Osaka University

Doctoral Thesis

# Interpretable Multivariate Analysis Procedures by Means of Their Matrix-Intensive Refinements

*Author:*
Naoto Yamashita

*Supervisor:*
Prof. Kohei Adachi

*A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy (Human Sciences)*

*in the*

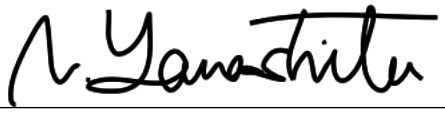
Division of Behavioral Statistics

March, 2020

# Declaration of Authorship

I, Naoto YAMASHITA, declare that this thesis titled, "Interpretable Multivariate Analysis Procedures by Means of Their Matrix-Intensive Refinements" and the work presented in it are my own. I confirm that:

- This work was done mainly while in candidature for a research degree at Osaka University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at Osaka University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: 

Date: Feb. 2. 2020

# Abstract

In empirical sciences, data are collected in various forms regardless of qualitative or quantitative, and structured or nonstructured. The ultimate goal of empirical sciences is to understand the laws of nature deeply, and successively, to predict its future. To accomplish the goal, researchers often conduct their researches in hypothetico-deductive manner, in which a hypothesis creation and its verification is repeated. For example, to understand a fundamental mechanism of psychological process of human, a psychologist establishes a research hypothesis based on the previous researches. Then, some psychological experiments are planned and conducted to verify the hypothesis. Because the verification process is totally based on data collected by experiments or observations, data play a critical role in today's scientific research and its data-based hypothesis verification.

The thesis firstly focuses on multivariate data analysis procedures and their exploratory usages, which is called exploratory data analysis. They are often used for extracting clues for hypothesis creation. For example, principal component analysis (PCA) can extract the essential components of a data matrix, excluding redundant information by compressing multivariate data into lower dimensional space.

In exploratory data analysis, it is considered that how easily a resulting solution can be interpreted, that is, the interpretability of a solution, is a very important property. Interpretability is also important in the context of prediction, because it assures the accountability of prediction models. However, in general, interpretability is not always considered in the most of multivariate data analysis procedures, while they sorely consider how well their solution fits to the given data.

Therefore, in this thesis, the author proposes a series of procedures to improve the interpretability of various multivariate data analysis procedures, addressing that interpretability is an essential property in their exploratory use cases. Two sub-concepts of interpretability are proposed, sparseness (how many a solution matrix contains zero elements) and simple structure. The existing procedures are modified by matrix-intensive refinements so as to their solutions are highly interpretable. Six procedures proposed in the thesis are categorized into the following four families: rotation of solution matrices, combination with clustering, sparse estimation with cardinality constraint, and some emerging techniques.

In Chapter 2, a new rotation technique is proposed, to overcome the critical drawback of the existing procedures for target rotation; the correspondence of the variables in a target matrix to a loading matrix is unknown. In the proposed procedure, a loading matrix is rotated simultaneously with a permutation of the rows of a target matrix,

so that the rotated loading matrix is optimally matched with the row-permuted target matrix. Its algorithm is presented, with Thurstone's definition of simple structure modified so as to specify a target matrix uniquely. Permutimin is illustrated with real data examples, and the relationships between Permutimin and Procrustes rotation are discussed.

In Chapter 3, rotation toward a simple structure is extended to canonical correlation analysis (CANO). There, a new formulation of CANO is firstly proposed, which is proved to be equivalent to the exiting one. Two canonical structure matrices have freedom with respect to orthogonal rotation under the new formulation. Thus, the study proposes an orthogonal rotation method for rotating two structure matrices individually for their simplicity and interpretability, while the existing formulation allows only simultaneous rotation of the matrices. It is illustrated that the proposed method facilitates the interpretation of solutions of CANO by a real data example.

Chapter 4 deals with biplot, a technique for obtaining a low-dimensional configuration of a data matrix. Biplot with a large number of objects and variables is known to be difficult to interpret. Therefore, this study proposes a new biplot procedure where objects and variables are classified into a small number of clusters by means of $K$-means clustering. The resulting clusters are simultaneously biplotted in lower-dimensional space. The resulting biplot is composed of fewer points and vectors and therefore easily captured. An extension of the proposed method to fuzzy $K$-means clustering is also proposed. A simulation study and real data example are also provided to demonstrate the effectiveness of the proposed procedures.

Chapter 5 considers the interpretability in $K$-means clustering because it does not allow any post-hoc transformation of solutions. The study proposes a new procedure for obtaining a centroid matrix, so that it has a number of exactly zero elements by cardinality constraint. This allows easier interpretation of the matrix, as we may focus on only the nonzero centroids. The development of an iterative algorithm for the constrained minimization is described. A special case of the proposed procedure is also proposed, in which some restrictions are imposed on the positions of nonzero elements. The behaviors of the proposed procedure were evaluated in simulation studies and are illustrated with three real data examples, which demonstrate that the performance of the procedure is promising.

In Chapter 6, a novel framework of multivariate analysis procedure is proposed that is called layered multivariate analysis (LMA), which includes layered multivariate regression (LMR) as a special case. In LMR, a regression coefficient matrix is assumed to be the sum of several sparse matrices, which is called layer. Therefore, the sparseness of the resulting coefficient matrix is controlled by how many layers are used. It is theoretically guaranteed that an LMR solution converges to the unconstrained solution as the number of layers increases. LMR is assessed in a simulation study and illustrated with a real data example. Further, Layered PCA is also proposed, in which a loading matrix is constrained to have a layered simple structure.

In Chapter 7, a new penalty function that can be used for penalized estimation

in various multivariate analysis procedures is proposed. The new penalty function shrinks a solution matrix towards a target matrix with a simple structure. The proposed function is a generalization of the existing ones, in that it includes LASSO and ridge penalties as special cases.

Chapter 8 is devoted for summarization and general discussions throughout the foregoing chapters. The thesis finally concludes that the studies successfully developed the procedures that are able to simplify solution matrices, by accomplishing sparsity and simplicity at the same time, which is the main contribution of the thesis to the studies on multivariate data analysis.

# Contents

**Declaration of Authorship** i

**Abstract** ii

**List of Figures** viii

**List of Tables** x

**1 Introduction** 1

1.1 Multivariate Statistical Analysis in Scientific Research . . . . . 1

1.2 Matrix-Based Formulations of Multivariate Analysis Procedures . . . . 5

1.2.1 Multivariate Regression Analysis . . . . . 5

1.2.2 Principal Component Analysis . . . . . 8

1.2.3 Factor Analysis . . . . . 15

1.2.4 $K$-means Clustering . . . . . 22

1.2.5 Canonical Correlation Analysis . . . . . 25

1.3 Interpretability in Multivariate Analysis Procedures . . . . . 28

1.3.1 Sub-concepts of Interpretability . . . . . 29

1.4 Purpose of the Thesis . . . . . 30

**2 Factor Rotation to Simple Structure with Permutation of Variables 33**

2.1 Introduction . . . . . 33

2.2 Proposed Method: Permutimin . . . . . 37

2.3 Relationships to the Existing Procedures . . . . . 38

2.4 Algorithm . . . . . 39

2.5 Target Matrices of Simple Structures . . . . . 41

2.5.1 Modification of Thurstone's Simple Structure . . . . . 41

2.5.2 Constructing Binary Target Matrices . . . . . 42

2.5.3 Selection of Target Cardinality . . . . . 43

2.6 Illustrations . . . . . 44

2.6.1 Thurstone's Box Problem . . . . . 44

2.6.2 Personality Loadings . . . . . 46

2.7 Discussion . . . . . 48

Appendix 2.A Details of Rotation Procedures in Section 6 . . . . . 50

Appendix 2.B Multiple Runs Procedure . . . . . 50

Appendix 2.C Uniqueness of the binary target matrix $\mathbf{B}$ for a given $N$ . . 51

**3 Rotation in Canonical Correlation Analysis as Maximizing Sum of Squared Correlations 52**
3.1 Introduction . . . 52
3.2 A New Formulation of Canonical Correlation Analysis . . . 53
3.3 Individual Rotation of Structure Matrices . . . 54
3.4 Real Data Example . . . 55
3.5 Discussion . . . 57
Appendix 3.A An Algorithm for Simultaneous Varimax Rotation . . . 58

**4 Biplot Procedures with Joint Classification Objects and Variables 60**
4.1 Introduction . . . 60
4.2 Proposed Method . . . 63
4.2.1 Crisp Cluster-wise Biplot Analysis . . . 63
4.2.2 Fuzzy $K$-means Clustering . . . 64
4.2.3 Fuzzy Cluster-wise Biplot Analysis . . . 64
4.2.4 Multiple Runs Procedure . . . 67
4.3 Simulation Study . . . 68
4.4 Real Data Example . . . 70
4.5 Concluding Remarks . . . 76
Appendix 4.A Joint Procrustes transformation . . . 77

**5 A Modified $K$-means clustering for obtaining a simple centroid matrix 79**
5.1 Introduction . . . 79
5.1.1 Related Method: Factor Rotation . . . 82
5.2 Algorithm . . . 82
5.2.1 Outline: Whole Algorithm . . . 82
5.2.2 Optimization in M-step and Y-step . . . 83
5.3 Cardinality Selection based on Information Criteria . . . 85
5.4 CCKM with Row/Column-wise Cardinality Constraint (RC-CCKM) . 87
5.5 Simulation Studies . . . 88
5.5.1 Accuracy of Cardinality Selection . . . 88
5.5.2 Accuracy of Parameter Estimation . . . 90
5.6 Real Data Examples . . . 91
5.6.1 Example 1: Fisher's Iris Data . . . 91
5.6.2 Example 2: Wine Data . . . 92
5.6.3 Example 3: Job Impression Data . . . 94
5.7 Concluding Remarks . . . 94

**6 Multivariate Analysis Procedures with Layered Simple Structure 96**
6.1 Introduction . . . 96
6.2 Proposed Method . . . 98
6.2.1 Layered Multivariate Regression (LMR) . . . 100

6.2.2 Layered Principal Component Analysis (LPCA) . . . 101
6.2.3 Determining number of layers $L$ . . . 101
6.2.4 Properties . . . 102
6.3 Real Data Examples . . . 104
6.3.1 Example of LMR: Tobacco Data . . . 104
6.3.2 Example of LPCA: Wine Data . . . 104
6.4 Concluding Remarks . . . 106

**7 Procrustes Penalty Function for Matching Matrices to Targets 108**
7.1 Introduction . . . 108
7.2 Proposed Method . . . 109
7.2.1 General Case: Multivariate Regression . . . 109
7.2.2 Applications in Machine Learning Problems . . . 111
7.3 Example . . . 112
7.3.1 Three-Mode Component Analysis to Multiple Personality Data 114
7.4 Conclusions . . . 114

**8 Discussions 116**
8.1 Summary of Chapters 2 to 7 . . . 116
8.2 Four Approaches towards Enhanced Simplicity . . . 117
8.2.1 Rotation of Solution Matrix . . . 118
8.2.2 Combination with Clustering . . . 119
8.2.3 Sparse Estimation with Cardinality Constraint . . . 120
8.2.4 Some Emerging Techniques . . . 121
8.3 Hierarchical Relationships of the Proposed Methods . . . 121

**A Short Paper (In Japanese) 123**

**Acknowledgements 143**

# List of Figures


1.1 Hierarchical relationship between phenomena, data, and model (cited from Takane (1980)) and modified). . . . . 2
1.2 Scree plot of variance explained for baseball data. . . . . 14
1.3 Biplot representation of baseball data. . . . . 15
1.4 Path diagram of factor analysis result. . . . . 16
1.5 Path diagram of Promax-rotated MDFA solution to personality data. . 21
1.6 Example of dendrogram . . . . . 22

2.1 Attained minimum of function value (A) and $\delta(N)$ (B) for $N = 15 \cdots, 36$ for the box problem. . . . . 46
2.2 Attained minimum of function value (A) and $\delta(N)$ for $N = 12 \cdots, 36$ for Yanai & Ichikawa's (2006) loading matrix . . . . . 48

3.1 Network representation of the independently rotated structure matrices. 57

4.1 PC biplots of two artificial datasets . . . . . 61
4.2 PC biplots of generated datasets under two different conditions . . . . 69
4.3 The scree plot of the singular values of the wine dataset. . . . . 72
4.4 Joint plot of the attained function value (solid line) and $FPI$ (dotted line) against the number of variable clusters $L$. . . . . 73
4.5 The estimated object memberships expressed as bar plots with $\alpha = 1.0, 1.1, 1.2, 1.3, 1.4$ and 1.5. . . . . 73
4.6 The PC and FC biplots of the wine dataset. . . . . 74

5.1 Boxplots of $SD(c_{AIC})$ and $SD(c_{BIC})$ as indices of discrepancy between true and estimated cardinality by AIC and BIC criterion. . . . . 88
5.2 Boxplot of $ARI$ (adjusted Rand index) and $AAE(\hat{\mathbf{Y}}, \mathbf{Y})$ values in the case with $(n, p) = (100, 30)$. . . . . 89
5.3 Boxplot of $ARI$ (adjusted Rand index) and $AAE(\hat{\mathbf{Y}}, \mathbf{Y})$ values in the case with $(n, p) = (30, 100)$. . . . . 90

6.1 Scree plot of variance explained against the number of layers; variance explained of unconstrained solution is shown as dotted line . . . . . 103
6.2 Scree plot of variance explained against the number of layers; variance explained of standard PCA is shown as dotted line . . . . . 105
6.3 Principal component biplots of standard PCA, LPCA, and CDPCA . . 106
6.4 Differences between reduced space of LPCA/CDPCA and standard PCA 107

7.1 Solution paths for three components . . . . . 113

8.1 Hierarchical relationship between the proposed and existing procedures. 122

A.1 提案手法，既存手法とスパース性/単純構造の関係. . . . . 126
A.2 既存手法/提案手法によるバイプロットの比較. . . . . 128

# List of Tables

1.1 Estimated matrix of regression coefficients. . . . . 7
1.2 Estimated coefficient vectors by sparse estimation. . . . . 8
1.3 Examples of a simple and not-simple structure. . . . . 13
1.4 Initial and Varimax-rotated loading matrix for baseball data. . . . . 14
1.5 Factor loading matrices for personality data. . . . . 21
1.6 Centroid matrix obtained by $K$-means clustering. . . . . 24
1.7 Structure matrices and canonical correlations obtained by canonical correlation analysis to Sports data. . . . . 28

2.1 Thurstone's (1947) definition of simple structure rewritten more clearly by Browne (2001, p. 115) . . . . . 34
2.2 Examples of $7 \times 3$ matrices for illustrating the definition of simple structure. . . . . 35
2.3 Binary target matrices $\mathbf{B}$ $(8 \times 3)$ satisfying the proposed simple structure conditions. . . . . 43
2.4 Rotated loading matrices and factor correlations for the box problem by seven rotation procedures. . . . . 49
2.5 Rotated loading matrix for Yanai & Ichikawa's (2006) loading matrix obtained by six rotation procedures. . . . . 49

3.1 Unrotated and individually/simultaneously rotated structured matrices obtained by CANO to Sports data. . . . . 56
3.2 Averaged within-column variance and between covariance of the structure matrices. . . . . 56

4.1 The values for the 1st, 2nd, and 3rd quartiles of $ARI$ and the index of the congruence of memberships ($CM$). . . . . 71
4.2 The values for 1st, 2nd and 3rd quartiles of $CC$ and $SS$, and average percentage of the frequency of the local minimum. . . . . 71
4.3 Contingency table for comparing object membership with the three types of wine. . . . . 75

5.1 Estimated centroid matrix by $k$-means with four clusters to job impression data. . . . . 81
5.2 Estimated centroid matrices by CCKM with $c = 10$ (left) and $k$-means (right). . . . . 92

5.3 Contingency table for species vs. the estimated partitions by CCKM and $k$-means. . . . . 92
5.4 Centroid matrix obtained by CCKM with cardinality of $r = 35$. . . . . 93
5.5 Contingency table for wine categories vs. the estimated partitions by CCKM and $k$-means. . . . . 93
5.6 Estimated centroid matrix by CCKM with row/column-cardinality constraint and four clusters for job impression data . . . . . 94

6.1 Estimated coefficient matrices by LMR with $L = 1, \cdots, 4$ and unconstrained solution with proportion of variance explained (Vexp.); element equaling to zero shows a blanc cell. . . . . 104

7.1 Estimated $\mathbf{A}$s for $\lambda = 20, 50, 100$ and the target matrix. . . . . 114
7.2 Estimated core arrays by Tucker3, PARAFAC, and the proposed method with three $\lambda$s. . . . . 115

A.1 性格データに関する因子分析の結果．太字は絶対値が 0.5 以上の要素． 124
A.2 Box データに対する Permutimin 回転，Oblimin 回転，Promax 回転の結果． . . . . 126
A.3 体力診断テスト/運動能力テストに対する正準相関分析の結果と提案手法による回転の結果． . . . . 127
A.4 $K$-means と RSKM により得られたセントロイド行列．空白のセルはゼロに等しい要素を表す． . . . . 129
A.5 Layered Multivariate Regression の適用例．空白のセルはゼロに等しい要素を表す． . . . . 130
A.6 提案手法によるスパース PCA の結果と，用いたターゲット行列．空白はゼロに等しい要素を表す． . . . . 131

# Chapter 1

# Introduction

**ABSTRACT**

Multivariate data analysis procedures are widely used in today's data-based scientific researches to extract useful knowledge from data for hypothesis creation. A formal definition of data analysis is presented, followed by a matrix-based introduction of the following six fundamental multivariate analysis procedures: multivariate regression analysis, principal component analysis, factor analysis, $K$-means clustering, and canonical correlation analysis, with small examples. Further, it is discussed that interpretable solutions are essential in the exploratory use of these procedures, and two sub-concepts of interpretability are introduced. As presented in the final section, the purpose of the thesis is to refine the multivariate analysis procedures towards their improved interpretability with their matrix-intensive modifications, which is presented in the final section.

## 1.1 Multivariate Statistical Analysis in Scientific Research

Following the classical but time-honored definition by Popper (1962), today's scientific studies are driven by the *hypothetico-deductive* method. In these studies, a series of hypotheses with falsifiability is made and verified to test the theory on a phenomena of interest. In empirical sciences including psychology, sociology, and human sciences, hypotheses are verified based on *data* collected by experiments and observations, and the data are collected in various forms; it may be qualitative or quantitative, and structural or nonstructural. For example, in psychology, for understanding the fundamental mechanism of the psychological and cognitive process of humans, a research hypothesis is established based on previous research or empirical facts, and some psychological experiments are planned and conducted to verify the hypothesis. Further, the result of the experiment is used to improve the research hypothesis in the following manner. It is assumed that the current hypothesis is true, and the assumption would be rejected if it is not consistent with the observed data. Once the current hypothesis is rejected, another more probable hypothesis would be made and the same verification

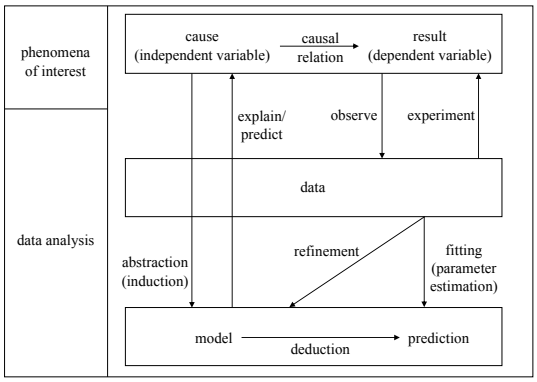



FIGURE 1.1: Hierarchical relationship between phenomena, data, and model (cited from Takane (1980)) and modified).

process is repeated. Similarly, in sociology, to understand a principle of dynamism of our society, researchers often conduct questionnaires or interview surveys that are designed to test their hypotheses, and determine whether they are true based on the results. Thus, data collected by experiments or observations play a critical role in today's scientific research driven by data-based hypothesis verification.

To understand what data tell us about a phenomena of interest, we have to transform it into an understandable form for researchers. In particular, *data analysis* is a tool for extracting some useful but non-apparent information from data, and its result would be used for hypothesis creation, verification, and its refinement. Takane (1980) described the relationship between phenomena, data, and model as shown in Figure1.1 and classified the data analysis methods into the following two families. The first one is *confirmatory data analysis* (CDA), which verifies a hypothesis by comparing the actual behavior of a phenomena of interest with a prediction derived from a model of the phenomena. Student's $t$-test and analysis of variance (ANOVA) are well known as examples of CDA (Heumann et al., 2016; Hogg et al., 2005). In CDA, a research hypothesis is expressed as a form of a mathematical model, and its fitness to given data is evaluated as a measure of the validity of the hypothesis; if the model fairly fits to the data, the hypothesis is considered true, and vise versa. Therefore, how well a model fits to data is of the greatest interest in CDA. The second one is *exploratory data analysis* (EDA), which dates back to Tukey (1970), and it aims to extract useful knowledge from data. For example, when a hypothesis has been rejected, there may be a need for insight or clues that help establish another hypothesis. EDA serves to extract such insight that would lead to another hypothesis. Especially, in recent years, EDA is called *data mining*, and it is also used for specifying a small number of variables that are useful for prediction. Variable selection is beneficial for building a robust prediction model. EDA is thus becoming increasingly important not only because it plays a key role in scientific research, but also because it is suitable for dealing with high-dimensional and large data, which is becoming common today. Thus the thesis focuses on EDA rather than CDA.

Above, the author discussed the role of multivariate analysis procedures in scientific researches. Here, as a preliminary introduction to the following discussion, a mathematical definition of data analysis is provided.

**Definition 1.1.1** (data analysis)**.** *Let $\mathcal{X}$ be the set of given data with $\mathcal{M}$ being the set of mathematical models to be fitted to $\mathcal{X}$. We define the minimization problem*

$$min.\ L(\mathcal{M}(\mathcal{P}|\mathcal{X}), \mathcal{X})\ over\ \mathcal{P} \tag{1.1}$$

*as the "data analysis of $\mathcal{X}$", where $\mathcal{M}(\mathcal{P}|\mathcal{X})$ denotes a prediction of $\mathcal{X}$ derived by $\mathcal{M}$, and $\mathcal{P}$ is the set of model parameters of $\mathcal{M}$. $L(\mathcal{M}(\mathcal{P}|\mathcal{X}), \mathcal{X})$ is a measure of the discrepancy between $\mathcal{M}(\mathcal{P}|\mathcal{X})$ and $\mathcal{X}$.*

In the above definition, data analysis is defined as a data-fitting procedure, in that the discrepancy between $\mathcal{X}$ and its prediction $\mathcal{M}(\mathcal{P}|\mathcal{X})$ is minimized over $\mathcal{P}$. As an example, consider the case with the multiple regression analysis where a single dependent variable is regressed on multiple independent variables. Multiple regression analysis is a basic technique of multivariate statistical data analysis, and it is used to make a prediction of a dependent variable or explain its variation by a linear combination of independent variables (Izenman, 2008). Let $\mathbf{y} = [y_1 \cdots, y_n]'$ be the $n$-dimensional vector of a dependent variable where $y_i$ $(i = 1, \cdots, n)$ is the observed value for the $i$-th observation, and $\mathbf{X} = \{x_{ij}\} = [\mathbf{x}_1, \cdots, \mathbf{x}_p]$ be the $n \times p$ matrix of $p$ independent variables with respect to the same observations in $\mathbf{y}$. We thus have the set of data $\mathcal{X} = \{\mathbf{X}, \mathbf{y}\}$. Note that $\mathbf{X}$ is assumed to be column standardized without the loss of generality. The mathematical model for the multiple regression analysis on $\{\mathbf{X}, \mathbf{y}\}$ is formally expressed as

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \mathbf{e} \tag{1.2}$$

where $\mathbf{w} = \{w_j\}$ is a $p$-dimensional vector of regression coefficients that corresponds to $p$ independent variables. Further, $\mathbf{e} = \{e_i\}$ is an $n$-dimensional vector of residuals. (1.2) is re-expressed as the scalar form

$$y_i = x_{i1}w_1 + x_{i2}w_2 + \cdots + x_{ip}w_p + e_i \tag{1.3}$$

indicating a linear relationship is assumed between the independent and the dependent variables.

Following Definition 1.1.1, the multiple regression analysis considered above seeks to find $\mathcal{P} = \{\mathbf{w}\}$ referring $\mathcal{X} = \{\mathbf{X}, \mathbf{y}\}$. Optimization criteria to be minimized can take various forms, and here, the least squares criterion

$$LS(\mathbf{w}) = ||\mathbf{e}||^2 = ||\mathbf{y} - \mathbf{X}\mathbf{w}||^2 \tag{1.4}$$

is considered, which is the sum of the squared elements of $\mathbf{e}$. For the minimization, we have the following normal equation

$$\frac{\partial LS(\mathbf{w})}{\partial \mathbf{w}} = -2\mathbf{X}'\mathbf{X}\mathbf{w} + 2\mathbf{X}'\mathbf{y} = \mathbf{0}_p \tag{1.5}$$

obtained by the partial derivative of (1.4) over $\mathbf{w}$, where $\mathbf{0}_p$ denotes the $p$-dimensional vector filled with 0s. Solving (1.5) with respect to $\mathbf{w}$, we have

$$\mathbf{w} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{1.6}$$

under the assumption that the inverse of $\mathbf{X}'\mathbf{X}$ exists, and it is the minimizer of (1.4). The estimate of the regression coefficient vector is obtained by substituting $\mathbf{X}$ and $\mathbf{y}$ with the observed data, and its validity is guaranteed by the optimality of minimizing the criterion (1.4).

So far, a least squares criterion is used for the parameter estimation in the multiple regression analysis, and other criteria are also available. The *maximum likelihood* criterion is known as such a criterion, in which the elements of $\mathbf{e}$ and $\mathbf{y}$ are assumed as random variables, each of which follows a specific probability distribution. In maximum likelihood, a distribution of $\mathbf{e}$ is called an *error model*, and a mathematical model $\mathcal{M}$ as a *representation model*, in that it is a parametric representation of data generation process of $\mathbf{y}$. As an error model, a normal distribution is often assumed.

$$e_i \overset{iid}{\sim} N(0, \sigma^2) \ for \ i = 1, \cdots, n. \tag{1.7}$$

Above, $\sigma^2$ is called *error variance* and treated as a parameter to be estimated. Under the above error model and the representation model in (1.4), each of the elements in $\mathbf{y}$ has the probability density function

$$\phi(y_i|\mathbf{w}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{(y_i - \mathbf{x}_{(i)}\mathbf{w})^2}{2\sigma^2}\right) \tag{1.8}$$

where $\mathbf{x}_{(i)}$ is the $p$-dimensional vector of the observed values for the $i$-th observation in $\mathbf{X}$. The probability distribution function stands for, by its definition, the probability that $y_i$ is observed with $\mathbf{w}$ and $\sigma^2$ being given. Conversely, it is considered as a function of $\mathbf{w}$ and $\sigma^2$ given $\mathbf{y}_i$, and it evaluates how likely $\mathbf{w}$ and $\sigma^2$ are when a specific $y_i$ is observed. In the maximum likelihood estimation, a probability density function (or, probability mass function if $y_i$ is discrete) is treated as a function of parameters with given data, and the likelihood function, which is the product of the probability density functions for all observations, is maximized over parameters. In other words, a parameter estimate is given by one that is the most probable within the parameter space. For example, the likelihood function in the multiple regression

analysis is given by

$$L(\mathbf{w}, \sigma^2) = \prod_{i=1}^{n} \left\{ (2\pi\sigma^2)^{-1/2} \exp\left( \frac{(y_i - \mathbf{x}_{(i)}\mathbf{w})^2}{2\sigma^2} \right) \right\}. \tag{1.9}$$

Often, for the ease of computation, the log-likelihood function $\log L(\mathbf{w}, \sigma^2)$ is maximized. It should be noted that, in multiple regression analysis, least squares and maximum likelihood estimation are proved to be equivalent. This is easily shown as follows; in $\log L(\mathbf{w}, \sigma^2)$, the term relevant to $\mathbf{w}$ is $-\sum_i (y_i - \mathbf{x}_{(i)}\mathbf{w})^2 = -||\mathbf{y} - \mathbf{X}\mathbf{w}||^2$, and it is equivalent to (1.4). The equivalence is, however, not always the case in other procedures or other error models. In EDA, it is considered to be beneficial to make no strong distributional assumptions on the data and error model because its purpose is information extraction and not evaluating the fitness to data. Therefore, in this thesis, least squares criterion, which does not require special distributional assumptions on the distribution of data, is considered as an optimization criterion.

## 1.2 Matrix-Based Formulations of Multivariate Analysis Procedures

In this section, six multivariate analysis procedures are introduced with their mathematical formulations, estimation procedures, numerical examples, and related methods. Note that notations in a subsection are not consistent with those in other subsections.

### 1.2.1 Multivariate Regression Analysis

Multivariate regression analysis is a natural generalization of multiple regression analysis, in which two or more dependent variables are regressed on multiple independent variables (Izenman, 2008). In multivariate regression, $q$ dependent variables are regressed on $p$ independent ones.

#### Formulation

Let $\mathbf{Y} = \{\mathbf{y}_1, \cdots, \mathbf{y}_q\}$ be an $n \times q$ matrix, which contains the observed values for $n$ observations with respect to $q$ dependent variables. The least squares criterion to be minimized is expressed as

$$L_{MVR}(\mathbf{W}) = ||\mathbf{Y} - \mathbf{X}\mathbf{W}||^2 \tag{1.10}$$

and the multivariate regression analysis of $\{\mathbf{X}, \mathbf{Y}\}$ thus minimizes $L_{MVR}(\mathbf{W})$ over $\mathbf{W}$, an unknown $p \times q$ matrix. The $(j,k)$-th element $(j = 1, \cdots, p,\ k = 1, \cdots, q)$ of $\mathbf{W}$ denoted as $w_{jk}$ stands for the regression coefficient of the $j$-th independent variable for the prediction of the $k$-th dependent variable.

### Parameter Estimation

It is known that there exists a closed form solution of $\mathbf{W}$ that minimizes (1.10). Here, by partially differentiating (1.10) with respect to $\mathbf{W}$ (Magnus and Neudecker, 2019), we have

$$\frac{\partial L_{MVR}(\mathbf{W})}{\partial \mathbf{W}} = -2\mathbf{X}'\mathbf{X}\mathbf{W} + 2\mathbf{X}'\mathbf{Y} = \mathbf{O}_{p\times q} \tag{1.11}$$

where $\mathbf{O}_{p\times q}$ denotes the $p\times q$ matrix filled with zeros. The normal equation is therefore solved by

$$\mathbf{W} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{1.12}$$

that is the minimizer of the loss function in (1.10). Note that the $k$-th column of $\mathbf{W}$ noted as $\mathbf{w}_k$ is expressed as $\mathbf{w}_k = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}_k$. Noting (1.6), multivariate regression analysis is thus equivalent to applying a multiple regression analysis to each of the $q$ dependent variables with the common $\mathbf{X}$ separately, and arranging the resulting $q$ regression coefficient vectors as the columns of $\mathbf{W}$. Also, as seen in (1.12), $\mathrm{rank}(\mathbf{X}) = p$ (i.e., $\mathbf{X}$ is of full column-rank) is required in order to obtain the inverse of $\mathbf{X}'\mathbf{X}$. In addition, if $\mathbf{X}$ contains two independent variables that are highly correlated, which is called *multicollinearity*, the computation of $(\mathbf{X}'\mathbf{X})^{-1}$ would be unstable because $\mathbf{X}$ has at least one singular value that is close to zero. The observation leads to another requirement to $\mathbf{X}$ that it should have a sufficiently large condition number, defined as the absolute ratio of the largest singular value to the smallest one.

### Example

An example of multivariate analysis is shown here using an artificial dataset. 100 university students were asked to evaluate their degree of interest in the following six academic fields from 0 (no interest) to 100 (highest interest); law, economics, mathematics, engineering, physics, and informatics. They also took three examinations named test A, B and C and their scores were recorded. Multivariate regression analysis is then applied to $\mathbf{X}(100 \times 6)$ (interest score) and $\mathbf{Y}(100 \times 3)$ (test score). The estimated regression coefficient matrix is shown in Table 1.1, and it can be interpreted as follows. The regression coefficient on "law" in the regression of "test A" is 3.909, which stands for the score of "test A" increases 3.909 on average, with the unit increment of "law". Because the scales of all independent variables are identical, regression coefficients are able to be compared mutually. $\mathbf{X}$ should be column-wise standardized if they are not identical. For instance, "law" has a higher impact on "test A" compared with "economics". It is rephrased as that "law" and "test A" are closely related. As seen here, one can capture the correspondence between independent and dependent variables by referring the elements in the coefficient matrix that are higher in absolute than others.

TABLE 1.1: Estimated matrix of regression coefficients.

| | test A | test B | test C |
|---|---|---|---|
| law | 3.909 | 0.499 | -0.106 |
| economics | 2.697 | 0.201 | -0.199 |
| mathematics | -0.204 | -5.098 | 0.100 |
| engineering | -1.397 | 6.799 | 0.300 |
| physics | 0.201 | -0.195 | 3.706 |
| informatics | -0.103 | 0.195 | 4.895 |

**Related Methods**

As seen above, a regression coefficient matrix expresses correspondences between independent and dependent variables. Here, consider the case with the $j$-th independent variable is not necessary for a prediction of the $k$-th dependent variable. In this case, if we set $w_{jk} = 0$, the independent variable is no longer used for the prediction. Specifying the subset of independent variables that is necessary for a prediction and removing the unnecessary variables, is called *variable selection*, and it is accomplished by setting several regression coefficients at 0 or close to 0. The potential benefits of variable selection are the following two points. The first point is that a prediction would be robust if a limited number of independent variables are used; the fluctuation of the variables that are unnecessary for the prediction is suppressed and thus the prediction would be stabilized. The second point is that the correspondence between two sets of variables is easier to be captured if fewer independent variables are used. The hypothesis derived from a result of variable selection would be appreciated in terms of parsimony, because a phenomenon of interest (dependent variable) is explained by a reduced number of causes (independent variables).

