matrix; the rotational freedom in the quantification parameters disappear when all variables are binary.

To estimate rotation matrices, we introduce gradient projection (GP) algorithm proposed by Jennrich (2001, 2002). Assume that $f(\mathbf{T})$ is a objective function. Then, the minimization of $\mathbf{T}$ subject to $diag(\mathbf{T}'\mathbf{T}) = \mathbf{I}$ can be achieved by the following steps:

1. Compute gradient $\mathbf{Gr} = \mathrm{d}f/\mathrm{d}\mathbf{T}$ (Gradient step).
2. Replace $\mathbf{T}$ by $\rho(\mathbf{T} - a\mathbf{Gr})$ and go to step 1 or stop (Projection step).

Here, a is a constant ($a > 0$) and $\rho(\mathbf{T})$ is the function of projecting $\mathbf{T}$ onto the manifold of all such matrices whose columns have unit length, given by

$$\rho(\mathbf{T}) = \mathbf{T}(diag(\mathbf{T}'\mathbf{T}))^{-1/2}. \tag{5.51}$$

Differenciate (5.47) with respect to $\mathbf{T}$ and $\mathbf{T}_{oblq_i}$ for each i respectively, the we have

$$\frac{\partial JS}{\partial T} = 2\mathbf{A}'\mathbf{T}_{oblq}^{-1\prime}(\mathbf{T}_{oblq}^{-1}\mathbf{A}\mathbf{T} - \mathbf{B}) \tag{5.52}$$

and

$$\frac{\partial JS}{\partial T_{oblq_i}} = -2\mathbf{T}'^{-1}_{oblq_i}(\mathbf{T}_{oblq_i}^{-1}\mathbf{A}_i\mathbf{T} - \mathbf{B}_i)\mathbf{T}'\mathbf{A}'_i\mathbf{T}_{oblq_i}^{-1\prime}. \tag{5.53}$$

The minimizations over $\mathbf{T}$ and $\mathbf{T}_{oblq_i}$ for each i can be achieved by iterating the following estimation steps:

Step 1: Set the number of zeros l.

Step 2: Initialize the parameters .

Step 3: Update $\mathbf{T}$ by solving oblique Procrustes problem using GP algorithm.

Step 4: Update $\mathbf{T}_{oblq}$ by solving oblique Procrustes problem using GP algorithm.

Step 5: Update $\mathbf{B}$ with (3.13) in ordinary simplimax rotation.

Step 6: Finish if convergence is reached; otherwise, back to Step 3.

5.5.2 Mathematical properties in reformulated MCA

The proposed reformulation (5.42) is discussed in more detail from the mathematical view-points.

Property 1. The proposed loss function (5.42) can be considered as the minimization of subspace distances between spanned by $\mathbf{F}$ and $\mathbf{G}_i\mathbf{Q}_i$ for each i.

Proof. Let $\mathbf{Z}_1$ and $\mathbf{Z}_2$ be arbitrary matrices. Zuccon, Azzopardi and Van Rijsbergen (2009) defined the subspace distance between $sp(\mathbf{Z}_1)$ and $sp(\mathbf{Z}_2)$ as

$$Max\{rank(\mathbf{Z}_1), rank(\mathbf{Z}_2)\} - tr(\mathbf{P}_{Z_1}\mathbf{P}_{Z_2}), \tag{5.54}$$

where $\mathbf{P}_{Z_1} = \mathbf{Z}_1(\mathbf{Z}'_1\mathbf{Z}_1)^{-1}\mathbf{Z}'_1$ and $\mathbf{P}_{Z_2} = \mathbf{Z}_2(\mathbf{Z}'_2\mathbf{Z}_2)^{-1}\mathbf{Z}'_2$ are the orthogonal projector matrices onto $sp(\mathbf{Z}_1)$ and $sp(\mathbf{Z}_2)$, respectively, with $sp(\mathbf{Z}_*)$ denoting the column space of $\mathbf{Z}_*$.

According to the definition (5.54), the loss function (5.42) can be viewed as the minimization of subspace distances between spanned by $\mathbf{F}$ and $\mathbf{G}_i\mathbf{Q}_i$ $(i = 1, \ldots, p)$ if c is equivalent to $Max\{rank(\mathbf{F}), rank(\mathbf{G}_i\mathbf{Q}_i)\}$. In (5.42), we have

$$Max\{rank(\mathbf{F}), rank(\mathbf{G}_i\mathbf{Q}_i)\} = Max\{c, R_i = min(c, K_i - 1)\}. \tag{5.55}$$

When $c \leq K_i - 1$,

$$Max\{c, R_i = min(c, K_i - 1)\} = Max(c, c) = c \tag{5.56}$$

is obtained. Next, assume the case when $c > K_i - 1$. then we have

$$Max\{c, R_i = min(c, K_i - 1)\} = Max(c, K_i - 1) = c. \tag{5.57}$$

The results above imply that the equation $Max\{rank(\mathbf{F}), rank(\mathbf{G}_i\mathbf{Q}_i)\} = c$ holds in each i. Thus, homogeneity analysis can be defined as the minimization of subspaces between spanned by the quantified data and by components. □

Property 1 implies that oblique rotation does not affect subspace distances and obliquely-rotated solutions still can be interpreted as the graphical representation.

Property 2. $\mathbf{P}_{gq_i}\mathbf{F} = \mathbf{G}_i\mathbf{Q}_i(\mathbf{Q'}_i\mathbf{G'}_i\mathbf{G}_i\mathbf{Q}_i)^{-1}\mathbf{Q'}_i\mathbf{G'}_i\mathbf{F}$ is eqivalent to $\mathbf{P}_{g_i}\mathbf{F} = \mathbf{G}_i(\mathbf{G'}_i\mathbf{G}_i)^{-1}\mathbf{G'}_i\mathbf{F}$ for each i.

Proof. $\mathbf{P}_{gq_i}\mathbf{F}$ can be descriced as

$$\begin{aligned}\mathbf{P}_{gq_i}\mathbf{F} &= n^{-1}\mathbf{G}_i\mathbf{Q}_i\mathbf{Q'}_i\mathbf{G'}_i\mathbf{F} \\ &= \mathbf{G}_i\mathbf{Q}_i\mathbf{A}_i \\ &= \mathbf{G}_i\mathbf{W}_i \\ &= \mathbf{G}_i(\mathbf{G'}_i\mathbf{G}_i)^{-1}\mathbf{G'}_i\mathbf{F} \\ &= \mathbf{P}_{g_i}\mathbf{F}. \end{aligned} \tag{5.58}$$

The equation above holds in each i. □

Using property 2, the loss function (5.42) can be expressed as

$$\begin{aligned}\sum_{i=1}^{p} tr(npc - n\mathbf{P}_{gq_i}\mathbf{P}_f) &= n\{pc - \sum_{i=1}^{p} tr(\mathbf{P}_{g_i}\mathbf{P}_f)\} \\ &= n\{pc - n^{-1}tr\mathbf{F}'(\sum_{i=1}^{p}\mathbf{G}_i(\mathbf{G}'_i\mathbf{G}_i)^{-1}\mathbf{G}'_i)\mathbf{F}\} \\ &= npc - tr\mathbf{F}'\mathbf{JGD}^{-1/2}\mathbf{D}^{-1/2}\mathbf{G}'\mathbf{JF} \\ &= npc - ntr\mathbf{K}'_c\mathbf{K}\mathbf{\Delta}^2\mathbf{K}'\mathbf{K}_c \\ &= n(pc - tr\mathbf{\Delta}_c^2). \end{aligned} \tag{5.59}$$

Property 3. The three expressions of the discrimination measures are equal when rotation is applied based on the proposed reformulation.

Proof. According to (5.28) and (5.18), rotated $\mathbf{W}_i$ can be described as

$$\begin{aligned}\mathbf{Q}_i\mathbf{T}_{oblq_i}(\mathbf{T}'_{oblq_i}\mathbf{Q}'_i\mathbf{G}'_i\mathbf{G}_i\mathbf{Q}_i\mathbf{T}_{oblq_i})^{-1}\mathbf{T}'_{oblq_i}\mathbf{Q}'_i\mathbf{G}'_i\mathbf{FT} &= n^{-1}\mathbf{Q}_i\mathbf{Q}'_i\mathbf{G}_i\mathbf{FT} \\ &= \mathbf{D}_i^{-1}\mathbf{G}_i\mathbf{FT} \\ &= \mathbf{W}_i\mathbf{T}. \end{aligned} \tag{5.60}$$

The equation (5.60) shows that matrices $\mathbf{W}_i$ for each i are rotated by the same matrix in $\mathbf{F}$. Thus, The three expressions are equal in the proposed reformulation. □

In contrast to Adachi (2004), oblique rotation is allowed in the proposed reformulation without affecting the discrimination measures.

5.6 Unpenalized sparse multiple correspondence analysis

Mori *et al.* (2016) extended unpenalized sparse PCA proposed by Adachi and Trendafilov (2016) to nonmetric data via the quantification technique. Unpenalized sparse MCA loss function can be obtained if the data matrix $\mathbf{X}$ is replaced to the quantified data $\mathbf{G}_i\mathbf{Q}_i$. The loss function

$$\sum_{i=1}^{p} ||\mathbf{G}_i\mathbf{Q}_i - \mathbf{FA}'_i||^2 \tag{5.61}$$

is minimized, subject to $card(\mathbf{A}) = l$ and ordinary MCA constraints $n^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_c$, $\mathbf{JF} = \mathbf{F}$, $\mathbf{1}'_n\mathbf{G}_i\mathbf{Q}_i = \mathbf{0}'_{K_i}$, and $n^{-1}\mathbf{Q}'_i\mathbf{G}'_i\mathbf{G}_i\mathbf{Q}_i = \mathbf{I}_{R_i}$.

The loss function (5.61) can be rewritten as

$$\sum_{i=1}^{p} ||\mathbf{G}_i\mathbf{Q}_i - \mathbf{F}\tilde{\mathbf{A}}'_i - \mathbf{F}\tilde{\mathbf{A}}'_i + \mathbf{F}\mathbf{A}'_i||^2 = \sum_{i=1}^{p} ||\mathbf{G}_i\mathbf{Q}_i - \mathbf{F}\tilde{\mathbf{A}}'_i||^2 + n||\mathbf{A} - \tilde{\mathbf{A}}||^2, \tag{5.62}$$

where $\tilde{\mathbf{A}} = n^{-1}\mathbf{B}'_Q\mathbf{G}'\mathbf{F}$. The loss function (5.62) can be solved by alternately performing three steps:

1. minimizing the loss function (5.62) over $\mathbf{Q}$ subject to $\mathbf{1}_n'\mathbf{G}_i\mathbf{Q}_i = \mathbf{0}_{K_i}'$, and $n^{-1}\mathbf{Q}_i'\mathbf{G}_i'\mathbf{G}_i\mathbf{Q}_i = \mathbf{I}_{R_i}$.
2. minimizing the loss function (5.62) over $\mathbf{F}$ subject to $\mathbf{J}\mathbf{F} = \mathbf{F}$ and $n^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_c$.
3. minimizing the loss function (5.62) over $\mathbf{A}$ subject to $card(\mathbf{A}) = l$.

Firstly, the estimations of $\mathbf{Q}_i$ for $i = 1, \ldots, p$ are considered, which are equivalent to the minimizations of $\sum_{i=1}^{p} ||\mathbf{G}_i\mathbf{Q}_i - \mathbf{F}\mathbf{A}'_i||^2$. The function $\sum_{i=1}^{p} ||\mathbf{G}_i\mathbf{Q}_i - \mathbf{F}\mathbf{A}'_i||^2$ can be re-expressed as $\sum_{i=1}^{p} ||\mathbf{G}_i\mathbf{D}_i^{-1/2}\mathbf{D}_i^{1/2}\mathbf{Q}_i - \mathbf{F}\mathbf{A}'_i||^2$. Thus, the minimization for the quantification parameters subject to the constraints $\mathbf{1}_n'\mathbf{G}_i\mathbf{Q}_i = \mathbf{0}'_{K_i}$ and $n^{-1}\mathbf{Q}_i'\mathbf{G}_i'\mathbf{G}_i\mathbf{Q}_i = \mathbf{I}_{R_i}$ reduces to the orthogonal Procrustes problem and the optimal $\mathbf{Q}_i$ can be obtained by the SVD of $\mathbf{D}_i^{-1/2}\mathbf{G}'_i\mathbf{F}\mathbf{A}'_i$.

In estimation of $\mathbf{F}$, minimizing the loss function (5.62) is equivalent to the minimization of $\sum_{i=1}^{p} ||\mathbf{G}_i\mathbf{Q}_i - \mathbf{F}\mathbf{A}'_i||^2$. Hence, the optimal $\mathbf{F}$ can be obtained by the SVD of $\sum \mathbf{G}_i\mathbf{Q}_i\mathbf{A}_i$.

For updating $\mathbf{A}$, the algorithm proposed by Adachi and Trendafilov (2016) can be used because the term $n||\tilde{\mathbf{A}} - \mathbf{A}||^2$ is only relevant to $\mathbf{A}$ in the loss function (5.62). Therefore, minimizing (5.62) is equivalent to the minimization of $n||\tilde{\mathbf{A}} - \mathbf{A}||^2$ subject to $card(A) = l$. The optimal $\mathbf{A}$ can be obtained by

$$a_{ij} = \begin{cases} 0, & iff\ \tilde{a}_{ij}^2 \leq \tilde{a}_{(l)}^2 \\ \tilde{a}_{ij}, & otherwise, \end{cases} \tag{5.63}$$

where $\tilde{a}_{(l)}^2$ is the l-th smallest value among all $\tilde{a}_{ij}^2$.

The whole algorithm of sparse MCA can be described as follows:

Step 1: Set $card(\mathbf{A}) = l$ and the number of components c.

Step 2: Initialize the parameters.

Step 3: Update $\mathbf{Q}_i$ for each i with the SVD of $\mathbf{D}_i^{-1/2}\mathbf{G}'_i\mathbf{F}\mathbf{A}'_i$.

Step 4: Update $\mathbf{F}$ with the SVD of $\sum_{i=1}^{p} \mathbf{G}_i\mathbf{Q}_i\mathbf{A}_i$.

Step 5: Update $\mathbf{A}$ with (5.63)

Step 6: Finish if convergence is reached; otherwise, back to Step 3.

In the same manner as PCA and sparse PCA, the following inequality holds in MCA and sparse MCA:

$$\sum_{i=1}^{p} ||\mathbf{G}_i\mathbf{Q}_i - \mathbf{F}\tilde{\mathbf{A}}'_i||^2 + n||\mathbf{A} - \tilde{\mathbf{A}}||^2 \geq \sum_{i=1}^{p} ||\mathbf{G}_i\mathbf{Q}_i - \mathbf{F}\tilde{\mathbf{A}}'_i||^2. \tag{5.64}$$

It implies that MCA fits better than sparse MCA and that common variances in MCA are greater than or equal to those in sparse MCA.

5.7 Joint dimension reduction and clustering in multiple correspondence analysis

In this section, joint PCA and k-means is extended to nonmetric data using the quantification. In the literature on qualitative variables, several methods for joint dimension reduction and clustering for qualitative data have been proposed. van Buuren and Heiser (1989) proposed a method for clustering and optimal scaling of variables, named GROUPALS, in which object scores are expressed as the product of the cluster membership and the cluster centroids matrices. In other techniques, Hwang, Dillon and Takane (2006) and Mitsuhiro and Yadohisa (2015) combined MCA with k-means clustering for simultaneous analysis, in which cluster centroids and variable categories are jointly displayed in a low-dimensional space. Here, we focus on techniques proposed by Hwang *et al* (2006) and Mitsuhiro and Yadohisa (2015) for joint dimension reduction and clustering of qualitative variables, which is simply extension of (4.14).

5.7.1 Hwang, Dillon and Takane's method

Hwang *et al* (2006) proposed joint MCA and k-means method, in which the loss function is defined as

$$\alpha_1 \sum_{i=1}^{p} ||\mathbf{F} - \mathbf{G}_i\mathbf{W}_i||^2 + \alpha_2||\mathbf{F} - \mathbf{PC}||^2, \tag{5.65}$$

and minimized over $\mathbf{F}$, $\mathbf{C}$, $\mathbf{W}$ and $\mathbf{P}$, subject to $\mathbf{F}'\mathbf{F} = \mathbf{I}_c$ and $\alpha_1 + \alpha_2 = 1$. It is obvious that the loss function (5.65) reduces to the standard homogeneity analysis criterion for MCA (Gifi, 1990) if $\alpha_1 = 1$, and to the standard k-means criterion if $\alpha_2 = 1$.

An alternating least squares algorithm is considered to minimize the loss function (5.65), which comprises three main steps. Firstly, $\mathbf{F}$ is updated for fixed $\mathbf{W}$, $\mathbf{P}$ and $\mathbf{C}$. The loss function (5.65) is expanded as

$$-2tr\{\mathbf{F}'(\alpha_1 \sum_{i=1}^{p} \mathbf{G}_i\mathbf{W}_i + \alpha_2\mathbf{PC})\} + c_{HDT}, \tag{5.66}$$

where c_{HDT} is a constant, which is irrelevant to the $\mathbf{F}$. The minimization of the loss function (5.65) with respect to $\mathbf{F}$ is equivalent to the maximization of (5.66). This optimization can be achieved by the SVD of $\alpha_1 \sum_{i=1}^{p} \mathbf{G}_i\mathbf{W}_i + \alpha_2\mathbf{PC}$.

Secondly, the optimizations of $\mathbf{P}$ and $\mathbf{C}$ are considered. This can be attained by minimizing the second term of the loss function (5.65) over $\mathbf{P}$ and $\mathbf{C}$. It suggests that the ordinary k-means algorithm can be used for optimization:

$$\mathbf{C} = (\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'\mathbf{F}, \tag{5.67}$$

and

$$p_{ik} = \begin{cases} 1 & \text{if } ||\mathbf{f}_{(i)} - \mathbf{c}_{(k)}||^2 \text{attains the minimum among (5.65)} \\ 0 & \text{otherwise.} \end{cases} \tag{5.68}$$

Finally, we consider the update of $\mathbf{W} = [\mathbf{W}_1, \ldots, \mathbf{W}_p]$ for fixed $\mathbf{F}$, $\mathbf{P}$ and $\mathbf{C}$. The optimal $\mathbf{W}_i$ can be accquired by solving the regression problem (ten Berge, 1993), because $\mathbf{W}_i$ is involved only in the first term of the loss function (5.65). Thus, $\mathbf{W}_i$ can be updated by

$$\mathbf{W}_i = (\mathbf{G}'_i\mathbf{G}_i)^{-1}\mathbf{G}'_i\mathbf{F} \tag{5.69}$$

for each i.

The algorithm is summarized as follows:

Step 1: Specify the number of components c and clusters γ.

Step 2: Initialize the parameters.

Step 3: Update the component score matrix $\mathbf{F}$ by the SVD of $\alpha_1 \sum_{i=1}^{p} \mathbf{G}_i\mathbf{W}_i + \alpha_2\mathbf{PC}$.

Step 4: Update the membership matrix $\mathbf{P}$ and the cluster centroid matrix $\mathbf{C}$ using the ordinary k-means algorithm.

Step 5: Update the category coordinate matrices $\mathbf{W}_i$ for each variable by $\mathbf{W}_i = (\mathbf{G}'_i\mathbf{G}_i)^{-1}\mathbf{G}'_i\mathbf{F}$.

Step 1: Finish if convergence is reached; otherwise, back to Step 3.

5.7.2 Mitsuhiro and Yadohisa's method

There is another type of joint MCA and clustering loss function proposed by Mitsuhiro and Yadohisa (2015, 2018). Their proposed loss function is expressed as

$$\alpha_1 \sum_{i=1}^{p} ||\mathbf{G}_i\mathbf{Q}_i - \mathbf{F}\mathbf{A}'_i||^2 + \alpha_2||\mathbf{F} - \mathbf{PC}||^2, \tag{5.70}$$

and minimized over $\mathbf{F}$, $\mathbf{P}$, $\mathbf{C}$, $\mathbf{A}_i$ and $\mathbf{Q}_i$ ($i = 1, \ldots, p$), subject to $n^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_c, \mathbf{JF} = \mathbf{F}, \mathbf{JG}_i\mathbf{Q}_i = \mathbf{G}_i\mathbf{Q}_i$, and $n^{-1}\mathbf{Q}'_i\mathbf{G}'_i\mathbf{G}_i\mathbf{Q}_i = \mathbf{I}_{R_i}$. Comparing the loss function (5.70) with (5.65), the first term that define the MCA criterion are different. MCA is formulated by the homogeneity analysis criterion in Hwang *et al* (2006), whereas in Mitsuhiro and Yadohisa (2015), it is formulated as approximating a data matrix by a lower rank matrix using the quantification technique (Murakami *et al*, 1999; Adachi *et al*, 2011; Mori *et al.*, 2016).

The optimizations of $\mathbf{P}$ and $\mathbf{C}$ can be achieved by (5.67) and (5.68) because they are relevant to the second term. Thus, the optimization of $\mathbf{Q}_i$ and $\mathbf{A}_i$ for each i are considered below. Firstly, the optimization of $\mathbf{A}_i$ is discussed. The optimal $\mathbf{A}_i$ can be given by (5.18) because the first term in (5.70) is only relevant to $\mathbf{A}_i$ for each i. Expanding the loss function (5.70), we have

$$\sum_{i=1}^{p} -2tr(\mathbf{Q'}_i\mathbf{G'}_i\mathbf{FA'}_i) + c_{MY_q} = \sum_{i=1}^{p} -2tr\{\mathbf{Q'}_i\mathbf{D}_i^{1/2}\mathbf{D}_i^{-1/2}\mathbf{G'}_i(\mathbf{FA'}_i)\} + cc_{MY_q}. \tag{5.71}$$

Here, c_{MY_q} is a constant which is irelevant to $\mathbf{Q}_i$ for each i. Let us define the SVD as $\mathbf{D}_i^{-1/2}\mathbf{G'}_i(\mathbf{FA'}_i) = \dot{\mathbf{K}}\dot{\mathbf{\Lambda}}\dot{\mathbf{L}}'$. The optimal $\mathbf{D}_i^{1/2}\mathbf{Q}_i$ is given by

$$\mathbf{D}_i^{1/2}\mathbf{Q}_i = \dot{\mathbf{K}}\dot{\mathbf{L}}' \tag{5.72}$$

(ten Berge, 1993), and then we have

$$\mathbf{Q}_i = n^{1/2}(\mathbf{G}_i'\mathbf{G}_i)^{-1/2}\dot{\mathbf{K}}\dot{\mathbf{L}}'. \tag{5.73}$$

Recall that the first term in (5.70) can be described as

$$\sum_{i=1}^{p} ||\mathbf{G}_i\mathbf{Q}_i - \mathbf{FA'}_i||^2 = \sum_{i=1}^{p} ||\mathbf{F} - \mathbf{G}_i\mathbf{W}_i||^2 + \sum_{i=1}^{p} ||\mathbf{D}_i^{1/2}\mathbf{W}_i - \mathbf{D}_i^{1/2}\mathbf{Q}_i\mathbf{A}_i||^2 + c_{NPCA_Q}. \tag{5.74}$$

The first terms in (5.65) and (5.70) are essentially equivalent when the quantification dimensions are set to be $min(K_i - 1, c)$ for each variable. Thus, (5.65) and (5.70) give the same solutions (Mori *et al.*, 2016). This result implies that $\mathbf{F}$ can be updated by replacing $\mathbf{W}_i$ to $\mathbf{Q}_i\mathbf{A}_i$ in (5.66).

The whole algorithm is summarized as follows:

Step 1: Specify the number of components c and clusters γ.

Step 2: Initialize the parameters.

Step 3: Update the component score matrix $\mathbf{F}$ by the SVD of $\alpha_1 \sum_{i=1}^{p} \mathbf{G}_i\mathbf{Q}_i\mathbf{A}_i + \alpha_2\mathbf{PC}$.

Step 4: Update the quantification matrices $\mathbf{Q}_i$ and the loading matrices $\mathbf{A}_i$ by (5.73) and solve the regression problem.

Step 5: Update the membership matrix $\mathbf{P}$ and the cluster centroid matrix $\mathbf{C}$ using the ordinary k-means algorithm.

Step 6: Finish if convergence is reached; otherwise, back to Step 3.

5.7.3 Rotational freedom in joint multiple correspondence analysis and clustering

Recall the loss function of joint MCA and clustering

$$\alpha_1 \sum_{i=1}^{p} ||\mathbf{G}_i\mathbf{Q}_i - \mathbf{F}\mathbf{A}'_i||^2 + \alpha_2||\mathbf{F} - \mathbf{P}\mathbf{C}||^2, \tag{5.75}$$

which can be rewriten as

$$\begin{aligned}&\alpha_1 \sum_{i=1}^{p} ||\mathbf{F} - \mathbf{G}_i\mathbf{Q}_i\mathbf{A}_i||^2 + \alpha_2||\mathbf{F} - \mathbf{P}\mathbf{C}||^2 + c_{HA}\\ &= -\alpha_1 \sum_{i=1}^{p} tr(\mathbf{P}_f\mathbf{P}_{gq_i}) - \alpha_2 tr(\mathbf{P}_f\mathbf{P}_p) + c_{HA}.\end{aligned} \tag{5.76}$$

Here, c_{HA} is a constant which is irrelevant to the paramters. As shown in section 5.5, the equation (5.76) suggests that the solutions of joint MCA and clustering can be transformed by nonsingular matrices; joint MCA and k-means solutions have rotational freedom and can be obliquely rotated.

5.7.4 Joint sparse multiple correspondence analysis and clustering

In the same manner as joint PCA and k-means, joint MCA and k-means can be combined with sparse method. The loss function of joint sparse MCA and clustering can be defined as

$$\alpha_1 \sum_{i=1}^{p} ||\mathbf{G}_i\mathbf{Q}_i - \mathbf{F}\mathbf{A}'_i||^2 + \alpha_2||\mathbf{F} - \mathbf{P}\mathbf{C}||^2 = \alpha_1 \sum_{i=1}^{p} ||\mathbf{G}_i\mathbf{Q}_i - \mathbf{F}\tilde{\mathbf{A}}'_i||^2 + \alpha_2||\mathbf{F} - \mathbf{P}\mathbf{C}||^2 + n\alpha_1||\mathbf{A} - \tilde{\mathbf{A}}||^2. \tag{5.77}$$

Here, $\mathbf{A}$ is assumed to have l zero elements ($card(\mathbf{A}) = l$) and $\tilde{\mathbf{A}} = n^{-1}\mathbf{B}'_Q\mathbf{G}'\mathbf{F}$. The parameters can be updated by combining Mitsuhiro and Yadohisa (2015) algorithm with sparse MCA one: the updates of $\mathbf{F}$, $\mathbf{Q}_i$, $\mathbf{P}$ and $\mathbf{C}$ can be achieved using Mitsuhiro and Yadohisa (2015) algorithm and the optimal $\mathbf{A}$ is given by sparse MCA algorithm (5.63).

The estimation algorithm can be described as follows:

Step 1: Set c, γ, and $card(\mathbf{A}) = l$.

Step 2: Initialize the parameters.

Step 3: Update $\mathbf{Q}_i$ for each i with the SVD of $\mathbf{D}_i^{-1/2}\mathbf{G}'_i\mathbf{F}\mathbf{A}'_i$.

Step 4: Update $\mathbf{F}$ with the SVD of $\alpha_1 \sum_{i=1}^{p} \mathbf{G}_i\mathbf{Q}_i\mathbf{A}_i + \alpha_2\mathbf{P}\mathbf{C}$.

Step 5: Update $\mathbf{A}$ with (5.63).

Step 6: Update $\mathbf{P}$ with (5.68).

Step 7: Update $\mathbf{C}$ with (5.67).

Step 8: Finish if convergence is reached; otherwise, back to Step 3.

6

Nonmetric exploratory factor analysis

6.1 From matrix decomposion factor analysis to nonmetric exploratory factor analysis

Here, we extend MDFA to handle nonmetric data via the quantification. Makino (2015) proposed the extended version of MDFA to nonmetric data, which is called generalized data-fitting factor analysis (GDFFA). Replacing i-th column vector of data matrix $\mathbf{x}_{(i)}$ in (2.40) to $\mathbf{G}_i\mathbf{Q}_i$, we have the nonmetric MDFA loss function:

$$\sum_{i=1}^{p} \|\mathbf{G}_i\mathbf{Q}_i - \mathbf{F}\mathbf{A'}_i - \mathbf{U}_i\mathbf{\Psi}_i\|^2 = \|\mathbf{G}\mathbf{B}_Q - \mathbf{S}\mathbf{\Lambda}'\|^2. \tag{6.1}$$

Here, we denote the each parameters as $\mathbf{B}_Q = b - diag(\mathbf{Q}_1, \ldots, \mathbf{Q}_p)$, $\mathbf{A} = [\mathbf{A'}_1, \ldots, \mathbf{A'}_p]'$, $\mathbf{U} = [\mathbf{U}_1, \ldots, \mathbf{U}_p]$, $\mathbf{\Psi} = b - diag(\mathbf{\Psi}_1, \ldots, \mathbf{\Psi}_p)$, $\mathbf{S} = [\mathbf{F}, \mathbf{U}]$ and $\mathbf{\Lambda} = [\mathbf{A}, \mathbf{\Psi}]$. The loss function is minimized over $\mathbf{F}$, $\mathbf{A}$, $\mathbf{U}$, $\mathbf{\Psi}$, and $\mathbf{Q}_i$ ($i = 1, \ldots, p$), subject to the following constraints:

$$\mathbf{J}\mathbf{G}_i\mathbf{Q}_i = \mathbf{G}_i\mathbf{Q}_i, n^{-1}\mathbf{Q}'\mathbf{G'}_i\mathbf{G}_i\mathbf{Q}_i = \mathbf{I}_{R_i}, \tag{6.2}$$

$$\mathbf{J}\mathbf{F} = \mathbf{F}, n^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_c, \mathbf{J}\mathbf{U} = \mathbf{U}, n^{-1}\mathbf{U}'\mathbf{U} = \mathbf{I}_p, \tag{6.3}$$

$$n^{-1}\mathbf{F}'\mathbf{U} = \mathbf{O}, \mathbf{\Psi} \text{ is diagonal.} \tag{6.4}$$

Note that the matrix-formed parameters $\mathbf{A}_i$, $\mathbf{U}_i$, $\mathbf{\Psi}_i$, and $\mathbf{Q}_i$ reduce to the vectors or the scalar $\mathbf{a}_i$, $\mathbf{u}_i$, ψ_i and $\mathbf{q}_i$ when R_i is set to one. Makino's GDFFA is considered as the generalization of MDFA to nonmetric data, and thus we call (6.1) as generalized matrix decomposition factor analysis (GMDFA) in this thesis.