*Penalized estimation* is widely used for obtaining regression coefficients including several elements equaling or close to 0 for the purpose of variable selection (Hastie et al., 2015; James et al., 2013). For instance, consider to minimize the composite loss function

$$||\mathbf{y} - \mathbf{X}\mathbf{w}||^2 + \lambda P(\mathbf{w}) \tag{1.13}$$

that is the sum of the loss function in multiple regression analysis, and the *penalty function* $P(\mathbf{w})$ on $\mathbf{w}$ weighted by the tuning parameter $\lambda(> 0)$. Variety of penalty functions had been proposed, such that LASSO penalty (Tibshirani, 1996)

$$P_{LASSO}(\mathbf{w}) = ||\mathbf{w}||_1 = \sum_j |w_j| \tag{1.14}$$

where $||\bullet||_1$ stands for the $l_1$ norm of a vector or matrix, Ridge penalty (Hoerl and Kennard, 1970)

$$P_{Ridge}(\mathbf{w}) = ||\mathbf{w}||^2 \tag{1.15}$$

TABLE 1.2: Estimated coefficient vectors by sparse estimation. Blanc cells express elements equal to 0.

| | LASSO penalty | | | | Ridge penalty | | | |
|---|---|---|---|---|---|---|---|---|
| | $\lambda$=0 | $\lambda$=2 | $\lambda$=4 | $\lambda$=12 | $\lambda$=0 | $\lambda$=50 | $\lambda$=70 | $\lambda$=100 |
| law | 3.900 | 3.815 | 3.726 | 3.412 | 3.900 | 2.831 | 2.555 | 2.231 |
| economics | 2.700 | 2.631 | 2.564 | 2.270 | 2.700 | 2.029 | 1.846 | 1.626 |
| mathematics | -0.199 | -0.121 | -0.043 | | -0.199 | -0.154 | -0.142 | -0.128 |
| engineering | -1.398 | -1.296 | -1.199 | -0.846 | -1.398 | -0.841 | -0.716 | -0.581 |
| science | 0.199 | 0.120 | 0.041 | | 0.199 | 0.056 | 0.030 | 0.006 |
| informatics | -0.098 | -0.019 | | | -0.098 | -0.066 | -0.060 | -0.053 |

or elastic net (Zou and Hastie, 2005)

$$P_{enet}(\mathbf{w}) = \frac{\alpha}{2}||\mathbf{w}||^2 + (1-\alpha)||\mathbf{w}||_1 \tag{1.16}$$

where $\alpha$ ($0 \leq \alpha \leq 1$) is the weight parameter. Note that LASSO and elastic-net penalties are able to estimate $\mathbf{w}$ including exact 0 elements, while ridge penalty does not have such property. $\lambda$ ($\geq 0$) controls the relative strength of a penalty function against a loss function, and it is often set in advance via cross-validation. Generally, when $\lambda$ is large, some elements in $\mathbf{w}$ shrink to 0, and therefore the resulting $\mathbf{w}$ gets to be sparse.

As an example, the estimated regression coefficient vectors via LASSO and Ridge penalty functions with several $\lambda$s are shown in Table 1.2. When $\lambda = 0$, two penalty functions vanish and therefore the results are equivalent to the one by the non-penalized estimation shown in Figure 1.1. We can see that some elements in $\mathbf{w}$ get closer to 0 as $\lambda$ gets to be large in both penalties. Further, the result by LASSO penalty have some elements equal to 0, while it is not in the case of Ridge penalty.
It should be noted that penalized estimation in multiple regression is easily extend to multivariate regression analysis;

$$||\mathbf{Y} - \mathbf{XW}||^2 + \sum_k \lambda_k P(\mathbf{w}_k) \tag{1.17}$$

is minimized where $\lambda_k$ is the tuning parameter corresponds to the $k$-th column of $\mathbf{W}$.

### 1.2.2 Principal Component Analysis

Multivariate regression analysis requires two sets of variables; independent and dependent variables. A data analysis procedure that requires at least one dependent variable is called *supervised learning*, while it is not required in *unsupervised learning*. Unsupervised learning is often used for exploratory purpose, in that it extracts potentially useful information hidden in multiple variables.

*Principal component analysis (PCA)* is known as an unsupervised learning procedure and often used for EDA (Jolliffe, 2011). PCA summarizes a multivariate data matrix by compressing the variables into several weighted scores called *principal components*, and the estimated principal components are interpreted as the essential part that is hidden in the matrix. PCA also serves to reduce dimensionality of data and thus used for data preprocessing in order to handle high-dimensional data by omitting its redundant information.

**Formulation**

Here, PCA is formulated as lower rank approximation of a data matrix. Consider the case with the a $n \times p$ matrix $\mathbf{X}$ are going to be compressed into $r(\leq p)$ principal components. Let $\mathbf{F} = \{f_{ij}\}$ be an $n \times r$ matrix of principal component scores, and $\mathbf{A} = \{a_{ij}\}$ be a $p \times r$ matrix called a *c*omponent loading matrix. As shown below, $\mathbf{A}$ is considered as the correlation matrix of the $p$ variables and the $r$ components, if $\mathbf{X}$ is a column-standardized matrix. A PCA on $\mathbf{X}$ with $r$ principal components is formulated as the minimization of

$$L_{PCA}(\mathbf{F}, \mathbf{A}) = ||\mathbf{X} - \mathbf{F}\mathbf{A}'||^2 \tag{1.18}$$

under the constraint that

$$\frac{1}{n}\mathbf{F}'\mathbf{F} = \mathbf{I}_r \tag{1.19}$$

where $\mathbf{I}_r$ denotes the $r \times r$ identity matrix. The constraint indicates that a component score has unit variance and is independent to other component scores. It also serves to eliminate the indeterminacy of PCA solution except for orthogonal rotation as discussed below.

Compared with (1.10), (1.18) is regarded as a multivariate regression's loss function in which the $r$ principal components in $\mathbf{F}$ are regressed on the $p$ variables in $\mathbf{X}$. In this case, the $i$-th data vector $\mathbf{x}_{(i)}$ having $p$ dimensionality is approximated by the linear combination of the $r$-dimensional vector of principal component score $\mathbf{f}_{(i)}$; $\mathbf{x}_{(i)} \approx \mathbf{f}_{(i)}\mathbf{A}$. There, the original data vector in $p$-dimensional space is embedded into $r$-dimensional space by $\mathbf{A}$, which indicates that PCA reduces dimensionality of original data matrix. One can apply PCA with $r = 2$ to high dimensional data matrix and make a scatter plot of the resulting $\mathbf{F}$, in order to visualize the data matrix instantly.

Denoting the matrix rank of $\mathbf{X}$ and its approximation $\mathbf{F}\mathbf{A}'$ as $\text{rank}(\mathbf{X})$ and $\text{rank}(\mathbf{F}\mathbf{A}')$, respectively, we have the following relationship;

$$\text{rank}(\mathbf{X}) \geq \text{rank}(\mathbf{F}\mathbf{A}'). \tag{1.20}$$

PCA is thus viewed as a lower rank approximation of $\mathbf{X}$, or, equivalently, a decomposition of the data matrix into a score matrix $\mathbf{F}$ and a loading matrix $\mathbf{A}$. Further, for the covariance matrix of $\mathbf{X}$ denoted as $\mathbf{\Sigma}_{\mathbf{X}} = n^{-1}\mathbf{X}'\mathbf{X}$ and the counterpart of $\mathbf{F}\mathbf{A}'$ as

$\mathbf{\Sigma_{FA'}} = n^{-1}\mathbf{AF'FA'} = \mathbf{AA'}$, we also have

$$\text{rank}(\mathbf{\Sigma_X}) \geq \text{rank}(\mathbf{\Sigma_{FA'}}) \tag{1.21}$$

meaning that the covariance structure of the data matrix is also approximated by the one with reduced rank.

### Parameter Estimation

It should be noted that the above formulation of PCA is based on Pearson (1901), while Hotelling (1933) proposed another formulation that is more popular than the former (Jolliffe, 2011). In the Hotelling's formulation, a $p$-dimensional coefficient vector $\mathbf{w}_1$ is obtained so as to maximize

$$n^{-1}\mathbf{w}_1'\mathbf{X'X}\mathbf{w}_1 \tag{1.22}$$

the variance of the weighted score vector $\mathbf{X}\mathbf{w}_1$. The constraint $\mathbf{w}_1'\mathbf{w}_1 = 1$ is often imposed on $\mathbf{w}_1$ in order to avoid a trivial solution that $\mathbf{w}_1 \to \infty$. The first principal component score is $\mathbf{X}\mathbf{w}_1$, and $\mathbf{w}_2$ that forms the second component is then obtained as a maximizer of

$$n^{-1}\mathbf{w}_2'\mathbf{X'X}\mathbf{w}_2 \tag{1.23}$$

subject to $n^{-1}\mathbf{w}_1'\mathbf{X'X}\mathbf{w}_2 = 0$ and $\mathbf{w}_2'\mathbf{w}_2 = 1$. In this manner, $r$ principal components are sequentially obtained. In the following, however, the principal components are simultaneously estimated based on Pearson's formulation, in order to emphasize the essential feature of PCA; it approximates a data matrix by a matrix with reduced rank.

First, consider to estimate $\mathbf{A}$ that minimizes (1.18) with $\mathbf{F}$ kept fixed. It is easily obtained as follows;

$$\mathbf{A} = \mathbf{X'F}(\mathbf{F'F})^{-1} = n^{-1}\mathbf{X'F}. \tag{1.24}$$

Note that it is a correlation matrix if $\mathbf{X}$ is column-wise centered.

Next, $\mathbf{F}$ that minimizes (1.18) subject to (1.19) is estimated with fixed $\mathbf{A}$. (1.18) is expanded as

$$L_{PCA}(\mathbf{F}, \mathbf{A}) = \text{tr}\mathbf{X'X} - 2\text{tr}\mathbf{X'FA'} + n\mathbf{AA'} \tag{1.25}$$

and only the second term is relevant to $\mathbf{F}$. It immediately leads the equivalence of the minimization of $L_{PCA}(\mathbf{F}|\mathbf{A})$ to the maximization of $\text{tr}\mathbf{X'FA'}$ over $\mathbf{F}$.

For this maximization, the following theorem by ten Berge (1993) is useful.

**Theorem 1.2.1** (ten Berge's theorem)**.** *Let* $\mathbf{M}$ *be an* $m \times m$ *sub-orthonormal matrix with rank* $r \leq m$*, which is defined as a matrix if it can be completed to an orthonormal matrix by adding row or columns, or both. Let* $\mathbf{C} = \text{diag}(c_1, \cdots, c_m)$ *be a diagonal and nonnegative matrix and its diagonal elements are arranged in descending order.*

*Then, the trace function* tr$\mathbf{GC}$ *has the upper bound*

$$tr\mathbf{GC} \leq \sum_{i}^{r} c_i \tag{1.26}$$

*and the upper bound is attained with*

$$\mathbf{G} = \begin{bmatrix} \mathbf{I}_r & \mathbf{O}_{r\times(m-r)} \\ \mathbf{O}_{(m-r)\times r} & \mathbf{O}_{(m-r)\times(m-r)} \end{bmatrix} \tag{1.27}$$

*Proof.* See ten Berge et al. (1987) for the proof. □

The theorem is a generalization of Kristof's theorem (Kristof, 1970), which provides an upper bound of a trace function of an orthonormal matrix. Here, denote the singular value decomposition (SVD) of $\mathbf{X}$ as

$$\mathbf{X} = \mathbf{UDV}' \tag{1.28}$$

where $\mathbf{D}$ is the diagonal matrix of the $p^*$ singular values of $\mathbf{X}$ arranged in descending order, and $\mathbf{U}(n \times p^*)$ and $\mathbf{V}(p \times p^*)$ denote the matrices of the left and right singular vectors, respectively. Using the above theorem and (1.24), tr$\mathbf{X}'\mathbf{FA}'$ is maximized as follows;

$$\begin{aligned} \text{tr}\mathbf{X}'\mathbf{FA}' &= n^{-1}\text{tr}\mathbf{X}'\mathbf{FF}'\mathbf{X} \\ &= n^{-1}\text{tr}\mathbf{F}'\mathbf{XX}'\mathbf{F} \\ &= n^{-1}\text{tr}\mathbf{F}'\mathbf{UDV}'\mathbf{VDU}'\mathbf{F} \\ &= \text{tr}(n^{-1}\mathbf{U}'\mathbf{FF}'\mathbf{U})\mathbf{D}^2 \\ &\leq \text{tr}\mathbf{D}_r^2. \end{aligned} \tag{1.29}$$

The upper bound is attained when

$$n^{-1}\mathbf{U}'\mathbf{FF}'\mathbf{U} = \begin{bmatrix} \mathbf{I}_r & \mathbf{O}_{r\times(p^*-r)} \\ \mathbf{O}_{(p^*-r)\times r} & \mathbf{O}_{(p^*-r)\times(p^*-r)} \end{bmatrix} \tag{1.30}$$

indicating

$$\mathbf{F} = n^{1/2}\mathbf{U}_r \tag{1.31}$$

where $\mathbf{U}_r$ denotes the matrix of the $r$ left singular vectors correspond to the $r$ largest singular values. (1.31) is proved to satisfy the constraint on $\mathbf{F}$; $n^{-1}\mathbf{F}'\mathbf{F} = n^{-1}n\mathbf{U}'\mathbf{U} = \mathbf{I}_r$.

Further, using (1.28) and (1.31), (1.24) is reduced to

$$\mathbf{A} = n^{-1/2}\mathbf{VDU}'\mathbf{U}_r = n^{-1/2}\mathbf{V}_r\mathbf{D}_r \tag{1.32}$$

where $\mathbf{V}_r$ is a matrix of the $r$ first column vectors of $\mathbf{V}$. Thus, the PCA solution is simply obtained by SVD the of $\mathbf{X}$;

$$\mathbf{X} = \mathbf{UDV}' \approx \mathbf{U}_r\mathbf{D}_r\mathbf{V}_r' = (n^{1/2}\mathbf{U}_r)(n^{-1/2}\mathbf{V}_r\mathbf{D}_r) = \mathbf{FA}'. \tag{1.33}$$

If $\mathbf{X}$ is column-wise centered, namely, we have $\mathbf{J}_n\mathbf{X} = \mathbf{X}$ with $\mathbf{J}_n = \mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n'$ where $\mathbf{1}_n$ the $n$-dimensional vector filled with 1s. Then, $\mathbf{F}$ is also column-wise centered shown as

$$\mathbf{J}_n\mathbf{F} = \sqrt{n}\mathbf{J}_n\mathbf{K}_r = \sqrt{n}\mathbf{J}_n\mathbf{X}\mathbf{L}_r\mathbf{\Lambda}_r^{-1} = \sqrt{n}\mathbf{X}\mathbf{L}_r\mathbf{\Lambda}_r^{-1} = \mathbf{F}. \tag{1.34}$$

Referring (1.24), $\mathbf{A}$ is the covariance matrix of $\mathbf{X}$ and $\mathbf{F}$ and turns to be a correlation matrix if $\mathbf{X}$ is column-wise standardized.

In (1.33) $\mathbf{X}$ is approximated by $\mathbf{U}_r\mathbf{D}_r\mathbf{V}_r'$, while the loss of the approximation is noted as $\mathbf{U}^\perp\mathbf{D}^\perp\mathbf{V}^{\perp\prime}$; namely

$$\mathbf{X} = \left[\mathbf{U}_r, \mathbf{U}^\perp\right] \begin{bmatrix} \mathbf{D}_r & \\ & \mathbf{D}^\perp \end{bmatrix} \begin{bmatrix} \mathbf{V}_r' \\ \mathbf{V}^{\perp\prime} \end{bmatrix} = \mathbf{U}_r\mathbf{D}_r\mathbf{V}_r' + \mathbf{U}^\perp\mathbf{D}^\perp\mathbf{V}^{\perp\prime}. \tag{1.35}$$

Noting $\mathbf{U}^\perp\mathbf{D}^\perp\mathbf{V}^{\perp\prime}$ as $\mathbf{X}^\perp$, we have

$$\begin{aligned} \mathrm{tr}\mathbf{\Sigma}_\mathbf{X} &= \mathrm{tr}\mathbf{\Sigma}_{\mathbf{FA}'} + \mathrm{tr}\mathbf{\Sigma}_{\mathbf{X}^\perp} \\ &= \mathrm{tr}\mathbf{D}_r^2 + \mathrm{tr}\mathbf{D}^{\perp 2} \\ &= \mathrm{tr}\mathbf{D}^2. \end{aligned} \tag{1.36}$$

Therefore, the total variance of the $p$ variables in $\mathbf{X}$ is equal to the sum of the squared singular values of $\mathbf{X}$, and the amount of variance explained by the $r$ principal components is equal to $\mathrm{tr}\mathbf{D}_r$, sum of the squared $r$ largest singular values. *Variance explained ratio* is thus defined as

$$\frac{\mathrm{tr}\mathbf{D}_r^2}{\mathrm{tr}\mathbf{D}^2} \tag{1.37}$$

and it is used to determine the value of $r$, how many principal components should be used.

A PCA's result is commonly interpreted as follows; a component loading matrix is referred at first in order to interpret what the components in its column represent, by abstracting the variables that highly load on the components. The interpretation would be easier if the correspondence between each of the components and the corresponding set of the variables is easily captured. For example, suppose that a loading matrix has a similar structure shown in the left part of Figure 1.3, compared with the one in the right. The former structure is called *simple structure*, and known as a desired property that solution matrices should be possessed, originally proposed in the context of factor analysis. In particular, in the simple structure in Table 1.3 one can abstract the meaning of the first and the second variable to interpret the first component, that would be straightforward. In the not-simple case, on the contrary, all variables are matched to the first component while nothing to the third component,

TABLE 1.3: Examples of a simple structure (left) and not-simple structure (right). $*$ stands for a element whose absolute value is sufficiently large, and blanc cells are elements close to 0.

| | Comp.1 | Comp.2 | Comp.3 |
|---|---|---|---|
| Var.1 | * | | |
| Var.2 | * | | |
| Var.3 | | * | |
| Var.4 | | * | |
| Var.5 | | | * |
| Var.6 | | | * |

| | Comp.1 | Comp.2 | Comp.3 |
|---|---|---|---|
| Var.1 | * | | |
| Var.2 | * | * | |
| Var.3 | * | * | |
| Var.4 | * | | |
| Var.5 | * | * | |
| Var.6 | * | * | |

and it would be difficult to interpret these components.

The challenge here is to obtain a loading matrix with simple structure. It is known that PCA has a property called *rotational indeterminacy*; a certain post-hoc transformation of a component loading matrix is allowed without losing its optimality with respect to the loss function (1.18). Here, (1.18) is rewritten as follows, using an arbitrary $r \times r$ orthonormal matrix $\mathbf{T}$ satisfying $\mathbf{T}'\mathbf{T} = \mathbf{T}\mathbf{T}' = \mathbf{I}_r$

$$L_{PCA}(\mathbf{F}, \mathbf{A}) = ||\mathbf{X} - \mathbf{F}\mathbf{A}'||^2 = ||\mathbf{X} - \mathbf{F}\mathbf{T}(\mathbf{A}\mathbf{T})'||^2 = L_{PCA}(\tilde{\mathbf{F}}, \tilde{\mathbf{A}}). \tag{1.38}$$

It leads that $\{\mathbf{F}, \mathbf{A}\}$ and $\{\mathbf{F}\mathbf{T}, \mathbf{A}\mathbf{T}\}$ are equivalent, in that the latter does not change the function value $L_{PCA}(\mathbf{F}, \mathbf{A})$. Note that left-multiplying $\mathbf{A}$ by $\mathbf{T}$ stands for a rotation of the space spanned $\mathbf{A}$'s column vectors, and it is called *orthogonal rotation* of $\mathbf{A}$ when $\mathbf{T}$ is an orthonormal matrix. Using this property, $\mathbf{T}$ can be determined so that it maximizes the simplicity of $\mathbf{A}\mathbf{T}$. According to how the simplicity of a matrix is defined, there exist various methods for rotation. For example, in Varimax rotation (Kaiser, 1958), a matrix $\mathbf{A}$ is said to be simple if the sum of the column variance of the squared matrix $\mathbf{A}^2 = \{a_{jk}^2\}$, and $\mathbf{T}$ is specified so that it maximizes the simplicity of $\mathbf{A}\mathbf{T}$.

**Example**

Thirty-three baseball players' batting average, number of home runs, hitting point, score, number of strike outs, and number of steals were collected, and thus we have a 33 (players) $\times$ 6 (scores) data matrix. First, variance explained defined in (1.37) for $r = 1$ to 6 were computed, and they were plotted in Figure 1.2. The plot is called *scree plot*, and used to determine how many principal components should be used for explaining the total information that a data matrix contains. In the figure, we can see that the cumulative variance explained does not changes significantly when $r \geq 3$. The point, $r = 3$, is called a *scree point*, and often used for the number of principal components. At the scree point, 94.1% of the variance of $\mathbf{X}$ is explained, and therefore $r$ is set at 3.

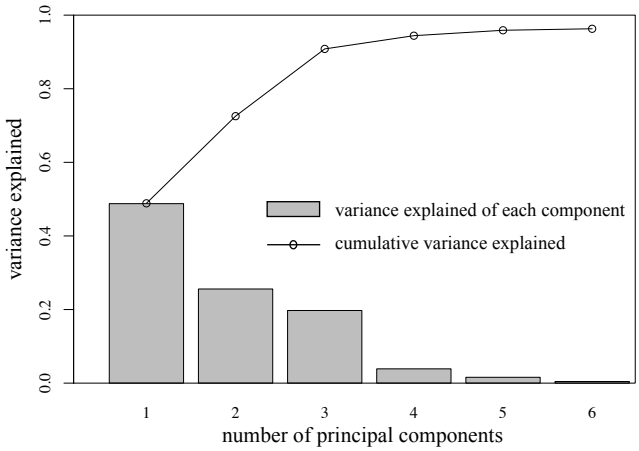



FIGURE 1.2: Scree plot of variance explained for baseball data.

TABLE 1.4: Initial and Varimax-rotated loading matrix for baseball data.

| | Initial | | | Varimax-rotated | | |
|---|---|---|---|---|---|---|
| | Comp.1 | Comp.2 | Comp.3 | Comp.1 | Comp.2 | Comp.3 |
| batting ave. | **0.429** | **-0.650** | **0.524** | 0.330 | -0.151 | **0.866** |
| home runs | **0.957** | 0.051 | -0.040 | **0.957** | -0.041 | 0.039 |
| hitting point | **0.965** | 0.059 | 0.084 | **0.961** | 0.048 | 0.126 |
| steals | **-0.481** | **-0.610** | **-0.571** | **-0.521** | **-0.809** | -0.068 |
| score | **0.543** | **-0.680** | **-0.407** | **0.481** | **-0.805** | 0.210 |
| strike outs | **0.527** | **0.472** | **-0.609** | **0.603** | -0.090 | **-0.708** |
| steals | **-0.481** | **-0.610** | **-0.571** | **-0.521** | **-0.809** | -0.068 |

The resulting loading matrix is shown in Table 1.4. A loading matrix is used for interpreting what the principal components stand for, and the rotated one is generally easier to interpret in that it approximates a certain simple structure. It is observed in the table that the contrast between the elements facilitates to capture the correspondence between the variables and the components. For example, "home runs", "hitting point", and "strike outs" highly load on the first principal component, and thus the component is interpreted as *batting power*. In the same manner, the second component matched to "score" and "steals" stands for *weakness in running*, and the third component that positively loads on "batting average" and negatively loads on "strike" outs is *batting accuracy*.

### Related Methods

In the above example, the three-dimensional solution of PCA was presented. *Biplot* is a graphical representation of a two-dimensional solution of PCA and it enables to interpret the result easily and intuitively (Gower and Hand, 1995). In order to obtain a biplot, PCA with $r = 2$ is applied to a given data matrix, followed by plotting the rows of $\mathbf{F}$ as points and those of $\mathbf{A}$ as vectors, as shown in Figure 1.3. The Biplot

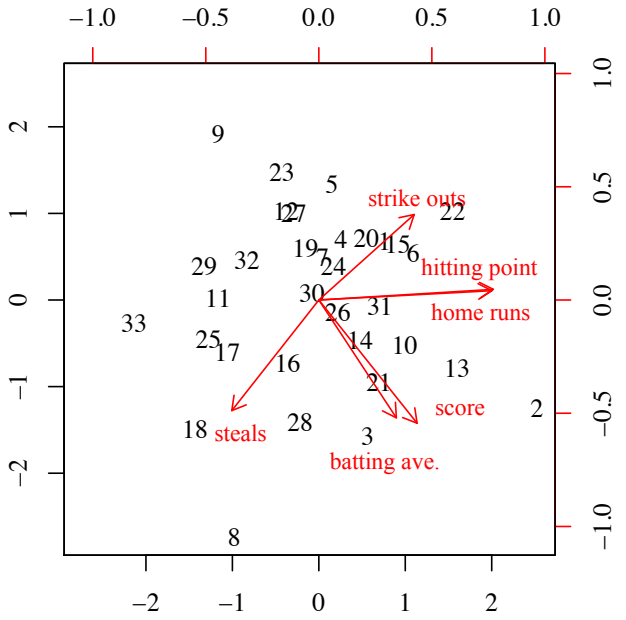



FIGURE 1.3: Biplot representation of baseball data.

of $\mathbf{X}$ is thus composed of thirty-three points and six vectors corresponding to the observations and variables, respectively.

Here, noting that $\mathbf{X}$ is column-wise centered, we have the following approximation;

$$\mathbf{\Sigma_X} = n^{-1}\mathbf{X}'\mathbf{X} \approx \mathbf{A}\mathbf{A}'. \tag{1.39}$$

It indicates that the $(j,k)$-th element of $\mathbf{\Sigma_X}$, the covariance of the $j$-th and $k$-th variable $(j \neq q)$, is approximated by the inner product of the $j$-th and $k$-th row vector of $\mathbf{A}$, $\mathbf{a}_{(j)}\mathbf{a}'_{(k)}$. The cosine of the vectors for these variables, expressed as $\frac{\mathbf{a}_{(j)}\mathbf{a}'_{(k)}}{||\mathbf{a}_{(j)}|| \times ||\mathbf{a}'_{(k)}||}$, is therefore approximates the correlation coefficient of the two variables. For instance, the vectors of "batting average" and "score" are almost in same direction, which indicates the cosine of these vectors is close to its maximum 1 and thus these variables are highly correlated. Conversely, the cosine of "score"'s and "strike out"'s vectors are around 0, and thus these variables are considered to be independent. In addition, the point numbered as 3 is higher in "hitting point" and "scores" compared with the point 30, because the direction of a vector is interpreted as the axis of the variable. It should be noted that a biplot is valid only when a two dimensional solution fits well to the original data matrix.

### 1.2.3 Factor Analysis

*Factor analysis (FA)* (Harman, 1976; Mulaik, 2010; Bartholomew et al., 2011) is known to be similar but completely different from PCA, and it aims to extract latent variables hidden in multivariate data matrix. The latent variables are called *common factors*, and hereafter they are simply called *factors*. As a model case, consider the

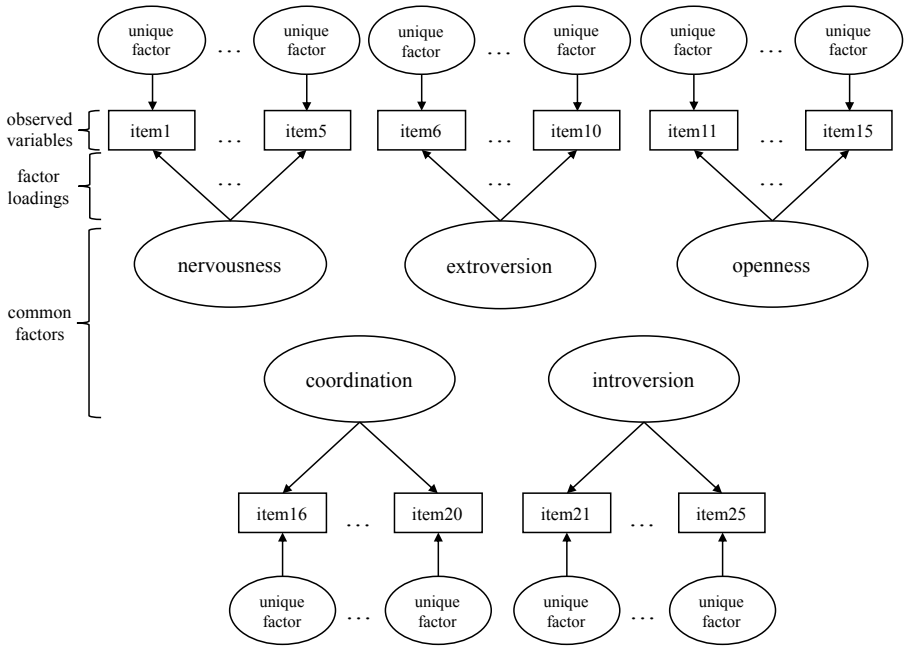



FIGURE 1.4: Path diagram of factor analysis result.

situation where FA is applied to a data matrix of Big5 personality test scores. Big5 personality test is developed to describe human personality based on the following five components: nervousness, extroversion, openness, coordination, and introversion. Each component is composed of multiple sub-scales called items. Figure 1.4 demonstrates a result of the FA to the Big5 test scores with five common factors as a path diagram. The extracted factors and variables (questionnaires, in this example) are connected by the paths, and they are obtained as a form of a matrix called the *factor loading matrix*. In addition, *unique factors* are also obtained, each of which corresponds to a single variable. As seen here, the FA explains the information that the data matrix contains using common and unique factors.

### Probabilistic Model

As the most popular formulation for FA, a probabilistic formulation of FA is firstly presented. Let $\mathbf{x}$ be a $p$-dimensional vector of the observed values for an observation. A FA model with $r$ factors is expressed as the following model, using an $r$-dimensional vector of the common factor score $\mathbf{f}$ and a $p$-dimensional vector of unique factor score $\mathbf{u}$.

$$\mathbf{x} \approx \mathbf{\Lambda f} + \mathbf{u} \tag{1.40}$$

where $\mathbf{\Lambda}$ is an unknown $p \times r$ matrix called a *factor loading matrix*. Noting the common and unique factor score vectors for the $i$-th observation as $\mathbf{f}_{(i)}$ and $\mathbf{u}_{(i)}$, respectively, (1.40) is rewritten as

$$\mathbf{X} \approx \mathbf{F\Lambda}' + \mathbf{U} \tag{1.41}$$

where $\mathbf{F} = \{\mathbf{f}'_{(1)}, \cdots, \mathbf{f}'_{(n)}\}'$ and $\mathbf{U} = \{\mathbf{u}'_{(1)}, \cdots, \mathbf{u}'_{(n)}\}'$. Here, it is assumed that $\mathbf{f}_{(i)}$ and $\mathbf{u}_{(i)}$ are random variables, which identically and independently follows the following distributions;

$$\mathbf{f}_{(i)} \sim N_r(\mathbf{0}_r, \mathbf{I}_r),\ \mathbf{u}_{(i)} \sim N_r(\mathbf{0}_p, \mathbf{\Psi})\ (i = 1, \cdots, n) \tag{1.42}$$

where $\mathbf{0}_r$ is the $r$-dimensional vector filled with 0s, and $\mathbf{\Psi} = \mathrm{diag}(\psi_1, \cdots, \psi_p)$ is the diagonal matrix of the variances of $p$ unique factors. This formulation of FA is thus called a probabilistic FA model. The covariance matrix of $\mathbf{f}$ is constrained to be $\mathbf{I}_r$, indicating that each factor has a unit variance and is uncorrelated to the others. It is expressed as

$$\mathrm{Cov}[\mathbf{f}, \mathbf{f}] = \mathbf{I}_r. \tag{1.43}$$

Further, common and unique factor scores are assumed to be independent.

$$\mathrm{Cov}[\mathbf{f}, \mathbf{u}] = \mathbf{O}_{r \times p}. \tag{1.44}$$

In a probabilistic FA model, different from the other multivariate analysis methods introduced so far, the model parameters $\mathbf{f}$ and $\mathbf{u}$ are treated as random variables. $\mathbf{x}$ is also a random variable following the model in (1.40) and $\mathbf{x}_{(1)}, \cdots, \mathbf{x}_{(n)}$ are identically and independently follow a multivariate normal distribution with mean vector $\mathbf{0}_p$, and the covariance matrix that is modeled as

$$\mathrm{Var}[\mathbf{x}] \approx \mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}. \tag{1.45}$$

Given a data matrix $\mathbf{X}$, one can evaluate the discrepancy between the observed covariance matrix $\mathbf{\Sigma}_\mathbf{X}$ and its theoretical counterpart derived by (1.40) and (1.45) by the following least squares loss function of $\mathbf{\Lambda}$ and $\mathbf{\Psi}$

$$||\mathbf{\Sigma}_\mathbf{X} - (\mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi})||^2. \tag{1.46}$$

In the least squares estimation under the probabilistic FA model, the goal is to find $\{\mathbf{\Lambda},\ \mathbf{\Psi}\}$ minimizing the loss function. Alternatively, one can define the following log-likelihood function

$$-\frac{n}{2}(\log|\mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}| + \mathrm{tr}\mathbf{\Sigma}_\mathbf{X}(\mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi})^{-1}) \tag{1.47}$$

and find its maximizer with respect to $\mathbf{\Lambda}$ and $\mathbf{\Sigma}_\mathbf{X}$. See Yanai and Ichikawa (2006) for the detailed procedures of the least squares and maximum likelihood estimation.

### Matrix Formulation of FA Model and Its Parameter Estimation

In the probabilistic formulation of FA, matrices $\mathbf{F}$ and $\mathbf{U}$ are not directly obtained by the minimization of (1.46) or maximization of (1.47), because $\mathbf{f}$ and $\mathbf{u}$ are treated as random variables. They are usually estimated by post-hoc estimation procedures as those found in Bartlett (1950), Anderson and Rubin (1956), and McDonald (1981).

This property is considered a critical drawback of FA for the following reasons. For example, in the case with the Big5 personality test, one might attempt to quantify the five personalty traits by the factor scores, and they classify the respondents as to the scores. Although a post-hoc estimation of factor scores provides the scores, optimization criteria considered in the estimation is not identical to the one in the parameter estimation of FA. In other words, joint usage of FA and a post-hoc estimation of factor scores does not satisfy the definition of data analysis in (1.1.1) because both independently optimize two different criteria.