The parameters in GMDFA can be estimated by alternate iterations of the operations that quantify the categorical variables and that obtain the EFA model parameters. Firstly, we consider the quantification estimation procedure. The loss function (6.1) is expanded as

$$\begin{aligned}
&\sum_{i=1}^{p} -2tr(\mathbf{Q}'_i\mathbf{G}'_i\mathbf{F}\mathbf{A}'_i) - 2tr(\mathbf{Q}'_i\mathbf{G}'_i\mathbf{U}_i\mathbf{\Psi}_i) + c_{FMDFA_{qi}} \\
&= -2tr\{\mathbf{Q}'_i\mathbf{G}'_i(\mathbf{F}\mathbf{A}'_i + \mathbf{U}_i\mathbf{\Psi}_i)\} + c_{FMDFA_{qi}} \\
&= -2tr\{\mathbf{Q}'_i\mathbf{D}_i^{1/2}\mathbf{D}_i^{-1/2}\mathbf{G}'_i(\mathbf{F}\mathbf{A}'_i + \mathbf{U}_i\mathbf{\Psi}_i)\} + c_{FMDFA_{qi}}. \qquad (6.5)
\end{aligned}$$

Here, $c_{FMDFA_{qi}}$ is a constant, which is irrelevant to $\mathbf{Q}_i$ for each i. Let us define the SVD as $\mathbf{D}_i^{-1/2}\mathbf{G}'_i(\mathbf{F}\mathbf{A}'_i + \mathbf{U}_i\mathbf{\Psi}_i) = \mathbf{K}\mathbf{\Delta}\mathbf{L}'$ with $\mathbf{K}'\mathbf{K} = \mathbf{L}'\mathbf{L} = \mathbf{L}\mathbf{L}' = \mathbf{I}_{R_i}$, and let $\mathbf{\Delta}$ be the $q \times q$ diagonal matrix whose diagonal elements are arranged in descending order. The optimal $\mathbf{D}_i^{1/2}\mathbf{Q}_i$ is given by

$$\mathbf{D}_i^{1/2}\mathbf{Q}_i = \mathbf{K}\mathbf{L}' \qquad (6.6)$$

(ten Berge, 1993), and then we have

$$\mathbf{Q}_i = n^{1/2}(\mathbf{G}_i^{'}\mathbf{G}_i)^{-1/2}\mathbf{K}\mathbf{L}'. \qquad (6.7)$$

Therefore, the optimal $\mathbf{Q}_i$ can be given by the SVD of $\mathbf{D}_i^{-1/2}\mathbf{G}'_i(\mathbf{F}\mathbf{A}'_i + \mathbf{U}_i\mathbf{\Psi}_i)$ for each i ($i = 1, \ldots, p$). The rank of $\mathbf{D}_i^{-1/2}\mathbf{G}'_i(\mathbf{F}\mathbf{A}'_i + \mathbf{U}_i\mathbf{\Psi}_i)$ at most equals the smaller of $K_i - 1$ or c. Consequently, the dimension of quantification R_i is the upper bounded by $\min(K_i - 1, c)$.

After updating the quantification parameters, the model parameters are updated using the MDFA algorithm; the model parameters can be obtained if the observed numerical data are replaced by quantified data in the MDFA algorithm. Thus, optimal $\mathbf{A}$ and $\mathbf{\Psi}$ are given by

$$\mathbf{A}_i = n^{-1}\mathbf{Q}'_i\mathbf{G}'_i\mathbf{F}, \qquad (6.8)$$

$$\mathbf{\Psi} = n^{-1}diag(\mathbf{Q}'_i\mathbf{G}'_i\mathbf{U}). \qquad (6.9)$$

Next, the minimization over $\mathbf{S}$ is considered. The loss function (6.1) can be expressed as

$$\sum_{i=1}^{p} -tr(\mathbf{S}'\mathbf{G}_i\mathbf{Q}_i[\mathbf{A}_i, \mathbf{\Psi}_i]) + c_{GMDFA_S} = -tr(\mathbf{S}'\mathbf{G}\mathbf{B}_Q\mathbf{\Lambda}) + c_{GMDFA_S}, \qquad (6.10)$$

where c_{GMDFA_S} is a constant which is irrelevant to $\mathbf{S}$. Minimizing the loss fucntion over $\mathbf{S}$ can be achieved by maximizing $tr(\mathbf{S}'\mathbf{G}\mathbf{B}_Q\mathbf{\Lambda})$ over $\mathbf{S}$. As shown in chapter 2, the maximization of the trace function $tr(\mathbf{S}'\mathbf{G}\mathbf{B}_Q\mathbf{\Lambda})$ is achieved by

$$\mathbf{S} = \sqrt{n}\hat{\mathbf{K}}\hat{\mathbf{L}}' = \sqrt{n}\hat{\mathbf{K}}_1\hat{\mathbf{L}}'_1 + \sqrt{n}\hat{\mathbf{K}}_2\hat{\mathbf{L}}'_2. \qquad (6.11)$$

Here, $\hat{\mathbf{K}} = [\hat{\mathbf{K}}_1, \hat{\mathbf{K}}_2]$ and $\hat{\mathbf{L}} = [\hat{\mathbf{L}}_1, \hat{\mathbf{L}}_2]$ are block matrices of left and right singular vectors with the SVD of $\mathbf{GB}_Q\mathbf{\Lambda}$. $\mathbf{S}$ is satisfied with $\mathbf{S} = \mathbf{JS}$ because $\mathbf{GB}_Q = \mathbf{JGB}_Q$ holds.

The algorithm is summarized as follows:

Step 1: Initialize the parameters.

Step 2: Update the quantification parameters by solving the orthogonal Procrustes problems.

Step 3: Update the EFA model parameters using the MDFA algorithm.

Step 4: Finish if convergence is reached; otherwise, back to Step 3.

6.2 FACTALS

Takane, Young, and de Leeuw (1979) proposed a nonmetric factor analysis procedure for analyzing categorical variables, which is named as FACTALS. In FACTALS, the quantification technique is used to deal with categorical variables in the same manner as GMDFA. FACTALS and GMDFA are both nonmetric factor analysis procedures, but the assumptions made for quantification are different. The difference between FACTALS and GMDFA is most clearly seen by expressing both models:

$$\text{GMDFA: } \mathbf{G}_i\mathbf{Q}_i \cong \mathbf{FA}_i + \mathbf{U}_i\Psi_i, \tag{6.12}$$

$$\text{FACTALS: } \mathbf{G}_i\mathbf{q}_i \cong \mathbf{Fa}_i + \psi_i\mathbf{u}_i. \tag{6.13}$$

GMDFA model reduces to FACTALS when the quantification dimension is restricted to be one for all variables. It implies that GMDFA model includes FACTALS model. According to van der Burg and de Leeuw (1988), multiple quantification may complicates interpretation, particularly when categories in variables have a clear ordinal interpretation. However, cases are likely exist for which categories in a nominal variable cannot be scaled unidimensionally (Adachi and Murakami, 2011; Makino, 2015).

There is another difference between FACTALS and GMDFA. In FACTALS, Takane *et al.* (1979) defined the loss function as the minimization between the correlation matrix of the quantified data $\mathbf{R}^*$ and the model parameters:

$$||\mathbf{R}^* - \mathbf{AA'} - \mathbf{\Psi}^2||^2. \tag{6.14}$$

The parameters are estimated by fitting the model part to the quantified correlation structure in FACTALS, although the model part is directly fitted to the quantified data part in GMDFA. That is, FACTALS can be considered as the generalization (2.36) to deal with nonmetric data. It implies that common and unique factor scores cannot be simultaneously estimated by Takane *et al*. (1979).

6.3 Differences between related methods

GMDFA is compared with MCA and FACTALS in this section. It has been previously shown that MCA is a generalization of principal component analysis to nonmetric data (Adachi and Murakami, 2011), and that the unique term distinguishes the EFA model from the PCA model. Another difference, however, occurs between GMDFA and MCA. In MCA, two rotational freedoms exist in the quantification parameter and loading matrices:

$$\sum_{i=1}^{p} ||\mathbf{G}_i\mathbf{Q}_i - \mathbf{FA'}_i||^2 = \sum_{i=1}^{p} ||\mathbf{G}_i\mathbf{Q}_i\mathbf{T}_{oblq_i} - \mathbf{FT'}^{-1}\mathbf{T'A'}_i\mathbf{T}_{oblq_i}||^2. \tag{6.15}$$

In contrast, the unique term in GMDFA eliminates rotational indeterminacy of the quantification matrix, as seen in the following loss function:

$$\sum_{i=1}^{p} ||\mathbf{G}_i\mathbf{Q}_i - \mathbf{S}_i\mathbf{\Lambda'}_i||^2 = \sum_{i=1}^{p} ||\mathbf{G}_i\mathbf{Q}_i - [\mathbf{FT}^{'-1}\ \mathbf{U}_i][\mathbf{A}_i\mathbf{T}\ \ \mathbf{\Psi}_i]^{'}||^2. \tag{6.16}$$

Consequently, pre- and post-multiplied rotation of the MCA solution are precluded in GMDFA. It suggests that existing rotation techniques can be applied to the rotation of the loading matrix in GMDFA. Although the two techniques are very close, the unique term in GMDFA creates different structure from that in MCA (Makino, 2015).

6.3.1 Relationships among GMDFA, MCA, and FACTALS

As described above, GDFFA, MCA, and FACTALS are chiefly distinguished by their model parameters and the quantification dimensions. In FACTALS, the quantification parameters are constrained to be unidimensional in the analysis. In MCA literature, MCA reduces to nonmetric principal component analysis if all categorical variables are unidimensionally quantified (NCA; Murakami *et al*, 1999; Adachi *et al*, 2011). When all variables are numerical, FACTALS and NCA are equivalent to MDFA and PCA respectively, because only standardization is performed in the quantification step. Single quantification is a also generalization of numerical cases as well as multiple quantification. Therefore, the relationships among GMDFA, MCA, FACTALS, NCA, PCA, and MDFA are summarized in Table 6.1 (Makino, 2015).

Table 6.1: Relationships among GMDFA, MCA, FACTALS, NCA, PCA, and MDFA

		Model	
		PCA model	EFA model
	Numerical	PCA	MDFA
Quantification	Single	NCA	FACTALS
	Multiple	MCA	GMDFA

6.4 Sparse generalized matrix decomposition factor analysis

The loss function of sparse GMDFA can be straightforwardly formulated as

$$\begin{aligned}\sum_{i=1}^{p} \|\mathbf{G}_i\mathbf{Q}_i - \mathbf{F}\mathbf{A}'_i - \mathbf{U}_i\mathbf{\Psi}_i\|^2 &= \sum_{i=1}^{p} \|\mathbf{G}_i\mathbf{Q}_i - \mathbf{F}\tilde{\mathbf{A}}'_i - \mathbf{U}_i\mathbf{\Psi}_i\|^2 + n\sum_{i=1}^{p} \|\tilde{\mathbf{A}}_i - \mathbf{A}_i\|^2, \\ &= \|\mathbf{G}\mathbf{B}_Q - \mathbf{F}\tilde{\mathbf{A}}' - \mathbf{U}\mathbf{\Psi}\|^2 + n\|\tilde{\mathbf{A}} - \mathbf{A}\|^2, \end{aligned} \tag{6.17}$$

in which the data matrix $\mathbf{X}$ is replaced to the quantified data $\mathbf{G}\mathbf{B}_Q$ in (4.44). Here, $\tilde{\mathbf{A}}$ is defined as $\tilde{\mathbf{A}} = n^{-1}\mathbf{B}'_Q\mathbf{G}'\mathbf{F}$.

For updating $\mathbf{A}$, the algorithm proposed by Adachi and Trendafilov (2014, 2016) can be used because the term $n\|\tilde{\mathbf{A}} - \mathbf{A}\|^2$ is only relevant to $\mathbf{A}$ in the loss function (6.17): the optimal $\mathbf{A}$ can be obtained by

$$a_{ij} = \begin{cases} 0, & iff\ \tilde{a}_{ij}^2 \le \tilde{a}_{(l)}^2 \\ \tilde{a}_{ij} & otherwise, \end{cases} \tag{6.18}$$

where $\tilde{a}_{(l)}^2$ is the l-th smallest value among all $\tilde{a}_{ij}^2$.

The other parameters can be estimated using the GMDFA algorithm. The estimation algorithm of sparse GMDFA can be described as follows:

Step 1: Set the number of zero elements l and factors c.

Step 2: Initialize the parameters.

Step 3: Update $\mathbf{B}_Q$, $\mathbf{F}$, $\mathbf{U}$, $\mathbf{\Psi}$ with MDFA procedures.

Step 4: Update $\mathbf{A}$ can be obtained by (6.18).

Step 5: Finish if convergence is reached; otherwise, back to Step 3.

In the same manner as MDFA and sparse MDFA, the following inequality holds in GMDFA and sparse GMDFA:

$$\sum_{i=1}^{p} ||\mathbf{G}_i\mathbf{Q}_i - \mathbf{F}\tilde{\mathbf{A}}'_i - \mathbf{U}_i\mathbf{\Psi}_i||^2 + n\sum_{i=1}^{p} ||\tilde{\mathbf{A}}_i - \mathbf{A}_i||^2 \geq \sum_{i=1}^{p} ||\mathbf{G}_i\mathbf{Q}_i - \mathbf{F}\tilde{\mathbf{A}}'_i - \mathbf{U}_i\mathbf{\Psi}_i||^2. \tag{6.19}$$

It implies that GMDFA fits better than sparse GMDFA and that common variances in GMDFA are greater than or equal to those in sparse GDMFA.

6.5 Joint dimension reduction and clustering in nonmetric exploratory factor analysis

6.5.1 Joint generalized matrix decomposition factor analysis and clustering

Mitsuhiro and Yadohisa (2018) proposed the loss function of joint GMDFA and clustering, which is defined as

$$\alpha_1||\mathbf{G}\mathbf{B}_Q - \mathbf{F}\mathbf{A}' - \mathbf{U}\mathbf{\Psi}||^2 + \alpha_2||\mathbf{S} - \mathbf{P}\mathbf{C}||^2.$$

In this loss function, clustering is performed to both of common and unique factors, not common factor. As discussed in chapter4, it may be difficult to have well-clustered $\mathbf{F}$ and display cluster centroid in a low-dimensional space. Thus, we generalize the loss function (4.34), which is proposed in this thesis, to deal with nonmetric data.

Replacing the data matrix $\mathbf{X}$ to $\mathbf{GB}_Q$ in (4.34), we have the loss function of joint modified GMDFA and clustering:

$$\alpha_1 \|\mathbf{GB}_Q - \mathbf{FA}' - \mathbf{U\Psi}\|^2 + \alpha_2 \|\mathbf{F} - \mathbf{PC}\|^2. \tag{6.20}$$

The loss function is minimized over $\mathbf{B}_Q$, $\mathbf{A}$, $\mathbf{U}$, $\mathbf{\Psi}$, $\mathbf{P}$ and $\mathbf{C}$, subject to (6.2), (6.3), (6.4), (4.2) and (4.3). The optimal $\mathbf{B}_Q$, $\mathbf{F}$, $\mathbf{A}$, $\mathbf{U}$, $\mathbf{\Psi}$, $\mathbf{P}$ and $\mathbf{C}$ are given by the joint modified MDFA and clustering algorithm; replacing $\mathbf{X}$ to $\mathbf{GB}_Q$ in the joint modified MDFA and clustering algorithm, we have the joint GMDFA and clustering estimation algorithm:

Step 1: Set the number of components c and that of cluster γ.

Step 2: Initialize the parameters.

Step 3: Update the quantification parameters by solving the orthogonal Procrustes problems with (6.7).

Step 4: Update $\mathbf{F}$ by the SVD of $(\alpha_1(\mathbf{I} - \mathbf{P}_u)\mathbf{GB}_Q\mathbf{A} + \alpha_2(\mathbf{I} - \mathbf{P}_u)\mathbf{PC})$.

Step 5: Update $\mathbf{U}$ by the SVD of $(\mathbf{I} - \mathbf{P}_f)\mathbf{GB}_Q\mathbf{\Psi}$.

Step 6: Update $\mathbf{\Lambda}$ with (6.8) and (6.9).

Step 7: Update $\mathbf{P}$ with (4.31).

Step 8: Update $\mathbf{C}$ with (4.30).

Step 9: Finish if convergence is reached; otherwise, back to Step 3.

6.5.2 Joint sparse generalized matrix decomposition factor analysis and clustering

Joint sparse MDFA and clustering can be extended to handle nonmetric data. The loss function of joint sparse GMDFA and clustering is defined as the combination of (6.17) and (6.20):

$$\alpha_1 \|\mathbf{GB}_Q - \mathbf{FA}' - \mathbf{U\Psi}\|^2 + \alpha_2 \|\mathbf{F} - \mathbf{PC}\|^2 + \alpha_1 n \|\mathbf{A} - \tilde{\mathbf{A}}\|^2. \tag{6.21}$$

The loss function (6.21) is minimized over $\mathbf{B}_Q$, $\mathbf{F}$, $\mathbf{A}$, $\mathbf{U}$, $\mathbf{\Psi}$, $\mathbf{P}$ and $\mathbf{C}$, subject to (6.2), (6.3), (6.4), (4.2), (4.3) and $Card(\mathbf{A}) = l$. The optimal parameters can be obtained using the joint GMDFA and clustering, and sparse GMDFA algorithms. We can update $\mathbf{B}_Q$, $\mathbf{U}$ and $\mathbf{\Psi}$, $\mathbf{P}$ and $\mathbf{C}$ using the joint GMDFA and clustering algorithm because the first term of (6.21) is relevant to these parameters. The loading matrix $\mathbf{A}$ is given by the sparse GMDFA algorithm.

The whole algorithm of joint sparse GMDFA and clustering can be described as follows:

Step 1: Set the number of components c and that of cluster γ.

Step 2: Initialize the parameters.

Step 3: Update the quantification parameters by solving the orthogonal Procrustes problems with (6.7).

Step 4: Update $\mathbf{F}$ by the SVD of $(\alpha_1(\mathbf{I} - \mathbf{P}_u)\mathbf{G}\mathbf{B}_Q\mathbf{A}' + \alpha_2(\mathbf{I} - \mathbf{P}_u)\mathbf{P}\mathbf{C})$.

Step 5: Update $\mathbf{U}$ by the SVD of $(\mathbf{I} - \mathbf{P}_f)\mathbf{G}\mathbf{B}_Q\mathbf{\Psi}$.

Step 6: Update $\mathbf{A}$ with (6.8).

Step 7: Update $\mathbf{\Psi}$ with (6.9).

Step 8: Update $\mathbf{P}$ with (4.31).

Step 9: Update $\mathbf{C}$ with (4.30).

Step 10: Finish if convergence is reached; otherwise, back to Step 3.

7

Real data analysis

In this chapter, we illustrate the methods discussed in chapter 5 and 6 using two real datasets.

First example is baseball dataset (Makino, 2015). This dataset describes the scores of 62 batters on 7 items in Japanese professional baseball in 2010. The variables consist of batting average, runs, doubles, home runs, runs batted in, strikeouts and batting order. The first six variables are binary: 0 is obtained when the batters are categorized as lower group and 1 is assigned when the batters belong to higher group. The batting order is a nominal variable in which the batters are categorized into three groups by their batting order: group 1 comprising first and second (Nos.1 and 2), group 2 (Nos. 3−5), and group 3 (Nos. 6−9).

Second example is mathematics assessment dataset. The dataset originally contains 664 students on 30 items, but we picked out 23 items whose domains are arithmetic or geometry for illustration. The dataset is available from sirt package in R (Robitzsch, 2019). The all variables in mathematics assessment dataset are binary: 0 is obtained when incorrect response is observed and 1 is assigned to correct response.

Before we enter demonstrations, we discuss the sign interminacy property. Then, the rotation and the sparse techniques are demonstrated in section 7.2. In section 7.3, we compare joint dimension reductions and clustering methods with sparse one.

7.1 Sign indeterminacy

It should be mentioned that the methods discussed in chapter 5 and 6 have a sigh indeterminacy. Let $\mathbf{Z} = b - diag(\mathbf{Z}_1, \ldots, \mathbf{Z}_p)$ be a $R \times R$ diagonal matrix whose diagonal elements are 1 or -1

($i = 1, \ldots, p$). Then the loss functions are not affected by $\mathbf{Z}$ because $\mathbf{Z}_i^2 = \mathbf{I}_{R_i}$:

$$\sum_{i=1}^{p} ||\mathbf{F} - \mathbf{G}_i\mathbf{Q}_i\mathbf{A}_i||^2 = \sum_{i=1}^{p} ||\mathbf{F}\mathbf{Z} - n^{-1}\mathbf{G}_i\mathbf{Q}_i\mathbf{Z}_i\mathbf{Z}_i\mathbf{Q}'_i\mathbf{G}'_i\mathbf{F}||^2, \tag{7.1}$$

$$\sum_{i=1}^{p} ||\mathbf{G}_i\mathbf{Q}_i - \mathbf{F}\mathbf{A}'_i - \mathbf{U}_i\mathbf{\Psi}_i||^2 = \sum_{i=1}^{p} ||\mathbf{G}_i\mathbf{Q}_i\mathbf{Z}_i - \mathbf{F}\mathbf{A}'_i\mathbf{Z}_i - \mathbf{U}_i\mathbf{\Psi}_i\mathbf{Z}_i||^2, \tag{7.2}$$

$$\sum_{i=1}^{p} ||\mathbf{G}_i\mathbf{Q}_i - \mathbf{F}\tilde{\mathbf{A}}_i||^2 + n||\mathbf{A} - \tilde{\mathbf{A}}||^2 = \sum_{i=1}^{p} ||\mathbf{G}_i\mathbf{Q}_i\mathbf{Z}_i - \mathbf{F}\tilde{\mathbf{A}}_i\mathbf{Z}_i||^2 + n||\mathbf{A}\mathbf{Z}_i - \tilde{\mathbf{A}}\mathbf{Z}_i||^2, \tag{7.3}$$

$$||\mathbf{G}\mathbf{B}_Q - \mathbf{F}\tilde{\mathbf{A}} - \mathbf{U}\mathbf{\Psi}||^2 + n||\tilde{\mathbf{A}} - \mathbf{A}||^2 = ||\mathbf{G}\mathbf{B}_Q\mathbf{Z} - \mathbf{F}\tilde{\mathbf{A}}\mathbf{Z} - \mathbf{U}\mathbf{\Psi}\mathbf{Z}||^2 + n||\tilde{\mathbf{A}}\mathbf{Z} - \mathbf{A}\mathbf{Z}_i||^2. \tag{7.4}$$

It implies that the sign of quantification parameters can be changed without violating the fitness of the loss functions. This equations also hold in the loss functions joint MCA/GMDFA and clustering bacause they are defined as the combinations of the loss functions above. In real data analysis, we adjust the sign of the parameters for interpreting the results.

7.2 Real data analysis 1

In this section, the rotation and sparse techniques are demonstrated by baseball and mathematics assessment datasets.

7.2.1 Rotated and sparse solutions in baseball data

We set the number of components and factors to two, which will be interpreted as expressing whether batters hit for average (table setters) or power (sluggers) in the same as Makino (2015). According to Makino (2015), the loading matrix seems to have approximate simple structure. Using this prior knowledge, the number of zeros in target matrix is set to eight in this analysis. Here, the number of the quantification dimension is set to $min(K_i - 1, c)$ for each variable: the quantification dimension is set to one in the first six variables and to two in batting order.

The results of the rotated loading matrix in MCA and GMDFA are reported in table 7.1. As shown in table 7.1, approximate simple structure is obtained by the proposed joint simplimax rotation. Component1/ Factor1 are positively correlated with batting average, runs, doubles and first quantification dimension in batting order, which are associated with reliable batters: players whose main concern is to reach bases and hit constantly. Component2/ Factor2 are positively correlated with home runs, runs batted in, strikeouts and second quantidication dimension in batting order. This second factor characterizes sluggers who frequently hit home runs and drive in a lot of runs, but whose long swings lead to many strikeouts. Thus, the same components/factors previously identified in Makino (2015) were extracted in this analysis.

Next, we applied sparse MCA and GMDFA to this dataset. In sparse techniques, the number of zeros l must be specified. The users would like to obtain the solutions with ideal simple structure as little loss comparing to the rotation as possible. The plot in figure 7.1 shows the common variances on the y-axis and the number of zeros on the x-axis. In the figure, we can see that common variances do not change significantly when $l \leq 8$. It indicates that loss of common variances is relatively small if l is set to eight. Thus, we adopted the solution of $l = 8$ in this analysis. Table 7.2 expresses the loading matrices of sparse MCA and sparse GMDFA. As shown in table 7.2, the ideal simple structures are accquired by sparse techniques.

Lastly, the loss of sparse solutions is discussed. Table 7.3 reports the proportion of the loss of common variances in sparse solutions, which defined as

$$LCV = 1 - \frac{tr(\mathbf{A}'\mathbf{A})}{tr(\tilde{\mathbf{A}}'\tilde{\mathbf{A}})}. \tag{7.5}$$

The sparse solutions may not be desirable when large LCV is obtained. The values of LCV in this dataset are reported in table 7.3. Table 7.3 suggests that the solutions with ideal simple structure is obtained with little loss of common variances in this application.

7.2.2 Rotated and sparse solutions in mathematics assessment dataset

We set the number of components/factors to two because it is assumed that two different components/factors (arithmetic and geometry) exist in the dataset. In this data, items are expected to load on either one of arithmetic or geometry components/factors. Using the prior knowledge on this dataset, the number of zeros in target matrix of simplimax was set to twenty-three. All items are binary and thus the dimension of quantification is set to one. Note that the rotational freedom in the quantification parameters disappears in this data because they are unidimensional; joint simplimax rotation reduces to ordinary simplimax rotation here.

Table 7.1: Rotated loading matrices and correlations in baseball data

	MCA		GMDFA		
	Component1	Component2	Factor1	Factor2	Uniqueness
BA	**0.782**	-0.129	**0.685**	-0.111	0.529
R	**0.903**	0.027	**0.964**	0.014	0.051
D	**0.796**	0.123	**0.685**	0.12	0.482
HR	-0.022	**0.848**	-0.031	**0.814**	0.326
RBI	0.202	**0.87**	0.151	**0.923**	0.077
S	-0.141	**0.751**	-0.13	**0.603**	0.625
BO1	**0.597**	0.041	**0.436**	0.055	0.811
BO2	0.077	**0.726**	0.072	**0.574**	0.659

Component correlation		Factor correlation	
1	-0.083	1	-0.115
-0.083	1	-0.115	1

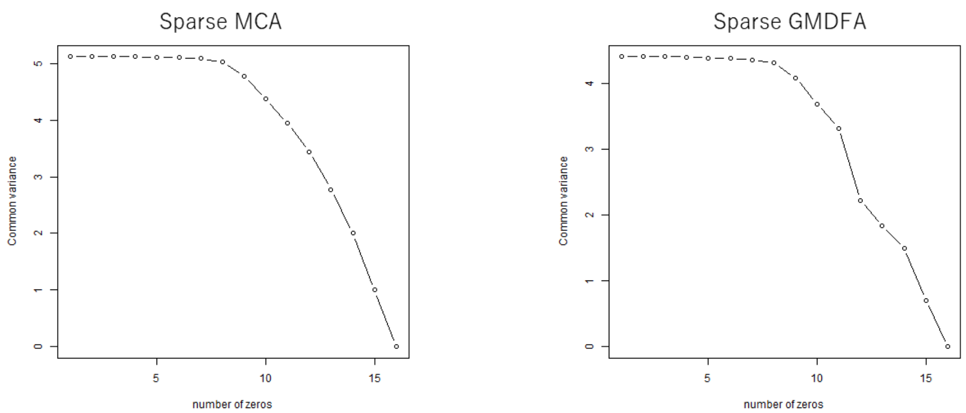


Figure 7.1: Plot of common variances in baseball dataset

Table 7.2: Sparse loading matrices and correlations in baseball data

	Sparse MCA		Sparse GMDFA		
	Component2	Component1	Factor1	Factor2	Uniqueness
BA	**0.784**	0	**0.676**	0	0.540
R	**0.906**	0	**0.99**	0	0.008
D	**0.799**	0	**0.679**	0	0.515
HR	0	**0.851**	0	**0.823**	0.315
RBI	0	**0.881**	0	**0.926**	0.097
S	0	**0.751**	0	**0.603**	0.625
BO1	**0.605**	0	**0.43**	0	0.811
BO2	0	**0.723**	0	**0.568**	0.670

Component correlation		Factor correlation	
1	0	1	0
0	1	0	1

Table 7.3: The loss of common variances in baseball dataset

	Sparse MCA/MCA	Sparse GMDFA/GMDFA
LCV	0.018	0.021

The rotated loadings via simplimax rotation are reported in table 7.4. As shown in table 7.4, approximate simple structure is obtained by the simplimax rotation. Component1/Factor1 are positively correlated with arithmetic items and Component2/Factor2 are positively connected to geometry items. It suggests that we have arithmetic and geometry components/factors are obtained as we expected before the analysis.