In recent years, a matrix formulation of FA has been proposed (Adachi and Trendafilov, 2018a; De Leeuw, 2004; Unkel and Trendafilov, 2010), and it assumes that $\mathbf{f}$ and $\mathbf{u}$ are not random but fixed parameters, and thus, it enables us to estimate $\mathbf{F}$ and $\mathbf{U}$ simultaneously. This is called the *matrix decomposition factor analysis (MDFA)*. The MDFA model is expressed as

$$\mathbf{X} \approx \mathbf{F}\boldsymbol{\Lambda}' + \mathbf{U}\boldsymbol{\Psi} = \mathbf{Z}\mathbf{B}' \tag{1.48}$$

with $\mathbf{Z} = [\mathbf{F}, \mathbf{U}]$ and $\mathbf{B} = [\boldsymbol{\Lambda}, \boldsymbol{\Psi}]$. There, the following two constraints are imposed;

$$\mathbf{1}_N'\mathbf{Z} = \mathbf{0}_{r+p}' \tag{1.49}$$

and

$$n^{-1}\mathbf{Z}'\mathbf{Z} = \begin{bmatrix} n^{-1}\mathbf{F}'\mathbf{F} & n^{-1}\mathbf{F}'\mathbf{U} \\ n^{-1}\mathbf{U}'\mathbf{F} & n^{-1}\mathbf{U}'\mathbf{U} \end{bmatrix} = \begin{bmatrix} \mathbf{I}_r & \mathbf{O}_{r\times p} \\ \mathbf{O}_{p\times r} & \boldsymbol{\Psi} \end{bmatrix} \tag{1.50}$$

and they are equivalent to those in the probabilistic model in (1.43) and (1.44). For parameter estimation, the least squares criterion

$$L_{FA}(\mathbf{Z}, \mathbf{B}) = ||\mathbf{X} - \mathbf{Z}\mathbf{B}'||^2 = ||\mathbf{X} - (\mathbf{F}\boldsymbol{\Lambda} + \mathbf{U}\boldsymbol{\Psi})||^2 \tag{1.51}$$

is minimized under (1.49) and (1.50).

The minimization of $L_{FA}(\mathbf{Z}, \mathbf{B})$ is accomplished by the following iterative algorithm (Adachi and Trendafilov, 2018a).

1. Initialize $\mathbf{B}$.

2. Update $\mathbf{Z}$ by

$$\mathbf{Z} = n^{1/2}\mathbf{K}\mathbf{L}' \tag{1.52}$$

   using the SVD

$$n^{-1/2}\mathbf{X}\mathbf{B} = \mathbf{K}\boldsymbol{\Delta}\mathbf{L}'. \tag{1.53}$$

3. Update $\mathbf{B}$ by

$$\mathbf{B} = [n^{-1}\mathbf{X}'\mathbf{F}, \mathrm{diag}(n^{-1}\mathbf{X}'\mathbf{U})] \tag{1.54}$$

4. Stop the algorithm if the decrement of $L_{FA}(\mathbf{Z}, \mathbf{B})$ is less than $\epsilon$, otherwise go back to 2.

Note that the algorithm for FA requires some iterative steps, and it is different from those employed in multivariate regression analysis and PCA, in which the parameters have closed-form solutions. This is because the update formulas (1.52) and (1.54) include other unknown parameter matrices. This indicates that there is no closed-form solution in FA, and the parameters must be estimated by iterating some steps where a parameter is updated with the other parameters fixed. This is called *Alternating least squares* (ALS), because parameter matrices are alternately updated. If the function value is guaranteed to decrease monotonically at each step and the function value is lower bounded (in a minimization problem), the values of the parameters after the convergence is reached reasonably accepted as the estimate of the parameters. Obviously, $L_{FA}(\mathbf{Z}, \mathbf{B})$ is lower bounded by 0. Further, note that if a function to be minimized is not a convex function, the estimated parameter is not guaranteed to be a global optima. To obtain a solution close enough to the global optima, a *multiple-starts strategy* is often used, wherein the algorithm starts from multiple initial values separately and the solution that minimizes the optimization criterion the most within the set of obtained solutions is accepted as the final solution.

**Some Properties**

It is known that there exists an indeterminacy in the MDFA model. This is called *rotational indeterminacy*. To express this, (1.40) is rewritten as

$$\mathbf{x} \approx \mathbf{\Lambda T T}'\mathbf{f} + \mathbf{u} = \tilde{\mathbf{\Lambda}}\tilde{\mathbf{f}} + \mathbf{u} \tag{1.55}$$

using an arbitrary $r \times r$ orthonormal matrix $\mathbf{T}$ satisfying

$$\mathbf{T}'\mathbf{T} = \mathbf{T}\mathbf{T}' = \mathbf{I}_r. \tag{1.56}$$

This implies that

$$\tilde{\mathbf{\Lambda}} = \mathbf{\Lambda T} \tag{1.57}$$

and

$$\tilde{\mathbf{F}} = \{\mathbf{f}'_{(1)}\mathbf{T}', \cdots, \mathbf{f}'_{(n)}\mathbf{T}'\}' = \mathbf{F}\mathbf{T} \tag{1.58}$$

are equally optimal as well as $\mathbf{\Lambda}$ and $\mathbf{F}$ because

$$||\mathbf{\Sigma_X} - (\tilde{\mathbf{\Lambda}}\tilde{\mathbf{\Lambda}}' + \mathbf{\Psi})||^2 = ||\mathbf{\Sigma_X} - (\mathbf{\Lambda T}'\mathbf{T}\mathbf{\Lambda}' + \mathbf{\Psi})||^2 = ||\mathbf{\Sigma_X} - (\mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi})||^2 \tag{1.59}$$

indicating the value of the optimization criterion is kept unchanged after the transformation by $\mathbf{T}$. It is shown that the property also holds in the maximum likelihood estimation

$$-\frac{n}{2}(\log|\tilde{\mathbf{\Lambda}}\tilde{\mathbf{\Lambda}}' + \mathbf{\Psi}| + \mathrm{tr}\mathbf{\Sigma_X}(\tilde{\mathbf{\Lambda}}\tilde{\mathbf{\Lambda}}' + \mathbf{\Psi})^{-1}) = -\frac{n}{2}(\log|\mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}| + \mathrm{tr}\mathbf{\Sigma_X}(\mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi})^{-1}) \tag{1.60}$$

and in the matrix formulation

$$L_{FA}(\tilde{\mathbf{Z}}, \tilde{\mathbf{B}}) = ||\mathbf{X} - (\mathbf{F}\mathbf{T}'\mathbf{T}\mathbf{\Lambda}' + \mathbf{U}\mathbf{\Psi})||^2 = L_{FA}(\mathbf{Z}, \mathbf{B}). \tag{1.61}$$

Further, using an arbitrary $r \times r$ matrix $\mathbf{S}$ satisfying

$$\mathrm{diag}(\mathbf{S}^{-1}\mathbf{S}^{-1\prime}) = \mathbf{I}_r \tag{1.62}$$

that includes (1.56) as a special case, the same property holds if we transform $\mathbf{F}$ and $\lambda$ as

$$\tilde{\mathbf{F}} = \mathbf{F}\mathbf{S}^{-1\prime}, \quad \tilde{\mathbf{\Lambda}} = \mathbf{\Lambda}\mathbf{S}. \tag{1.63}$$

We thus have

$$L_{FA}(\tilde{\mathbf{Z}}, \tilde{\mathbf{B}}) = ||\mathbf{X} - (\mathbf{F}\mathbf{S}'^{-1}\mathbf{S}'\mathbf{\Lambda}' + \mathbf{U}\mathbf{\Psi})||^2 = L_{FA}(\mathbf{Z}, \mathbf{B}). \tag{1.64}$$

The transformation by $\mathbf{T}$ is called an *orthogonal rotation* and the one by $\mathbf{S}$ is called an *oblique rotation*. Note that after the transformation by $\mathbf{S}$, we have

$$n^{-1}\tilde{\mathbf{F}}'\tilde{\mathbf{F}} = \mathbf{S}^{-1}\mathbf{S}'^{-1} \neq \mathbf{I}_r \tag{1.65}$$

and thus, the rotated factors are correlated.

Using the above properties, we can freely choose $\mathbf{T}$ or $\mathbf{S}$ and rotate $\mathbf{F}$ and $\mathbf{\Lambda}$. It is common to choose a rotation matrix such that the resulting $\tilde{\mathbf{\Lambda}}$ has a simple structure as introduced before, which is called *factor rotation*. Procedures for factor rotation had been researched for decades in Psychometrics, and a variety of rotation procedures have been proposed. The details of the factor rotation are discussed in Chapter 2.

**Example**

An example of FA is presented using an artificial dataset called Personality data. One hundred participants evaluated their daily behaviors with respect to the following eight aspects from 1 (not applicable at all) to 9 (completely applicable); "actively do anything (active)," "being cheerful (cheerful)," "leading others (leading)," "being unfriendly (unfriendly)," "being talkative (talkative)," "being motivated (motivated)," "being hesitative (hesitative)," and "being popular (popular)." A two-factors FA model was fitted to the data matrix that consists of 100 participants and 8 questionnaires. To compare the three formulations of FA, [1] the probabilistic model with the least squares estimation, [2] the probabilistic model with the maximum likelihood estimation, and [3] the matrix formulation were employed. Further, Varimax and Promax rotations were applied to the initial loading matrices. The results of FA are shown in Table 1.5. Compared with the unrotated matrices, it is clear that the rotated loading matrices are more interpretable in that variables (rows) and factors (columns) are clearly matched. The Promax-rotated loading matrix indicates that the first factor

TABLE 1.5: Factor loading matrices for personality data. MDFA: initial solution of MDFA, MDFA+Varimax: Varimax-rotated MDFA solution, MDFA+Promax: Promax-rotated MDFA solution, ML+Promax: Promax-rotated solution by probabilistic formulation based FA with maximum likelihood, LS+Promax: Promax-rotated solution by probabilistic formulation based FA with least squares.

| | MDFA | | MDFA+Varimax | | MDFA+Promax | | ML+Promax | | LS+Promax | |
|---|---|---|---|---|---|---|---|---|---|---|
| active | **0.824** | **0.241** | **0.814** | **0.273** | **0.827** | 0.058 | 0.060 | **-0.826** | **0.836** | 0.058 |
| cheerful | 0.107 | **0.776** | 0.076 | **0.779** | -0.151 | **0.848** | **0.852** | 0.153 | -0.152 | **0.854** |
| leading | **0.734** | 0.193 | **0.726** | **0.222** | **0.744** | 0.028 | 0.022 | **-0.754** | 0.743 | 0.033 |
| unfriendly | **-0.274** | **-0.785** | **-0.243** | **-0.795** | -0.030 | **-0.816** | **-0.822** | 0.028 | -0.035 | **-0.816** |
| talkative | **0.287** | **0.853** | **0.253** | **0.864** | 0.022 | **0.889** | **0.894** | -0.023 | 0.019 | **0.893** |
| motivated | **0.858** | 0.190 | **0.850** | **0.224** | **0.883** | -0.008 | -0.002 | **-0.882** | **0.894** | -0.014 |
| hesitative | **-0.770** | -0.085 | **-0.766** | -0.115 | **-0.822** | 0.103 | 0.104 | **0.829** | **-0.818** | 0.102 |
| popular | **0.390** | **0.611** | **0.366** | **0.626** | **0.220** | **0.589** | **0.587** | **-0.225** | **0.218** | **0.594** |
| factor correlation | 1.000 | 0.000 | 1.000 | 0.000 | 1.000 | 0.505 | 1.000 | -0.504 | 1.000 | 0.506 |
| | 0.000 | 1.000 | 0.000 | 1.000 | 0.505 | 1.000 | -0.504 | 1.000 | 0.506 | 1.000 |

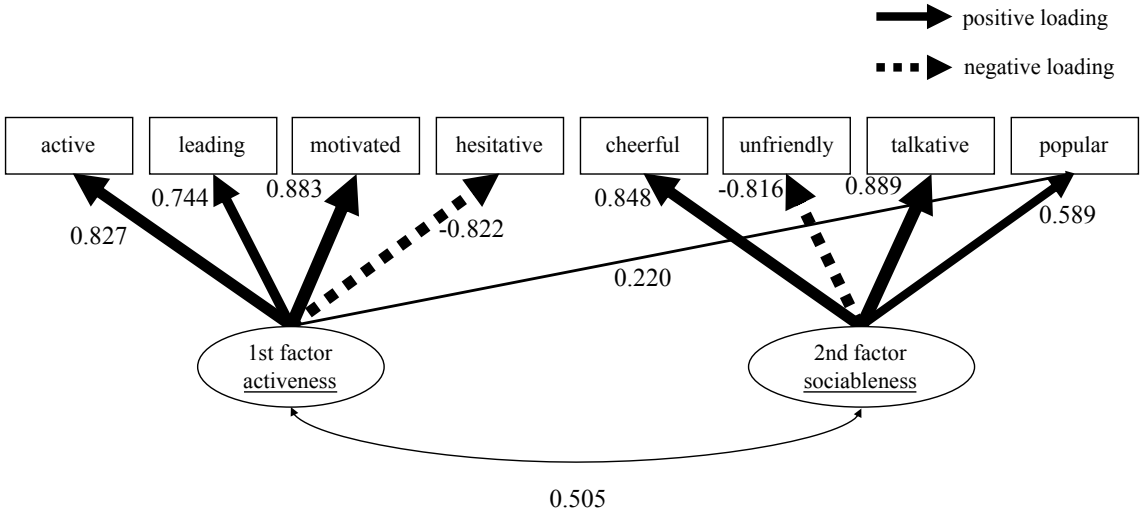



FIGURE 1.5: Path diagram of Promax-rotated MDFA solution to personality data. Width of the paths are proportional to the corresponding loadings, and those for loading less than 0.2 were omitted.

is matched to the variables "active," "leading," "motivated," and "hesitative" and interpreted as *activeness*, while the second factor is *sociableness* in that it is matched to "cheerful," "unfriendly," "talkative," and "popular." The interpretation is consistent with the path diagram shown in Figure 1.5. In general, oblique rotation tends to achieve simper results than the orthogonal one because the rotational constraint in the former in (1.62) is less restrictive than (1.56). In addition, the loading matrices after rotation are almost equivalent for the three formulations, and thus, all formulations result in similar solutions even if their assumptions on the factor scores and estimation procedures are different.

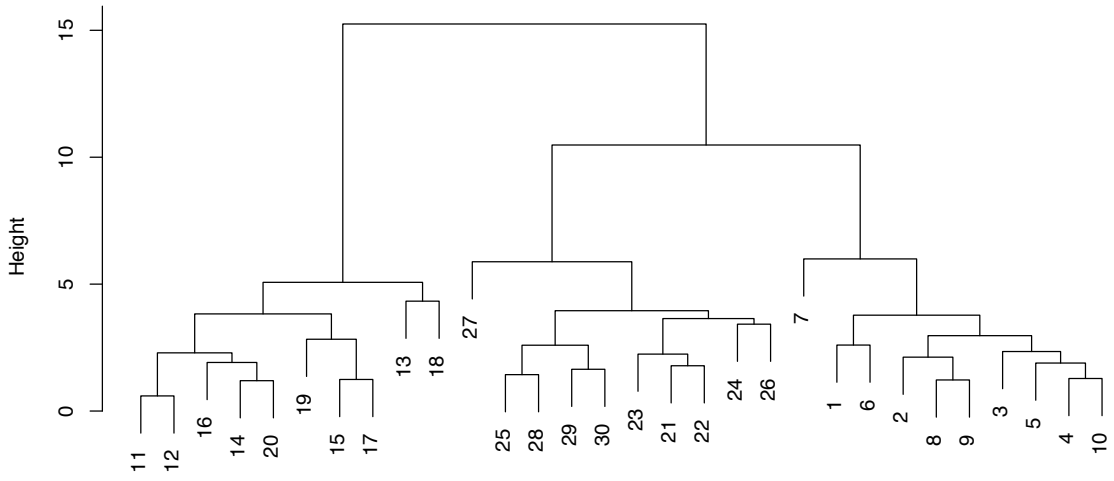



FIGURE 1.6: Example of dendrogram

### 1.2.4 $K$-means Clustering

Classification is often employed at the early stage of EDA and helps to grasp homogeneity of observations, which serves to understand data structure roughly and intuitively. *Clustering* is a class of multivariate analysis procedures for classification and aims to classify multiple observations into reduced the number of groups called *clusters*. There exist a number of clustering procedures and they are classified into the following two families. The first one is *hierarchical clustering*. It creates clusters by merging similar observations or clusters according to a certain aggregation rule. In other words, the number of clusters is equal to the number of observations at the initial step, and it is reduced sequentially by repeating cluster agglomerations. Group-average method and Ward's method are known as such procedures (Ward, 1963). A dendrogram shown in Figure 1.6 is provided by a hierarchical clustering applied to an artificial data. Based on the figure, one might decide that three clusters are sufficient. Note that hierarchical clustering yields nested solutions; a $K$-clusters solution includes another $(K+1)$-clusters solution.

The second family is *non-hierarchical clustering* represented by $K$-means clustering (Aggarwal, 2014). $K$-means clustering requires the number of clusters as its input, and seeks a partition of observations that minimizes a certain loss function. Thus, two solutions of $K$-means clustering with different $K$s are not nested. Its algorithm is detailed below.

#### Formulation

Consider the case with the $n$ observations in a $n \times p$ data matrix $\mathbf{X}$ are classified into $K(< n)$ clusters by $K$-means clustering. It is formally defined as the minimization of

the least squares loss function

$$L_{KM}(\mathbf{M}, \mathbf{C}) = ||\mathbf{X} - \mathbf{M}\mathbf{C}'||^2 \tag{1.66}$$

over two unknown matrices $\mathbf{M}(n \times K)$ and $\mathbf{C}(p \times K)$. $\mathbf{M} = \{m_{ik}\}$ is called a *membership matrix* satisfying $\sum_k m_{ik} = 1$ for all $k$s and $m_{ik} \in \{0, 1\}$ and $m_{ik} = 1$ stands for the $i$-th observation is a member of the $k$-cluster. Because its row-sum is equal to 1, each observation always belongs to a single cluster and this type of partitioning is called *crisp partition*. $\mathbf{C} = \{\mathbf{c}_1, \cdots, \mathbf{c}_K\}$ is called a *centroid matrix*, and its $k$-th column stands for the averaged vector of the observations that belong to the $k$-th cluster, as shown later. The matrix therefore represents the centroids of the $K$ clusters in $p$-dimensional space.

**Parameter Estimation**

First, consider to minimize $L_{KM}(\mathbf{M}, \mathbf{C})$ with respect to $\mathbf{M}$ with $\mathbf{C}$ kept fixed. The objective function can be rewritten as

$$L_{KM}(\mathbf{M}|\mathbf{C}) = \sum_{i}^{n} \sum_{k}^{K} m_{ik} ||\mathbf{x}_{(i)} - \mathbf{c}_k||^2 \tag{1.67}$$

that is the sum of $\sum_k m_{ik}||\mathbf{x}_{(i)} - \mathbf{c}_k||^2$ for all observations. This indicates that $\mathbf{x}_{(i)}$ is approximated by $\mathbf{c}_k$ if the $i$-th observation belongs to the $k$-th cluster. An observation is therefore classified into a cluster minimizing $||\mathbf{x}_{(i)} - \mathbf{c}_k||$ for all $k$s, and it is accomplished by assigning the observation to a cluster whose centroid vector is closest to $\mathbf{x}_{(i)}$ in $p$-dimensional space. The elements of $\mathbf{M}$ is thus updated by

$$m_{ik} = \begin{cases} 1 & if\ k = \underset{l}{\operatorname{argmin}} ||\mathbf{x}_{(i)} - \mathbf{c}_l|| \\ 0 & otherwise \end{cases} . \tag{1.68}$$

It is easily verified that the updated $\mathbf{M}$ satisfies a series of constraints on it. Notice that the update rule in (1.68) requires the $K$-times computations for each of the observations in order to specify the cluster that is closest to it, and thus it is computationally inefficient when $K$ is large.

Next, $L_{KM}(\mathbf{M}, \mathbf{C})$ is minimized over $\mathbf{C}$ with fixed $\mathbf{M}$. This is simply accomplished by using the same manner as in multivariate regression, assuming that $\mathbf{X}$ is regressed on $\mathbf{M}$ with an unknown matrix of regression coefficients $\mathbf{C}$. The update formula for $\mathbf{C}$ is thus obtained as

$$\mathbf{C} = \mathbf{X}'\mathbf{M}(\mathbf{M}'\mathbf{M})^{-1}. \tag{1.69}$$

Noting that $\mathbf{M}'\mathbf{M}$ is the $K \times K$ diagonal matrix having the size of the $k$-th cluster as its $k$-th diagonal element, the $k$-th column of $\mathbf{C}$ is expressed as $n_k^{-1}\mathbf{X}'\mathbf{m}_k$, where $\mathbf{m}_k$

TABLE 1.6: Centroid matrix obtained by $K$-means clustering.

| | 1st Cluster | 2nd Cluster | 3rd cluster |
|---|---|---|---|
| Nervous | -0.492 | -0.809 | 1.300 |
| Extroversion | -0.330 | -0.747 | 1.077 |
| Openness | -1.258 | 0.568 | 0.690 |
| Cooperative | 1.196 | -1.109 | -0.087 |
| Honesty | -0.654 | 1.161 | -0.507 |

denotes the $k$-th column vector of $\mathbf{M}$. This immediately leads that the columns of $\mathbf{C}$ is regarded as the centroid vectors of the $K$ clusters.

The two parameter matrices $\mathbf{M}$ and $\mathbf{C}$ are iteratively updated starting from some initial values, in the same manner as in FA; the objective function is minimized over the one parameter with the others kept fixed, and it is repeated until convergence. The whole algorithm is described as follows;

1. Randomly initialize $\mathbf{M}$.
2. Update $\mathbf{C}$ by (1.69) using the current $\mathbf{M}$.
3. Update $\mathbf{M}$ by (1.68) using the current $\mathbf{C}$.
4. Terminate the algorithm if the decrement of the function value is less than $\epsilon$, otherwise go back to 2.

**Example**

Thirty participants were required to score their personality traits in terms of the following five scales; "nervous", "extroversion", "openness", "cooperative", and "honesty". Note that the dataset treated here is artificial. A $K$-means clustering with three clusters was applied to the data matrix having 30 observations and 5 variables. In order to avoid local minima, a hundred of initial values were employed. As a result, three clusters consist of ten observations were obtained, and the estimated centroid matrix is shown in Table 1.6. As discussed above, column vectors of a centroid matrix are seen as representative observations of the obtained clusters. In other words, it serves to interpret what kind of homogeneity were extracted from the dataset. In the example, one might notice that the first cluster is less *nervous* and much *cooperative* than the other clusters, which is considered as a common characteristic of the observations in the cluster. Often, a centroid matrix is interpreted by finding correspondences between rows and columns, as the same way in PCA and FA. Therefore, it would be beneficial for interpretation that a centroid matrix also possesses a simple structure. $K$-means clustering has, however, no indeterminacy but column permutations of $\mathbf{M}$, and it means that no post-hoc transformation toward simplicity is not allowed.

### Related Method

As noted above, $K$-means clustering provides a crisp partition, in which observations are classified into non-over-wrapping clusters. As a natural extension, *fuzzy $K$-means clustering* has been proposed Zadeh (1965); Bezdek (2013). There, observations are classified into $K$ over-wrapping clusters; an observation is allowed to belong to multiple clusters. In other words, the observations are fuzzily classified. It is achieved by relaxing the constraints on the membership matrix $\mathbf{M}$, such that

$$\sum_{k}^{K} m_{ik} = 1,\ m_{ij} \in [0, 1]. \tag{1.70}$$

Therefore, $m_{ik}$ is interpreted as the probability that the $i$-th observation belongs to the $K$-th cluster. The objective function of fuzzy $K$-means clustering is

$$L_{FCM}(\mathbf{M}, \mathbf{C}) = \sum_{i} \sum_{k} m_{ik}^{q} ||\mathbf{x}_{(i)} - \mathbf{c}_k||^2 \tag{1.71}$$

where $q \geq 1$ is called a *fuzziness parameter* that controls how fuzzy the resulting clusters are. (1.71) is reduced to the one of $K$-means clustering when $q = 1$. The optimization procedure for minimizing (1.71) is detailed in Chapter 4.

## 1.2.5 Canonical Correlation Analysis

The unsupervised learning methods introduced so far (PCA, FA, and $K$-means clustering) require a single set of variables as their input. Often, however, dataset is composed of two or more variable sets. *Canonical correlation analysis* (CANO) can deal with two sets of variables in order to summarize and extract the relation between the sets (Hotelling, 1936). Specifically, it composes linear combinations for each of the variable sets so that the correlation between the combinations is maximized, and the linear combinations are called *canonical variates*.

### Formulation

Suppose we have two sets of variables with their sizes of $p_1$ and $p_2$, respectively, and let $\mathbf{X}_1(n \times p_1)$ and $\mathbf{X}_2(n \times p_2)$ be matrices of these variable sets with regard to the same $n$ observations. CANO is formulated by minimizing

$$L_{CA}(\mathbf{W}_1, \mathbf{W}_2) = ||\mathbf{X}_1\mathbf{W}_1 - \mathbf{X}_2\mathbf{W}_2||^2 \tag{1.72}$$

over unknown weight matrices $\mathbf{W}_1(p_1 \times r)$ and $\mathbf{W}_2(p_2 \times r)$ of two variable sets, with $r \leq \min(p_1, p_2)$. In order to avoid the trivial solution $\{\mathbf{W}_1, \mathbf{W}_2\} = \{\mathbf{O}_{p_1 \times r}, \mathbf{O}_{p_1 \times r}\}$, the constraints

$$n^{-1}\mathbf{W}_1'\mathbf{X}_1'\mathbf{X}_1\mathbf{A}_1 = n^{-1}\mathbf{W}_2'\mathbf{X}_2'\mathbf{X}_2\mathbf{A}_2 = \mathbf{I}_r. \tag{1.73}$$

are often imposed on the weight matrices. CANO creates two sets of $r$ canonical variates corresponding to two given variable sets in order to explain the correlation structure between them.

The loss function in (1.72) is expanded as

$$L_{CA}(\mathbf{W}_1, \mathbf{W}_2) = -2\mathrm{tr}\mathbf{W}_2'\mathbf{X}_2'\mathbf{X}_1\mathbf{W}_1 + 2n \propto -\mathrm{tr}(n^{-1}\mathbf{W}_2'\mathbf{X}_2'\mathbf{X}_1\mathbf{W}_1) \tag{1.74}$$

and thus the minimization of $L_{CA}(\mathbf{W}_1, \mathbf{W}_2)$ is proved to be equivalent to maximizing $\mathrm{tr}(n^{-1}\mathbf{W}_2'\mathbf{X}_2'\mathbf{X}_1\mathbf{W}_1)$, the sum of the correlation coefficients between the two canonical variates called *canonical correlations*. Therefore, CANO is regarded as a procedure that construct two sets of canonical variates so that the sum of their inter-sets correlations is maximized.

**Parameter Estimation**

CANO has a closed-form solution for $\mathbf{W}_1$ and $\mathbf{W}_2$ unlike FA and $K$-means clustering as shown below. Here, consider to maximize $\tilde{L}_{CA}(\mathbf{W}_1, \mathbf{W}_2) = \mathrm{tr}(n^{-1}\mathbf{W}_2'\mathbf{X}_2'\mathbf{X}_1\mathbf{W}_1)$ under (1.73). For the maximization, $\mathbf{W}_1$ and $\mathbf{W}_2$ are reparametrized as

$$n^{-1/2}(\mathbf{X}_1'\mathbf{X}_1)^{1/2}\mathbf{W}_1 = \mathbf{W}_1^*,\ n^{-1/2}(\mathbf{X}_2'\mathbf{X}_2)^{1/2}\mathbf{W}_2 = \mathbf{W}_2^* \tag{1.75}$$

where $(\mathbf{M})^{1/2}$ is the square-root matrix of a symmetric and positive-semi-definite matrix $\mathbf{M}$; $(\mathbf{M})^{1/2} = \mathbf{U}_M\mathbf{D}_M^{1/2}\mathbf{U}_M'$ with $\mathbf{M} = \mathbf{U}_M\mathbf{D}_M\mathbf{U}_M'$ being the eigenvalue decomposition of $\mathbf{M}$. Using (1.75), (1.73) is re-expressed as

$$\mathbf{W}_1^{*\prime}\mathbf{W}_1^* = \mathbf{W}_2^{*\prime}\mathbf{W}_2^* \tag{1.76}$$

and the loss function is

$$\tilde{L}_{CA}(\mathbf{W}_1, \mathbf{W}_2) = \mathrm{tr}\mathbf{W}_2^{*\prime}(\mathbf{X}_2'\mathbf{X}_2)^{-1/2}\mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1/2}\mathbf{W}_1^*. \tag{1.77}$$

The maximizer of $\tilde{L}_{CA}(\mathbf{W}_1, \mathbf{W}_2)$ is therefore obtained by maximizing (1.77) over orthonormal matrices $\mathbf{W}_1^*$ and $\mathbf{W}_2^*$.

The maximization of the trace function is simply attained by Theorem 1.2.1 as

$$L_{CA}^*(\mathbf{W}_1, \mathbf{W}_2) = \mathrm{tr}\mathbf{W}_2^{*\prime}\mathbf{U}\mathbf{D}\mathbf{V}'\mathbf{W}_1^* \leq \mathrm{tr}\mathbf{D}_r \tag{1.78}$$

using the SVD

$$(\mathbf{X}_2'\mathbf{X}_2)^{-1/2}\mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1/2} = \mathbf{U}\mathbf{D}\mathbf{V}'. \tag{1.79}$$

The equality holds when

$$\mathbf{W}_1^* = \mathbf{V}_r,\ \mathbf{W}_2^* = \mathbf{U}_r \tag{1.80}$$

where $\mathbf{U}_r$ and $\mathbf{V}_r$ are the first $r$ columns of $\mathbf{U}$ and $\mathbf{V}$, respectively, and $\mathbf{D}_r$ is the diagonal matrix of the $r$-th largest singular values of $(\mathbf{X}_2'\mathbf{X}_2)^{-1/2}\mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1/2}$ arranged in descending order. Putting (1.80) into (1.75), we have

$$\mathbf{W}_1 = n^{1/2}(\mathbf{X}_1'\mathbf{X}_1)^{-1/2}\mathbf{V}_r, \ \mathbf{W}_2 = n^{1/2}(\mathbf{X}_2'\mathbf{X}_2)^{-1/2}\mathbf{U}_r. \tag{1.81}$$

It is shown that (1.81) satisfies (1.73);

$$\begin{aligned} n^{-1}\mathbf{W}_1'\mathbf{X}_1'\mathbf{X}_1\mathbf{W}_1 &= \mathbf{V}_r'(\mathbf{X}_1'\mathbf{X}_1)^{-1/2}(\mathbf{X}_1'\mathbf{X}_1)(\mathbf{X}_1'\mathbf{X}_1)^{-1/2}\mathbf{V}_r & (1.82)\\ &= \mathbf{V}_r'\mathbf{V}_r & (1.83)\\ &= \mathbf{I}_r & (1.84) \end{aligned}$$

that also holds true for $\mathbf{W}_2$.

A matrix of canonical correlation coefficients are obtained as

$$\begin{aligned} n^{-1}(\mathbf{W}_2'\mathbf{X}_2'\mathbf{X}_1\mathbf{W}_1) &= \mathbf{U}_r'(\mathbf{X}_2'\mathbf{X}_2^{-1/2}(\mathbf{X}_2'\mathbf{X}_1)(\mathbf{X}_1'\mathbf{X}_1)^{-1/2}\mathbf{V}_r & (1.85)\\ &= \mathbf{U}_r'\mathbf{U}\mathbf{D}\mathbf{V}'\mathbf{V}_r & (1.86)\\ &= \mathbf{D}_r & (1.87) \end{aligned}$$

and it is proved to be the diagonal matrix of the $r$ largest singular values in (1.79).

**Example**

CANO was applied to Sports dataset by Tanaka et al. (1984). Thirty nine high school students took the two types of athletic ability tests named test A and B, and the former is composed of "side jump", "back strength", "vertical jump", "grip strength", and "step aerobics", while the latter is made of "long-distance run", "ball throw", "long jump", "chin-up", and "50 meters run". We therefore have $\mathbf{X}_1$ and $\mathbf{X}_2$ both having the dimension of $39 \times 5$. A CANO with three canonical variates was applied to the two data matrices after their column-wise scaling, and the solution matrices $\mathbf{W}_1$ and $\mathbf{W}_2$ were obtained. Table 1.7 shows *structure matrices*, correlation matrices of the original variables and the canonical variates, which are obtained as

$$\mathbf{S}_1 = n^{-1}\mathbf{X}_1'\mathbf{X}_1\mathbf{W}_1, \ \ \mathbf{S}_2 = n^{-1}\mathbf{X}_2'\mathbf{X}_2\mathbf{W}_2. \tag{1.88}$$

The table shows that the first canonical variate is strongly characterized by almost all of the variables. On the other hand, few of the variables are associated to the rest of canonical variates, and it is not clear what the canonical variate stand for. It is well known that CANO often yields difficult-to-interpret solutions as seen in the example.

TABLE 1.7: Structure matrices and canonical correlations obtained by canonical correlation analysis to Sports data.

| | canonical variates | | |
|---|---|---|---|
| | 1st | 2nd | 3rd |
| test A | | | |
| side jump | **-0.670** | -0.278 | -0.187 |
| back strength | **-0.744** | 0.167 | -0.333 |
| vertical jump | **-0.670** | **0.605** | **0.307** |
| grip strength | -0.251 | -0.006 | **0.365** |
| step aerobics | -0.172 | **-0.773** | **0.526** |
| testB | | | |
| long-distance run | **0.667** | 0.117 | **0.441** |
| ball throw | **-0.775** | 0.168 | -0.294 |
| long jump | **-0.789** | **0.497** | 0.038 |
| chin-up | **-0.649** | 0.214 | **0.565** |
| 50m run | **0.676** | **0.362** | **-0.497** |
| | 0.849 | | |
| canonical correlation | | 0.715 | |
| | | | 0.553 |

## 1.3 Interpretability in Multivariate Analysis Procedures

In Section 1.1, the author discussed how multivariate data analysis procedures work for hypotheses creations and their validations in scientific researches by means of CDA and EDA; the author also introduced some important procedures in EDA. In particular, interpretable solutions are of greater importance in EDA, in that they are used for data-based hypotheses creations.