Next, sparse MCA and GMDFA were performed to this dataset. According to figure 7.2, the plots seem to be sharply dropped when l is around twenty-three. If l is greater than twenty-three, we have a solution in which a item loads on neither arithmetic nor geometry components/factors. Thus, we choose $l = 23$ solutions from viewpoint of interpretation of the result in this analysis. Table 7.5 expresses the loading matrices in sparse techniques. In the same manner as the rotated solutions, first component/factor are connected to arithmetic items and second ones are related to geometry items. It indicates that the expected components/factors were obtained by the sparse techniques as well as the rotation. The values of loss of common variances are reported in table 7.6. Table 7.6 shows that we have the sparse solutions with little loss of common variances compared with the rotated solutions.

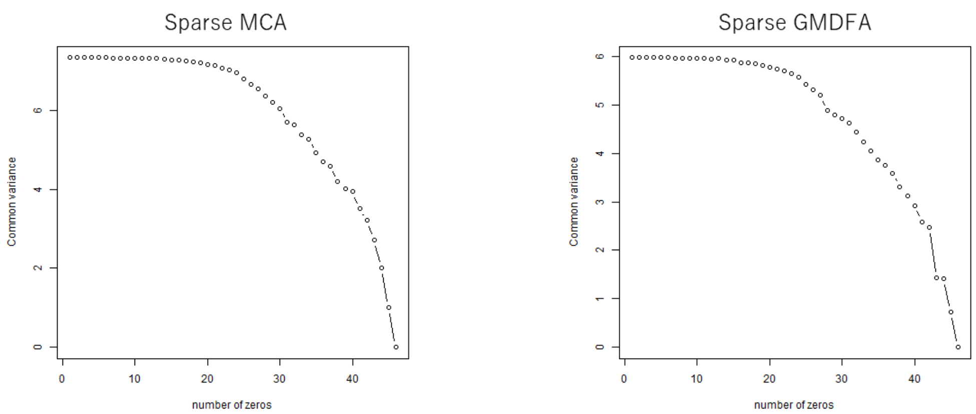


Figure 7.2: Plot of common variances in mathematics asssessment dataset

Table 7.4: Rotated loading matrices and correlations in mathematics assesment dataset

		MCA		GMDFA		
Item	Domain	Component1	Component2	Factor1	Factor2	Uniqueness
MA1	arithmetic	**0.331**	-0.093	**0.264**	-0.052	0.916
MA2	arithmetic	**0.481**	-0.089	**0.405**	-0.051	0.798
MA3	arithmetic	**0.508**	-0.006	**0.439**	0.014	0.731
MA4	arithmetic	**0.495**	0.01	**0.431**	0.023	0.735
MB1	arithmetic	**0.612**	-0.069	**0.568**	-0.064	0.607
MB2	arithmetic	**0.563**	-0.056	**0.516**	-0.049	0.661
MB3	arithmetic	**0.495**	0.11	**0.455**	0.091	0.666
MB4	arithmetic	**0.522**	0.12	**0.482**	0.103	0.621
MC1	arithmetic	**0.434**	0.003	**0.368**	0.017	0.813
MC2	arithmetic	**0.518**	-0.013	**0.453**	-0.003	0.73
MC3	arithmetic	**0.493**	-0.043	**0.426**	-0.025	0.774
MC4	arithmetic	**0.501**	0.026	**0.438**	0.032	0.724
MD1	arithmetic	**0.56**	-0.076	**0.494**	-0.055	0.687
MD2	arithmetic	**0.521**	-0.129	**0.45**	-0.096	0.746
MG3	arithmetic	**0.349**	0.218	**0.312**	0.195	0.666
MG4	arithmetic	**0.319**	0.251	**0.287**	0.22	0.671
MH1	geometry	-0.161	**0.696**	-0.158	**0.631**	0.578
MH2	geometry	0.019	**0.605**	0.015	**0.534**	0.631
MH3	geometry	-0.049	**0.674**	-0.054	**0.611**	0.563
MH4	geometry	-0.031	**0.629**	-0.032	**0.555**	0.629
MI1	geometry	-0.03	**0.433**	0.007	**0.308**	0.858
MI2	geometry	0.048	**0.468**	0.069	**0.353**	0.794
MI3	geometry	0.012	**0.475**	0.047	**0.344**	0.809

Component correlation		Factor correlation	
1	-0.382	1	-0.438
-0.382	1	-0.438	1

Table 7.5: Sparse loading matrices and correlations in mathematics asssessment dataset

		Sparse MCA		Sparse GMDFA		
Item	Domain	Component1	Component2	Factor1	Factor2	Uniqueness
MA1	arithmetic	**0.315**	0	**0.269**	0	0.918
MA2	arithmetic	**0.477**	0	**0.423**	0	0.802
MA3	arithmetic	**0.541**	0	**0.49**	0	0.731
MA4	arithmetic	**0.535**	0	**0.486**	0	0.734
MB1	arithmetic	**0.629**	0	**0.593**	0	0.616
MB2	arithmetic	**0.577**	0	**0.539**	0	0.67
MB3	arithmetic	**0.577**	0	**0.542**	0	0.669
MB4	arithmetic	**0.607**	0	**0.576**	0	0.624
MC1	arithmetic	**0.464**	0	**0.41**	0	0.815
MC2	arithmetic	**0.548**	0	**0.496**	0	0.732
MC3	arithmetic	**0.513**	0	**0.459**	0	0.776
MC4	arithmetic	**0.545**	0	**0.495**	0	0.726
MD1	arithmetic	**0.561**	0	**0.509**	0	0.699
MD2	arithmetic	**0.501**	0	**0.447**	0	0.759
MG3	arithmetic	**0.49**	0	**0.459**	0	0.671
MG4	arithmetic	**0.47**	0	**0.44**	0	0.686
MH1	geometry	0	**0.679**	0	**0.607**	0.611
MH2	geometry	0	**0.647**	0	**0.58**	0.623
MH3	geometry	0	**0.697**	0	**0.639**	0.559
MH4	geometry	0	**0.664**	0	**0.595**	0.615
MI1	geometry	0	**0.484**	0	**0.361**	0.842
MI2	geometry	0	**0.531**	0	**0.415**	0.784
MI3	geometry	0	**0.512**	0	**0.386**	0.808

Component correlation		Factor correlation	
1	0	1	0
0	1	0	1

Table 7.6: The loss of common variances in mathematics dataset

	Sparse MCA/MCA	Sparse GMDFA/GMDFA
LCV	0.0416	0.0537

7.3 Real data example 2

We demonstrate joint dimension reduction and clustering method using the same data in real data example 1.

7.3.1 Joint dimension reductions and clustering in baseball data

We set $c = 2$, $R = 8$ and $l = 8$ in the same way as real data example 1. In joint MCA/GMDFA and clustering, the number of clusters γ, and the scale parameters must be pre-specified. We used $\alpha_1 = \alpha_2 = 0.5$, which indicates that dimension reduction and clustering are equally treated in the analysis. The plot in figure 7.3 and 7.4 show the sum of squared errors (SSE) in the k-means term on the y-axis and the number of clusters on the x-axis. The SSE in k-means term expresses the dispersion inside the clusters and the smaller dispersion is desirable. We can see that SSE do not decrease significantly when $\gamma \geq 4$. Thus, the elbow is at $\gamma = 4$ indicating the optimal the number of cluster for this dataset.

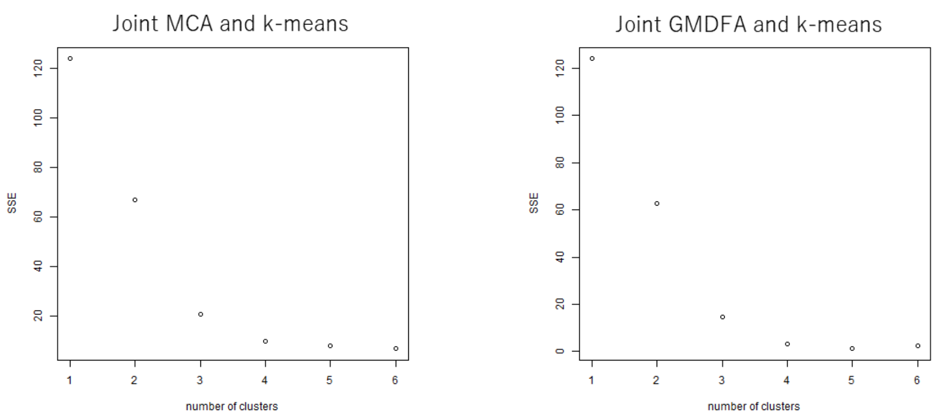


Figure 7.3: SSE plot of joint MCA/GMDFA and k-means in baseball dataset

Firstly, the estimated loadings are discussed. Table 7.7 and 7.8 report the obtained loading matrices in each analysis. As shown in chapter 5 and 6, the solutions of joint MCA/GMDFA and

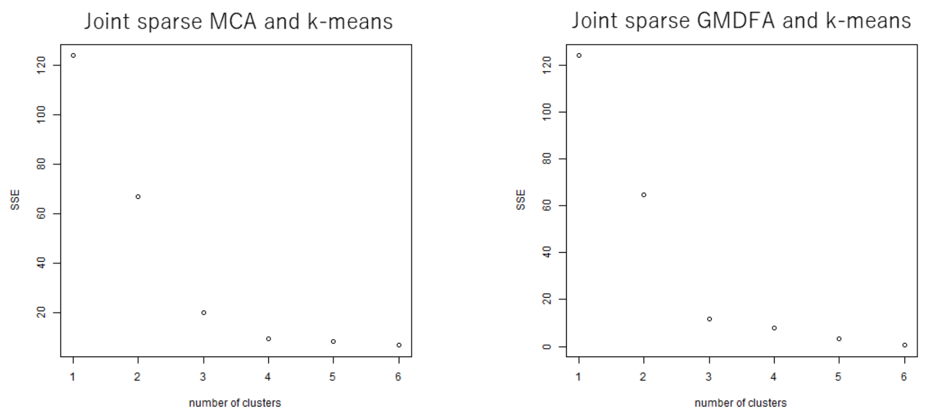


Figure 7.4: SSE plot of joint sparse MCA/GMDFA and k-means in baseball dataset

clustering can be rotated; the simplimax rotation was performed in this analysis. The table shows that we obtained components/factors for eliable batters and sluggers in the same manner as real data example 1.

Next, we examine the clustering solutions. Estimated cluster centroids are shown in table 7.9, and we display object common scores in figure 7.5 and 7.6. The clusters can be distinguished by the high and low eliable batters and sluggers common scores in each method. The clusters are separated enough and we can interpret the solutions easily. Although similar results are obtained, there is a difference between MCA and GMDFA; a dispersion in GMDFA methods tend to be smaller than that in MCA (table 7.10). It suggests that GMDFA methods give well-classified result in this analysis.

7.3.2 Joint dimension reductions and clustering in mathematics assessment data

We set $c = 2$, $R = 23$ and $l = 23$ in the same way as real data example 1. In this analysis, the dimension reduction and clustering are equally treated ($\alpha_1 = \alpha_2 = 0.5$). Figure 7,7 and 7.8 show the SSE plot in mathematics dataset. We can see that SSE do not decrease significantly when

Table 7.7: Loadings of joint MCA/sparse MCA and k-means in baseball data

	Joint MCA and k-means		Joint sparse MCA and k-means	
	Component1	Component2	Component1	Component2
BA	**0.774**	-0.151	0	**0.782**
R	**0.908**	0.005	0	**0.94**
D	**0.792**	0.105	0	**0.783**
HR	-0.042	**0.855**	**0.868**	0
RBI	0.178	**0.898**	**0.91**	0
S	-0.163	**0.715**	**0.719**	0
BO1	**0.569**	0.044	**0.678**	0
BO2	-0.022	**0.704**	0	**0.547**

Component correlation		Component correlation	
1	-0.127	1	0
-0.127	1	0	1

Table 7.8: Loadings of joint GMDFA/sparse GMDFA and k-means in baseball data

	Joint GMDFA and k-means			Joint sparse GMDFA and k-means		
	Factor1	Factor2	Uniqueness	Factor1	Factor2	Uniqueness
BA	**0.743**	0.175	0.428	**0.681**	0	0.495
R	**0.838**	0.012	0.276	**0.814**	0	0.31
D	**0.635**	-0.167	0.516	**0.747**	0	0.403
HR	-0.053	**0.737**	0.386	0	**0.801**	0.291
RBI	-0.245	**0.693**	0.377	0	**0.753**	0.351
S	-0.025	**0.531**	0.693	0	**0.513**	0.708
BO1	**0.473**	-0.103	0.746	**0.4**	0	0.835
BO2	0.265	**0.808**	0.297	0	**0.733**	0.437

Factor correlation		Factor correlation	
1	-0.11	1	0
-0.11	1	0	1

Table 7.9: Cluster centroid in baseball data

	Joint MCA and k-means		Joint sparse MCA and k-means	
	Component1	Component2	Component1	Component2
Cluster1	-0.752	-0.804	-1.043	-1.046
Cluster2	-0.941	1.139	-0.84	0.893
Cluster3	1.011	-1.338	1.007	-0.723
Cluster4	1.132	0.613	0.959	1.303

	Joint GMDFA and k-means		Joint sparse GMDFA and k-means	
	Factor1	Factor2	Factor1	Factor2
Cluster1	-0.637	-1.516	-1.196	-0.483
Cluster2	-1.109	0.768	-0.607	1.571
Cluster3	1.289	-0.558	0.932	-1.203
Cluster4	0.623	1.002	0.891	0.759

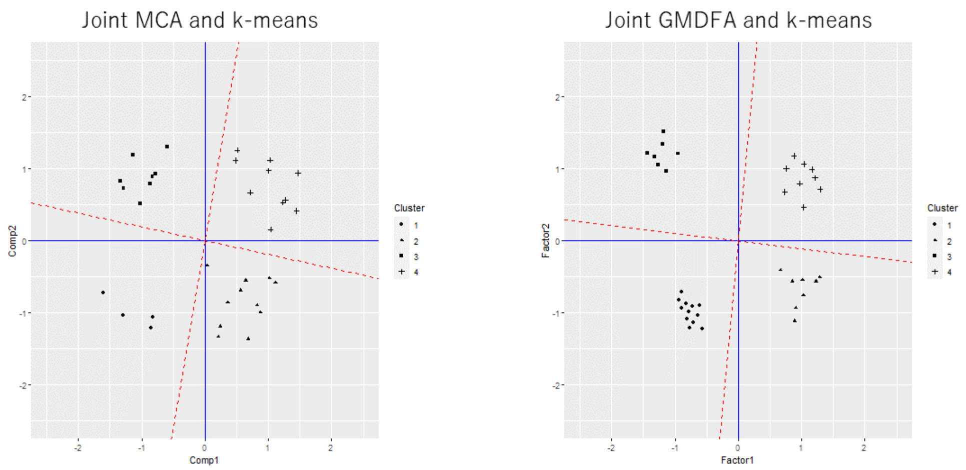


Figure 7.5: Common scores obtained from joint MCA/GMDFA and k-means with rotation in baseball dataset. The axes after rotation are expressed as dashed lines.

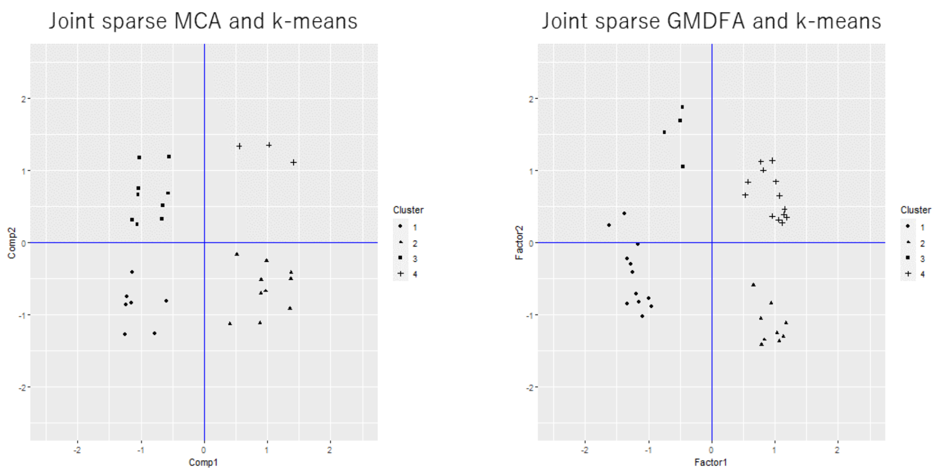


Figure 7.6: Common scores obtained from sparse MCA/GMDFA and k-means in baseball dataset

Table 7.10: SSE in baseball data ($\gamma = 4$)

	Joint MCA and k-means	Joint GMDFA and k-means	Joint sparse MCA and k-means	Joint sparse GMDFA and k-means
SSE	9.787	3.152	9.494	7.873

$\gamma \geq 5$. Thus, we set $\gamma = 5$ for this dataset.

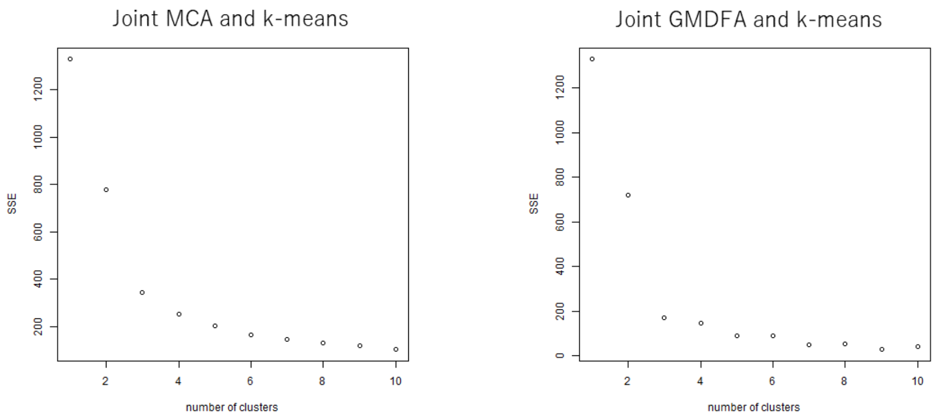


Figure 7.7: SSE in joint MCA/GMDFA and k-means in mathematics dataset

Table 7.11 and 7.12 report the obtained loading matrices in each analysis. The first component/factor are connected to arithmetic, and the second ones are related to geometry. It suggests that we have the expected components/factors by joint dimension reduction and k-means techniques.

Next, the clustering results are examined. Common scores plots are displayed in figure 7.9 and 7.10, and cluster centroids are reported in table 7.13. The clusters seem to be distinguished by academic achievement evaluated by arithmetic and geometry scores. Cluster centroids in four results show similar tendency, but the cluster structures in joint MCA/sparse MCA and k-means are ambiguous, comparing with those in joint GMDFA/sparse GMDFA and k-means. Table 7.14 shows SSE of k-means term in each method. The values of SSE in joint MCA/sparse MCA and k-means are greater than those of joint GMDFA and sparse GMDFA and k-means. In this analysis, GMDFA methods are more preferable from the interpretation viewpoint.

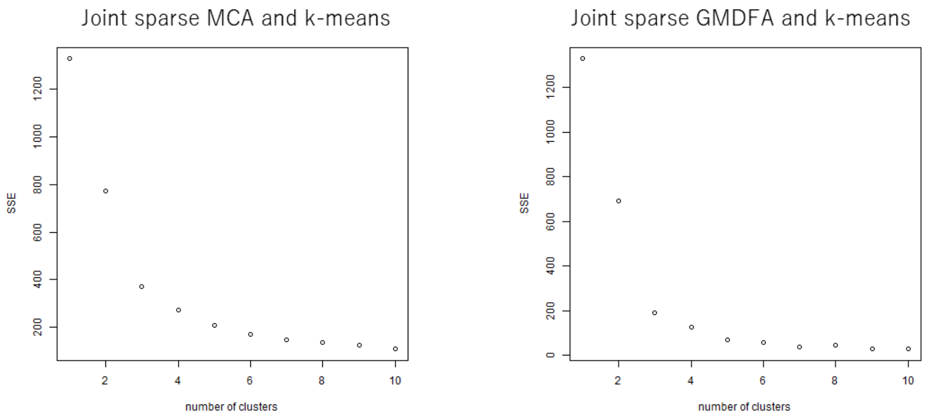


Figure 7.8: SSE in joint sparse MCA/GMDFA and k-means in mathematics dataset

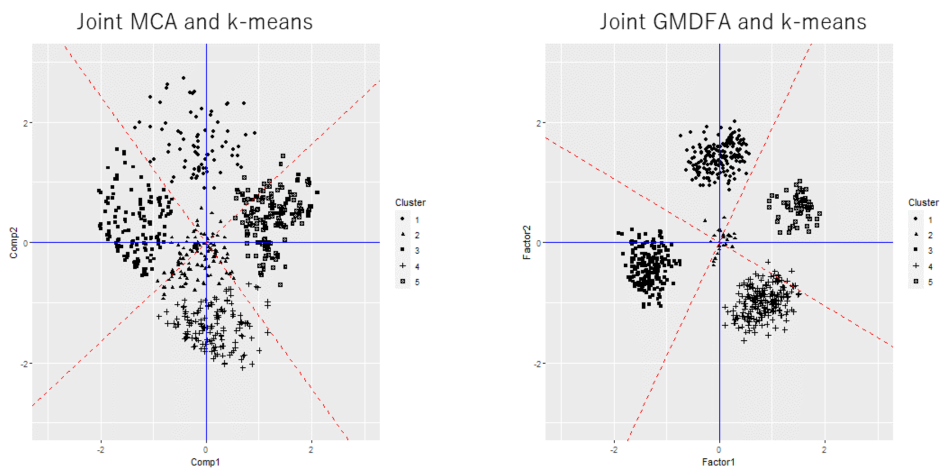


Figure 7.9: Common scores obtained from joint MCA/GMDFA and k-means with rotation in mathematics assessment dataset. The axes after rotation are expressed as dashed lines.

Table 7.11: Loadings of joint MCA/sparse MCA and k-means in mathematics dataset

		Joint MCA and k-means		Joint sparse MCA and k-means	
Item	Domain	Component1	Component2	Component1	Component2
MA1	arithmetic	**0.342**	0.074	**0.302**	0
MA2	arithmetic	**0.503**	0.069	**0.471**	0
MA3	arithmetic	**0.551**	0.008	**0.535**	0
MA4	arithmetic	**0.537**	-0.013	**0.537**	0
MB1	arithmetic	**0.675**	0.093	**0.639**	0
MB2	arithmetic	**0.622**	0.072	**0.596**	0
MB3	arithmetic	**0.549**	-0.109	**0.583**	0
MB4	arithmetic	**0.586**	-0.107	**0.615**	0
MC1	arithmetic	**0.456**	-0.01	**0.458**	0
MC2	arithmetic	**0.554**	0.008	**0.553**	0
MC3	arithmetic	**0.522**	0.05	**0.503**	0
MC4	arithmetic	**0.538**	-0.028	**0.544**	0
MD1	arithmetic	**0.595**	0.082	**0.554**	0
MD2	arithmetic	**0.541**	0.136	**0.493**	0
MG3	arithmetic	**0.379**	-0.239	**0.469**	0
MG4	arithmetic	**0.344**	-0.277	**0.447**	0
MH1	geometry	0.113	**0.728**	0	**0.658**
MH2	geometry	-0.079	**0.625**	0	**0.656**
MH3	geometry	-0.012	**0.701**	0	**0.707**
MH4	geometry	-0.012	**0.685**	0	**0.661**
MI1	geometry	-0.015	**0.432**	0	**0.481**
MI2	geometry	-0.1	**0.471**	0	**0.512**
MI3	geometry	-0.062	**0.475**	0	**0.49**

Component correlation		Component correlation	
1	-0.318	1	0
-0.318	1	0	1

Table 7.12: Loadings of joint GMDFA/sparse GMDFA and k-means in mathematics dataset

		Joint GMDFA and k-means			Joint sparse GMDFA and k-means		
Item	Domain	Factor1	Factor2	Uniqueness	Factor1	Factor2	Uniqueness
MA1	arithmetic	**0.223**	0.009	0.939	**0.311**	0	0.861
MA2	arithmetic	**0.552**	0.19	0.729	**0.399**	0	0.804
MA3	arithmetic	**0.604**	0.204	0.672	**0.454**	0	0.75
MA4	arithmetic	**0.537**	0	0.715	**0.308**	0	0.828
MB1	arithmetic	**0.419**	-0.113	0.719	**0.605**	0	0.593
MB2	arithmetic	**0.426**	-0.187	0.647	**0.564**	0	0.619
MB3	arithmetic	**0.417**	-0.152	0.718	**0.391**	0	0.775
MB4	arithmetic	**0.572**	0.002	0.629	**0.5**	0	0.684
MC1	arithmetic	**0.302**	-0.202	0.786	**0.453**	0	0.748
MC2	arithmetic	**0.35**	-0.062	0.815	**0.391**	0	0.807
MC3	arithmetic	**0.428**	-0.019	0.789	**0.584**	0	0.631
MC4	arithmetic	**0.453**	0.06	0.769	**0.458**	0	0.756
MD1	arithmetic	**0.497**	0.142	0.737	**0.536**	0	0.651
MD2	arithmetic	**0.453**	0.271	0.753	**0.442**	0	0.751
MG3	arithmetic	**0.475**	-0.158	0.606	**0.279**	0	0.763
MG4	arithmetic	**0.523**	-0.127	0.558	**0.271**	0	0.767
MH1	geometry	-0.002	**0.438**	0.759	0	**0.434**	0.729
MH2	geometry	-0.174	**0.33**	0.771	0	**0.572**	0.575
MH3	geometry	-0.105	**0.433**	0.713	0	**0.392**	0.769
MH4	geometry	-0.087	**0.399**	0.76	0	**0.708**	0.43
MI1	geometry	0.08	**0.409**	0.826	0	**0.418**	0.778
MI2	geometry	-0.022	**0.437**	0.773	0	**0.298**	0.851
MI3	geometry	0.001	**0.447**	0.773	0	**0.507**	0.695

Factor correlation			Factor correlation	
1	-0.424		1	0
-0.424	1		0	1

Table 7.13: Cluster centroid in mathematics dataset

	Joint MCA and k-means		Joint sparse MCA and k-means	
	Component1	Component2	Component1	Component2
cluster1	1.173	1.498	0.891	1.088
cluster2	-0.029	-0.326	0.012	0.177
cluster3	1.216	-0.235	-1.238	0.576
cluster4	-1.004	-1.269	-0.938	-1.313
cluster5	-0.711	0.857	0.937	-1.143

	Joint GMDFA and k-means		Joint sparse GMDFA and k-means	
	Factor1	Factor2	Factor1	Factor2
cluster1	0.659	1.425	1.289	0.688
cluster2	-0.022	0.03	-0.023	0.522
cluster3	0.98	-0.49	-0.919	1.047
cluster4	-1.194	-0.946	-1.007	-0.922
cluster5	-1.036	0.658	0.725	-1.336

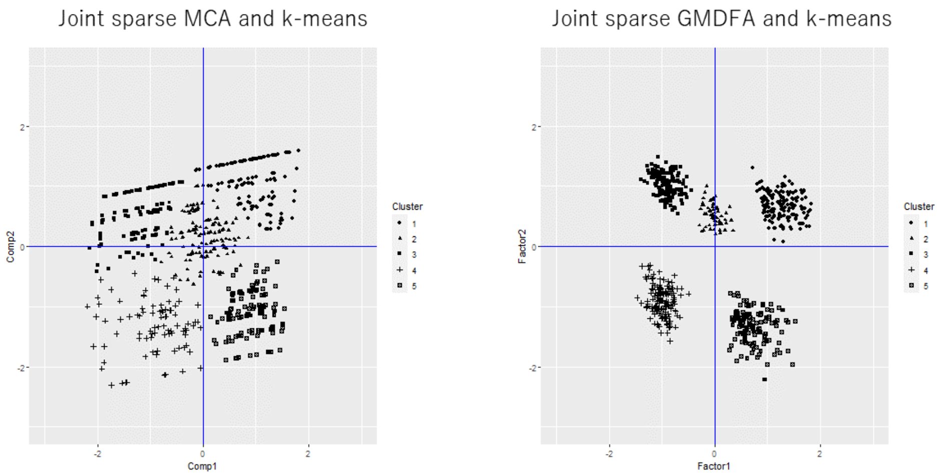


Figure 7.10: Common scores obtained from sparse MCA/GMDFA and k-means in mathematics assessment

Table 7.14: SSE in mathematics assessment data ($\gamma = 5$)

	Joint MCA and k-means	Joint GMDFA and k-means	Joint sparse MCA and k-means	Joint sparse GMDFA and k-means
SSE	204.362	88.984	208.631	67.068

8

Discussion

The aim of this study was twofold. First, PCA and MDFA were generalized to handle nonmetric multivariate data using the quantification technique. The estimation algorithms for nonmetric verision of PCA and MDFA were developed and they were compared with each other from mathematical and emprical viewpoints. Second, the methods to facilitate an interpretation of the solutions were developed. These methods were classified into two types in this thesis: the method for obtaining interpretable structures between variables and components/factors, and the method to facilitate an interpretation of objects charactalistics with their scores. We will begin with this chapter by summarizing the finding of this thesis, and the conclude by pointing out some remarks and suggesting problems for the future research.