The thesis considers interpretability as an important property that should be possessed by the solutions of multivariate analysis procedures. To illustrate that interpretability is important for hypotheses creations, the example of PCA would be worth revisiting here. Of the two component loading matrices exhibited in Table 1.4, the one before the rotation noted as "Initial" shows ambiguous correspondences between the rows and columns, and thus, the components are difficult to interpreted. This is because the first component highly loads all variables in the matrix. On the other hand, the other matrix approximates a simple structure by Varimax rotation, and thus, it is easier to capture what the components stand for. For example, the easy-to-interpret components allow building the following hypothesis; a baseball player's capability is composed of *batting power*, *running power*, and *batting accuracy*. Further, one can hypothesize that *batting power*, which cannot be measured directly, can be numerically evaluated by the two measurable scales "number of home runs" and "batting score", in that the first component highly loads to the two variables.

Interpretability is also important for prediction because it makes the prediction model more accountable. Mathematical models for prediction are widely used today, and being able to explain clearly how a certain prediction is made is essential in addition to its prediction accuracy and generalizability, and not only for maintaining

its accuracy but also ensuring its social credibility. For instance, in Section 1.2.1, the regression coefficient vector estimated by the penalized estimation with the LASSO penalty uses fewer independent variables when $\lambda$ is large. This means that the scores of "test A" can be explained or predicted by fewer predictors compared with the other results in the table that use all six predictors. Based on the observation, the solution by LASSO penalty presented there is more interpretable. Further, a model that predicts the score of the test by linear combinations of small number predictors are said to be more parsimonious compared with non-parsimonious models, such as the one with more variables and higher order terms or nonlinear functions. In parsimonious models, it is easy to explain its prediction mechanism. For example, the solution vector with $\lambda = 12$ indicates the following simple explanation of the prediction model; the score of "test A" increases by 3.412 points on average if the interest on "law" increases by 1 point, given the other interest scores are fixed. Further, the test score is only affected by the interest scores of "law", "economics", and "engineering". The observation implies that a prediction model is explainable in general if the model is parsimonious and interpretable.

In addition, it is essential to explain clearly how a prediction is generated, especially when the prediction highly affects on our society or a mis-prediction is potentially hazardous. Imagine a prediction model for consumer credit risk and one's future illness as examples of such models. It would be easy to explain how the models predict the future, if they possesses high interpretability and explainability.

### 1.3.1 Sub-concepts of Interpretability

Considering the examples of the multivariate analysis procedures presented above, the thesis addresses the following two sub-concepts of interpretability, which the procedures proposed in the following chapters aim to accomplish.

#### Simple Structure

The concept of a simple structure introduced in the examples of PCA and FA can be rephrased as a property that helps to find which rows are associated to each columns. Often, in multivariate analysis procedures, a solution matrix has variables in its rows and components or clusters in its columns, and the latter need to be named by abstracting the names of the associated variables. Thus, columns and rows are required to be paired. The necessary conditions of the simple structure were first proposed by Thurstone (1947):

- Each row should contain at least one zero.
- Each column should contain at least $r$ zeros.
- Every pair of columns should have several rows with a zero in one column but not the other.

- If $r \geq 4$, every pair of columns should have several rows with zeros in both columns.
- Every pair of columns should have few rows with nonzero loadings in both columns.

These conditions are further elaborated in the next chapter. A factor rotation procedure usually defines a measure of simplicity following these conditions, and it finds a rotation matrix maximizing the measure.

**Sparseness**

A vector or matrix having a number of elements equaling to zero is said to be *sparse*. For instance, the LASSO estimate with $\lambda = 12$ in Table 1.2 is sparse, in that it contains three zeros. Sparseness generally improves explainability, because, in the case of the multiple regression, the number of variables used for a prediction is reduced. Sparsity also facilitates interpretation in that only non-zero elements in a solution matrix need to be interpreted, and one can concentrate on interpreting only few non-zero elements. For instance, consider the case of a multiple regression model with 100 independent variables; the estimated regression coefficient vector contains only 10 non-zero elements and the rest of the elements are zero. The coefficient vector is much easier to interpret, compared with the one with no zero elements, because the variables that explain the dependent variable are remarkably reduced.

## 1.4 Purpose of the Thesis

Thus far, it has been discussed that the interpretability of solutions in multivariate analysis is necessary, and that a simple structure and sparseness are key properties that a highly interpretable solution should possess. In this thesis, the author proposes multivariate analysis procedures that produce easy-to-interpreted solutions. The five procedures introduced in Section 1.2 are going to be refined by means of matrix-intensive modifications. Each introduced procedure is formulated as a minimization of a discrepancy function between the observed values and their prediction over its model parameters. The estimated parameter certainly minimizes the discrepancy; however whether the solutions are interpretable is not always taken into account. The ignorance of interpretability results in solutions that fit data fairly well but are difficult to interpret. The proposed procedures consider not only the solutions' fitness to data but also their interpretability. More precisely, the proposed methods estimate their model parameters so that they possess simplicity and/or sparseness, balancing it with their fitness to the given data.

The remaining parts of the thesis are organized as follows.

In Chapters 2 to 7, a series of multivariate data analysis procedures are proposed, and they are classified into the following four categories according to their methodological characteristics: rotation of solution matrices (Chapter 2 and 3), combination

with clustering (Chapter 4), sparse estimation with cardinality constraint (Chapter 5), and some emerging techniques (Chapter 6 and 7). The details of the procedures are presented below.

In Chapter 2, a new rotation procedure is proposed, which directly approximates the modified Thurstone's simple structure. It can give a $p$-variables $\times$ $r$-factors target matrix of zero and nonzero elements, which stands for the properties to be possessed by the rotated loading matrix. However, the correspondence of the variables in the target matrix to the loading matrix is unknown. In the proposed procedure, a loading matrix is rotated simultaneously with the permutations of the rows of the target matrix, so that the rotated loading matrix is optimally matched with the permuted target matrix. Its novel feature is the use of permutation, and thus the technique is called Permutimin. Its algorithm is presented, with Thurstone's definition of simple structure modified so as to ensure that the target matrix is uniquely specified. Permutimin is illustrated with real data examples, and finally, the relationships between Permutimin and Procrustes rotation are discussed.

Rotation toward a simple structure is extended to CANO in Chapter 3. There, a new formulation of CANO is proposed, which is proved to be equivalent to the exiting one. Two canonical structure matrices are shown to have freedom with respect to orthogonal rotation under the new formulation. The author thus proposes an orthogonal rotation method for rotating two structure matrices individually for their simplicity and interpretability. The proposed method is different from the one derived from the existing formulation, in that it allows to rotate the matrices by a common rotation matrix. This illustrates that the proposed method facilitates the interpretation of the solutions of CANO applied to a real dataset.

The next chapter focuses on a biplot, a technique for obtaining a low-dimensional configuration of a data matrix in that both the objects and variables of the matrix are jointly represented as points and vectors, respectively. However, a biplot with a large number of objects and variables is considered difficult to interpret. Therefore, this study proposes a new biplot procedure that allows us to interpret a large data matrix with ease. In particular, the objects and variables are classified into a small number of clusters using $K$-means clustering. The resulting clusters are simultaneously biplotted in lower-dimensional space. It thus is allowed to understand the configuration easily and grasp the homogeneity of the objects and variables. Further, an extension of the procedure to fuzzy $K$-means clustering is proposed. A simulation study and a real data example are provided to demonstrate the effectiveness of the proposed procedures.

Chapter 5 considers the interpretability in $K$-means clustering because it does not allow any post-hoc transformation of solutions. As noted in Section 1.2, a $p$(variables) $\times$ $K$(clusters) centroid matrix is noted for interpreting variables that characterize clusters, but often, the between-clusters differences are not always clearly captured. The study addresses this problem and proposes a new procedure for obtaining an centroid matrix with sparseness and a simple structure. This makes the interpretation of the clusters easier, as one may focus on only the nonzero centroids. The development

of an iterative algorithm for the constrained minimization is described. A cardinality selection procedure for identifying the optimal cardinality is presented, as well as a modified version of the proposed procedure, in which some restrictions are imposed on the positions of nonzero elements. The behaviors of the proposed procedure were evaluated in simulation studies and are illustrated with three real data examples, which demonstrate that the performances of the procedure are promising.

Chapters 6 and 7 focus on multivariate regression analysis.

Recently, various procedures for sparse multivariate regression (SMR) have been proposed, and they can be classified into the following two types: penalized procedures and cardinality-constrained procedures. Among them, the resulting number of zeros in the regression coefficient matrix is controlled/constrained by a certain penalty function or a prespecified cardinality. The study proposes a novel framework of a multivariate data analysis procedure with sparsity called layered multivariate analysis (LMA), including layered multivariate regression (LMR) in which SMR is elaborated. In LMR, a regression coefficient matrix is assumed to be the sum of several sparse matrices, which is called layers. Therefore, the sparseness of a coefficient matrix is controlled by how many layers are used. In the same way, layered PCA (LPCA), where a component loading matrix is constrained to have a layered structure, is also proposed. It is theoretically guaranteed that a LMR solution converges to the unconstrained solution as the number of layers increases. LMA is assessed by means of a simulation study and illustrated with a real data example.

In Chapter 7, a new penalty function is proposed, which can be used for penalized estimation in various multivariate analysis procedures. The proposed function is a generalization of the existing ones, in that it includes LASSO and ridge penalties as special cases. The proposed function shrinks solutions to a prespecified target matrix that possesses a certain simple structure. The resulting solution is therefore simple and easy to interpret, and its simplicity is controlled by some tuning parameters. We present the two applications of the proposed method in sparse principal component analysis and three-way component analysis. The effectiveness of the proposed method is demonstrated by real data examples.

The final chapter is devoted for summarization and general discussions of the foregoing chapters.

# Chapter 2

# Factor Rotation to Simple Structure with Permutation of Variables

**ABSTRACT**

In this study, we propose a new rotation technique, which is directly based on Thurstone (1947)'s definition. The proposed method can give a $p$-variables $\times$ $r$-factors target matrix of zero and nonzero elements, which stands for the properties to be possessed by the rotated loading matrix, and a loading matrix is obliquely rotated to the target. Further, it is rotated jointly with permuting the rows of the target matrix, in order to match its rows to the ones of the loading matrix optimally, which has not been proposed in the context of factor rotation. The proposed method is therefore called *Permutimin*. An iterative algorithm for Permutimin is proposed, with a refinement of Thurstone's definition of simple structure so as to specify the target matrix uniquely. Two real data examples are presented as illustrations of Pemutimin, and the relationships between Permutimin and Procrustes rotation is discussed.

## 2.1 Introduction

In exploratory factor analysis (EFA), the correlations among $p$ observed variables are assumed to be explained by $r(< p)$ common factors. Their relationships to the variables are described by the factor loadings in the resulting $p$-variables $\times r$-factors matrix $\mathbf{A}$. However, the matrix $\mathbf{A}$ is not uniquely determined, and we call this property *rotational freedom*. The rotational freedom is formally expressed as $\mathbf{AU}$, which can also be regarded as a loading matrix, where $\mathbf{U}(r \times r)$ is called a rotation matrix and satisfies a certain condition.

This chapter is based on the published article Yamashita and Adachi (2019b) with some modifications and extensions.

The rotational freedom is exploited by specifying **U** for rotating **A** so that the resulting loading matrix **AU** has a desired property. This is considered as interpretability, but what matrix is interpretable? A belief answer to the question is the one having the two features; being sparse and well-clustered (e.g., Adachi (2016)). The first feature implies that the loading matrix includes a number of zero elements. It facilitates interpreting what factors are expressed by the columns of the matrix, as only nonzero elements may be focused on for the interpretation. The second well-clustered feature implies that different rows (variables) are associated with different columns (factors). It clarifies how factors differ mutually, which also facilitates interpretation. The matrix with such features is said to have simple structure (Thurstone, 1947; Browne, 2001).

However, the above explanations for the features are not sufficient for specifying the simple structure. The definition that can be considered sufficient has been presented by Thurstone (1947), as in Table 2.1; he has defined simple structure by rules specifying how zero elements are arranged in a matrix.

For elucidating what Thurstone (1947) means by simple structure, we illustrate the matrices provided by the rules in Table 2.1. The first two rules [T1] and [T2] can provide the left matrix in Table 2.2, which is undesirable in that the third factor (column) does not work and three variables (rows) do not load any factor. But, the addition of [T3] to [T1] and [T2] excludes such a matrix and provide the matrices of simple structure. One of them is the matrix (B) in Table 2.2. Here, different factors are mutually distinctive and also different variables are so: the former explain different variables and the latter load different factors. This property is coincident to the EFA aim of exploring the factors which fulfill distinctive roles. Whereas the set of [T1], [T2], and [T3] specifies the necessity for a matrix to have simple structure, the remaining [T4] and [T5] require the resulting matrix to be further simpler. That is, by considering [T5], we can replace the matrix (B) with (C) in Table 2.2. The latter is simpler than the former in that the matrix (C) has more zero elements and can be interpreted more easily. In [T4], it is argued that such simplicity is required in particular for cases with a greater number of factors.

The Thurstone's definition of simple structure illustrated above is still relevant today, as the explanations of factor rotation start with introducing the definition in the influential literature such as Browne (2001), Harman (1976), and Mulaik (2010).

TABLE 2.1: Thurstone's (1947) definition of simple structure rewritten more clearly by Browne (2001, p. 115)

| No. | Rule |
|---|---|
| [T1] | Each row should contain at least one zero. |
| [T2] | Each column should contain at least $r$ zeros. |
| [T3] | Every pair of columns should have several rows with a zero in one column but not the other. |
| [T4] | If $r \geq 4$, every pair of columns should have several rows with zeros in both columns. |
| [T5] | Every pair of columns should have few rows with nonzero loadings in both columns. |

TABLE 2.2: Examples of $7 \times 3$ matrices for illustrating the definition of simple structure, where # and blank indicate nonzero and zero elements, respectively.

(A)

| | |
|---|---|
| # | # |
| # | # |
| # | # |
| # | # |

(B)

| | | |
|---|---|---|
| # | | |
| # | # | |
| # | # | |
| # | # | # |
| | # | # |
| | | # |
| | | # |

(C)

| | | |
|---|---|---|
| # | | |
| # | | |
| # | # | |
| | # | |
| | # | # |
| | | # |
| | | # |

However, the arrangement of zero elements in Table 2.1 is not directly considered in the existing rotation methods. Rather, the value of a function of the elements in $\mathbf{AU}$ is considered, whose maximization or minimization is expected to make the loading matrix $\mathbf{AU}$ satisfy the definition in Table 2.1. For example, in the Varimax rotation (Kaiser, 1958), such a function is defined by summing the column variances of the squared elements of $\mathbf{AU}$ and maximized over $\mathbf{U}$, whereas in the Oblimin rotation (Carroll, 1953), the function is defined using the inter-column covariances of the squared elements and minimized over $\mathbf{U}$. Although these methods may eventually produce $\mathbf{AU}$ that approximately meets the conditions in Table 2.1, the objective functions such as the variances and covariances considered in them are not directly related to the arrangement of zero elements.

In this paper, a new rotation method is proposed, which is directly related to Thurstone's definition of simple structure, i.e., procedure to rotate the loading matrix in order to approximate the structure defined in Table 2.1. To introduce the procedure, we start with an example of the target matrix

$$\mathbf{T} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \left[\begin{array}{ccc} \# & & \\ \# & & \\ \# & \# & \\ & \# & \\ & \# & \# \\ & & \# \\ & & \# \end{array}\right] \end{array}, \tag{2.1}$$

which is the matrix (C) of simple structure in Table 2.2, with # and blank indicating nonzero and zero elements, respectively. This matrix is called a target matrix and used as the target of simple structure to be approximated by the rotated loading

matrix. Let us consider the problem of rotating the artificial loading matrix

$$\mathbf{A} = \begin{array}{l} \text{1. } \textit{Agreeableness} \\ \text{2. } \textit{Conscientiousness} \\ \text{3. } \textit{Extraversion} \\ \text{4. } \textit{Neuroticism} \\ \text{5. } \textit{Agressiveness} \\ \text{6. } \textit{Depressiveness} \\ \text{7. } \textit{Openness} \end{array} \begin{bmatrix} -0.6 & -0.6 & -0.9 \\ -0.3 & -0.3 & 0.4 \\ -0.6 & 0.1 & -0.4 \\ -0.4 & 0.2 & 0.7 \\ 0.4 & 0.4 & -0.5 \\ -0.1 & 0.5 & 0.3 \\ -0.8 & 0.2 & -0.5 \end{bmatrix} \tag{2.2}$$

so that the resulting matrix has a simple structure indicated by (2.1). Here, we should notice that the correspondence of the rows in (2.1) to the variables in (2.2) is not specified. Thus, the rows of $\mathbf{T}$ in (2.1) must be permuted, for the reason illustrated as follows: for example, if the second row *Durable* in (2.2) is to load the third factor (after rotation), then, the fifth, sixth, or seventh row of (2.1) (loading the third factor) must be moved to the second row, through permutation, so as to be paired to the row of *Durable* in (2.2). On the other hand, if the third row *Durable* in (2.1) is to load the second factor, the third, fourth, or fifth row of (2.1) (loading the second factor) must be moved to the third row through permutation. We therefore consider a rotation procedure, in which $\mathbf{A}$ is rotated simultaneously with the rows of $\mathbf{T}$, which are permuted so that the two resulting matrices are optimally matched mutually. Using $\mathbf{P}$ a $p \times p$ permutation matrix $\mathbf{P}$, the matrices $\mathbf{P}$ and $\mathbf{U}$ are obtained, so that the rotated matrix $\mathbf{AU}$ approximates permuted target matrix $\mathbf{T}$ in the proposed method. Indeed, it provides the following results:

$$\mathbf{P} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}, \quad \mathbf{PT} = \begin{array}{c} 3 \\ 6 \\ 4 \\ 1 \\ 7 \\ 2 \\ 5 \end{array} \begin{bmatrix} \# & \# & \\ & & \# \\ & \# & \\ \# & & \\ & & \# \\ \# & & \\ & \# & \# \end{bmatrix}, \tag{2.3}$$

$$\mathbf{AU} = \begin{array}{l} \text{1. } \textit{Agreeableness} \\ \text{2. } \textit{Conscientiousness} \\ \text{3. } \textit{Extraversion} \\ \text{4. } \textit{Neuroticism} \\ \text{5. } \textit{Agressiveness} \\ \text{6. } \textit{Depressiveness} \\ \text{7. } \textit{Openness} \end{array} \begin{bmatrix} \mathbf{0.91} & \mathbf{-0.91} & 0.12 \\ -0.05 & -0.01 & \mathbf{0.57} \\ 0.04 & \mathbf{-0.74} & -0.16 \\ \mathbf{-0.66} & -0.01 & 0.36 \\ 0.04 & 0.04 & \mathbf{-0.74} \\ \mathbf{-0.61} & 0.01 & 0.36 \\ -0.03 & \mathbf{-0.98} & \mathbf{-0.25} \end{bmatrix}. \tag{2.4}$$

Here, the values of nonzero elements # in (2.4) are also estimated, as described in Section 2. By comparing the rows in (2.1) with those in (2.3), we can see that the permutation of the former rows gives (2.3); the third row in (2.1) has been moved

to the first row in **PT**. The comparison between (2.3) and (2.4) shows that the boldfaced elements in (2.4) corresponding to # in (2.3) have large absolute values, while the remaining ones are close to zero, except for a few elements. That is, the rotated loadings in (2.4) attain simple structure fairly well, in that each factor is manifested by the reduced number of variables; e.g., the first factor is manifested by *Agreeableness*, *Neuroticism*, and *Depressiveness*.

The remaining parts of the paper are organized as follows: in the next section, we formulate our proposed method, which we call Permutimin. How it can be located among the existing methods are discussed in Section 3. Its algorithm for a specified target **T** is presented in Section 4. Here, **T** is merely one matrix in a set of target matrices (target set), which has a simple structure as that indicated by Thurstone's definition in Table 2.1. However, the definition includes vague aspects as discussed in Section 5.1, and therefore it does not allow us to specify a target set; this is discussed in Section 5. There, the definition is modified so that the target set can be specified in an objective manner. Two examples are given to illustrate Permutimin in Section 6. Finally, discussions are made for reconsidering Permutimin.

## 2.2 Proposed Method: Permutimin

A target matrix can be expressed as

$$\mathbf{T} = \mathbf{B} \bullet \mathbf{C}. \tag{2.5}$$

Here, $\bullet$ denotes Hadamard element-wise product, $\mathbf{C}$ is a $p \times r$ unknown and unconstrained matrix, and $\mathbf{B}$ is the binary matrix with its elements taking one and zero corresponding to the zero and nonzero ones in $\mathbf{T}$, respectively; for example, if $\mathbf{T}$ is given as in (2.1), then

$$\mathbf{B} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}. \tag{2.6}$$

It shows that the simple structure expressed as the zero/nonzero elements in $\mathbf{T}$ is also represented by the zero/ones in $\mathbf{B}$; they specify the simple structure in the target by the arrangement of nonzero elements. Therefore, we call $\mathbf{B}$ a binary target matrix.

For the loading matrix $\mathbf{A}$ $(p \times r)$ resulting in EFA, the optimal permutation matrix $\mathbf{P}$ $(p \times p)$, rotation matrix $\mathbf{U}$ $(r \times r)$, and $\mathbf{C}$ $(p \times r)$ are obtained in Permutimin, so that the rotated loading matrix $\mathbf{AU}$ is optimally matched with the permuted target

matrix $\mathbf{PT} = \mathbf{P}(\mathbf{B} \bullet \mathbf{C})$ in a least square sense. It is formally formulated as minimizing

$$f(\mathbf{P}, \mathbf{U}, \mathbf{C}) = ||\mathbf{AU} - \mathbf{P}(\mathbf{B} \bullet \mathbf{C})||^2 \tag{2.7}$$

over $\mathbf{P}$, $\mathbf{U}$, and $\mathbf{C}$ subject to two constraints. One constraint is that $\mathbf{P}$ is a permutation matrix. The other is that $\mathbf{U}$ is a rotation matrix. This constraint differs depending on whether the rotation is orthogonal or oblique. We choose the latter, as the latter includes the former as a special case. In this paper, we suppose the following: the matrix $\mathbf{A}$ to be rotated has been obtained so that the matrix of the inter-factor correlations is equal to the $r \times r$ identity matrix $\mathbf{I}_r$ which indicates that the FA solution is obtained as an orthogonal solution. Then, in our chosen oblique rotation, $\mathbf{U}$ is constrained as

$$\text{diag}(\mathbf{U}^{-1}\mathbf{U}^{-1\prime}) = \mathbf{I}_r. \tag{2.8}$$

Here, diag($\mathbf{M}$) denotes the diagonal matrix whose diagonal elements are the same as those of the square matrix $\mathbf{M}$.

The loss function in (2.7) can be rewritten as

$$f(\mathbf{P}, \mathbf{U}, \mathbf{C}) = ||\mathbf{P}'\mathbf{AU} - \mathbf{B} \bullet \mathbf{C}||^2 \tag{2.9}$$

using the property of permutation matrices $\mathbf{P}'\mathbf{P} = \mathbf{P}\mathbf{P}' = \mathbf{I}_r$. As $\mathbf{P}'$ is also a permutation matrix, (2.9) shows that Permutimin can also be formulated as a rotating matrix $\mathbf{A}$ simultaneously by permuting its rows.

It should be noted that the binary target matrix $\mathbf{B}$ in (2.5) is merely an example:

$$\mathbf{B} \in \mathbb{S} \tag{2.10}$$

where $\mathbb{S}$ is a set of binary target matrices having simple structure. Thus, we must consider all $\mathbf{B}$ contained in $\mathbb{S}$. The overall procedure in Permutimin is as follows:

1. Perform the constrained minimization of (2.7) for each $\mathbf{B}$ in $\mathbb{S}$.

2. Compare the solutions across different $\mathbf{B}$s to select the best solution.

In the next section, we present the algorithm for [1] without considering which $\mathbf{B}$ in $\mathbb{S}$ is to be used; we describe the algorithm for a given $\mathbf{B}$. The set $\mathbb{S}$ is considered in Section 4. For the selection in [2], we use a scree criterion as illustrated in Section 5.

## 2.3 Relationships to the Existing Procedures

Permutimin can be included in the family of target rotation procedures which include a step of rotation a loading matrix to be matched with a target matrix. In this section, we discuss hierarchical relationships among the procedures including Permutimin.

The simplest procedure in the family is Procrustes rotation (Green, 1952; Huley and Cattell, 1962; Mosier, 1939; Gower and Dijksterhuis, 2004). It can be formulated

as minimizing $||\mathbf{AU}-\mathbf{T}||^2$ over the rotation matrix $\mathbf{U}$ subject to (2.8) or its orthogonal version for a given target $\mathbf{T}$. This matrix is reparametrized as $\mathbf{T} = \mathbf{B} \bullet \mathbf{C}$ in Browne's (1972a, 1972b) variant of Procrustes rotation, where $\mathbf{B}$ is a given binary matrix and $\mathbf{C}$ is an unknown unconstrained matrix; $||\mathbf{AU}-\mathbf{B}\bullet\mathbf{C}||^2$ is minimized over $\mathbf{U}$ and $\mathbf{C}$. This function is the Permutimin's loss function (2.7) with $\mathbf{P}$ fixed to $\mathbf{I}_p$. Thus, Permutimin can be viewed as an extension of Browne's variant of Procrustes rotation. Another one of its extensions is Kiers's (1994) Simplimax, which is formulated as minimizing $||\mathbf{AU} - \mathbf{B} \bullet \mathbf{C}||$ over $\mathbf{U}$ and $\mathbf{C}$ as well as $\mathbf{B}$, subject to the number of zero elements in $\mathbf{B}$ being specified. Though the function does not include $\mathbf{P}$, the locations of zeros in $\mathbf{B}$ are not fixed, but estimated. Thus, the Simplimax function may be rewritten as $||\mathbf{AU} - \mathbf{P}(\mathbf{B} \bullet \mathbf{C})||^2$. It shows that Permutimin can be viewed as a constrained version of Simplimax with the locations of zeros fixed. These discussions give the following hierarchy:

$$\text{Procrustes} \prec \text{Browne's variant} \prec \text{Permutimin} \prec \text{Simplimax} \tag{2.11}$$

where the symbol $\prec$ stands for the method preceding it being more restrictive than the following one.

As described above, Browne's rotation is the restricted version of Permutimin with its permutation matrix $\mathbf{P}$ fixed to $\mathbf{I}_p$. It implies that any rows of $\mathbf{B}$ are not permuted. Thus, Browne's rotation only works for the cases where the correspondences of the rows in the rotated matrix $\mathbf{AU}$ to those of target $\mathbf{T} = \mathbf{B} \bullet \mathbf{C}$ are known a priori. However, such correspondences are often unknown. In such cases, Browne's rotation is not useful, but Permutimin can be utilized since reasonable correspondences can be found with estimating $\mathbf{P}$ optimally, if a suitable value is selected as target's cardinality, i.e., the number of nonzero elements in $\mathbf{T}$. This selection can be attained with the scree procedure introduced in Section 5.

The Procrustes rotation is the simplest and most restrictive as found in (2.11). However, even if it is unknown how the rows in $\mathbf{A}$ correspond to those of the target, the rotation can be used by specifying $\mathbf{T}$ according to the result of another rotation procedure. This version of Procrustes rotation is Hendrickson & White's (1964) Promax rotation, in which $\mathbf{T}$ is specified using the Varimax-rotated loading matrix. Obviously, the result of Promax depends on that of the Varimax rotation. A drawback of the Promax is that it fails to provide a simple structure, if the preceding Varimax rotation is misleading.

## 2.4 Algorithm

In this section, we present an algorithm for the constrained minimization of (2.7) or (2.9) over $\mathbf{C}$, $\mathbf{U}$, and $\mathbf{P}$ for a given $\mathbf{B}$. For the minimization, the three steps called C-, U-, and P-step are alternately iterated until convergence is reached.

In C-step, (2.7) is minimized over $\mathbf{C}$ for fixed $\mathbf{U}$ and $\mathbf{P}$. For this minimization, we note that $\mathbf{B}$ in (2.9) is a binary matrix. It shows that the optimal $\mathbf{C}$ is simply given by

$$\mathbf{C} = \mathbf{P}'\mathbf{A}\mathbf{U}. \tag{2.12}$$

It should be noted that $\mathbf{C}$ in (2.12) cannot be determined uniquely, since it is filtered by $\mathbf{B}$ as $\mathbf{B} \bullet \mathbf{C}$ and the elements in $\mathbf{C}$ corresponding to the zeros in $\mathbf{B}$ do not affect on the function value in (2.7). However, the target matrix is uniquely determined as $\mathbf{T} = \mathbf{B} \bullet \mathbf{C}$, to which the initial loading matrix is rotated.

In U-step, (2.7) is minimized over $\mathbf{U}$ subject to (2.8) with $\mathbf{C}$ and $\mathbf{P}$ being fixed. For this minimization, we use Jennrich's (2002) gradient projection (GP) algorithm in which the update of $\mathbf{U}$ is iterated, and the iteration is finished when the decrease in (2.7) with the update is less than $10^{-5}$, or the number of iterations reaches 1000.

In P-step, the permutation matrix $\mathbf{P}$ is updated so that (2.7) decreases. For this update, we find the permutation of two rows in $\mathbf{P}_0\mathbf{T}$ which leads to the greatest decrease in (2.7), with $\mathbf{P}_0$ denoting $\mathbf{P}$ before update. This problem is formulated as finding $[I, J] = \underset{[i,j],1\le i<j\le m}{\operatorname{argmin}} D(\mathbf{Q}_{ij})$. Here, $[i, j]$ denotes a pair of row numbers; with $\mathbf{e}_i$ being the $i$-th column of $\mathbf{I}_p$, the $p \times p$ permutation matrix for exchanging the $i-th$ and $j-th$ rows of $\mathbf{P}_0\mathbf{T}$ is expressed as $\mathbf{Q}_{ij} = \mathbf{I}_p - \mathbf{e}_i\mathbf{e}_i' - \mathbf{e}_j\mathbf{e}_j' + \mathbf{e}_i\mathbf{e}_j' + \mathbf{e}_j\mathbf{e}_i'$; and $D(\mathbf{Q}_{ij}) = ||\mathbf{A}\mathbf{U} - \mathbf{Q}_{ij}\mathbf{P}_0\mathbf{T}||^2 = ||\mathbf{A}\mathbf{U} - \mathbf{P}_0\mathbf{T}||^2$ expresses the decrease in (2.7) by the exchange. To specify $[I, J]$, $D(\mathbf{Q}_{ij})$ for all possible pairs of $(i, j)$ are evaluated. The update of $\mathbf{P}$ is given by

$$\mathbf{P} = \begin{cases} \mathbf{Q}_{IJ}\mathbf{P}_0 & iff \quad D(\mathbf{Q}_{IJ}) < 0 \\ \mathbf{P}_0 & otherwise \end{cases} \tag{2.13}$$

which ensures that (2.7) decreases if $D(\mathbf{Q}_{IJ}) < 0$. Otherwise, $\mathbf{P}$ remains $\mathbf{P}_0$, and the value of (2.7) also remains unchanged, implying that the iteration is over.

The above steps allow us to form the following algorithm for the minimization of (2.7):

1. Initialize $\mathbf{P}$ and $\mathbf{U}$ randomly.

2. Update $\mathbf{C}$ with (2.12).

3. Update $\mathbf{U}$ with the GP algorithm.

4. Update $\mathbf{P}$ with (2.13).

5. Finish if convergence is reached; otherwise go back to Step [2].

In Step [1], $\mathbf{P}$ is initialized at a permutation matrix chosen randomly, and the initial $\mathbf{U}$ is obtained by transforming an $r \times r$ matrix $\mathbf{U}_0$ so as to satisfy (2.8) by adjusting the column length of $\mathbf{U}_0^{-1}$ to 1. Here, each element of $\mathbf{U}_0$ is drawn from the uniform distribution $U(-1, 1)$. The convergence of Step [5] is defined as that the decrease in (2.7) from the previous round is less than $10^{-7}$ or $\mathbf{P} = \mathbf{P}_0$ in (2.13). In order to reduce the possibility of selecting a local minimum as the optimal solution, we run the above algorithm multiple times and consider the solution with the lowest function value of (2.7) as the optimal one. See Appendix 2 for details of the multiple run procedure.

## 2.5 Target Matrices of Simple Structures

In Section 4.1, we consider the definition of simple structure which allows the set $\mathbb{S}$ in (2.10) to be specified. In Section 4.2, we describe how $\mathbb{S}$ is specified and how a suitable $\mathbf{B}$ can be selected among $\mathbf{S}$.

### 2.5.1 Modification of Thurstone's Simple Structure

The definition of the simple structure to be possessed by the target matrix (2.5) was originally presented by Thurstone (1947), and is shown in Table 2.1. However, the rules [T3] to [T5] in the table include vague adjectives such as "several" and "few", which do not allow the set in (2.10) to be identified objectively, although [T1] and [T2] in Table 2.1 are clearer. Here, "objectively" can be rewritten as "uniquely": the set must be uniquely specified so that a target matrix can be computationally formed. Therefore, we modify Thurstone's rule such that its modified version allows $\mathbb{S}$ to be specified easily.

Simple structure can be defined by the zeros/ones in $\mathbf{B}$ which is equivalent to the zero/nonzero elements in $\mathbf{T}$ as found in (2.5). Thus, in our modified version, simple structure is defined with zeros/ones in $\mathbf{B}$, as follows:

[M1] Each row has at least one zero and at least one element being one.