8.1 Summary of findings

As discussed in chapter 1, PCA and EFA are popular methods for depicting hidden structures of quantitative observed data in a low-dimensional space. One of the limitations of PCA and EFA is that only quantitative variables can be handled although qualitative variables are often encountered in human science research process. To address such limitation, PCA and EFA were extended to handle nonmetric multivariate data using the quantification technique in the proceeding chapters.

In chapter 2, we breifly introduced the models and estimation algorithms of PCA and EFA. Although there are various estimation methods in EFA, MDFA was presented among them. It was shown that PCA and MDFA can be integratedly represented via the SVD, which is called as the matrix decomposition approaches in this thesis. The quantification technique was also introduced in order to extend PCA and MDFA.

In chapter 3 and 4, we discussed two types of the methods to facilitate an interpretation of the solutions in PCA and MDFA: one is the interpretability of relationships between variables and the other is the interpretability of objects scores. Chapter 3 discussed the methods for obtaining the interpretable loading matrix, which expresses the correlations between components/factors and observations in PCA and MDFA. To consider the interpretability in the loading matrix, we firstly discussed ideal and approximate simple structure based on Thurstone's five rules. Then, the rotation and sparse techniques were introduced to estimate interpretable loading matrix: the rotation gives the approximate simple structure without affecting the fitness of PCA and MDFA solutions, and the ideal simple structure can be obtained by using sparse PCA and MDFA. In chapter 4, SARC proposed by Mitsuhiro and Yadohisa (2018) was introduced as a simultaneous framework, in which dimension reduction methods with matrix decomposition approaches are integrated with k-means clustering. The advantages of joint dimension dimension reduction and clustering are that both object clustering and a low-dimenisional subspace that reflects the cluster structure are simultaneously obtained. In MDFA literature, SARC fails to estimate interpretable solutions, and thus the modified SARC was proposed to overcome its drawback. It was also shown that the rotation and sparse techniques can be combined with joint dimension reduction and clustering methods.

In chapter 5 and 6, PCA and MDFA were extended to handle nonmetric multvariate data via the quantification technique. Chapter 5 discussed the nonmetric verion of PCA. It was shown that the nonmetric version of PCA is equivalent to MCA. Then, we investigated the rotation problems in MCA. There are three options in MCA rotation and we discussed differences among these options and limitations of MCA rotation in the previous studies. To tackle the limitations, the reformulation of MCA was proposed, in which oblique rotation can be permitted without affecting homogeneity assumption. It implies that MCA solutions can be graphically interpreted after the oblique rotation. The pre- and post- rotational freedom exists in the proposed reformulation, and joint simplimax was developed to deal with two-sided rotation. Some mathematical properties in the proposed reformulation were also examined. Lastly, we developed sparse MCA, joint MCA and k-means clustering, and joint sparse MCA and k-means clustering. In chapter 6, MDFA were extened to deal with nonmetric data via the quantification technique, which is called as GMDFA in this thesis. Then, we compared GMDFA with MCA and FACTALS (Takane *et al*., 1979). Through the comparison, it was shown that GMDFA and MCA differs from the viewpoint of rotational freedom, which indicates that existing rotation technique can be applied to the rotation in GMDFA. It was also discovered that GMDFA and FACTALS are different from the constraints for quantification dimension. Lastly, we developed sparse GMDFA, joint GMDFA and k-means clustering, and joint sparse GMDFA and k-means clustering. In chapter 7, the proposed methods in chapter 5 and 6 were illustrated through two real data examples.

8.2 Interpretation of the category score in multiple correspondence analysis

We focused on the interpretations of relationships between components and variables in this thesis. In this section, the interpretability of category score $\mathbf{W}$ is deiscussed. In the formulation proposed by Adachi (2004), the solutions can be obliquely rotated, but the rotated solutions cannot be graphically interpreted because the homogeneity assumptions are violated by the oblique rotation. It implies that roteted component score $\mathbf{F}$ and category score $\mathbf{W}$ cannot be graphically represented. This limitation is a major drawback because one of the main purpose of MCA is to obtain the configuration to interpret the solutions. However, the MCA solutions can be obliquely rotated without affecting the homogeneity assumption as shown in chapter 5; the rotated solutions can be interpreted as graphical representation based on the proposed MCA reformulation. It indicates that the category score matrix $\mathbf{W}$ can be rotated towards simple structure to interpret the configuration easier. As will be shown below, the sparse technique can also be applied to $\mathbf{W}$.

8.2.1 Rotation for the category score

Recall that the category score can be expressed as

$$\mathbf{W} = [\mathbf{Q}_1\mathbf{A}_1, \ldots, \mathbf{Q}_p\mathbf{A}_p] = [n^{-1}\mathbf{Q}_1\mathbf{Q}'{}_1\mathbf{G}'_1\mathbf{F}, \ldots, n^{-1}\mathbf{Q}_p\mathbf{Q}'{}_p\mathbf{G}'_p\mathbf{F}] = \mathbf{D}^{-1}\mathbf{G}'\mathbf{F}. \tag{8.1}$$

It indicates that $\mathbf{W}$ can be rotated by $\mathbf{T}$ in the proposed reformulation. The loading matrix has pre- and post-multiplied rotational freedom in the proposed reformulation, but the category score can be rotated by a single matrix. It should be noted that the rotation matrix differs between Adachi (2004) and the proposed reformulation: $\mathbf{T}^{-1}$ is used for the rotation matrix fo $\mathbf{W}$ if $\mathbf{F}$ is rotated by $\mathbf{T}$ in Adachi (2004), but we can use the same rotation matrix $\mathbf{T}$ for the rotation of component and category score in the proposed reformulation. Therefore, the various existing rotation methods can be applied to the rotation of the category score matrix.

There are some cases that the component and category score matrices are simultaneously rotated so that the solutions can be interpretable. Adachi (2009) proposed Joint Procrustes analysis (JPA) to deal with such cases. The loss function of JPA based on the proposed reformulation can be defined as

$$\eta_1||\mathbf{WT} - \mathbf{B}_w||^2 + \eta_2||\mathbf{FT} - \mathbf{B}_f||^2. \tag{8.2}$$

Here, η_1 and η_2 are positive constants, which is determined by the users in advance, and $\mathbf{B}_w$ and $\mathbf{B}_f$ are target matrices. The users can predetermine the target matrices if they would like to reflect knowledge as to what the component and category matrices should be. It they have such knowledge on the component and category score matrices, simplimax or promax techniques can be used for estimating target matrices (Adachi, 2009). It should be noted that the loading matrix in the JPA solution via (8.2) is not guaranteed to be have simple structure because rotation matrix is not optimized so that the loading matrix has the simple structure.

We can also consider JPA for the category score and loading matrices, and for the category score and compenent score matrices. These loss functions of JPA can be described as

$$\eta_1||\mathbf{WT} - \mathbf{B}_w||^2 + \eta_2 \sum_{i=1}^{p} ||\mathbf{T}_{oblq_i}^{-1}\mathbf{A}_i\mathbf{T} - \mathbf{B}_a||^2, \tag{8.3}$$

and

$$\eta_1||\mathbf{FT} - \mathbf{B}_f||^2 + \eta_2 \sum_{i=1}^{p} ||\mathbf{T}_{oblq_i}^{-1}\mathbf{A}_i\mathbf{T} - \mathbf{B}_a||^2, \tag{8.4}$$

where $\mathbf{B}_a$ is the target matrix of $\mathbf{A}$. The purpose of these loss functions seems to be similar to that of the joint MCA and clustering method. However, it differs from the viewpoint of the fitness of the loss function: the solutions can be estimated without affecting the value of MCA loss function via JPA.

8.2.2 Sparse multiple correspondence analysis for category score

In chapter 5, sparse MCA was considered to obtain the interpretable loading matrix $\mathbf{A}$. Sparse multiple correspondence analysis for category score can be also considered. The loss function can be defined as

$$||\mathbf{F} - \mathbf{GW}||^2 \tag{8.5}$$

and it is minimized over $\mathbf{F}$ and $\mathbf{W}$, subject to $card(\mathbf{W}) = l$, $n^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_c$ and $\mathbf{JF} = \mathbf{F}$. The loss function (8.5) is regarded as homogeneity analysis with sparse constraints, and thus obtained solutions can be graphically interpreted.

The loss fucntion (8.5) can be rewritten as

$$||\mathbf{F} - \mathbf{G}\tilde{\mathbf{W}}||^2 + ||\mathbf{D}^{1/2}(\mathbf{W} - \tilde{\mathbf{W}})||^2, \tag{8.6}$$

where $\tilde{\mathbf{W}} = \mathbf{D}^{-1}\mathbf{G}'\mathbf{F}$. The decomposition above shows that the second term in (8.6) is only relevant to $\mathbf{W}$. Thus, the minimization with respect to $\mathbf{W}$ reduces to that of the second term $||\mathbf{D}^{1/2}(\mathbf{W} - \tilde{\mathbf{W}})||^2$ for a fixed $\mathbf{F}$. This result suggests that the solutions can be obtained by using the USLPCA algorithm:

$$w_{ij} = \begin{cases} 0 & (\text{iff } \tilde{w}_{ij}^2 \leq \tilde{w}_l^2) \\ \tilde{w}_{ij} & (\text{otherwise}). \end{cases} \tag{8.7}$$

Here, $\tilde{w}_l^2$ expresses the l smallest values of $\tilde{w}_{ij}^2$.

The minimization over $\mathbf{F}$ with a fixed $\mathbf{W}$ is equivalent to the maximization of $tr(\mathbf{F}'\mathbf{GW})$. As discussed in chapter 2, this maximization can be achieved by the SVD of $\mathbf{GW}$.

The decomposition (8.6) implies that the fitness of sparse MCA for category score is less than that of homogeneity analysis:

$$||\mathbf{F} - \mathbf{G}\tilde{\mathbf{W}}||^2 + ||\mathbf{D}^{1/2}(\mathbf{W} - \tilde{\mathbf{W}})||^2 \geq ||\mathbf{F} - \mathbf{G}\tilde{\mathbf{W}}||^2. \tag{8.8}$$

8.3 Sparse constraints for each variable

In this thesis, the number of zeros in the simplimax target or sparse techniques were determined for $\mathbf{A}$. However, the loss functions is defined as the sum of loss of each variable. It indicates that it is possible to specify the number of zeros for each variable and to imposed sparse constraints on the loading matrix to have a single nonzero element in each row and zeros elsewhere when the quantification dimension is unidimensional. Such a loading matrix is the sparsest possible for certain number of variables and components/factors (Adachi and Trendafilov, 2018). We discuss sparse constraints for each variables in the following subsections.

8.3.1 Sparse constraints for each variable in simplimax target

Recall the loss function of joint simplimax rotation:

$$\sum_{i=1}^{p} ||\mathbf{T}_{oblq_i}^{-1}\mathbf{A}_i\mathbf{T} - \mathbf{B}_i||^2$$

Let l_i denote the number of zeros in each variables, $b_{i_{jk}}$ denote the elements of $\mathbf{B}_i$ and $\dot{a}_{i_{jk}}$ denote the elements of $\mathbf{T}_{oblq_i}^{-1}\mathbf{A}_i\mathbf{T}$. The target matrices can be updated by

$$b_{i_{jk}} = \begin{cases} 0 & (\text{if } \dot{a}_{i_{jk}} \leq \dot{a}_{l_i}) \\ 1 & (\text{otherwise}), \end{cases} \tag{8.9}$$

where $\dot{a}_{l_i}$ is the l_i smallest values of $\dot{a}_{i_{jk}}$. The update algorithm above can be applied to the rotation of $\mathbf{W}$.

8.3.2 Sparse constraints for each variable in sparse MCA and GMDFA

The sparse constraints for each variable can be applied in sparse MCA and GMDFA, because the loss function can be described as

$$\sum_{i=1}^{p} ||\mathbf{G}_i\mathbf{Q}_i - \mathbf{F}\mathbf{A'}_i||^2 + n\sum_{i=1}^{p} ||\tilde{\mathbf{A}}_i - \mathbf{A}_i||^2, \tag{8.10}$$

$$\sum_{i=1}^{p} ||\mathbf{G}_i\mathbf{Q}_i - \mathbf{F}\mathbf{A'}_i - \mathbf{U}_i\mathbf{\Psi}_i||^2 + n\sum_{i=1}^{p} ||\tilde{\mathbf{A}}_i - \mathbf{A}_i||^2. \tag{8.11}$$

It suggests that sparse constraints can be imposed on $\mathbf{A}_i$ for each i. The optimal $\mathbf{A}_i$ can be obtained in the same manner as (8.9).

It should be noted that we can impose the sparse constraints for each variable to sparse MCA for category score:

$$\sum_{i=1}^{p} ||\mathbf{F} - \mathbf{G}_i\tilde{\mathbf{W}}_i||^2 + \sum_{i=1}^{p} ||\mathbf{D}_i^{1/2}(\mathbf{W}_i - \tilde{\mathbf{W}}_i)||^2.$$

8.4 Redefinition of SARC framework

Mitsuhiro and Yadohisa (2018) proposed a unified representation of joint dimension reduction and clustering analysis called as SARC. However, as discussed in chapter 4, there is the limitation in SARC: it may be difficult to acquire an interpretable common factor scores in SARC when joint MDFA and clustering is performed. To address this drawback, we proposed modified joint MDFA and clustering loss function in this thesis.

Let $\mathbf{H}_1$ and $\mathbf{H}_2$ denote the objects scores, $\mathbf{X}^*$ denote the data part, Ω denote the loadings, $\mathbf{P}$ denote the membership marix and $\mathbf{C}$ denote the cluster centroid matrix. Using the SARC framework, the methods discussed in this thesis can be unified in a single loss function:

$$\alpha_1||\mathbf{H}_1 - \mathbf{X}^*\mathbf{\Omega}||^2 + \alpha_2||\mathbf{H}_2 - \mathbf{PC}||^2, \tag{8.12}$$

where α_1 and α_2 are constants that satisfied with $\alpha_1 + \alpha_2 = 1$.