[M2] Each column has at least $r$ zeros.

[M3] The maximum of inter-column inner products is the smallest among possible $\mathbf{B}$s.

[M4] The maximum of the inter-column differences in the number of ones is the smallest among possible $\mathbf{B}$s.

Note that the above rules are made for the binary matrix $\mathbf{B}$, and [M1] would be "Each row has at least one zero and at least large element in magnitude" for other matrices. Let us compare this version with the original one in Table 2.1. Rule [M2] is equivalent to [T2]. Rule [M1] is simply given by adding the condition that each row has at least one nonzero elements to [T1]. Rules [M3] and [M4] are detailed in the next paragraphs.

Rule [M3] is related to [T3], [T4], and [T5]. They imply that an element must be zero, if the corresponding element in a different column takes a nonzero value. This implication is expressed using inner-products in [M3] so as to be assessed numerically.

Rule [M4] is newly introduced, by which the number of nonzero elements in each column can be specified. This rule can be rewritten as that homogeneous numbers of variables load each common factor, which implies that all common factors are equivalently significant. For example, three variables in rows of $\mathbf{B}$ in (2.6) are heavily load each of the three factors in columns. The maximum of the inter-column differences in the number of ones in $\mathbf{B}$ attains its minimum 0 among all $\mathbf{B}$s, and [M4] is therefore satisfied. This condition, though it may be restrictive, serve to avoid undesirable factors that contribute to few variables. It is contradictory to the concept of common factors in that they explain the variations in multiple variables.

As shown in the next subsection, $\mathbf{B}$ can be uniquely specified, when the number of non-zero elements is given.

### 2.5.2 Constructing Binary Target Matrices

Let $N = Card(\mathbf{B}) = Card(\mathbf{T})$ be the cardinality of $\mathbf{B}$ or $\mathbf{T}$; i.e., the number of nonzero elements in the target matrix. Rules [M1] and [M2] imply that $N$ is an integer satisfying

$$p \leq N \leq r(p-r). \tag{2.14}$$

Here, the lower bound is set to $p$ for avoiding an empty row, in order to satisfy [M1] which is necessary to specify $\mathbf{B}$ uniquely. The upper bound is set to $r(p-r)$, as $N$ exceeding it would not allow [M1] to be satisfied.

Table 2.3 illustrates how the binary target matrix $\mathbf{B}$ meeting [M1] to [M4] is specified for each $N$ in (2.14) for $p = 8$ and $r = 3$. Its leftmost panel (A) shows $\mathbf{B}$ for $N$ being the minimal 8. Obviously, it follows the simple structure rules. Let us note the change from Panel (A) to (B). In the latter, $N$ is increased by one so that the bold faced one is added in $(1, 3)$. It can be seen that the addition ensures that the rules are followed, with the number of ones in each column and the inter-column inner products (shown at the bottom of Table 2.1) satisfying [M3] and [M4]. Next, let us note the change from (B) to (C), where one is added to the cell $(2, 1)$ so that the rules are followed. We can find that the addition of one is replicated in similar manners from (C) to (H).

The algorithm for forming the the binary target matrices that satisfy the modified Thurstone's rule can be described as next, using $\mathcal{N}$ for a set of indexes $(i, j)$ and $O$ for the $p \times r$ zero matrix:

1. Set $i = j = 1$, $n = 0$, $\mathcal{N} = \emptyset$, and $\mathbf{B} = \mathbf{O}$.
2. Set $b_{ij} = 1$ with $b_{ij}$ is the $(i, j)$-th element of $\mathbf{B}$.
3. $j = j + 1$ if $j < r$; otherwise set $j = 1$.

4. $i = i + 1$ if $i < p$; otherwise set $i = 1$.

5. If the current $(i, j)$ is found in the set $\mathcal{N}$, go back to [4]; otherwise add the current $(i, j)$ to $\mathcal{N}$ and go to [6].

6. Finish if $n = N$; otherwise set $n = n + 1$ and go back to [1].

Above, an integer $n$ is used for counting the current number of zeros in $\mathbf{B}$.

As discussed in Appendix 3, the optimal binary target matrix $\mathbf{B}$ that satisfies [M1]-[M4] in Section 4.1 can be uniquely determined for a given $N$. That is, the modified Thurstone's rule gives the specified $\mathbf{B}$ through the above algorithm, in contrast to its original version with unclear properties. This justifies the modification of the rule and the proposed procedure based on the modified rule.

TABLE 2.3: Binary target matrices $\mathbf{B}$ ($8 \times 3$) satisfying the proposed simple structure conditions.

| | (A) | | | (B) | | | (C) | | | (D) | | | (E) | | | (F) | | | (G) | | | (H) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $N$ | 8 | | | 9 | | | 10 | | | 11 | | | 12 | | | 13 | | | 14 | | | 15 | | |
| $R(N/p)$ | 8 | | | 1 | | | 2 | | | 3 | | | 4 | | | 5 | | | 6 | | | 7 | | |
| $R(N/r)$ | 2 | | | 3 | | | 1 | | | 2 | | | 3 | | | 1 | | | 2 | | | 3 | | |
| | 1 | | | 1 | | **1** | 1 | | 1 | 1 | | 1 | 1 | | 1 | 1 | | 1 | 1 | | 1 | 1 | | 1 | 1 |
| | | 1 | | | 1 | | **1** | 1 | | 1 | 1 | | 1 | 1 | | 1 | 1 | | 1 | 1 | | 1 | 1 | | |
| | | | 1 | | | 1 | | | 1 | | **1** | 1 | | 1 | 1 | | 1 | 1 | | 1 | 1 | | | 1 | 1 |
| | 1 | | | 1 | | | 1 | | | 1 | | | 1 | | **1** | 1 | | 1 | 1 | | 1 | 1 | | 1 |
| **B** | | 1 | | | 1 | | | 1 | | | 1 | | | 1 | | **1** | 1 | | 1 | 1 | | 1 | 1 | | |
| | | | 1 | | | 1 | | | 1 | | | 1 | | | 1 | | | 1 | | **1** | 1 | | | 1 | 1 |
| | 1 | | | 1 | | | 1 | | | 1 | | | 1 | | | 1 | | | 1 | | | 1 | | **1** |
| | | 1 | | | 1 | | | 1 | | | 1 | | | 1 | | | 1 | | | 1 | | | 1 | |
| Number of 1 | 3 | 3 | 2 | 3 | 3 | 3 | 4 | 3 | 3 | 4 | 4 | 3 | 4 | 4 | 4 | 5 | 4 | 4 | 5 | 5 | 4 | 5 | 5 | 5 |
| Product* | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 2 | 1 | 2 | 2 | 1 | 2 | 2 | 2 | 2 | 3 | 2 |

*In each panel, the 1st, 2nd, and 3rd cells show inner products $\mathbf{b}_1'\mathbf{b}_2$, $\mathbf{b}_1'\mathbf{b}_3$ , and $\mathbf{b}_2'\mathbf{b}_3$, respectively, with $\mathbf{b}_j$ the $j$th column of $\mathbf{B}$.

We can ascertain that the algorithm gives Table 2.3, when $p = 8$ and $r = 3$. By increasing $n$ as $n = 1, 2, \cdots, N$, with $N = 8$, we have $(R(n/p), R(n/r)) = (1, 3), (2, 1), \cdots, (7, 3)$. They indicate the bold faced elements in Panels (B), (C), $\cdots$, (H), which are the ones added with an increase in $N$. Obviously, $\mathbf{B}$ can also be defined for other $p$ and $r$ values.

### 2.5.3 Selection of Target Cardinality

Here, we consider the selection of a suitable $N$ (the cardinality of the target matrix) from its range (2.14). For this consideration, it must be noted that a trade-off relationship exists between the interpretability of the rotated loading matrix $\mathbf{AU}$ and how well it approximates the permuted target $\mathbf{PT} = \mathbf{P}(\mathbf{B} \bullet \mathbf{C})$. As $N$ increases, the approximation is improved; i.e., the resulting value of the loss function (2.7) is decreased, since a larger $N$ implies a greater number of parameters being estimated. On the other hand, a decrease in $N$ facilitates the interpretation, as a smaller $N$ causes a greater number of loadings to be close to zero, and they may not be noted for the interpretation.

A reasonable approach to selecting $N$ is using a scree plot, in which $\tilde{f}_N = \alpha^{-1} f_N$ is plotted against $N$, where $\alpha = ||\mathbf{A}||^2$ is used for the normalization. Here, $f_N$ denotes the resulting (2.7) value as a function of $N$ and $\alpha$ is a suitable normalizing factor. A typical procedure with the plot is to note the amounts of decreases in the (2.7) values, $\tilde{f}_{N-1} - \tilde{f}_N$ and $\tilde{f}_N - \tilde{f}_{N+1}$, where the two differences indicate the decrements before the cardinality reaching $N$ and after it exceeding $N$, respectively. Then, an $N$ value with $\tilde{f}_{N-1} - \tilde{f}_N$ great but $\tilde{f}_N - \tilde{f}_{N+1}$ small can be selected as a suitable one. However, inspecting the scree plot does not allow such an $N$ value to be chosen easily. Thus, we propose to use the numerical index

$$\delta(N) = \frac{\tilde{f}_{N-1} - \tilde{f}_N}{\tilde{f}_N - \tilde{f}_{N+1}} \tag{2.15}$$

in place of the inspection: the $N$ value maximizing (2.15) can be reasonably chosen as a suitable cardinality. The scree plot procedure is illustrated in the next section. However, the above procedure with (2.15) is not guaranteed to give optimal $N$ in terms of resulting simplicity of loading matrix. Thus, one may try to obtain the loading matrices for some different $N$s and compare the resulting ones to find an $N$ value which yields a preferable result.

## 2.6 Illustrations

In this section, we illustrate Permutimin with the selection of a suitable value of cardinality $N$, using two examples. In the first one, prior knowledge can be given for the correspondences of the rows of a rotated loading matrix to the target's counterparts, but the knowledge is not given in the second example.

### 2.6.1 Thurstone's Box Problem

As an example, we use Cureton & Mulaik's (1975, p. 184, Table 1) 15 variables $\times$ 3 factors loading matrix given by Thurstone's (1947) box problem. In this problem, the variables are defined using the functions of factors $(x, y, z)$ as shown in Table 2.4 (left column), and the data simulated on the basis of the definition are analyzed by using EFA to provide the loading matrix to be rotated. The rotated loadings can be regarded as successful, if the variables show large loadings for the associated factors, for example, the variable $2y + 2z$ gives a large loading for factors $y$ and $z$. Thus, the

suitable target matrix is

$$\mathbf{T} = \begin{array}{r} \\ x \\ y \\ z \\ xy \\ (xy)^2 \\ 2(x+y) \\ \sqrt{x^2+y^2} \\ xz \\ (xz)^2 \\ 2(x+z) \\ \sqrt{x^2+z^2} \\ yz \\ (yz)^2 \\ 2(y+z) \\ \sqrt{y^2+z^2} \end{array} \begin{bmatrix} \# & & \\ & \# & \\ & & \# \\ \# & \# & \\ \# & \# & \\ \# & \# & \\ \# & \# & \\ \# & & \# \\ \# & & \# \\ \# & & \# \\ \# & & \# \\ & \# & \# \\ & \# & \# \\ & \# & \# \\ & \# & \# \end{bmatrix} \tag{2.16}$$

with Card($\mathbf{T}$) = 27. It can be given by $\mathbf{T} = \mathbf{P}(\mathbf{B} \bullet \mathbf{C})$, where the binary target $\mathbf{B}$ chosen for Card($\mathbf{T}$) = 27 by the algorithm in Section 5.2.

Without using this prior knowledge, i.e., supposing that $\mathbf{B}$, $\mathbf{U}$, and $\mathbf{U}$ are unknown, we performed Permutimin for Cureton and Mulaik's matrix, with $N$ set at each integer with the interval (2.14), i.e., from 15 to 36, and the corresponding target matrix obtained by the procedure described in Section 5.2. Figure 2.1(A) shows the scree plot of the loss function (2.7) values. There, it is obvious that the values hardly change for $N < 31$. For $N \leq 31$, we obtained the ratios (2.15)) and their plot against $N$ is presented in Figure 2.1(B), where $N = 27$ is to be selected as the cardinality of the target. This is the suitable value described above indicating that the scree test procedure appropriately specified the reasonable $N$.

We performed Permutimin for Cureton and Mulaik's matrix, with $N$ set at each integer within the interval (2.14), i.e., from 15 to 36, and the corresponding target matrix obtained by the procedure described in Section 4.2. Figure 2.1(A) shows the scree plot of the loss function (2.7) values. There, it is obvious that the values hardly change for $N > 31$. For $N \leq 31$, we obtained the ratios (2.15) and their plot against $N$ is presented in Figure 2.1(B), where $N = 27$ is to be selected as the cardinality of the target. is consistent to the assumption that the 9 variables are associated to each of the three factors. Therefore, it can be said that the scree test procedure appropriately specifies reasonable $N$.

Panel (A) in Table 2.4 shows the loading matrix by Permutimin with $N = 27$. It is seen that the loadings can be split into those greater than 0.3 and those close to zero with their absolute values not exceeding 0.1. The former loadings, which are boldfaced, demonstrate that the Permutimin solution can be considered to be successful in that the variables load the factor(s) used for defining the variables. For comparison, we also

rotated Cureton and Mulaik's matrix by five other oblique rotation methods, namely Oblimin, Promax, Geomin (Yates, 1987), Simplimax (Kiers, 1994), and component loss rotation (Jennrich, 2004, 2006), where their detailed procedures are described in Appendix 1. Panels (B) to (F) in Table 2.4 show the resulting loadings, with bold font used for the ones whose absolute values are greater than 0.3. We can find that those five methods did not attain the ideal structure as good as Permutimin did. Varimax rotation has also been found to have failed to give the structure by Cureton and Mulaik (1975). These comparisons demonstrate a promising behavior of Permutimin.

The procedures treated so far are those for the cases without prior knowledge. On the other hand, Browne's rotation is the procedure that works when the knowledge given. We performed the Browne's rotation with $\mathbf{T}$ set at (2.16). The resulting loading matrix shown in Table 2.4(G) was identical to the one by Permutimin. It demonstrates that Browne's rotation works as fairly well as Permutimin if the target is specified correctly, but Permutimin can also find the ideal structure even if only the cardinality of target is specified.

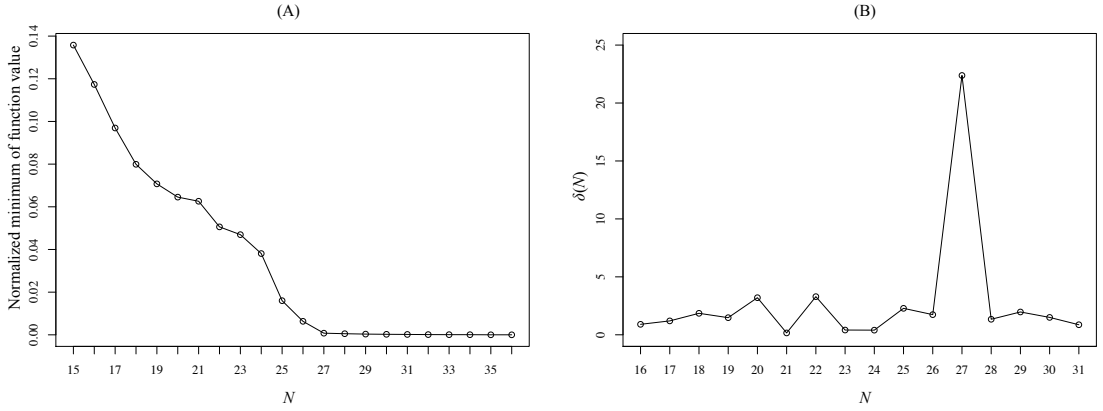



FIGURE 2.1: Attained minimum of function value (A) and $\delta(N)$ (B) for $N = 15 \cdots, 36$ for the box problem.

### 2.6.2 Personality Loadings

As the second example, we use Yanai & Ichikawa's (2006) 12 personality traits $\times$ 3 factors loading matrix obtained for 12 scales of New Personality Test (Yanai et al., 1987). For the matrix, we performed Permutimin in the same procedures as in the last example. Existing rotation procedures were also applied for comparison But, Browne's rotation was not performed, as the prior knowledge to be required by the rotation can not not be given. Figure 2.2(A) shows the scree plot of the resulting loss function values. There, the values hardly change for $N > 28$. For $N \leq 28$, we obtained the ratio (2.15) and their plot against $N$ is presented in Figure 2.2(B), where $N = 17$ is to be selected as the cardinality of the target.

The loading matrix resulted in Permutimin with $N = 17$ is shown in Panel (A) of Table 2.5, together with the results for the matrices obtained with the other procedures in Panels (D) - (H). Note that the loadings with their absolute values greater than 0.3 boldfaced. The matrix for Permutimin in Panel (A) is similar to those of other four procedures excluding Simplimax. That is, the five procedures including Permutimin form a majority. Their loading matrices can be interpreted reasonably as follows. The first factor, showing the large positive loadings for *Extroversion*, *Activity*, *Novelty*, *Self-revelation*, and *Aggressiveness*, can be interpreted as standing for "general activity". The second one with large loadings for *Aggressiveness*, *Lack of Cooperativeness*, *Inferiority Feeling*, *Nervousness*, and *Depression* can be labeled an "emotional instability" factor. The third factor, which shows remarkably positive loadings for *empathy*, *durability*, and *regularity*, can be interpreted as a "conscientiousness" factor.

For the potential users of Permutimin, we show two results with $N$s different from the above 17 which might satisfy some special interests. The result of Permutimin with $N = 14$ is shown in Panel (B) in Table 2.5. The target cardinality is close to its minimum and such a target would be useful for the purpose of variable clustering. In this example, except for two variables, each variable highly loads on a single factor and they can be classified into three clusters. One the other hand, one might be interested in comparing the solutions for some candidates of $N$ values. We thus performed Permutimin with $N = 22$ in which second largest value of (2.15) was observed. The rotation result is shown in Panel (C). This result is slightly different from the one in Panel (A); *Inferiority Feeling* shows significantly negative loading to the first factor and the correspondence between *Activity* and the third factor no longer observed. The two loading matrices are equally simple and interpretable, but which is preferred depends on users' interest. Users may choose the results matched with their purposes. In other words, while the scree test procedure suggests single $N$ as a best, users can try to run Permutimin with setting $N$ to different values and find a better rotation result from the stand point of interpretability and theoretical validity. Permutimin provides such freedom for users, which satisfies broader needs for FA.

Permutimin's solution is also useful for ordering (or permuting) the variables in a covariance matrix, so that those loading to the same factor are arranged mutually close. It facilitates interpreting the loading matrix. It is accomplished by referring

the estimated **PB**. For example, in this illustration, the estimated **PB** is

$$\begin{array}{ll} 1. & \textit{Extraversion} \\ 2. & \textit{Activity} \\ 3. & \textit{Empathy} \\ 4. & \textit{Novelty} \\ 5. & \textit{Durability} \\ 6. & \textit{Regularity} \\ 7. & \textit{Self-revelation} \\ 8. & \textit{Aggressiveness} \\ 9. & \textit{Lack of Cooperativeness} \\ 10. & \textit{Inferiority Feeling} \\ 11. & \textit{Nervousness} \\ 12. & \textit{Depression} \end{array} \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix},$$

which shows that the six variables *Extroversion*, *Activity*, *Novelty*, *Durability*, *Self-revelation*, and *Aggressiveness* load the first factor. Thus, we can permute the rows and columns of the corresponding covariance matrix, so that those six variables occupy the first six rows in resulting the matrix. It allows us to easily capture what variables characterize the first factor.

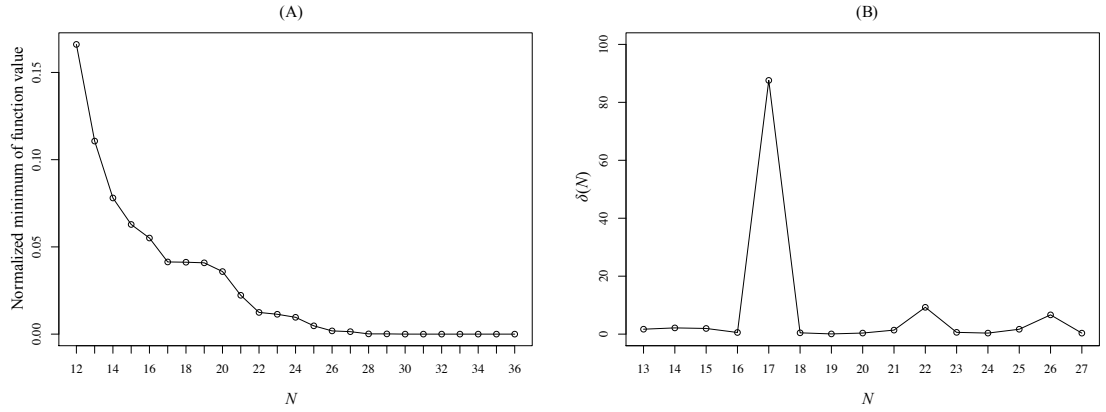



FIGURE 2.2: Attained minimum of function value (A) and $\delta(N)$ for $N = 12 \cdots, 36$ for Yanai & Ichikawa's (2006) loading matrix

## 2.7 Discussion

We proposed the Permutimin rotation, in which the loading matrix **A** is rotated jointly with permutations of the rows of the target matrix **T**, so that the rotated loading matrix optimally approximates the permuted target matrix. This procedure is formulated as minimizing (2.7), i.e., $f(\mathbf{P}, \mathbf{U}, \mathbf{C}) = ||\mathbf{AU} - \mathbf{PT}||^2 = ||\mathbf{AU} - \mathbf{P}(\mathbf{B} \bullet \mathbf{C})||^2$ over the rotation matrix **U**, permutation matrix **P**, and unconstrained **C**, with **T** =

TABLE 2.4: Rotated loading matrices and factor correlations for the box problem by seven rotation procedures; Permutimin (Pmin), Oblimin, Promax, Simplimax (Smax), Geomin, Component Loss Rotation (CLF), and Browne's (1972a) Procrustes Rotation (Browne).

| | | (A) Pmin | | | (B) Oblimin | | | (C) Promax | | | (D) Smax | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | F1 | F2 | F3 | F1 | F2 | F3 | F1 | F2 | F3 | F1 | F2 | F3 |
| | $x$ | **0.98** | 0.03 | 0.03 | **0.44** | **0.39** | **0.73** | **-0.52** | **0.48** | **0.88** | **0.74** | 0.32 | **0.61** |
| | $y$ | 0.02 | **0.96** | 0.03 | -0.17 | **0.98** | **-0.33** | **0.46** | **0.82** | **-0.37** | **0.44** | **0.69** | **-0.45** |
| | $z$ | 0.03 | 0.04 | **0.96** | **0.97** | -0.16 | **-0.34** | **0.84** | **-0.32** | **0.43** | **0.82** | **-0.32** | **-0.46** |
| | $xy$ | **0.61** | **0.67** | 0.00 | 0.11 | **0.91** | 0.23 | -0.04 | **0.86** | 0.26 | **0.72** | **0.68** | 0.08 |
| | $xy^2$ | **0.35** | **0.84** | 0.03 | 0.00 | **0.98** | -0.03 | 0.22 | **0.87** | -0.03 | **0.63** | **0.71** | -0.18 |
| | $2x+2y$ | **0.54** | **0.75** | -0.04 | 0.02 | **0.98** | 0.17 | 0.01 | **0.91** | 0.15 | **0.67** | **0.73** | 0.02 |
| | $(x^2+y^2)^0.5$ | **0.51** | **0.74** | -0.02 | 0.03 | **0.95** | 0.15 | 0.03 | **0.88** | 0.14 | **0.66** | **0.71** | 0.00 |
| loadings | $xz$ | **0.61** | 0.01 | **0.67** | **0.93** | 0.09 | 0.21 | 0.24 | 0.04 | **0.83** | **0.99** | -0.05 | 0.07 |
| | $xz^2$ | **0.42** | -0.03 | **0.82** | **1.02** | -0.05 | 0.03 | **0.46** | -0.13 | **0.75** | **0.97** | -0.20 | -0.10 |
| | $2x+2z$ | **0.58** | -0.02 | **0.71** | **0.97** | 0.04 | 0.19 | 0.27 | -0.01 | **0.84** | **0.99** | -0.10 | 0.05 |
| | $(x^2+z^2)^0.5$ | **0.55** | -0.02 | **0.71** | **0.95** | 0.03 | 0.17 | **0.30** | -0.02 | **0.81** | **0.96** | -0.11 | 0.03 |
| | $yz$ | -0.03 | **0.60** | **0.63** | **0.50** | **0.46** | **-0.46** | **0.85** | 0.26 | 0.00 | **0.75** | 0.19 | **-0.60** |
| | $yz^2$ | -0.01 | **0.43** | **0.77** | **0.68** | 0.26 | **-0.44** | **0.88** | 0.07 | 0.14 | **0.80** | 0.02 | **-0.58** |
| | $2y+2z$ | -0.01 | **0.61** | **0.63** | **0.49** | **0.48** | **-0.45** | **0.84** | 0.28 | 0.00 | **0.76** | 0.20 | **-0.59** |
| | $(y^2+z^2)^0.5$ | 0.01 | **0.61** | **0.59** | **0.47** | **0.50** | **-0.42** | **0.79** | **0.30** | 0.01 | **0.75** | 0.22 | **-0.56** |
| correlation | F1 | 1.00 | | | 1.00 | | | 1.00 | | | 1.00 | | |
| | F2 | 0.19 | 1.00 | | 0.55 | 1.00 | | 0.47 | 1.00 | | 0.00 | 1.00 | |
| | F3 | 0.20 | 0.31 | 1.00 | -0.07 | -0.08 | 1.00 | 0.47 | 0.43 | 1.00 | 0.00 | -0.12 | 1.00 |

| | | (E) Geomin | | | (F) CLF | | | (G) Browne | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | F1 | F2 | F3 | F1 | F2 | F3 | F1 | F2 | F3 |
| | $x$ | **1.49** | 0.00 | **0.88** | **1.30** | 0.01 | **0.63** | **0.98** | 0.03 | 0.03 |
| | $y$ | 0.03 | **0.58** | **-0.75** | 0.04 | **0.55** | **-0.77** | 0.02 | **0.96** | 0.03 |
| | $z$ | 0.07 | **-0.50** | **-0.79** | 0.08 | **-0.53** | **-0.77** | 0.03 | 0.04 | **0.96** |
| | $xy$ | **0.93** | **0.42** | 0.06 | **0.82** | **0.40** | -0.11 | **0.61** | **0.67** | 0.00 |
| | $xy^2$ | **0.54** | **0.50** | **-0.35** | **0.48** | **0.48** | **-0.45** | **0.35** | **0.84** | 0.03 |
| | $2x+2y$ | **0.82** | **0.49** | -0.04 | **0.72** | **0.47** | -0.20 | **0.54** | **0.75** | 0.04 |
| | $(x^2+y^2)^0.5$ | **0.78** | **0.47** | -0.07 | **0.70** | **0.45** | -0.22 | **0.51** | **0.74** | 0.02 |
| loadings | $xz$ | **0.94** | **-0.36** | 0.01 | **0.83** | **-0.38** | -0.12 | **0.61** | 0.01 | **0.67** |
| | $xz^2$ | **0.67** | **-0.47** | -0.25 | **0.60** | **-0.49** | **-0.34** | **0.42** | 0.03 | **0.82** |
| | $2x+2z$ | **0.91** | **-0.40** | -0.02 | **0.80** | **-0.42** | -0.14 | **0.58** | 0.02 | **0.71** |
| | $(x^2+z^2)^0.5$ | **0.85** | **-0.40** | -0.05 | **0.76** | **-0.42** | -0.17 | **0.54** | 0.02 | **0.71** |
| | $yz$ | -0.02 | 0.02 | **-1.01** | 0.00 | -0.01 | **-0.99** | 0.03 | **0.60** | **0.63** |
| | $yz^2$ | 0.00 | -0.16 | **-0.98** | 0.01 | -0.20 | **-0.96** | 0.01 | **0.43** | **0.77** |
| | $2y+2z$ | 0.00 | 0.03 | **-1.00** | 0.02 | 0.00 | **-0.99** | 0.01 | **0.61** | **0.62** |
| | $(y^2+z^2)^0.5$ | 0.04 | 0.05 | **-0.95** | 0.04 | 0.02 | **-0.95** | 0.01 | **0.61** | **0.59** |
| correlation | F1 | 1.00 | | | 1.00 | | | 1.00 | | |
| | F2 | 0.02 | 1.00 | | 0.00 | 1.00 | | 0.19 | 1.00 | |
| | F3 | -0.77 | -0.02 | 1.00 | -0.69 | | | | | |

TABLE 2.5: Rotated loading matrix for Yanai & Ichikawa's (2006) loading matrix obtained by six rotation procedures.

| | | (A) Pmin ($N=17$) | | | (B) Pmin ($N=14$) | | | (C) Pmin ($N=22$) | | | (D) Oblimin | | | (E) Promax | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | F1 | F2 | F3 | F1 | F2 | F3 | F2 | F3 | F1 | F1 | F2 | F3 | F2 | F3 | F1 |
| | Extroversion | **0.47** | -0.19 | 0.14 | **0.44** | **-0.33** | 0.16 | **0.47** | **-0.30** | 0.13 | **0.44** | **-0.33** | 0.16 | **0.47** | **-0.30** | 0.13 |
| | Activity | **0.63** | -0.07 | **0.33** | **0.60** | -0.25 | **0.34** | **0.64** | -0.20 | **0.31** | **0.60** | -0.25 | **0.34** | **0.64** | -0.20 | **0.31** |
| | Empathy | -0.05 | -0.01 | **0.59** | -0.05 | 0.04 | **0.60** | 0.00 | 0.08 | **0.60** | -0.05 | 0.04 | **0.60** | 0.00 | 0.08 | **0.60** |
| | Novelty | **0.61** | 0.13 | 0.03 | **0.58** | -0.07 | 0.01 | **0.58** | -0.04 | 0.00 | **0.58** | -0.07 | 0.01 | **0.58** | -0.04 | 0.00 |
| | Durability | -0.05 | -0.06 | **0.68** | -0.05 | -0.01 | **0.69** | 0.01 | 0.05 | **0.69** | -0.05 | -0.01 | **0.69** | 0.01 | 0.05 | **0.69** |
| loadings | Regularity | 0.08 | 0.11 | **0.73** | 0.08 | 0.12 | **0.72** | 0.14 | 0.19 | **0.73** | 0.08 | 0.12 | **0.72** | 0.14 | 0.19 | **0.73** |
| | Self-revelation | **0.66** | **0.34** | 0.01 | **0.64** | 0.12 | -0.04 | **0.63** | 0.15 | -0.04 | **0.64** | 0.12 | -0.04 | **0.63** | 0.15 | -0.04 |
| | Aggressiveness | **0.60** | **0.55** | -0.11 | **0.59** | **0.34** | -0.20 | **0.56** | **0.36** | -0.18 | **0.59** | **0.34** | -0.20 | **0.56** | **0.36** | -0.18 |
| | Lack of Cooperativeness | 0.20 | **0.51** | -0.10 | 0.22 | **0.42** | -0.17 | 0.18 | **0.43** | -0.14 | 0.22 | **0.42** | -0.17 | 0.18 | **0.43** | -0.14 |
| | Inferiority Feeling | -0.29 | **0.56** | -0.03 | -0.25 | **0.63** | -0.10 | **-0.30** | **0.62** | -0.04 | -0.25 | **0.63** | -0.10 | **-0.30** | **0.62** | -0.04 |
| | Nervousness | -0.02 | **0.73** | **0.37** | 0.01 | **0.74** | 0.27 | 0.00 | **0.77** | **0.33** | 0.01 | **0.74** | 0.27 | 0.00 | **0.77** | **0.33** |
| | Depression | 0.05 | **0.85** | 0.09 | 0.08 | **0.82** | -0.02 | 0.04 | **0.84** | 0.04 | 0.08 | **0.82** | -0.02 | 0.04 | **0.84** | 0.04 |
| correlation | F1 | 1.00 | | | 1.00 | | | 1.00 | | | 1.00 | | | 1.00 | | |
| | F2 | -0.32 | 1.00 | | -0.02 | 1.00 | | -0.03 | 1.00 | | -0.02 | 1.00 | | -0.03 | 1.00 | |
| | F3 | 0.26 | -0.24 | 1.00 | 0.24 | -0.17 | 1.00 | 0.19 | -0.31 | 1.00 | 0.24 | -0.17 | 1.00 | 0.19 | -0.31 | 1.00 |

| | | (F) Smax | | | (G) Geomin | | | (H) CLF | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | F1 | F2 | F3 | F1 | F2 | F3 | F1 | F2 | F3 |
| | Extroversion | **0.66** | -0.13 | 0.06 | **0.48** | -0.25 | 0.12 | **0.48** | -0.28 | 0.18 |
| | Activity | **0.68** | 0.03 | 0.24 | **0.65** | -0.15 | **0.31** | **0.64** | -0.19 | **0.36** |
| | Empathy | 0.00 | -0.03 | **0.59** | -0.01 | -0.02 | **0.58** | -0.01 | 0.00 | **0.59** |
| | Novelty | **0.42** | 0.24 | -0.03 | **0.59** | 0.06 | 0.02 | **0.62** | 0.00 | -0.01 |
| | Durability | 0.06 | -0.08 | **0.66** | 0.00 | -0.07 | **0.66** | 0.00 | -0.05 | **0.68** |
| loadings | Regularity | 0.00 | 0.13 | **0.71** | 0.12 | 0.09 | **0.72** | 0.12 | 0.10 | **0.71** |
| | Self-revelation | 0.23 | **0.49** | -0.04 | **0.63** | 0.27 | 0.01 | **0.62** | 0.20 | 0.00 |
| | Aggressiveness | -0.07 | **0.71** | -0.14 | **0.55** | **0.49** | -0.11 | **0.54** | **0.42** | -0.14 |
| | Lack of Cooperativeness | **-0.39** | **0.59** | -0.08 | 0.16 | **0.49** | -0.09 | 0.16 | **0.46** | -0.14 |
| | Inferiority_Feeling | **-0.89** | **0.55** | 0.06 | **-0.32** | **0.59** | 0.00 | **-0.32** | **0.60** | -0.09 |
| | Nervousness | **-0.81** | **0.79** | **0.43** | -0.04 | **0.72** | **0.39** | -0.03 | **0.72** | 0.29 |
| | Depression | **-0.90** | **0.94** | 0.16 | 0.01 | **0.84** | 0.12 | 0.01 | **0.83** | 0.01 |
| correlation | F1 | 1.00 | | | 1.00 | | | 1.00 | | |
| | F2 | 0.57 | 1.00 | | -0.17 | 1.00 | | -0.05 | 1.00 | |
| | F3 | 0.35 | 0.09 | 1.00 | 0.19 | -0.23 | 1.00 | 0.13 | -0.13 | 1.00 |

$\mathbf{B} \bullet \mathbf{C}$ where $\mathbf{B}$ denotes a known binary matrix. $\mathbf{B}$ can be selected from the set of possible $\mathbf{B}$s, which is specified by the modified version of Thurstone's (1947) definition of simple structure in Section 4, so as to obtain **PAU** of that structure. For selecting a suitable $N$, we can use the scree-plot of the attained function value of (2.7), as illustrated in Section 5.