The first term in (8.12) describes the loss of homogeneity in dimension reduction. Note that this term is defined as approximating a data matrix by a lower rank matrix in original SARC loss function. As shown in chapter 5, the loss of homogeneity is preferable from the viewpoint of rotational freedom. Therefore, we adopt the loss of homogeneity in (8.12). The second term in (8.12) expresses the loss of k-means clustering. In original SARC loss function, $\mathbf{H}_1$ is constrained to be equal to $\mathbf{H}_2$. However, interetable solutions in joint MDFA and clustering cannot be estimated due to the this constraint as discussed in chapter 4. Thus, the constraint are relaxed to describe modified SARC loss functions.

When we set $\mathbf{H}_1$ as $\mathbf{F}$ and $\mathbf{\Omega}$ as $\mathbf{A}$, (8.12) reduces to the loss functions of the PCA models. We have the loss functions of the MDFA models if $\mathbf{H}_1$ and $\mathbf{\Omega}$ are constrained to be $\mathbf{S} = [\mathbf{F}, \mathbf{U}]$ and $\mathbf{\Lambda} = [\mathbf{A}, \mathbf{\Psi}]$, respectively. Sparse constraints are additionally imposed on $\mathbf{A}$, and then (8.12) expresses the sparse models.

8.5 Empirical comparison between MCA and GMDFA

In this thesis, we mainly compared MCA with GMDFA from the view of interpretability in the solutions. Here, their differences are discussed from another viewpoint presented by Adachi and Trendafilov (2019). Adachi and Trendafilov (2019) examined the mathematical relationships between PCA and MDFA expressed in figure 8.1, and made the following suggestions:

1. PCA seems to be preferable when a large common part is wished to be found in data.
2. FA seems to be preferable when data is wished to be better explained.

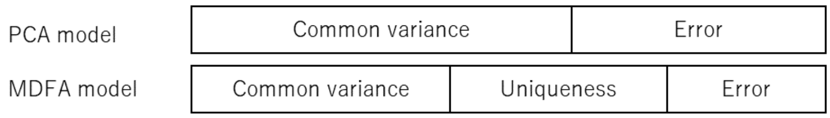


Figure 8.1: Relative largeness of common parts and residuals in PCA and MDFA solutions (Adachi and Trendafilov, 2019)

Table 8.1 shows common variances between MCA and GMDFA in real data examples. Common variances of MCA is greater than that of GMDFA in each real data example. Next, we consider the residuals of MCA and GMDFA. The residuals in real data examples are reported in table 8.2. The residuals of GMDFA are smaller than these of MCA.

Table 8.1: Common variances between MCA and GMDFA

	MCA	GMDFA
Baseball data	5.119	4.397
Mathmatics assessment data	7.331	5.970

The empirical results seem that the mathematical relationships discussed in Adachi and Trendafilov (2019) are possible to be generalized to nonmetric data. We need further studies on the relationships between MCA and GMDFA from this viewpoints in future.

Table 8.2: Residuals between MCA and GMDFA

	MCA	GMDFA
Baseball data	178.57	2.63
Mathematics assessment data	10404.5	412.79

References

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

Adachi, K. (2009). Joint Procrustes analysis for simultaneous nonsingular transformation of component score and loading matrices. *Psychometrika*, **74**, 667.

Adachi, K. and Murakami, T. (2011). *Nonmetric Multivariate Analysis. - MCA, NCA, and PCA -* Tokyo: Asakura Shoten (In Japanese).

Adachi, K. (2012). Some contributions to data-fitting factor analysis with empirical comparisons to covariance-fitting factor analysis. *Journal of the Japanese Society of Computational Statistics*.

Adachi, K. (2013). Factor analysis with EM algorithm never gives improper solutions when sample covariance and initial parameter matrices are proper. *Psychometrika*, **78**, 380-394.

Adachi, K. (2015). A Matrix-Intensive Approach to Factor Analysis, *Journal of the Japan Statistical Society, Japanese Issue*, **44**, 363-382.

Adachi, K. (2016). *Matrix-based introduction to multivariate data analysis*. Singapore: Springer Singapore.

Adachi, K., and Trendafilov, N. T. (2014). Sparse orthogonal factor analysis. In *Advances in Latent Variables*, pp. 227-239. Springer, Cham.

Adachi, K., and Trendafilov, N. T. (2016). Sparse principal component analysis subject to prespecified cardinality of loadings. *Computational Statistics*, **31**, 1403-1427.

Adachi, K., and Trendafilov, N. T. (2018). Sparsest factor analysis for clustering variables: a matrix decomposition approach. *Advances in Data Analysis and Classification*, **12**, 559-585.

Adachi, K., and Trendafilov, N. T. (2019). Some inequalities contrasting principal component and factor analyses solutions. *Japanese Journal of Statistics and Data Science*, **2**, 31-47.

Anderson, T. W., and Rubin, H. (1956). Statistical inference in factor analysis. In J. Neyman (Ed). *Proceedings of the third Berkeley symposium on mathematical statistics and probability, vol 5*. pp111-150. Berkeley, CA: University of California Press.

Arabie, P. and Hubert, L. (1994). Cluster analysis in marketing research, In R. P. Bagozzi (Ed.), *Handbook of Marketing Research*, Oxford: Blackwell.

Beh, E. J., and Lombardo, R. (2014). *Correspondence analysis: theory, practice and new strategies*. John Wiley & Sons.

Browne, M,W.(1972a). Orthogonal rotation to a partially specified target. *British Journal of Mathematical and Statistical Psychology*, **25**, 115-120.

Browne, M,W.(1972b). Oblique rotation to a partially specified target. *British Journal of Mathematical and Statistical Psychology*, **25**, 207-212

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

Chavent, M., Kuentz-Simonet, V. and Saracco J. (2012). Orthogonal rotation in PCAMIX. *Advances in Data Analysis and Classifications*, **6**, 131-146.

Choi, J., Zou, H., and Oehlert, G. (2010). A penalized maximum likelihood approach to sparse factor analysis. *Statistics and its Interface*, **3**, 429-436.

Crawford, C. B., and Ferguson, G. A. (1970). A general rotation criterion and its use in orthogonal rotation. *Psychometrika*, **35**, 321-332.

de Leeuw, J. (2004). Least squares optimal scaling of partially observed linear systems. In: van Montfort K, Oud J, Satorra A (eds) *Recent developments on structural equation models: theory and applications*. Kluwer, Dordrecht, 121-134.

de Leeuw, J. (2008). Factor analysis as matrix decomposition. Preprint series: Department of Statistics, University of California, Los Angeles.

de Leeuw, J. and van Rijckevorsel, J.L.A. (1988). Beyond homogeneity analysis. In J.L.A van Rijckevorsel & J. de Leeuw (Eds.) *Component and Correspondence Analysis: Dimension Reduction by Functional Approximation*. New York: Wiley. pp55-80.

De Soete, G. and Caroll, J. D. (1994) K-means clustering in a low-dimensional Euclidean space. In E. Diday, Y. Lechevallier, M. Schader, P. Bertrand, and B. Burtschy (Eds.) *New approaches in classification and data analysis*(pp212-219). Heidelberg: Springer.

Gifi, A. (1990). *Nonlinear multivariate analysis*. Chichester: Wiley.

Greenacre, M, J. (1989). Correspondence analysis of multivariate categorical data. *Biometrika*, **75**, 457-467.

Harman, H.H. (1976). *Mordan Factor Analysis*, Chicago: University of Chicago Press, 3rd ed.

Hendrickson, A. E., and White, P. O. (1964). Promax: A quick method for rotation to oblique simple structure. *British journal of statistical psychology*, **17**, 65-70.

Hirose, K., and Yamamoto, M. (2014). Estimation of an oblique structure via penalized likelihood factor analysis. *Computational statistics & data analysis*, **79**, 120-132.

Hirose, K., and Yamamoto, M. (2015). Sparse estimation via nonconcave penalized likelihood in factor analysis model. *Statistics and Computing*, **25**, 863-875.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, **24**, 417-441.

Hwang. H., Dillon. W. R., and Takane. Y. (2006) An extension of multiple correspondence analysis for identifying heterogeneous subgroups of respondents. *Psychometrika*, **71**, 161-171.

Hwang. H., Dillon. W. R., and Takane. Y. (2010). Fuzzy cluster multiple correspondence analysis. *Behaviormetrika*, **37**, 111-133.

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

Jennrich, R. I. (2002). A simple general method for oblique rotation. *Psychometrika*, **67**, 7-19.

Joreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika*, **34**, 443-482.

Jolliffe, L.T. (2002). *Principal Component Analysis*. Second edition. New York: Springer.

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

Kiers, H. A. (1994). SIMPLIMAX: Oblique rotation to an optimal target with simple structure. *Psychometrika*, **59**, 567-579.

Kiers, H. A. (1998). Three-way SIMPLIMAX for oblique rotation of the three-mode factor analysis core to simple structure. *Computational Statistics and Data Analysis*, **28**, 307-324.

Makino, N. (2015). Generalized data-fitting factor analysis with multiple quantification of categorical variables. *Computational Statistics,* **30**, 279-292.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* **1**, 281-297.

Mitsuhiro. M, and Yadohisa, H. (2015). Reduced k-means clustering with MCA in a low-dimensional space. *Computational Statistics*, **30**, 463-475.

Mitsushiro. M, and Yadohisa, H. (2018). A unified representation of simultaneous analysis methods of reduction and clustering. *Japanese Journal of Statistics and Data Sciences*. **1**, 393-412.

Mori, Y., Kuroda, M., and Makino, N. (2016). *Nonlinear Principal Component Analysis and Its Applications*. Springer, Singapore.

Mulaik, S.A. (2010). *Foundation of factor analysis*. Second edition. Boca Raton: Chapman and Hall/CRC.

Murakami, T. (1999). A psychometrics study on principal component analysis of categorical data. Technical report.

Murakami, T., Kiers, H,A.L., and ten Berge, J,M,F.(1999). Non-metric principal component analysis for categorical variables with multiple quantifications. Unpublished manuscript.

Ning, L., and Georgiou, T. T. (2011). Sparse factor analysis via likelihood and $l1$-regularization. In *50th IEEE Conference on Decision and Control and European Control Conference* pp. 5188-5192.

Nishisato, S. (2006). *Multidimensional Nonlinear Descriptive Analysis.* Chapman and Hall/CRC

OECD (2019), PISA 2018 Questionnaire Framework, in *PISA 2018 Assessment and Analytical Framework*, OECD Publishing, Paris, https://doi.org/10.1787/850d0ef8-en.

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, **2**, 559-572.

Rao, C. R. (1973). *Linear Statistical Inference and its Applications*. New York: Wiley.

Robitzsch, A. (2019). Package sirt, version in November 4, 2019.

Rubin, D. B. and Thayer, D. T. (1982). EM algorithm for ML factor analysis. *Psychometrika*, **47**, 69-76.

Shen, H., and Huang, J. Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *Journal of multivariate analysis*, **99**, 1015-1034.

Takane, Y., Young, F,W., and de Leeuw, J. (1979). Nonmetric common factor anaysis: an alternating least square method with optimal scaling features. *Behaviormetrika*, **6**, 45-56.

ten Berge, J,M,F. (1983). A generalization of Kristof's theorem on the trace if certain matrix products. *Psychometrika*, **48**, 519-523.

ten Berge, J,M,F. (1993). *Least Squares Optimization in Multivariate Analysis*. Leiden: DSNO Press.

ten Berge, J. M., and Kiers, H. A. (1996). Optimality criteria for principal component analysis and generalizations. *British Journal of Mathematical and Statistical Psychology*, **49**, 335-345.

ten Berge, J. M., and Nevels, K. (1977). A general solution to Mosier's oblique Procrustes problem. *Psychometrika*, **42**, 593-600.

Tenenhaus, M., and Young., Y. W. (1985). An analysis and synthesis of multiple correspondence analysis, optimal scaling, dualscaling, homogeneity analysis and other methods for quantifying categorical multivariate data. *Psychometrika*, **50**, 91-119.

Timmerman, M. E., Ceulmans. E., Kiers. H. A. L., and Vichi, M. (2010). Factorial and reduced k-means reconsidered. *Computational Statistics and Data Analysis*, **54**, 1858-1871.

Tipping, M. E., and Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **61**, 611-622.

Trendafilov, N. T. (2014). From simple structure to sparse components: a review. *Computational Statistics*, **29**, 431-454.

Trendafilov, N, T. and Unkel, S.(2010). Exploratory factor analysis of data matrices with more than observations. *Journal of Computational and Graphical Statistics*, **20**, 874-891.

Trendafilov, N, T., Unkel, S., and Krzanowski, W.(2011). Exploratory factor and principal component analyses: some new aspects. *Statistics and Computing*, online first.

Unkel, S., and Trendafilov, N, T. (2010). Simultaneous parameter estimation in exploratory factor analysis: An expository reviews. *International Statistical Reviews*, **78**, 363-382.

Unkel, S., and Trendafilov, N, T. (2013). Zig-zag exploratory factor analysis with more variables than observations. *Computational Statistics*, **28**, 107-125.

van der Burg, E., and de Leeuw, Y. (1988). Homogeneity analysis with k sets of variables: An alternating least squares method with optimal scaling features. *Psychometrika*, **53**, 177-197.

van Buuren, S., and Heiser, W. J. (1989). Clustering n objects into k groups under optimal scaling of variables. *Psychometrika*, **54**, 699-706.

Vichi, M., and Kiers. H. A. L. (2001). Factorial k-means analysis for two-way data. *Computational Statistics and Data analysis*, **37**, 49-64.

Yamamoto, M., and Hwang, H. (2014). A general formulation of cluster analysis with dimension reduction and subspace separation. *Behaviormetrika*, **41**, 115-129.

Young, F.W. (1981). Quantitative analysis of qualitative date. *Psychometrika*, **46**, 357-388.

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of computational and graphical statistics*, **15**, 265-286.

Zuccon, G., Azzopardi, L. A., and Van Rijsbergen, C. J. (2009). Semantic spaces: Measuring the distance between different subspaces. In *International Symposium on Quantum Interaction* (pp. 225-236). Springer, Berlin, Heidelberg.

Acknowledgement

本博士論文執筆にあたり、まずはじめに指導教官である足立浩平教授に心から感謝申し上げます. 研究に対するご助言だけではなく, 足立先生の研究に対する姿勢、特に学術論文執筆に関する考え方から非常に多くのことを学ぶことができました. この経験は今後の研究生活においても非常に重要な資産です. また, 研究室の皆さまからは, 研究に関して様々なご指摘を賜りました. この場を借りて, 御礼申し上げます. 最後に, 博士課程までの研究をあたたかく応援してくれた、父・長男と母・清美に感謝します.

2020, May, 14