As found in (2.11) Simplimax is found to be the least restrictive; however, it is not necessarily desirable, as found in Section 5.1, where Permutimin was shown to outperform Simplimax in rotating the box-problem loadings. Simplimax can give a target matrix including an empty column filled with zeros, which may make the corresponding factor irrelevant to any variable in the target. Such a case can be avoided in Permutimin by using $\mathbf{B}$ without an empty column.

## Appendix 2.A Details of Rotation Procedures in Section 6

We carried out Oblimin, Promax, Geomin, and Browne's rotation using R packages with their default settings. For Simplimax, the number of zero elements in the target matrix was set at $15 \times 3 - 27 = 18$ in Section 5.1 and $12 \times 3 - 17 = 19$ using the scree test procedure in which the N maximizing (15) is selected. The component loss rotation was performed using our our own FORTRAN program, in which the linear component loss function (Jennrich, 2006) was minimized with the GP algorithm.

The R program for Permutimin and the compiled FORTRAN program for the component loss rotation can be obtained from the authors upon request.

## Appendix 2.B Multiple Runs Procedure

In order to reduce the possibility of selecting a local minimum as the optimal solution, we use Adachi's (2011) multiple runs procedure in which an algorithm is run multiple times, starting with random initial values until two equivalently best solutions are found. Using $\mathbb{R}_m = \{\mathbf{C}_m, \mathbf{U}_m, \mathbf{P}_m\}$ for the resulting solution from the $m-$th run and $f(\mathbb{R}_m)$ for the attained value of $f(\mathbf{C}_m, \mathbf{U}_m, \mathbf{P}_m)$, the multiple run procedure for Permutimin is stated as follows.

1. Set $M = 50$ and obtain $\mathbb{R}_m$ for $m = 1, \cdots, M$; find $m^* = \underset{1 \leq m \leq M}{\operatorname{argmin}} f(\mathbb{R}_m)$.
2. Select $\mathbb{R}_m$ as the optimal solution to finish if $\mathbf{R}_{m^*}$ is equivalent to $\mathbb{R}_{m^*}$ is equivalent to $\mathbb{S}_l$, resulting from the $l-$th run with $l \neq m^*$; otherwise, go to [3].
3. Set $M := M + 1$ to obtain $\mathbb{R}_M$; exchange $\mathbb{R}_{m^*}$ for $\mathbb{R}_M$ if $f(\mathbb{R}_M) < f(\mathbb{R}_{m^*})$.
4. Select $\mathbb{R}_{m^*}$ as the optimal solution to finish if $\mathbb{R}_{m^*}$ is equivalent to $\mathbb{R}_M$ or $M = 400$; otherwise, go back to [3].

Above, the equivalence of $\mathbb{R}_{m^*}$ to $\mathbb{R}_l$ is defined as $|f(\mathbb{R}_{m^*}) - f(\mathbb{R}_l)| < 10^{-7}$.

## Appendix 2.C Uniqueness of the binary target matrix B for a given $N$

In this appendix, it is explained why [I] the algorithm (Steps[1] - [6]) in Section 3 uniquely specifies $\mathbf{B}$ with an arbitrary $N$ satisfying (2.14) and [II] such $\mathbf{B}$ satisfies the modified Thurstone's rule in Section 4.1.

We firstly explain [I]. The algorithm fails to form the unique $\mathbf{B}$ with a given $N$ if and only if the current $(i, j)$ corresponding to the current $n$ is already included in the set $\mathcal{N}$, at a repetition of the algorithm. In such a case, the current $\mathbf{B}$ is not updated at Step [2] because $b_{ij}$ is already equal to 1, and therefore $\mathbf{B}$ is not uniquely specified. This situation is avoided by Step [5], however, and therefore the algorithm uniquely specifies $\mathbf{B}$ for a given $N$.

Next, we explain [II]. All rows of $\mathbf{B}$ must have at least one and at least one zero as a result of Step [4] and (2.14), and therefore [M1] is satisfied.

Step [3] provides the binary matrix $\mathbf{B}$ in which the number of zero elements is expressed as

$$p - N/r \quad if \ \mathrm{mod}(N/r) = 0 \tag{2.17}$$

or

$$p - \lfloor N/r \rfloor \ or \ p - \lfloor N/r \rfloor + 1 \quad if \ \mathrm{mod}(N/r) > 0, \tag{2.18}$$

where $\mathrm{mod}(N/r)$ and $\lfloor \bullet \rfloor$ denote the remainder resulting from the division of integer $N$ by $r$ and the floor function, respectively. It can be found that those numbers are not less than $r$ if $N$ satisfies (2.14). For the case with (2.17), we have $p - N/r \geq r$ follows from $N \leq r(p - r)$. For the case with (2.18), we also have $p - \lfloor N/r \rfloor \geq r$ and $p - \lfloor N/r \rfloor + 1 \geq r$. This is because $\lfloor N/r \rfloor$ attains its maximum when $N = r(p - r)$. Then, we have $\lfloor N/r \rfloor = p - r$. [M2] is thus satisfied.

The maximum of inter-column products in $\mathbf{B}$ resulting in Steps [3] and [4] is given by 0 if $N \leq p$; otherwise $\lceil (N - p)/r \rceil$, where $\lceil \bullet \rceil$ denotes the ceiling function. These numbers are obviously the minimum of inter-column inner products for all possible $\mathbf{B}$s, and therefore $\mathbf{B}$s formed by the algorithm satisfy [M3].

Finally, the ones in $\mathbf{B}$ given through Step [3] is $\lfloor N/r \rfloor$ or $\lfloor N/r \rfloor + 1$, and the difference between those possible integers is one at maximum. It indicates that $\mathbf{B}$ satisfies [M4].

# Chapter 3

# Rotation in Canonical Correlation Analysis as Maximizing Sum of Squared Correlations

**ABSTRACT**

In this study, canonical correlation analysis, which is known to be difficult to interpret, is re-formulated as maximization of sum of squared correlations, and its equivalence to the existing formulation is shown. Under the new formulation, two structure matrices that are often used in interpretation are allowed to be rotated orthogonally and individually. The study thus proposes an orthogonal rotation method for rotating two structure matrices individually to simple structures, and it is different from the one allowed in the existing formulation. The existing rotation method only allows the simultaneous rotation of two structure matrices, and thus the proposed method further simplifies the matrices. It is illustrated that the proposed method facilitates the interpretation of solutions of canonical correlation analysis by a real data example.

## 3.1 Introduction

Canonical Correlation Analysis (CANO) is a multivariate analysis procedure that composes linear combinations of two sets of observed variables so as to maximize their correlation. The resulting linear combinations are called *canonical variates*, while the correlation between the canonical variates is *canonical correlation*.

While CANO is a classical and fundamental procedure dealing with multiple sets of variables, it has not been applied in real data analysis, mainly because its result is difficult to interpret, as empirically shown in Section 1.2.5.

This chapter is based on the published article Yamashita (2012a) with some modifications and extensions.

Related to the issue, Cliff and Krus (1976) and Reynolds and Jackosfsky (1981) proposed to rotate CANO's solution matrices to simplicity. It is certainly useful and improves the interpretability of CANO's result, but it has a critical limitation; structure matrices corresponding to two sets of variables need to be rotated by a common rotation matrix. In other words, their individual rotation are not allowed. Therefore, it is quite rare that those two matrices are simultaneously simplified.

In order to solve the above problem in CANO and the existing rotation methods, the study firstly proposes a new formulation of CANO that is equivalent to the existing one. The existing formulation of CANO is the maximization of sum of correlation of canonical variates over two weight matrices subject to some suitable constraints, as discussed in Section 1.2.5. The new formulation is to maximize sum of *squared* correlation of the variates under the same constraints, and it is proved to be equivalent to the existing formulation. Under the new formulation, it is shown that two weight matrices have a indeterminacy with respect to independent rotation. The author thus proposes an independent rotation procedure and it is expected to considerably improves the interpretability of CANO's results compared with simultaneous rotation under the existing formulation.

## 3.2 A New Formulation of Canonical Correlation Analysis

The function to be maximized in the existing formulation of CANO is expressed as

$$F_{CA}(\mathbf{W}_1, \mathbf{W}_2) = \mathrm{tr}\mathbf{R} = \mathrm{tr}n^{-1}\mathbf{W}_1'\mathbf{X}_1'\mathbf{X}_2\mathbf{W}_2. \tag{3.1}$$

where $\mathbf{R}$ is the $r \times r$ correlation matrix of canonical variates, and it is called *canonical correlation matrix*. Note that $\mathbf{X}_1\mathbf{W}_1$ and $\mathbf{X}_2\mathbf{W}_2$ are the matrices of $r$ canonical variates in 1st and 2nd variable sets, respectively. It is maximized under the constraint

$$n^{-1}\mathbf{W}_1'\mathbf{X}_1'\mathbf{X}_1\mathbf{A}_1 = n^{-1}\mathbf{W}_2'\mathbf{X}_2'\mathbf{X}_2\mathbf{A}_2 = \mathbf{I}_r \tag{3.2}$$

and the solution is given by given

$$\mathbf{W}_1 = n^{1/2}(\mathbf{X}_1'\mathbf{X}_1)^{-1/2}\mathbf{V}_r,\ \mathbf{W}_2 = n^{-1/2}(\mathbf{X}_2'\mathbf{X}_2)^{1/2}\mathbf{U}_r \tag{3.3}$$

using the SVD

$$\mathbf{M} = (\mathbf{X}_2'\mathbf{X}_2)^{-1/2}\mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1/2} = \mathbf{U}\mathbf{D}\mathbf{V}'. \tag{3.4}$$

See Section 1.2.5 for the derivation of the solution.

In interpreting CANO's result, rather than the weight matrices $\mathbf{W}_1$ and $\mathbf{W}_2$, matrices called *structure matrices* defined as

$$\mathbf{S}_1 = n^{-1}\mathbf{X}_1'\mathbf{X}_1\mathbf{W}_1,\ \mathbf{S}_2 = n^{-1}\mathbf{X}_2'\mathbf{X}_2\mathbf{W}_2 \tag{3.5}$$

is often used (Levine, 1977), because they are the correlation matrix of original variables and canonical variates and thus ranges from -1 to 1.

Here, we have the following theorem with regard to the formulation of CANO.

**Theorem 3.2.1.** *Under the constraint (3.2), we have the inequality*

$$F^*_{CA}(\mathbf{W}_1, \mathbf{W}_2) = ||\mathbf{R}||^2 = ||n^{-1}\mathbf{W}_1'\mathbf{X}_1'\mathbf{X}_2\mathbf{W}_2||^2 \leq \mathrm{tr}\mathbf{D}_r^2 \tag{3.6}$$

*and the inequality holds when (3.3).*

*Proof.* $F^*_{CA}(\mathbf{W}_1, \mathbf{W}_2)$ is expanded as

$$\begin{aligned} F^*_{CA}(\mathbf{W}_1, \mathbf{W}_2) &= \mathrm{tr}\mathbf{W}_1^{*\prime}\mathbf{M}'\mathbf{W}_2^*\mathbf{W}_2^{*\prime}\mathbf{M}\mathbf{W}_1^* \\ &= \mathrm{tr}\mathbf{W}_1^{*\prime}\mathbf{V}\mathbf{D}\mathbf{U}'\mathbf{W}_2^*\mathbf{W}_2^{*\prime}\mathbf{U}\mathbf{D}\mathbf{V}'\mathbf{W}_1^* \\ &= \mathrm{tr}(\mathbf{U}'\mathbf{W}_1^*\mathbf{W}_1^{*\prime}\mathbf{U})\mathbf{D}(\mathbf{V}'\mathbf{W}_2^*\mathbf{W}_2^{*\prime}\mathbf{V})\mathbf{D} \end{aligned} \tag{3.7}$$

with $\mathbf{W}_1^* = (\mathbf{X}_1'\mathbf{X}_1)^{1/2}\mathbf{W}_1$, $\mathbf{W}_2^* = (\mathbf{X}_2'\mathbf{X}_2)^{1/2}\mathbf{W}_2$, and $\mathbf{M}$ having the rank of $r^*(\leq \min(p_1, p_2))$. Since $\mathbf{W}_1^{*\prime}\mathbf{W}_1^* = \mathbf{W}_2^{*\prime}\mathbf{W}_2^* = \mathbf{I}_r$, $\mathbf{W}_1^*$ and $\mathbf{W}_2^*$ are column-orthogonal matrices and their transpose are row-orthogonal. The matrix products $\mathbf{U}'\mathbf{W}_1^*\mathbf{W}_1^{*\prime}\mathbf{U}$ and $\mathbf{V}'\mathbf{W}_2^*\mathbf{W}_2^{*\prime}\mathbf{V}$ are thus sub-orthogonal matrices by Lemma 4 in Ten Berge (1983). See Ten Berge (1983) for the definition of sub-orthogonality. Finally, using Theorem 2 in the same article, it follows that

$$F^*_{CA}(\mathbf{W}_1, \mathbf{W}_2) \leq \mathbf{E}_r\mathbf{D}^2 = \mathbf{D}_r^2 \tag{3.8}$$

where $\mathbf{E}_r$ denotes the $r^* \times r^*$ diagonal matrix whose first $r$ diagonal elements are 1 and zeros elsewhere. The equality holds when

$$\mathbf{U}'\mathbf{W}_1^*\mathbf{W}_1^{*\prime}\mathbf{U} = \mathbf{V}'\mathbf{W}_2^*\mathbf{W}_2^{*\prime}\mathbf{V} = \mathbf{E}_r \tag{3.9}$$

that leads (3.3). □

The theorem indicates that the maximization of $L^*_{CA}(\mathbf{W}_1, \mathbf{W}_2)$, which was newly introduced here, is maximized by the same $\mathbf{W}_1$ and $\mathbf{W}_2$ that also maximizes $L_{CA}(\mathbf{W}_1, \mathbf{W}_2)$, the function maximized in the existing formulation. It leads that CANO can be formulated by the maximization of sum of squared canonical correlations.

## 3.3 Individual Rotation of Structure Matrices

In the existing objective function of CANO, there exists an indeterminacy of simultaneous and orthogonal rotation of solution matrices expressed as

$$F_{CA}(\mathbf{W}_1, \mathbf{W}_2) = \mathrm{tr}n^{-1}\mathbf{W}_1'\mathbf{X}_1'\mathbf{X}_2\mathbf{W}_2 = \mathrm{tr}n^{-1}\mathbf{T}'\mathbf{W}_1'\mathbf{X}_1'\mathbf{X}_2\mathbf{W}_2\mathbf{T} \tag{3.10}$$

where $\mathbf{T}$ denotes an arbitrary $r \times r$ orthonormal matrix that satisfies $\mathbf{T}'\mathbf{T} = \mathbf{T}\mathbf{T}' = \mathbf{I}_r$. The weight matrices are rotated as $\mathbf{W}_1\mathbf{T}$ and $\mathbf{W}_2\mathbf{T}$ toward their simplicity to help

interpretation, by a common rotation matrix $\mathbf{T}$. As to be exemplified in the next section, the simultaneous rotation often fails to simplify $\mathbf{W}_1$ and $\mathbf{W}_2$ at the same time, because the rotation matrices used for rotating them are restricted to be identical.

A rotation matrix $\mathbf{T}$ can be specified so as to maximize sum of the simplicity attained after rotation. Such maximization criterion can be formulated as

$$S(\mathbf{T}) = \alpha s(\mathbf{W}_1\mathbf{T}) + \beta s(\mathbf{W}_2\mathbf{T}) \tag{3.11}$$

where $s(\mathbf{N})$ is a measure of how simple the matrix $\mathbf{N}$ is, and $\alpha$ and $\beta$ are the weight parameters. In this study, as such measure, Varimax criterion (Kaiser, 1958) is considered, and the simultaneous rotation by $\mathbf{T}$ is called *simultaneous Varimax rotation*. An algorithm to maximize $S(\mathbf{T})$ in (3.11) is described in Appendix.

The new formulation of CANO also has another rotational indeterminacy, but it allows to rotate the weight matrices individually;

$$L^*_{CA}(\mathbf{W}_1, \mathbf{W}_2) = ||n^{-1}\mathbf{T}_1'\mathbf{W}_1'\mathbf{X}_2\mathbf{W}_2\mathbf{T}^2||^2 \tag{3.12}$$

with $\mathbf{T}_1$ and $\mathbf{T}_2$ are both satisfy the same constraint as $\mathbf{T}$, and they are used for rotate $\mathbf{W}_1$ and $\mathbf{W}_2$, respectively. The individual rotation is called *individual Varimax rotation*, in that $\mathbf{T}_1$ and $\mathbf{T}_2$ are specified by applying Varimax rotation to $\mathbf{W}_1$ and $\mathbf{W}_2$ in parallel. Individual Varimax rotation is expected to simplify $\mathbf{W}_1$ and $\mathbf{W}_2$ better than the simultaneous one.

As noted above, the structure matrices defined in (3.5) are often used for the interpretation. They also allowed to be rotated individually and simultaneously since

$$\mathbf{S}_1\mathbf{T} = n^{-1}\mathbf{X}_1'\mathbf{X}_1(\mathbf{W}_1\mathbf{T}),\ \mathbf{S}_2\mathbf{T} = n^{-1}\mathbf{X}_2'\mathbf{X}_2(\mathbf{W}_2\mathbf{T}) \tag{3.13}$$

in simultaneous rotation, and

$$\mathbf{S}_1\mathbf{T}_1 = n^{-1}\mathbf{X}_1'\mathbf{X}_1(\mathbf{W}_1\mathbf{T}_1),\ \mathbf{S}_2\mathbf{T}_2 = n^{-1}\mathbf{X}_2'\mathbf{X}_2(\mathbf{W}_2\mathbf{T}_2) \tag{3.14}$$

in individual rotation. It indicates that simultaneous and individual rotations are equally applied to structure matrices $\mathbf{S}_1$ and $\mathbf{S}_2$ as well $\mathbf{W}_1$ and $\mathbf{W}_2$.

## 3.4 Real Data Example

Simultaneous and individual rotations were both applied to Sports data introduced in Section 1.2.5. CANO was applied to the dataset and the obtained structure matrices for the two variables sets were rotated by simultaneous and individual Varimax rotation. The dimension of canonical variates was set at 3, since the singular values of $(\mathbf{X}_2'\mathbf{X}_2)^{-1/2}\mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1/2}$ in was 0.849, 0.715, 0.553, 0.339, and 0.038 in descending order, and a scree point can bet seen at the third singular value. For simultaneous rotation, $(\alpha, \beta) = (1/p_1, 1/p_2) = (1/5, 1/5)$ was used, and it is recommended in Kiers (1998) for standard weights for simultaneous rotation of two or more matrices.

TABLE 3.1: Unrotated and individually/simultaneously rotated structured matrices obtained by CANO to Sports data.

| | Unrotated | | | Individual Varimax | | | Simultaneous Varimax | | |
|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 |
| Test A | | | | | | | | | |
| Side Jump | **-0.670** | 0.278 | -0.187 | **-0.697** | 0.117 | 0.246 | **-0.663** | -0.112 | 0.330 |
| Back Strength | **-0.744** | -0.167 | -0.333 | **-0.733** | 0.359 | -0.160 | -0.183 | **-0.935** | -0.009 |
| Vertical Jump | **-0.670** | **-0.605** | 0.307 | -0.202 | **0.926** | -0.105 | **-0.734** | **-0.388** | -0.049 |
| Grip Strength | **-0.251** | 0.006 | **0.365** | 0.013 | **0.334** | 0.29 | 0.055 | **-0.309** | **0.312** |
| Step Aerobics | -0.172 | **0.773** | **0.526** | -0.015 | -0.103 | **0.944** | 0.084 | 0.179 | **0.929** |
| Test B | | | | | | | | | |
| Long-distance Run | **0.667** | -0.117 | 0.441 | **0.784** | 0.149 | -0.129 | **0.799** | -0.102 | -0.062 |
| Ball Throw | **-0.775** | -0.168 | -0.294 | **-0.688** | **-0.473** | 0.193 | **-0.732** | **0.424** | -0.016 |
| Long Jump | **-0.789** | **-0.497** | 0.038 | -0.387 | **-0.827** | 0.136 | **-0.466** | **0.807** | -0.049 |
| Chin-up | **-0.649** | -0.214 | **0.565** | 0.020 | **-0.640** | **0.614** | -0.066 | **0.779** | **0.418** |
| 50m Run | **0.676** | **-0.362** | **-0.497** | 0.174 | 0.163 | **-0.882** | 0.232 | -0.375 | **-0.800** |
| canonical correlation | 0.849 | | | 0.706 | 0.13 | -0.235 | 0.717 | | |
| matrix | | 0.715 | | -0.002 | 0.261 | 0.612 | -0.099 | 0.675 | |
| | | | 0.553 | -0.020 | -0.687 | 0.229 | -0.106 | -0.016 | 0.725 |

TABLE 3.2: Averaged within-column variance and between covariance of the structure matrices.

| | Averaged Within-column Variance | | Averaged Between-column Covariance | |
|---|---|---|---|---|
| | Test A | Test B | Test A | Test B |
| Unrotated | 0.045 | 0.012 | -0.008 | -0.004 |
| Individual Varimax | 0.116 | 0.089 | -0.043 | -0.043 |
| Simultaneous Varimax | 0.112 | 0.081 | -0.041 | -0.039 |

The resulting two structure matrices are rotated by individual and simultaneous Varimax rotation, and their results are shown in Table 3.1. An element is bolded when its absolute value exceeds 1) 0.2 or 2) half of the largest-in-absolute value in its row. The unrotated structured matrices are difficult to interpret in that the first canonical variates highly correlate with all variables. Such canonical variate can be named *general athletic ability*, but it is useless when the purpose of CANO is to extract sub-concepts of the general athletic ability.

The rotated matrices are highly interpretable compared with the unrotated ones, because the canonical variates highly loads to the fewer variables, and thus these variates are understood as the sub-concepts of the athletic ability that the two tests aim to measure. Their relationships are expressed in the canonical correlation matrices.

Further, the attained simplicity of the individually and simultaneously-rotated structure matrices was evaluated by averaged within-column variance and averaged between-column covariance. The two metrics stand for simplicity and complexity, respectively, and their values are listed in Table 3.2. For both rotation, it is observed that the within-column variance was increased while the between-column covariance decreased, and thus the structure matrices are simplified. In addition, the increment of simplicity and the decrement of complexity is larger in individual rotation. It indicates that the individual rotation performs better than the simultaneous one, because it can freely rotate each of the matrices. Also, the successful result implies that potential effectiveness of the new formulation of CANO that enables individual rotation.

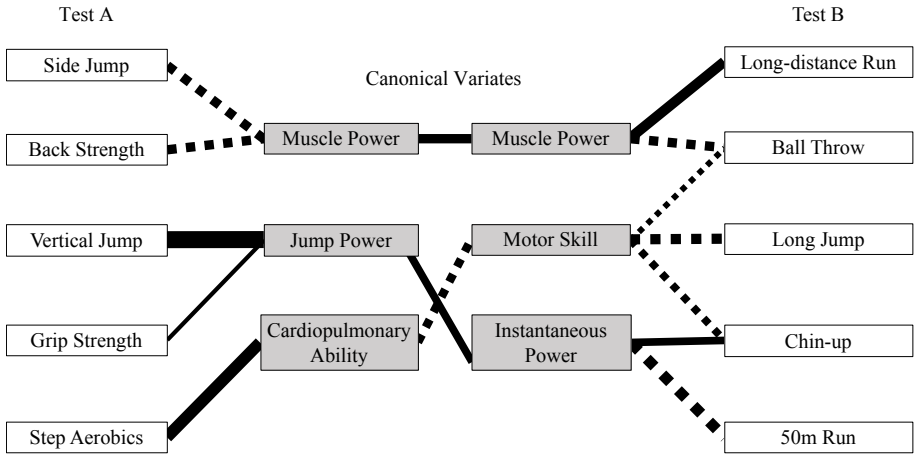



FIGURE 3.1: Network representation of the independently rotated structure matrices.

After rotation, a canonical correlation matrix is no longer diagonal. It is shown as, after simultaneous rotation

$$\mathbf{R} = \mathbf{T}'\mathbf{A}'(n^{-1}\mathbf{X}_1'\mathbf{X}_2)\mathbf{W}_2\mathbf{T} \tag{3.15}$$

it is a symmetric matrix, and after individual rotation

$$\mathbf{R} = \mathbf{T}_1'\mathbf{A}'(n^{-1}\mathbf{X}_1'\mathbf{X}_2)\mathbf{W}_2\mathbf{T}_1 \tag{3.16}$$

a regular square matrix. The correlation structure between two sets of canonical variates is relaxed after rotation and it is need to be interpreted carefully. To interpret it, the following network representation of CANO's result would be of great help. It is shown in the Figure 3.1 and width of the paths is proportional to the absolute value of the corresponding elements. Negative paths are shown in dashed lines. The elements less than 0.3 in absolute are omitted in the figure. The correspondence between the canonical variates and the variables are clearly captured because of the individual rotation, and the variates can be named shown in the figure. Further, as noted above, the correlation structure between the canonical variates are not diagonal nor symmetric. Taking in to account that column-wise flipping of signs in structure matrices is allowed, the resulting correlation structure expressed in the figure seems to be appropriate, in that *Jump Power* and *Instantaneous Power*, and *Cardiopulmonary Ability* and *Instantaneous Power* are both positively correlated.

## 3.5 Discussion

The study proposes a new formulation of CANO as maximization of sum of squared canonical correlations with its equivalence to the existing one, in which sum of the correlations is maximized. Under the new formulation, it is allowed to orthogonally

rotate structure matrices toward their simplicity individually, while only simultaneous rotation is allowed under the existing formulation. The superiority of individual rotation over simultaneous rotation is exemplified by a real data example. The result shows that the individual rotation increases the simplicity of structure matrices and helps to interpret the canonical variates.

As demonstrated in Section 3.4, individual rotation leads a non-diagonal and non-symmetric canonical correlation matrix. The property allows various correlation structure between the canonical variates, but sometimes difficult to interpret, because the rotation only considers the simplicity of the structure matrices. Despite the problem, rotation would be preferred especially in the exploratory use of CANO, because it greatly improves the interpretability of CANO's solution. To remedy the difficulty, it is recommended to use small $r$ as possible in order to reduce the dimensionality of $\mathbf{R}$ and the number of correlations that have be interpreted.

## Appendix 3.A An Algorithm for Simultaneous Varimax Rotation

An algorithm for minimizing (3.11) to specify the rotation matrix $\mathbf{T}$.

For the function $s(\cdot)$, a measure for simplicity of a matrix, the Varimax criterion is considered here. The Varimax criterion for a $p \times r$matrix $\mathbf{A} = \{a_{ij}\}$ is expressed as (Kaiser, 1958)

$$v(\mathbf{A}) = \sum_{i=1}^{p} \left\{ \sum_{j=1}^{r} a_{ij}^4 - \frac{1}{r} \left( \sum_{j=1}^{r} a_{ij} \right)^2 \right\} \tag{3.17}$$

and it is further rewritten as follows (Sherin, 1966; Neudecker, 1981);

$$v(\mathbf{A}) = \mathrm{tr}(\mathbf{A} \bullet \mathbf{A})' \mathbf{P}_p (\mathbf{A} \bullet \mathbf{A}) \tag{3.18}$$

where $\bullet$ denotes the Hadamard product and $\mathbf{P}_p = \mathbf{I}_p - p^{-1}\mathbf{1}_p\mathbf{1}_p'$. The minimization criterion $S(\mathbf{T})$ in (3.11) is thus rewritten as follows, using Result 2 in Neudecker (1981);

$$\begin{aligned} S &= \alpha \mathrm{tr}(\tilde{\mathbf{W}}_1 \bullet \tilde{\mathbf{W}}_1)' \mathbf{P}_{p_1} (\tilde{\mathbf{W}}_1 \bullet \tilde{\mathbf{W}}_1) + \beta \mathrm{tr}(\tilde{\mathbf{W}}_2 \bullet \tilde{\mathbf{W}}_2)' \mathbf{P}_{p_2} (\tilde{\mathbf{W}}_2 \bullet \tilde{\mathbf{W}}_2) \\ &= \mathrm{tr} \frac{\alpha}{r} [\{(\tilde{\mathbf{W}}_1 \bullet \tilde{\mathbf{W}}_1)' \mathbf{P}_{p_1}\} \bullet \tilde{\mathbf{W}}_1'] \mathbf{W}_1 \mathbf{T} + \mathrm{tr} \frac{\beta}{r} [\{(\tilde{\mathbf{W}}_2 \bullet \tilde{\mathbf{W}}_2)' \mathbf{P}_{p_2}\} \bullet \tilde{\mathbf{W}}_2'] \mathbf{W}_2 \mathbf{T} \\ &= \mathrm{tr}(\mathbf{B}_1 + \mathbf{B}_2)\mathbf{T} \end{aligned} \tag{3.19}$$

where $\tilde{\mathbf{W}}_1$ and $\tilde{\mathbf{W}}_2$ are the rotated matrices by $\mathbf{T}$, and $\mathbf{B}_1 = \frac{\alpha}{r}[\{(\tilde{\mathbf{W}}_1 \bullet \tilde{\mathbf{W}}_1)' \mathbf{P}_{p_1}\} \bullet \tilde{\mathbf{W}}_1'] \mathbf{W}_1$ and $\mathbf{B}_2 = \frac{\alpha}{r}[\{(\tilde{\mathbf{W}}_2 \bullet \tilde{\mathbf{W}}_2)' \mathbf{P}_{p_2}\} \bullet \tilde{\mathbf{W}}_2'] \mathbf{W}_2$. The maximum of $S(\mathbf{T})$ is attained by using the SVD $\tilde{\mathbf{B}}_1 + \tilde{\mathbf{B}}_2 = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}'$ as follows

$$S(\mathbf{T}) = \mathrm{tr}\mathbf{U}\boldsymbol{\Delta}\mathbf{V}' \leq \mathrm{tr}\boldsymbol{\Delta}. \tag{3.20}$$

The equality holds when

$$\mathbf{T} = \mathbf{V}\mathbf{U}' \tag{3.21}$$

Since the matrix $\tilde{\mathbf{B}}_1 + \tilde{\mathbf{B}}_2$ includes $\mathbf{W}_1$ and $\mathbf{W}_2$ before rotation, $S(\mathbf{T})$ is maximized by the following iterative algorithm.

1. Initialize $\mathbf{T}$.
2. Rotate $\mathbf{W}_1$ and $\mathbf{W}_2$ as $\tilde{\mathbf{W}}_1 = \mathbf{W}_1\mathbf{T}$ and $\tilde{\mathbf{W}}_2 = \mathbf{W}_2\mathbf{T}$.
3. Update $\mathbf{T}$ by (3.21).
4. Repeat 2. and 3. until convergence.

Note that in the example in Section 3.4, the structure matrices $\mathbf{S}_1$ and $\mathbf{S}_2$ were rotated, and the above algorithm was run with $\mathbf{W}_1 \leftarrow \mathbf{S}_1$ and $\mathbf{W}_2 \leftarrow \mathbf{S}_2$.

# Chapter 4

# Biplot Procedures with Joint Classification Objects and Variables

**ABSTRACT**

Biplot is a technique for obtaining a low-dimensional configuration of the data matrix but biplots with a large number of objects and variables remain difficult to interpret. The study proposes a new biplot procedure with high interpretability. In particular, the objects and variables are classified into a small number of clusters by using $K$-means clustering and the resulting clusters are simultaneously biplotted in lower-dimensional space. Further, it is extended to fuzzy clustering, where the objects and variables are fuzzily classified. A simulation study and real data example are also provided to demonstrate the effectiveness of the proposed procedure.

## 4.1 Introduction

Biplot (Gabriel, 1971; Gower and Hand, 1995; Gower et al., 2011) is a multivariate technique for visualizing a data matrix that provides low-dimensional configurations and where objects and variables are represented as points and vectors, respectively. In this research, a row vector is denoted with a subscript with parentheses and a column vector is expressed by a subscript with no parentheses. That is, $\mathbf{f}_{(i)}$ is the $i$-th row vector of $\mathbf{F}$, while $\mathbf{f}_j$ denotes its $j$-th column vector. Suppose that we have an $I$ (objects) $\times$ $J$ (variables) data matrix $\mathbf{X}$. One of the ways in which to obtain the biplot of $\mathbf{X}$, often referred to as a PC biplot, is to decompose $\mathbf{X}$ into the product of two low-rank matrices, say $\mathbf{F}$ and $\mathbf{A}$. After performing principal component analysis on $\mathbf{X}$, plot each row of the principal component matrix $\mathbf{F} = \{\mathbf{f}'_{(1)}, \cdots, \mathbf{f}'_{(I)}\}'(I \times r)$

This chapter is based on the published article Yamashita and Mayekawa (2012) and Yamashita and Mayekawa (2015) with some modifications and extensions.

and the loadings matrix $\mathbf{A} = \{\mathbf{a}'_{(1)}, \cdots, \mathbf{a}'_{(J)}\}'(J \times r)$ in $r$-dimensional space to obtain the biplot of $\mathbf{X}$, where

$$\mathbf{F} = \mathbf{K}_r, \ \mathbf{A} = \mathbf{L}_r \mathbf{\Lambda}_r \tag{4.1}$$

Here, $\mathbf{\Lambda}_r$ is an $r \times r$ diagonal matrix whose diagonal elements are the $r$ largest singular values of $\mathbf{X}$ arranged in descending order and $\mathbf{K}_r$ and $\mathbf{L}_r$ are the matrices of the corresponding left and right singular vectors, respectively. The dimension of the solution is denoted as $r$, which is often set to two or three to improve the convenience of interpretation. The rows of $\mathbf{F}$ and $\mathbf{A}$ represent the coordinates of the corresponding objects and variables and the $(i, j)$ element of $\mathbf{X}$, $x_{ij}$, is approximated by the inner product $\mathbf{f}_{(i)}\mathbf{a}'_{(j)}$. In other words, a PC biplot can be interpreted as a method that decomposes the data matrix $\mathbf{X}$ by minimizing the least squares criterion:

$$L_{PCA}(\mathbf{F}, \mathbf{A}) = ||\mathbf{X} - \mathbf{F}\mathbf{A}'||^2 \tag{4.2}$$

with respect to F and A and the plots of their rows in r-dimensional space.

Biplot is a useful technique for visualizing the relationships between the objects and variables in a dataset, resulting in many data visualization applications (Greenacre, 2010). An example of the PC biplot of an artificial dataset composed of 15 objects and five variables is presented in Figure 4.1a. In the PC biplot obtained by using the method above, the correlation between the $m$-th and $j$-th ($m \neq j$) variables is approximated by the cosine of the angle between the vectors corresponding to those variables. In addition, in Figure 4.1a, the length of the $m$-th vector is set proportional to the standard deviation of the $m$-th variable. From these properties of PC biplots, the following interpretation of the biplot shown in Figure 4.1a is possible: since the vectors of the 1st, 2nd, and 3rd variables and those of the 5th and 6th variables are almost orthogonal, the data consist of two almost independent groups of variables. As exemplified here, a biplot provides us with a visual interpretation of the relationships between objects and variables. In this research, without loss of generality, we assume that the data matrix $\mathbf{X}$ is of full-column rank, that is, $\text{rank}(\mathbf{X}) = J$, where $\text{rank}(\cdot)$ denotes the rank of a parenthesized matrix.

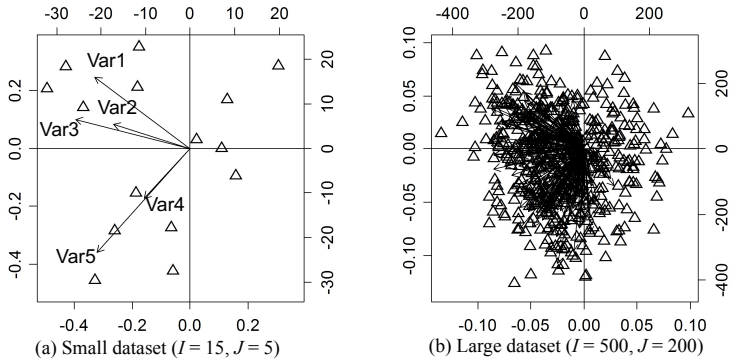



FIGURE 4.1: PC biplots of two artificial datasets

The difficulty of interpreting biplots often arises when $\mathbf{X}$ contains a large number of objects and variables. The PC biplot of a large dataset that has 500 objects and 200 variables is illustrated in Figure 4.1b. Compared with the biplot of the relatively small dataset shown in Figure 4.1a, the PC biplot of the large dataset is quite difficult to understand since its configuration is composed of many points and vectors. As explained here, the existing biplot procedure for obtaining a biplot is thought to be unsuitable for datasets consisting of many objects and variables. One way in which to overcome this difficulty with large datasets is to perform cluster analysis on $\mathbf{F}$ and $\mathbf{A}$ and obtain a biplot of the resulting object and variable clusters. Specifically, if we use the $K$-means clustering procedure, $\mathbf{F}$ and $\mathbf{A}$ are approximated as

$$\mathbf{F} \approx \mathbf{UP}, \ \mathbf{A} \approx \mathbf{VQ} \tag{4.3}$$

with $\mathbf{U}$ $(I \times K)$ and $\mathbf{V}$ $(J \times L)$ being dummy matrices indicating the inclusion/exclusion of row objects in each of the row clusters and $\mathbf{P}$ $(K \times r)$ and $\mathbf{Q}$ $(L \times r)$ matrices of row cluster means. Therefore, instead of $(\mathbf{F}, \mathbf{A})$, the biplot of $(\mathbf{P}, \mathbf{Q})$ should provide us with a more concise and interpretable result. However, this type of approach, called tandem analysis by Arabie and Hubert (1994), in which clustering is performed according to the results of the preceding analysis, has been criticized in some studies. First, this approach often fails to detect the hidden but important cluster of the data, as exemplified by Vichi and Kiers (2001). Second, the solution obtained by using a certain technique is not guaranteed to optimize two different criteria (Hwang et al., 2010).

This research proposes a new biplot procedure termed as *crisp cluster-wise biplot analysis* (CCBA), as an alternative to tandem analysis. It is defined as the minimization of the least squares criterion:

$$L_{CCBA}(\mathbf{U}, \mathbf{V}, \mathbf{P}, \mathbf{Q}) = ||\mathbf{X} - \mathbf{UPQ}'\mathbf{V}'||^2 \tag{4.4}$$

with respect to all the parameters, namely, $\mathbf{U}, \mathbf{V}, \mathbf{P}$, and $\mathbf{Q}$subject to certain constraints. The $I$ objects and $J$ variables are crisply classified into $K$ and $L$ clusters, respectively, and resulting $K$ and $L$ clusters are jointly plotted as a biplot. The configuration is achieved by plotting the rows of $\mathbf{P}$ and $\mathbf{Q}$. The biplot is thus composed of lesser points and vectors, and therefore easier to interpret compared with the existing biplot.

CCBA has the following two potential disadvantages. First, a local minimum frequently occurs in the practical use of CCBA because the CCBA algorithm involves updating the binary matrices $\mathbf{U}$ and $\mathbf{V}$. Second, the binary constraint imposed on the memberships seems to be too restrictive for dealing with real-world problems. These disadvantages of CCBA are dramatically remedied by generalizing CCBA to fuzzy clustering, where the memberships of objects and variables are represented in a stochastic manner. To overcome the problems of CCBA, this research also proposes an extension of CCBA where objects and variables are fuzzily classified by fuzzy $K$-means

clustering.

## 4.2 Proposed Method

### 4.2.1 Crisp Cluster-wise Biplot Analysis

CCBA is formulated by the minimization of the loss function in (4.4) subject to the constraint that

$$I^{-1}\mathbf{P}'\mathbf{U}'\mathbf{U}\mathbf{P} = \mathbf{I}_r,\ I^{-1}\mathbf{Q}'\mathbf{V}'\mathbf{V}\mathbf{Q} = \text{diag}. \tag{4.5}$$

First, consider to minimize (4.4) over $\mathbf{U}$ with other parameters kept fixed. The loss function is rewritten as

$$L_{CCBA}(\mathbf{U}|\mathbf{P},\mathbf{Q},\mathbf{V}) = ||\mathbf{X} - \mathbf{U}(\mathbf{P}\mathbf{Q}'\mathbf{V}')||^2 = L_{KM}(\mathbf{U},\mathbf{P}\mathbf{Q}'\mathbf{V}') \tag{4.6}$$

which is equivalent to the loss function of $K$-means clustering to $\mathbf{X}$. $L_{KM}(\mathbf{U},\mathbf{P}\mathbf{Q}'\mathbf{V}')$ is minimized over the membership matrix $\mathbf{U}$ by (1.68) with fixed centroid matrix $\mathbf{P}\mathbf{Q}'\mathbf{V}'$. In the similar manner, (4.4) is rewritten as

$$L_{CCBA}(\mathbf{U}|\mathbf{P},\mathbf{Q},\mathbf{V}) = ||\mathbf{X}' - \mathbf{V}(\mathbf{Q}\mathbf{P}'\mathbf{U}')||^2 = L_{KM}(\mathbf{V},\mathbf{Q}\mathbf{P}'\mathbf{U}'). \tag{4.7}$$

Thus, (4.4) is minimized over $\mathbf{V}$ by the $K$-means clustering to $\mathbf{X}'$.

Next, (1.68) is minimized over $\mathbf{P}$ and $\mathbf{Q}$ with fixed $\mathbf{U}$ and $\mathbf{V}$. $\mathbf{V}$ is obtained by (1.68) with fixed centroid matrix $\mathbf{Q}\mathbf{P}'\mathbf{U}'$. The loss function is expanded as

$$\begin{aligned} f_{CCBA}(\mathbf{P},\mathbf{Q}|\mathbf{U},\mathbf{V}) &= \text{tr}\mathbf{X}'\mathbf{X} - 2\text{tr}\mathbf{X}'\mathbf{U}\mathbf{P}\mathbf{Q}'\mathbf{V}' + \text{tr}\mathbf{P}'\mathbf{U}'\mathbf{U}\mathbf{P}\mathbf{Q}'\mathbf{V}'\mathbf{V}\mathbf{Q} && (4.8)\\ &= ||\mathbf{H} - (\mathbf{U}'\mathbf{U})^{1/2}\mathbf{P}\mathbf{Q}'(\mathbf{V}'\mathbf{V})^{1/2}||^2 + ||\mathbf{X}||^2 - ||\mathbf{H}||^2 && (4.9)\end{aligned}$$

using $\mathbf{H} = (\mathbf{U}'\mathbf{U})^{-1/2}\mathbf{U}'\mathbf{X}\mathbf{V}(\mathbf{V}'\mathbf{V})^{-1/2}$ (ten Berge, 1993). Thus, minimizing (4.4) over $\mathbf{P}$ and $\mathbf{Q}$ is equivalent to minimizing the first term of (4.9). It is accomplished by

$$\mathbf{P} = I^{1/2}(\mathbf{U}'\mathbf{U})^{-1/2}\mathbf{K}_r^\sharp\mathbf{T},\ \mathbf{Q} = I^{1/2}(\mathbf{V}'\mathbf{V})^{-1/2}\mathbf{L}_r^\sharp\mathbf{\Lambda}_r^\sharp\mathbf{T} \tag{4.10}$$

with the singular value decomposition

$$\mathbf{H} = \mathbf{K}^\sharp\mathbf{\Lambda}^\sharp\mathbf{L}^{\sharp\prime} \tag{4.11}$$

(Ten Berge, 1983). where $\mathbf{T}$ is an $r \times r$ orthonormal matrix. The matrix $\mathbf{U}'\mathbf{U}$ and $\mathbf{V}'\mathbf{V}$ are the diagonal matrices of the size of $K$ object clusters and $L$ variables clusters, respectively. Therefore, $(\mathbf{U}'\mathbf{U})^{1/2}$ and $(\mathbf{V}'\mathbf{V})^{1/2}$ exist if no empty cluster exists. Properties and recommendations for implementation are detailed in Yamashita and Mayekawa (2012).

In the next section, CCBA is therefore extended to fuzzy classification of objects and variables.

### 4.2.2 Fuzzy $K$-means Clustering

As noted in Chapter 1, fuzzy $K$-means clustering is formulated as minimizing (1.71). Here, an algorithm for fuzzy $K$-means clustering, which is a generalization of $K$-means clustering, is presented in detail. The membership matrix $\mathbf{M}$ and centroid matrix $\mathbf{C}$, which minimize the loss function, can be obtained through the following algorithm (Gan et al., 2007):

1. Given the number of clusters $K$, randomly generate $\mathbf{M}$.
2. With $\mathbf{M}$ fixed, update the $k$-th row of $\mathbf{C}$ with $c_{(k)} = \frac{\sum_i m_{ik}^q \mathbf{x}_{(i)}}{\sum_i m_{ik}^q}$, for $k = 1, \cdots, K$.
3. With $\mathbf{C}$ fixed, update $\mathbf{M}$ with

$$m_{ik} = \left\{ \sum_{l=1}^{K} \left( \frac{||\mathbf{x}_{(i)} - \mathbf{c}_{(k)}||}{||\mathbf{x}_{(i)} - \mathbf{c}_{(l)}||} \right)^{\frac{2}{q-1}} \right\}^{-1} \tag{4.12}$$

4. Repeat 2. and 3. until convergence.

The above formulation can be extended to a matrix-based one. Here, let $\mathbf{D}_\mathbf{M} = \mathrm{diag}(\mathrm{vec}(\mathbf{M}^{\frac{q}{2}}))$ be an $IK \times IK$ matrix with $\mathbf{M}^{\frac{q}{2}} = \{m_{ik}^{\frac{q}{2}}\}$ and $\mathrm{vec}(\cdot)$ being the column-wise vectorization of the parenthesized matrix by stacking its column vectors vertically. (1.71) can be rewritten as the following form:

$$L_{FKM}(\mathbf{M}, \mathbf{C}|\mathbf{X}) = ||\mathbf{D}_\mathbf{M}\{(\mathbf{1}_K \otimes \mathbf{I}_I)\mathbf{Y} - (\mathbf{I}_K \otimes \mathbf{1}_I)\mathbf{C}\}||^2 \tag{4.13}$$

where $\mathbf{1}_I$ is the $I$-dimensional vector filled with 1s, and $\otimes$ denotes the Kronecker product. In fuzzy $K$-means clustering, we may interpret the $(i, k)$ element of the $\mathbf{M}$ matrix as the posterior probability of object $i$ belonging to the $k$-th cluster in the finite mixture model based clustering. In maximum likelihood clustering, the clusters are crisp but the posterior probabilities are not. In fuzzy $K$-means clustering, the fuzziness is directly expressed as the model parameter $\mathbf{M}$.

### 4.2.3 Fuzzy Cluster-wise Biplot Analysis

In this subsection, *fuzzy cluster-wise biplot analysis (FCBA)* as a generalization of CCBA is proposed. FCBA for the data matrix $\mathbf{X}$ is formulated as minimizing the following least squares loss function:

$$L_{FCBA}(\mathbf{U}, \mathbf{V}, \mathbf{P}, \mathbf{Q}) = \sum_{i}^{I} \sum_{j}^{J} \sum_{k}^{K} \sum_{l}^{L} u_{ik}^{\alpha} v_{jk}^{\beta} (x_{ij} - \mathbf{p}_{(k)} \mathbf{q}_{(l)}')^2 \tag{4.14}$$

with respect to the parameter matrices $\mathbf{U} = \{u_{ik}\}$, $\mathbf{V} = \{v_{jl}\}$, $\mathbf{P} = \{\mathbf{p}_{(1)}', \cdots, \mathbf{p}_{(K)}'\}'$, and $\mathbf{Q} = \{\mathbf{q}_{(1)}', \cdots, \mathbf{q}_{(K)}'\}'$ subject to $\mathrm{rank}(\mathbf{PQ}') = r \leq \min(K, L)$. Note that $\mathbf{U}$ and $\mathbf{V}$ are the fuzzy membership matrices of the objects and variables, and thus they

satisfies that

$$\sum_k u_{ik} = 1,\ 0 \leq u_{ik} \leq 1 \tag{4.15}$$

$$\sum_l v_{jl} = 1,\ 0 \leq v_{jl} \leq 1. \tag{4.16}$$

$\mathbf{P}$ and $\mathbf{Q}$ coordinate matrices whose rows are plotted in $r$-dimensional space associated with the object and variable clusters, respectively. In addition, $\alpha(>1)$ and $\beta(>1)$ are the constants that control the fuzziness of the resulting object and variable clusters, respectively. In the proposed procedure, $I$ objects and $J$ variables are jointly classified into $K(\leq I)$ object clusters and $L(\leq J)$ variable clusters, making the resulting biplots of those clusters simpler and easier to understand than the PC biplot of $\mathbf{X}$, especially when the size of $\mathbf{X}$ is quite large. Hereafter the resulting biplot is called as a *fuzzy cluster-wise (FC) biplot*.

Below, ALS procedure for estimating FCBA's parameters is derived. First, the loss function is minimized over $\mathbf{U}$ with other parameters kept fixed. The loss function is rewritten as

$$L_{FCBA}(\mathbf{U},\mathbf{V},\mathbf{P},\mathbf{Q}|\mathbf{X}) = ||\mathbf{D_U}\{(\mathbf{1}_K\otimes\mathbf{I}_I)\mathbf{X}(\mathbf{1}_L\otimes\mathbf{I}_J)'\mathbf{D_V}-(\mathbf{I}_K\otimes\mathbf{1}_I)\mathbf{P}\mathbf{Q}'(\mathbf{I}_L\otimes\mathbf{1}_J)'\mathbf{D_V}\}||^2 \tag{4.17}$$

where $\mathbf{D_U}$ $(IK\times IK)$ and $\mathbf{D_V}$ $(JL\times JL)$ are diagonal matrices whose diagonal elements are $\mathrm{vec}(\mathbf{U}^{\frac{\alpha}{2}})$ and $\mathrm{vec}(\mathbf{V}^{\frac{\beta}{2}})$ with $\mathbf{U}^{\frac{\alpha}{2}} = \{u_{ik}^{\frac{\alpha}{2}}\}$ and $\mathbf{V}^{\frac{\beta}{2}} = \{u_{ik}^{\frac{\beta}{2}}\}$, respectively. Compared with (4.13), (4.17) can be regarded as the least squares criterion of the fuzzy $K$-means clustering of $\mathbf{X}(\mathbf{1}_L\otimes\mathbf{I}_J)'\mathbf{D_V}$, say

$$L_{FCBA}(\mathbf{U}|\mathbf{V},\mathbf{P},\mathbf{Q},\mathbf{X}) = L_{FKM}(\mathbf{U},\mathbf{P}\mathbf{Q}'(\mathbf{I}_L\otimes\mathbf{1}_J)'\mathbf{D_V}|\mathbf{X}(\mathbf{1}_L\otimes\mathbf{I}_J)'\mathbf{D_V}) \tag{4.18}$$

and thus, the fuzzy membership matrix $\mathbf{U}$ that minimizes (4.14) can be obtained by applying fuzzy $K$-means clustering to $\mathbf{X}(\mathbf{1}_L\otimes\mathbf{I}_J)'\mathbf{D_V}$ with the fixed centroid $\mathbf{P}\mathbf{Q}'(\mathbf{I}_L\otimes\mathbf{1}_J)'\mathbf{D_V}$, according to the third step of the algorithm described above.

Likewise, since (4.14) can also be rewritten as

$$L_{FCBA}(\mathbf{V}|\mathbf{U},\mathbf{P},\mathbf{Q},\mathbf{X}) = L_{FKM}(\mathbf{V},\mathbf{Q}\mathbf{P}'(\mathbf{I}_K\otimes\mathbf{1}_I)'\mathbf{D_U}|\mathbf{X}'(\mathbf{1}_K\otimes\mathbf{I}_I)'\mathbf{D_U}). \tag{4.19}$$

$\mathbf{V}$ minimizing the loss function can be obtained by the fuzzy $K$-means to $\mathbf{X}'(\mathbf{1}_K\otimes\mathbf{I}_I)'\mathbf{D_U}$ with the fixed centroid $\mathbf{Q}\mathbf{P}'(\mathbf{I}_K\otimes\mathbf{1}_I)'\mathbf{D_U}$.

Finally (4.14) can be rewritten as

$$\begin{aligned} & L_{FCBA}(\mathbf{U},\mathbf{V},\mathbf{P},\mathbf{Q}|\mathbf{X}) \\ = \ & ||\mathbf{H}^* - \{(\mathbf{I}_K\otimes\mathbf{1}_I)'\mathbf{D}_\mathbf{U}^2(\mathbf{I}_K\otimes\mathbf{1}_I)\}^{1/2}\mathbf{P}\mathbf{Q}'\{(\mathbf{I}_J\otimes\mathbf{1}_L)'\mathbf{D}_\mathbf{V}^2(\mathbf{I}_J\otimes\mathbf{1}_L)\}^{1/2}||^2 \\ - \ & ||\mathbf{H}^*||^2 + ||\mathbf{D_U}(\mathbf{1}_K\otimes\mathbf{I}_I)\mathbf{X}(\mathbf{1}_L\otimes\mathbf{I}_J)'\mathbf{D_V}||^2 \end{aligned} \tag{4.20}$$

using

$$\mathbf{H}^* = \{(\mathbf{I}_K \otimes \mathbf{1}_I)'\mathbf{D}_\mathbf{U}^2(\mathbf{I}_K \otimes \mathbf{1}_I)\}^{-1/2}(\mathbf{I}_K \otimes \mathbf{1}_I)'\mathbf{D}_\mathbf{U}^2(\mathbf{1}_K \otimes \mathbf{I}_I)\mathbf{X} \\ (\mathbf{1}_L \otimes \mathbf{I}_J)'\mathbf{D}_\mathbf{V}^2(\mathbf{I}_L \otimes \mathbf{1}_J)\{(\mathbf{I}_L \otimes \mathbf{1}_J)'\mathbf{D}_\mathbf{V}^2(\mathbf{I}_L \otimes \mathbf{1}_J)\}^{-1/2}. \tag{4.21}$$

(ten Berge, 1993) The minimization of (4.14) over $\mathbf{P}$ and $\mathbf{Q}$ are therefore equivalent to minimizing $||\mathbf{H}^* - \{(\mathbf{I}_K \otimes \mathbf{1}_I)'\mathbf{D}_\mathbf{U}^2(\mathbf{I}_K \otimes \mathbf{1}_I)\}^{1/2}\mathbf{P}\mathbf{Q}'\{(\mathbf{I}_J \otimes \mathbf{1}_L)'\mathbf{D}_\mathbf{V}^2(\mathbf{I}_J \otimes \mathbf{1}_L)\}^{1/2}||^2$ over the parameters. Let us define the singular value decomposition of the $K \times L$ matrix $\mathbf{H}^*$ having rank $h(\leq min(K, L))$ as

$$\mathbf{H}^* = \mathbf{K}^*\mathbf{\Lambda}^*\mathbf{L}^{*\prime}. \tag{4.22}$$

By using this decomposition, P and Q that minimize g(P,Q) are obtained as below (Ten Berge, 1983):

$$\mathbf{P} = \{(\mathbf{I}_K \otimes \mathbf{1}_I)'\mathbf{D}_\mathbf{U}^2(\mathbf{I}_K \otimes \mathbf{1}_I)\}^{-1/2}\mathbf{K}_r^*\mathbf{\Lambda}_r^{*1/2}\mathbf{D}\mathbf{T}, \tag{4.23}$$

$$\mathbf{Q} = \{(\mathbf{I}_L \otimes \mathbf{1}_J)'\mathbf{D}_\mathbf{V}^2(\mathbf{I}_L \otimes \mathbf{1}_J)\}^{-1/2}\mathbf{L}_r^*\mathbf{\Lambda}_r^{*1/2}\mathbf{D}^{-1}\mathbf{T} \tag{4.24}$$

where $\mathbf{K}_r^*$ and $\mathbf{L}_r^*$ are the matrices of the first $r$ columns of $\mathbf{K}^*$ and $\mathbf{L}^*$, respectively. $\mathbf{\Lambda}_r^*$ is a diagonal matrix of the $r$ largest singular values of $\mathbf{H}^*$, and $\mathbf{D}$ is an $r \times r$ arbitrary diagonal matrix with non-zero diagonal elements. In this chapter, we set $\mathbf{T} = \mathbf{D} = \mathbf{I}_r$ without loss of generality.

From the foregoing results, we propose the following ALS algorithm for obtaining $\mathbf{U}$, $\mathbf{V}$, $\mathbf{P}$, and $\mathbf{Q}$ that minimizes (4.14):

1. Given the dimension of solution $r$ and numbers of clusters $K$ and $L$, randomly generate $\mathbf{U}$ and $\mathbf{V}$.
2. Update $\mathbf{P}$ and $\mathbf{Q}$ by (4.23) and (4.24), respectively.
3. Apply fuzzy $K$-means on $\mathbf{X}(\mathbf{1}_L \otimes \mathbf{I}_J)'\mathbf{D}_\mathbf{V}$ with the fixed centroid $\mathbf{P}\mathbf{Q}'(\mathbf{I}_L \otimes \mathbf{1}_J)'\mathbf{D}_\mathbf{V}$ to update $\mathbf{U}$.
4. Apply fuzzy $K$-means on $\mathbf{X}'(\mathbf{1}_K \otimes \mathbf{I}_I)'\mathbf{D}_\mathbf{U}$ with the fixed centroid $\mathbf{Q}\mathbf{P}'(\mathbf{I}_K \otimes \mathbf{1}_I)'\mathbf{D}_\mathbf{U}$ to update $\mathbf{V}$.
5. If the reduction of the function value is less than $||\mathbf{X}||^2 \times 10^{-7}$, stop the procedure; otherwise, return to 2.

As noted above, $\alpha$ and $\beta$ must be given before applying FCBA, and these parameters should not be set too high (e.g., higher than 3) because of the following property. As $\alpha$ and $\beta$ increase, the resulting membership of objects and variables, $u_{ik}$ and $v_{jl}$, respectively, become close to $K^{-1}$ and $L^{-1}$, respectively, indicating that all object/variable clusters are almost identical. As a result, the row vectors of $\mathbf{P}$ and $\mathbf{Q}$ corresponding to the centroids of the object and variable clusters become identical, which leads to rank-deficient $\mathbf{H}^*$, $\mathbf{P}$, and $\mathbf{Q}$. In addition, it can be shown that the

function value of (4.14) attained after convergence with $\alpha > 0$ and $\beta > 0$ does not exceed the function value with $\alpha = \beta = 1$ since CCBA is a special case of FCBA. However, as $\alpha$ and $\beta$ increase, it is not guaranteed that the attained function value monotonically decreases or approaches a certain value asymptotically. This means that $\alpha$ and $\beta$ must be chosen in a subjective manner. However, while choosing $\alpha$ and $\beta$ may be a problem in the proposed procedure, setting $\alpha = \beta = 1.2$ has never failed in the following numerical simulation or in application to real datasets, as determined by the procedure introduced below.

Note that our approach to obtaining the biplot is different from tandem analysis, where the fuzzy $K$-means clustering procedure is performed to solve the PC biplot. Note also that the proposed biplot procedure is a generalization of the CCBA proposed in Yamashita and Mayekawa (2012), which classifies objects and variables by using K-means clustering, because if we impose the additional constraint on the membership matrices of $u_{ik}, v_{jl} \in [0, 1]$ and set $\alpha = \beta = 1$ both procedures are identical. Further, under this binary constraint on U and V, the objective function of FCBA reduces to the one of CCBA in (4.4). Therefore, the resulting membership of the objects and variables in FCBA provides a more insightful interpretation than the counterpart of CCBA, since membership values are fuzzily expressed in the former.

### 4.2.4 Multiple Runs Procedure

Although the objective function $f_{FCB}(\mathbf{U}, \mathbf{V}, \mathbf{P}, \mathbf{Q}|\mathbf{X})$ monotonically decreases according to the steps 2 to 4 of the algorithm above, the solution obtained after convergence is not guaranteed to be the global minimizer. To avoid the local minimum and obtain a solution as close as possible to the global minimizer, we use the procedure proposed by Adachi (2011) in which the FCBA algorithm is run multiple times starting from different initial values until the two equivalently best solutions are found. By noting $\mathbb{S}_m = \{\mathbf{U}_m, \mathbf{V}_m, \mathbf{P}_m, \mathbf{Q}_m\}$ as the solution obtained by the $m$-th run and $L_{FCBA}(\mathbb{S}_m)$ for the attained value of $L_{FCBA}(\mathbf{U}_m, \mathbf{V}_m, \mathbf{P}_m, \mathbf{Q}_m)$, the multiple runs procedure for FCBA is stated as follows:

1. Set $M = 15$ and obtain $\mathbb{S}_m$s for $m = 1, \cdots, M$; find $m^* = \underset{1 \leq m \leq M}{\operatorname{argmin}} L_{FCBA}(\mathbb{S}_m)$.
2. Select $\mathbb{S}_m$ as the optimal solution to finish if $L_{FCBA}(\mathbb{S}_{m^*})$ is equivalent to $LFCBA(\mathbb{S}_l)$ resulting from the $l$-th run with $l \neq m^*$; otherwise, proceed to 3.
3. Set $M = M + 1$ and let $\mathbb{S}_M$ be the output from the $M$-th run.
4. Exchange $\mathbb{S}_{m^*}$ for $\mathbb{S}_M$ if $L_{FCBA}(\mathbb{S}_M) < L_{FCBA}(\mathbb{S}_{m^*})$.
5. Select $\mathbb{S}_m$ as the optimal solution to finish if $L_{FCBA}(\mathbb{S}_{m^*})$ is equivalent to $L_{FCBA}(\mathbb{S}_M)$ or $M = 15$; otherwise, return to 3.

Here, the equivalence of $L_{FCBA}(\mathbb{S}_{m^*})$ to $L_{FCBA}(\mathbb{S}_l)$ is defined as $|L_{FCBA}(\mathbb{S}_{m^*}) - L_{FCBA}(\mathbb{S}_l)| < ||\mathbf{X}||2 \times 10^{-7}$.

## 4.3 Simulation Study

A simulation study was performed in order to investigate the performance of FCBA because it include CCBA as a special case. In particular, this study aimed to evaluate the recovery of the model parameter matrices and correctness of classification through a comparison with the predetermined true values of them, varying the inter-cluster variance and error level. In this section, artificial datasets with four object clusters and four variable clusters are considered, namely $K = L = 4$.

A data matrix $\mathbf{X}$ with $I = 100$ (objects), $J = 80$ (variables), $K = 4$ (object clusters), and $L = 4$ (variable clusters) was constructed as follows: [1] By randomly associating objects and variables with clusters, the crisp membership matrices $\mathbf{U}_T$ and $\mathbf{V}_T$ were generated as the true values of $\mathbf{U}$ and $\mathbf{V}$, respectively. [2] The true values of $\mathbf{P}$ and $\mathbf{Q}$, expressed as $\mathbf{P}_T$ and $\mathbf{Q}_T$, were then set as below:

$$\mathbf{P}_T = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 0 & -1 \\ -1 & 0 \end{bmatrix}, \ \mathbf{Q}_T = \begin{bmatrix} 1 & 1 \\ -1 & 1 \\ 1 & -1 \\ -1 & -1 \end{bmatrix}. \tag{4.25}$$

[3] For $i = 1, \cdots, I$ and $j = 1, \cdots, J$, the two-dimensional vectors $\mathbf{f}(i)$ and $\mathbf{a}_{(j)}$ that correspond to the coordinates of the objects and variables in two-dimensional space, respectively, were randomly drawn from $N_2(\mathbf{u}_{T(i)}\mathbf{P}_T, \sigma^2\mathbf{I})$ and $N_2(\mathbf{v}_{T(j)}\mathbf{Q}_T, \sigma^2\mathbf{I})$, respectively, where $\mathbf{u}_{T(i)}$ and $\mathbf{v}_{T(j)}$ denote the $i$-th and $j$-th rows of $\mathbf{U}_T$ and $\mathbf{V}_T$ and $\sigma^2$ is a constant that indicates the variance within each cluster. [4] By using $\mathbf{F} = \{\mathbf{f}'_{(1)}, \cdots, \mathbf{f}'_{(I)}\}'$ and $\mathbf{A} = \{\mathbf{a}'_{(1)}, \cdots, \mathbf{a}'_{(J)}\}'$, the data matrix $\mathbf{X}$ was constructed with $\mathbf{X} = \mathbf{F}\mathbf{A}' + \theta(\rho)\mathbf{E}$, where $\mathbf{E}$ was filled with standard normal errors and $\theta(\rho)$ given as follows (Adachi, 2011):

$$\theta(\rho) = \sqrt{\frac{1-\rho}{\rho} \times \frac{||\mathbf{F}\mathbf{A}'||^2}{||\mathbf{E}||^2}}. \tag{4.26}$$

Note that it is rewritten as

$$\rho = \frac{||\mathbf{F}\mathbf{A}'||^2}{||\mathbf{F}\mathbf{A}'||^2 + \theta^2(\rho)||\mathbf{E}||^2} \tag{4.27}$$

and thus, it approximately expresses the amount of variance explained by the model part $\mathbf{F}\mathbf{A}'$; in other words, if $\mathrm{tr}\mathbf{E}'\mathbf{F}\mathbf{A}'$ is close enough to zero, we have the approximation $||\mathbf{X}||^2 \simeq ||\mathbf{F}\mathbf{A}'||^2 + \theta^2(\rho)||\mathbf{E}||^2$, that is, $\rho = ||\mathbf{F}\mathbf{A}'||2/||\mathbf{X}||^2$ approximately expresses the variance of $\mathbf{X}$ accounted for by the model part $\mathbf{F}\mathbf{A}'$. We tried the proportion $\rho$ in a range from 0 to 1 in order to control the smallness of the error contained in the resulting dataset $\mathbf{X}$. The PC biplots of two datasets generated under two different conditions with $(\sigma^2, \rho) = (0.01, 0.85)$ and $(\sigma^2, \rho) = (0.10, 0.85)$ are shown in Figure 4.2.

For each of $3 \times 3 = 9$ combinations of $\sigma^2 = 0.01, 0.10$, and 0.30 and $\rho = 1.00, 0.85$,

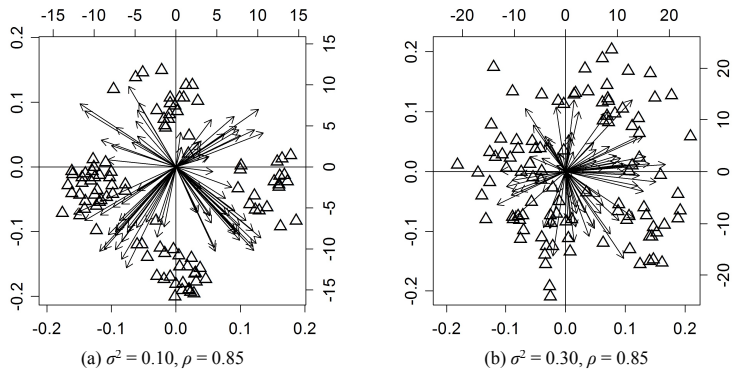



FIGURE 4.2: PC biplots of generated datasets under two different conditions

and 0.70, we generated 100 data matrices by repeating the foregoing steps and then applied FCBA to the resulting $100 \times 9 = 900$ data matrices. The results were evaluated from the following three points of view: [1] the rate of the correct classification of objects and variables, [2] the recovery of $\mathbf{P}$ and $\mathbf{Q}$, and [3] how often a local minimum occurred. To measure how correctly objects and variables were classified, we used the adjusted Rand index (ARI) (Hubert and Arabie, 1985), which ranges from 0 to 1, and the function $CM(\mathbf{M}, \mathbf{M}_T) = 100 \times m^{-1}\text{tr}\mathbf{M}'\mathbf{M}_T$ that ranges from 0 to 100. Both evaluate the congruence of the two $m \times n$ membership matrices $\mathbf{M}$ and $\mathbf{M}_T$. Note that they were computed after transforming the fuzzy membership matrices $\mathbf{U}$ and $\mathbf{V}$ into crisp ones whose elements are binary, with $\max_k(u_{i1}, \cdots, u_{iK}) = 1(i = 1, \cdots, I)$, $\max_l(v_{j1}, \cdots, v_{kL}) = 1(j = 1, \cdots, J)$, and zeros elsewhere. In order to evaluate how well $\mathbf{P}$ and $\mathbf{Q}$ are recovered, we used the congruence coefficient (CC) (Tucker, 1951) and the following function that evaluates the closeness of the two $m \times n$ matrices $\mathbf{W}$ and $\mathbf{W}_T$:

$$SS(\mathbf{W}, \mathbf{W}_T) = 1 - 0.5\frac{||\mathbf{W} - \mathbf{W}_T||^2}{||\mathbf{W} - \mathbf{W}^{\#}||^2 + ||\mathbf{W}_T - \mathbf{W}^{\#}||^2} \tag{4.28}$$

where $\mathbf{W}^{\#} = (2mn)^{-1}\mathbf{1}_m\mathbf{1}_m'(\mathbf{W} + \mathbf{W}_T)\mathbf{1}_n\mathbf{1}_n'$ (Adachi, 2011). These indices both range from 0 to 1, and attain their maximum 1 when $\mathbf{W} = \mathbf{W}_T$.

Note that it is not guaranteed that the estimated $k$-th cluster corresponds to the true $k$-th cluster, since $L_{FCBA}(\mathbf{U}, \mathbf{V}, \mathbf{P}, \mathbf{Q}|\mathbf{X}) = L_{FCBA}(\mathbf{U}\mathbf{R}_\mathbf{U}, \mathbf{V}\mathbf{R}_\mathbf{V}, \mathbf{R}_\mathbf{U}'\mathbf{P}, \mathbf{R}_\mathbf{V}'\mathbf{Q}|\mathbf{X})$, with $\mathbf{R}_\mathbf{U}(K \times K)$ and $\mathbf{R}_\mathbf{U}(L \times L)$ being arbitrary permutation matrices. To eliminate this indeterminacy with respect to the permutation of clusters and evaluate the degree of recovery of the parameter matrices appropriately, we chose $\mathbf{R}_\mathbf{U}$ and $\mathbf{R}_\mathbf{V}$ within $\mathbb{R}_\mathbf{U}$ and $\mathbb{R}_\mathbf{V}$ to maximize $CM(\mathbf{U}\mathbf{R}_\mathbf{U}, \mathbf{U}_T)$ and $CM(\mathbf{V}\mathbf{R}_\mathbf{V}, \mathbf{V}_T)$, where $\mathbb{R}_\mathbf{U}$ and $\mathbb{R}_\mathbf{V}$ denote the set of all possible permutation matrices having $K!$ and $L!$ elements, respectively.

In addition, in order to appropriately evaluate the recovery of $\mathbf{P}$ and $\mathbf{Q}$, it is necessary to eliminate the indeterminacy as to the rotation by an orthonormal matrix $\mathbf{T}$ and scaling by a diagonal matrix $\mathbf{D}$, as mentioned in (4.23) and (4.24). For this purpose, we developed an algorithm called the *joint Procrustes transformation* to

obtain $\mathbf{T}$ and $\mathbf{D}$ that minimizes the following least squares criterion:

$$f_{JPT}(\mathbf{T}, \mathbf{D}) = K^{-1}||\mathbf{P}_T - \mathbf{R}'_{\mathbf{U}}\mathbf{P}\mathbf{D}\mathbf{T}||^2 + ||\mathbf{Q}_T - \mathbf{R}'_{\mathbf{V}}\mathbf{Q}\mathbf{D}^{-1}\mathbf{T}||^2 \tag{4.29}$$

The details of the algorithm to determine $\mathbf{T}$ and $\mathbf{D}$ are presented in the Appendix A.

1. For any given $\sigma^2$ and $\rho$, generate a data matrix $\mathbf{X}$.
2. Carry out FCBA with 10 different random starts and obtain the best solution $\hat{\mathbf{S}} = \{\mathbf{U}, \mathbf{V}, \mathbf{P}, \mathbf{Q}\}$ that minimizes $L_{FCBA}(\mathbf{U}, \mathbf{V}, \mathbf{P}, \mathbf{Q}|\mathbf{X})$ the most.
3. Compute the percentage of the occurrence of the local minimum.
4. Obtain $\mathbf{R}_{\mathbf{U}}$ and $\mathbf{R}_{\mathbf{V}}$ and permute the columns of $\mathbf{U}$ and $\mathbf{V}$ by $\mathbf{R}_{\mathbf{U}}$ and $\mathbf{R}_{\mathbf{V}}$ and the rows of $\mathbf{P}$ and $\mathbf{Q}$ by $\mathbf{R}_{\mathbf{U}}$ and $\mathbf{R}_{\mathbf{V}}$, resulting in $\mathbf{U}\mathbf{R}_{\mathbf{U}}$,$\mathbf{V}\mathbf{R}_{\mathbf{V}}$, $\mathbf{R}'_{\mathbf{U}}\mathbf{P}$ and $\mathbf{R}'_{\mathbf{V}}\mathbf{Q}$.
5. Transform $\mathbf{U}\mathbf{R}_{\mathbf{U}}$ and $\mathbf{V}\mathbf{R}_{\mathbf{V}}$ into crisp membership matrices and evaluate $CM$ and $ARI$ with respect to $\mathbf{U}_T$ and $\mathbf{V}_T$.
6. Carry out the joint Procrustes transformation for $\mathbf{R}'_{\mathbf{U}}\mathbf{P}$ and $\mathbf{R}'_{\mathbf{V}}\mathbf{Q}$ to obtain $\mathbf{T}$ and $\mathbf{D}$ and evaluate Adachi's standardized similarity ($SS$) and $CC$ with respect to $\mathbf{P}_T$ and $\mathbf{Q}_T$.

Table 4.1 shows the 1st, 2nd, and 3rd quartiles of $CM(\mathbf{U}\mathbf{R}_{\mathbf{U}}, \mathbf{U}_T)$ and $CM(\mathbf{V}\mathbf{R}_{\mathbf{V}}, \mathbf{V}_T)$, namely the rates of correct classification in these nine situations. There, we observe that the objects and variables were correctly classified in many situations, even when $\sigma^2$ was large and $\rho$ was small (i.e., even if the data had an ambiguous cluster structure and was error-contaminated). Further, the 1st, 2nd and 3rd quartiles of the $SS$ and $CC$ values and average percentages of the frequency of the local minimum are listed in Table 4.2. This table shows that the recovery of $\mathbf{P}$ and $\mathbf{Q}$ was found to be fairly good in most cases in spite of the occurrence of the local minimum. From these results, we thus conclude that our proposed method and algorithm are suitable for practical use because the objects and variables are correctly classified and true configuration appropriately estimated even when the data had an obscure cluster structure and were error-contaminated.

## 4.4 Real Data Example

In this section, in order to demonstrate that the FCBA procedure provides a biplot that is easy to interpret, we present an application of the proposed procedure on a real dataset. Further, the result of FCBA is compared with that of CCBA. The dataset analyzed in this example is available at UCI Machine Learning Repository (Dua and Graff, 2019) as *Wine dataset*. In total, 178 samples were drawn from three types of wines, termed A, B, and C, with each sample evaluated with respect to the 13 chemical features listed in the caption of Figure 4.6.

TABLE 4.1: The values for the 1st, 2nd, and 3rd quartiles of *ARI* and the index of the congruence of memberships ($CM$). Each index was evaluated after appropriately permuting the columns of **U** and **V**, followed by the crisp transformation of them.

| $\sigma^2$ | | | 0.01 | | | 0.10 | | | 0.30 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\rho$ | | | 1.00 | 0.85 | 0.70 | 1.00 | 0.85 | 0.70 | 1.00 | 0.85 | 0.70 |
| | | index | | | | | | | | | |
| Objects (**U**) | 25% | ARI | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.80 | 0.80 | 0.82 |
| | | CM | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 90.00 | 90.00 | 91.00 |
| | 50% | ARI | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.84 | 0.84 | 0.86 |
| | | CM | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 92.00 | 92.00 | 93.00 |
| | 75% | ARI | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.88 | 0.88 | 0.88 |
| | | CM | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 94.00 | 94.00 | 94.00 |
| Variables (**V**) | 25% | ARI | 1.00 | 1.00 | 1.00 | 0.92 | 0.91 | 0.92 | 0.55 | 0.53 | 0.51 |
| | | CM | 100.00 | 100.00 | 100.00 | 96.30 | 95.90 | 96.30 | 76.30 | 75.00 | 73.80 |
| | 50% | ARI | 1.00 | 1.00 | 1.00 | 0.95 | 0.95 | 0.95 | 0.62 | 0.60 | 0.59 |
| | | CM | 100.00 | 100.00 | 100.00 | 97.50 | 97.50 | 97.50 | 80.00 | 78.80 | 78.80 |
| | 75% | ARI | 1.00 | 1.00 | 1.00 | 0.97 | 0.97 | 0.97 | 0.68 | 0.66 | 0.66 |
| | | CM | 100.00 | 100.00 | 100.00 | 98.80 | 98.80 | 98.80 | 83.80 | 82.50 | 82.50 |

TABLE 4.2: The values for 1st, 2nd and 3rd quartiles of $CC$ and $SS$, and average percentage of the frequency of the local minimum.

| $\sigma^2$ | | | 0.01 | | | 0.10 | | | 1.00 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\rho$ | | | 1.00 | 0.85 | 0.70 | 1.00 | 0.85 | 0.70 | 1.00 | 0.85 | 0.70 |
| | | index | | | | | | | | | |
| **P** | 25% | CC | 1.00 | 1.00 | 1.00 | 1.00 | 0.98 | 0.62 | 0.99 | 0.92 | 0.55 |
| | | SS | 0.99 | 1.00 | 1.00 | 0.99 | 0.99 | 0.80 | 0.99 | 0.96 | 0.76 |
| | 50% | CC | 1.00 | 1.00 | 1.00 | 1.00 | 0.99 | 0.70 | 0.99 | 0.95 | 0.59 |
| | | SS | 1.00 | 1.00 | 1.00 | 1.00 | 0.99 | 0.83 | 0.99 | 0.97 | 0.78 |
| | 75% | CC | 1.00 | 1.00 | 1.00 | 1.00 | 0.99 | 0.77 | 1.00 | 0.98 | 0.62 |
| | | SS | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.88 | 1.00 | 0.99 | 0.81 |
| **Q** | 25% | CC | 1.00 | 1.00 | 1.00 | 0.99 | 0.98 | 0.60 | 0.96 | 0.89 | 0.53 |
| | | SS | 0.99 | 1.00 | 1.00 | 0.99 | 0.98 | 0.78 | 0.97 | 0.93 | 0.74 |
| | 50% | CC | 1.00 | 1.00 | 1.00 | 1.00 | 0.99 | 0.68 | 0.98 | 0.93 | 0.57 |
| | | SS | 1.00 | 1.00 | 1.00 | 1.00 | 0.99 | 0.83 | 0.98 | 0.96 | 0.76 |
| | 75% | CC | 1.00 | 1.00 | 1.00 | 1.00 | 0.99 | 0.75 | 0.99 | 0.96 | 0.62 |
| | | SS | 1.00 | 1.00 | 1.00 | 1.00 | 0.99 | 0.86 | 0.99 | 0.97 | 0.78 |
| Local minimum | | | 0.00 | 0.00 | 0.00 | 5.00 | 0.00 | 0.00 | 12.70 | 12.20 | 14.90 |

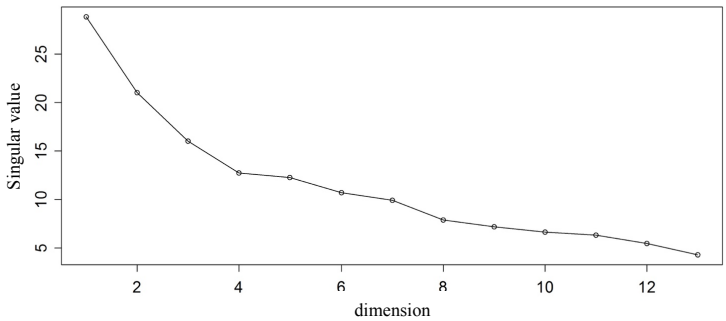



FIGURE 4.3: The scree plot of the singular values of the wine dataset.

Before carrying out CCBA and FCBA on this dataset, the number of dimensions $r$ was set to two in order to allow us to interpret the resulting biplots. Actually, this setting was not consistent with the fact that the "elbow point" was not $r = 2$ in Figure 4.3, even though the eduction of the singular values of the data matrix between the first and second dimensions was largest of all. This is, however, certainly optimum for visualization.

The number of object and variable clusters $K$ and $L$ was determined as follows. Since the objects were samples drawn from one of the three types of wine, it was reasonable to assume that three heterogeneous subgroups of objects existed in the dataset, and we thus set $K = 3$. For the number of variable clusters, we use $L = 3$, which can be validated as follows. For the validation, the author employed the fuzziness performance index ($FPI$) (Bezdek, 2013; Roubens, 1982) as an index of the validity of the given fuzzy membership matrix $\mathbf{M}(I \times K)$, defined as

$$FPI(\mathbf{M}) = 1 - \frac{K \times I^{-1}||\mathbf{M}||^2 - 1}{K - 1}. \tag{4.30}$$

Hwang et al. (2010) used $FPI$ in order to determine the number of clusters in the context of cluster-wise MCA, with small $FPI$ values indicating that the objects or variables are clearly separated from each other. In addition, the attained function value was also considered to be an index of model fit, which decreases as $L$ increases. In Figure 4.4, the joint plot of the $FPI$ values and attained function values for $L = 2, 3, \cdots, 12$ are presented. This plot shows that that $L = 3$ is appropriate because of the lowest $FPI$ value and largest reduction of the function value. The setting $K = L = 3$ is thus used for FCBA and it is also used in CCBA for comparability of their results.

In FCBA, as $\alpha$ and $\beta$ increase, the resulting membership becomes fuzzier, while the membership becomes crisp as they decrease. As pointed out above, FCBA and CCBA are identical when $\alpha = \beta = 1$, and thus the objects and variables are crisply assigned to object and variable clusters. On the contrary, if we use large values of $\alpha$ and $\beta$, we cannot differentiate these clusters. Therefore, we must choose certain appropriate values for $\alpha$ and $\beta$, avoiding extreme cases and obtaining clearly separated clusters.

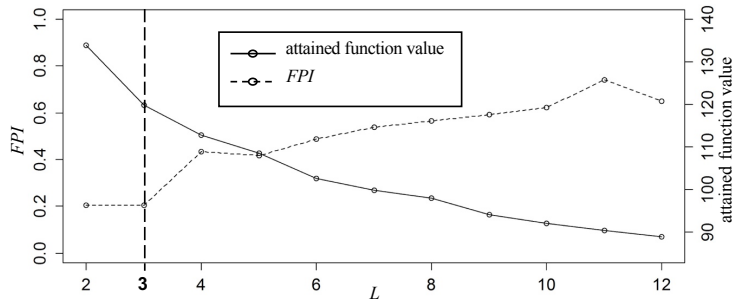



FIGURE 4.4: Joint plot of the attained function value (solid line) and $FPI$ (dotted line) against the number of variable clusters $L$.

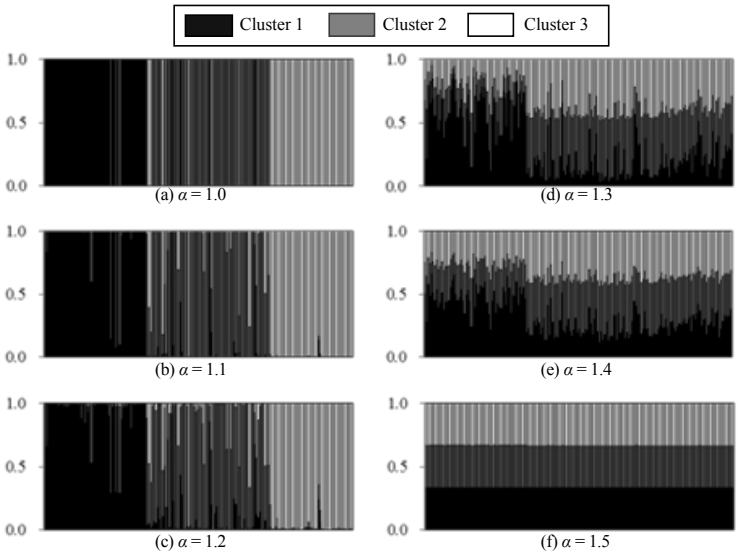



FIGURE 4.5: The estimated object memberships expressed as bar plots with $\alpha = 1.0, 1.1, 1.2, 1.3, 1.4$ and 1.5.

Generally, setting these parameters to their maximum values is recommended as long as the resulting clusters are clearly separated, because this improves interpretation and, at the same time, maintains the characteristic of the proposed method (i.e., fuzzily expressed clusters). In Figure 4.5, the estimated object membership values with $\alpha = 1.0, 1.1, 1.2, 1.3, 1.4$, and 1.5 are expressed as bar plots, in which, for each observation located on the horizontal axis, the vertical bar is divided into $K$ parts according the value of $u_{ik}$ $(k = 1, 2, \cdots, K)$. This figure indicates that the estimated clusters become fuzzier as the fuzziness parameter increases, while $\alpha = 1.2$ is considered to be reasonable since the object clusters in Figures 4.5d to 4.5f are not clearly separated. The clusters in Figure 4.5b are similar to the result of CCBA in Figure 4.5a, whereas the clusters in Figure 4.5c are clearly separated and they appropriately express the fuzziness of the object clusters. The value of parameter $\beta$ was also determined to be $\beta = 1.2$ by using the same procedure.

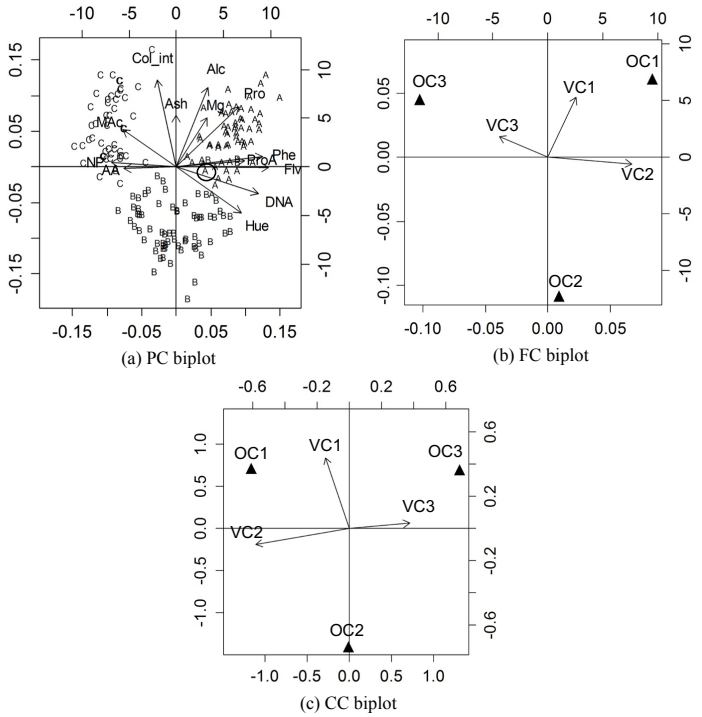



FIGURE 4.6: The PC and FC biplots of the wine dataset.

FCBA and CCBA were applied to this dataset with the settings mentioned above and the obtained biplots are shown in Figure 4.6b, where the symbols OC and VC represent the object and variable clusters, respectively. In the figure, the 13 variables are abbreviated as follows; alcohol (Alc), proline (Pro), magnesium (Mg), ash (Ash), DNA OD280/D315 of diluted wines (DNA), flavonoids (Flv), hue (Hue), total phenols (Phe), proanthocyanidins (ProA), non-flavonoid phenols (NP), malic acid (MAc), alkalinity of ash (AA), color intensity (Col_int). Let us compare Figure 4.6a (PC biplot) with Figure 4.6b (FC biplot) with respect to the conciseness of their interpretation. In the PC biplot, each object point is labeled A, B, or C according to the type of wine from which it was sampled. This biplot is composed of many points and vectors that hamper its straightforward interpretation. For example, recognizing that there are three variable clusters is difficult simply by inspecting the directions of the vectors. On the contrary, the FC biplot provides an intuitive interpretation of the relationships of within/between objects and variables, since it consists of a reduced number of points and vectors. Further, it is observed that the PC biplot is correctly summarized by the FC biplot in that [1] the resulting 1st, 2nd, and 3rd object clusters correspond to wines A, B, and C, respectively and [2] the variable vectors in the PC biplot are summarized in the FC biplot based on their directions in the PC biplot, indicating that both biplots are consistent.

It is clear that such fuzzily expressed membership cannot be obtained in CCBA,

TABLE 4.3: Contingency table for comparing object membership with the three types of wine.

| object cluster | 1st | 2nd | 3rd |
|---|---|---|---|
| type of wine | | | |
| A | 56 | 5 | 0 |
| B | 3 | 58 | 0 |
| C | 0 | 8 | 48 |

which enables the more insightful interpretation of the object/variable clusters mentioned below. The variable clusters are composed of four, five, and four variables, respectively, while the cluster structure extracted by using the proposed procedure is reasonable and suitable compared with the PC biplot (i.e., the directions of the vectors representing the 13 variables are appropriately summarized by the three variable clusters). In addition, since membership is expressed in a stochastic manner in the proposed method, some variables are not crisply classified into one of these three clusters. For example, the variable "color intensity" has the chemical properties of both variable cluster 1 (Alc, Pro, Mg, Ash) and variable cluster 3 (NP, Mac, AA). Similarly, the variable "ash" seems to be in between the 1st and 2nd variable clusters. Importantly, CCBA cannot provide such an interpretation as the objects and variables are crisply classified.

Table 4.3 is the $3 \times 3$ contingency table whose row and column categories are associated with the three types of wine from which each sample is drawn and estimated membership of each sample, respectively. Here, the fuzzy membership matrices of the objects are transformed into crisp ones whose memberships are represented as 0 or 1 with $\max_l(u_{j1}, \cdots, u_{jL}) = 1$ for $j = 1, \cdots, J$ and zeros elsewhere. This table shows that 178 objects were correctly classified according to the type of wine from which they were sampled. It should be noted that the correct classification rate of objects obtained by dividing the number of correct classifications by the total number of objects was 0.91, which was equal to the rate in CCBA. However, in contrast to CCBA, the memberships in FCBA are obtained in a fuzzy fashion and thus the risk of misclassification is considered to be lower than that in CCBA, where memberships must be expressed crisply, while fuzzy expression is allowed in FCBA. In addition, the frequency of the local minimum in CCBA was 0.8, while no local minimum occurred in FCBA, which is considered to be owing to the computational efficiency of the proposed method.

Considering the preceding results on memberships and configuration, the FCBA results for the wine dataset can be interpreted as follows. Given the relative positions of the variable vectors, it is assumed that the horizontal axis of the FC biplot marks the contrast between the variables classified into the 2nd and 3rd variable clusters, while the vertical axis represents the degree of chemical feature evaluated by the 1st variable cluster. Further, the variables organizing the 1st variable cluster are considered to be independent of the other variables, since the vector of VC1 is almost orthogonal to the other vectors. Moreover, since the object clusters are separate from each other, the

objects are clearly classified into three different clusters, each of which corresponds to a type of wine having different chemical features. The difference between wine A (OC1) and wine C (OC3) is considered to be defined by the variables classified into the 2nd and 3rd variable clusters, which represent almost completely opposite chemical features, because those object clusters are located in the opposite position to the horizontal axis. Those types of wine have the intermediate chemical features characterized by the 1st variable cluster. Further, the difference between wine B (OC2) and the other wines is negatively characterized by the 1st variable cluster, since OC2 is located in the opposite direction to VC1. For the 2nd and 3rd variable clusters, wine B is considered to be in between the other object clusters. Such biplot interpretations empirically exemplify the effectiveness of the proposed procedure.

The summarized biplot of the wine dataset obtained by using CCBA assuming three object/variable clusters is presented in Figure 4.6c. Even though the CC biplot is similar to the FC biplot in Figure 4.6b, the major difference between CCBA and FCBA is whether the resulting object/variable membership is crisply or fuzzily expressed. It is clear that the fuzzily expressed membership of the latter method leads to the more precise interpretation of clusters than that of the former, which is exemplified as follows. In Figure 4.6a, since the object point A enclosed by the circle is located on the boundary of wines A and B, this object should have a property in between these wine types (associated with OC1 and OC2, respectively). However, this object is crisply classified into the 2nd object cluster in CCBA. On the contrary, in FCBA, the row of the $\mathbf{U}$ matrix, indicating membership of the 28th object to three clusters, was $\{0.538, 0.462, 0.008\}$, expressing the intermediate property of this object appropriately.

## 4.5 Concluding Remarks

In this study, we proposed two new biplot procedures to obtain an interpretable biplot of a given data matrix, which simultaneously classifies objects and variables into a small number of clusters and plots these clusters in low-dimensional space. A simulation study was then performed in order to evaluate the performance of the proposed method for the recovery of parameters. The result showed that the proposed method recovered all the parameters fairly well and correctly classified objects and variables even when the data were error-contaminated. A real data example empirically verified that the proposed procedure can effectively summarize the PC biplot and correctly classify objects and variables. We thus conclude that the proposed procedure is suitable for addressing real-world problems.

In the simulation study and real data example, we set the fuzziness parameters $\alpha$ and $\beta$ subjectively from the perspective of improving interpretability as well as convenience when comparing the resulting membership matrices. However, how to determine these hyper parameters appropriately in FCBA remains a matter to be discussed. Suk and Hwang (2010) proposed a method for obtaining the optimal values

of fuzziness parameters based on entropy maximization, and their method seems to be applicable to FCBA.

How to choose the number of clusters is one of the most important issues in clustering, since this choice represents the researcher's hypothesis on the homogeneity of the data. In the demonstration of the procedure proposed herein, we determined the number of object and variable clusters $K$ and $L$ on the basis of prior information on the dataset and $FPI$, as the validity of the resulting fuzzy memberships depends on the number of clusters. Specifically, the validity of these memberships does not always lead to the suitability of the hypothesis of the latent homogeneity of the data, and vice versa. Taking this point into account, $K$ and $L$ should be determined based on not only the validity of the resulting fuzzy memberships but also the use of external or prior knowledge on the origin of the data, which can be used to make assumptions about the number of hidden homogeneities and support the decision based on $FPI$ value.

## Appendix 4.A Joint Procrustes transformation

In order to eliminate the indeterminacy as to an orthonormal matrix $\mathbf{T}$ and a diagonal matrix $\mathbf{D}$ and suitably evaluate the recovery of parameters $\mathbf{P}$ and $\mathbf{Q}$ in Section 4.3, an algorithm called the joint Procrustes transformation is developed, where $\mathbf{P}$ and $\mathbf{Q}$ are jointly transformed by $\mathbf{T}$ and $\mathbf{D}$ to maximize the agreement with $\mathbf{P}_T$ and $\mathbf{Q}_T$ in the least squares sense. This problem is equivalent to obtaining $\mathbf{T}$ and $\mathbf{D}$ that minimize

$$L_{JPT}(\mathbf{T},\mathbf{D}) = K^{-1}||\mathbf{P}_T - \mathbf{R}'_{\mathbf{U}}\mathbf{PDT}||^2 + L'||\mathbf{Q}_T - \mathbf{R}'_{\mathbf{V}}\mathbf{QD}^{-1}\mathbf{T}||^2 \tag{4.31}$$

subject to $\mathbf{T}'\mathbf{T} = \mathbf{T}\mathbf{T}' = \mathbf{I}$ and $\mathbf{D}$ being a diagonal matrix. The criterion is expanded as follows:

$$\begin{aligned} L_{JPT}(\mathbf{T},\mathbf{D}) &= \mathrm{tr}K^{-1}\mathbf{P}'_T\mathbf{P}_T + \mathrm{tr}L^{-1}\mathbf{Q}'_T\mathbf{Q}_T \\ &- 2\mathrm{tr}K^{-1}\mathbf{P}'_T\mathbf{R}'_{\mathbf{U}}\mathbf{PDT} - 2\mathrm{tr}L^{-1}\mathbf{Q}'_T\mathbf{R}'_{\mathbf{V}}\mathbf{QD}^{-1}\mathbf{T} \\ &+ \mathrm{tr}K^{-1}\mathbf{P}'\mathbf{PD}^2 + \mathrm{tr}L^{-1}\mathbf{Q}'\mathbf{QD}^{-2}. \end{aligned} \tag{4.32}$$

The minimization of (4.31) over $\mathbf{T}$ with fixed $\mathbf{D}$ is thus equivalent to maximize

$$\begin{aligned} L^{\mathbf{T}}(\mathbf{T}|\mathbf{D}) &= \mathrm{tr}K^{-1}\mathbf{P}'_T\mathbf{R}'_{\mathbf{U}}\mathbf{PDT} + \mathrm{tr}L^{-1}\mathbf{Q}'_T\mathbf{R}'_{\mathbf{V}}\mathbf{QD}^{-1}\mathbf{T} \\ &= \mathrm{tr}\{K^{-1}\mathbf{P}'_T\mathbf{R}'_{\mathbf{U}}\mathbf{PD} + L^{-1}\mathbf{Q}'_T\mathbf{R}'_{\mathbf{V}}\mathbf{QD}^{-1}\}\mathbf{T}. \end{aligned} \tag{4.33}$$

Here, define the singular value decomposition of $K^{-1}\mathbf{P}'_T\mathbf{R}'_{\mathbf{U}}\mathbf{PD} + L^{-1}\mathbf{Q}'_T\mathbf{R}'_{\mathbf{V}}\mathbf{QD}^{-1}$ as

$$K^{-1}\mathbf{P}'_T\mathbf{R}'_{\mathbf{U}}\mathbf{PD} + L^{-1}\mathbf{Q}'_T\mathbf{R}'_{\mathbf{V}}\mathbf{QD}^{-1} = \mathbf{K}^*\mathbf{\Lambda}^*\mathbf{L}^{*\prime} \tag{4.34}$$

where $\mathbf{K}^*$ and $\mathbf{L}^*$ are orthonormal matrices and $\mathbf{\Lambda}^*$ denotes the diagonal matrix whose diagonal elements are singular values arranged in descending order. The maximum of

$L^{\mathbf{T}}(\mathbf{T}|\mathbf{D})$ is attained by

$$\mathbf{T} = \mathbf{L}^{*\prime}\mathbf{K}^{*} \tag{4.35}$$

by using Theorem 1.2.1.

Next, by setting the partial derivative of $L_{JPT}(\mathbf{T},\mathbf{D})$ with respect to the $i$-th diagonal element of $\mathbf{D} = \mathrm{diag}\{d_1, \cdots, d_r\}$ equal to zero, we have the following normal equation for the elements of $\mathbf{D}$:

$$\begin{aligned} \frac{\partial L_{JPT}(\mathbf{T},\mathbf{D})}{\partial d_i} &= \phi(d_i) \\ &= 2\mathbf{p}_i'\mathbf{p}_i' d_i - 2\mathbf{t}_i'\mathbf{P}_T'\mathbf{R}_{\mathbf{U}}'\mathbf{p}_i - 2\mathbf{t}_i'\mathbf{Q}_T'\mathbf{R}_{\mathbf{V}}'\mathbf{q}_i d_i^{-2} - 2\mathbf{q}_i'\mathbf{q}_i d_i^{-3} = 0 \\ &\Leftrightarrow \quad \mathbf{p}_i'\mathbf{p}_i' d_i^4 - \mathbf{t}_i'\mathbf{P}_T'\mathbf{R}_{\mathbf{U}}'\mathbf{p}_i d_i^3 + \mathbf{t}_i'\mathbf{Q}_T'\mathbf{R}_{\mathbf{V}}'\mathbf{q}_i d_i + \mathbf{q}_i'\mathbf{q}_i = 0 \end{aligned} \tag{4.36}$$

where $\mathbf{p}_i$ and $\mathbf{q}_i$ denote the $i$-th column of $\mathbf{P}_T$ and $\mathbf{Q}_T$, respectively. Because of the rank constraint imposed on $\mathbf{P}$ and $\mathbf{Q}$, it is guaranteed that $\mathbf{P}$ satisfies $\mathbf{p}_i'\mathbf{p}_i > 0$; in other words, $\mathbf{p}_i$ has at least one non-zero element for all $i$. The equation $\phi(d_i) = 0$ is thus a quartic equation with respect to $d_i$ that has at most four real roots, expressed as $\gamma_t$ $(t = 1, \cdots, 4)$ with each of these roots able to be obtained explicitly by using Ferrari' s solution (Weisstein, 2002). Therefore, the optimal $d_i$ that minimizes $L_{JPT}(\mathbf{T},\mathbf{D})$ for fixed $\mathbf{T}$ and $d_k(k \neq i)$ is obtained by

$$d_i = \underset{1 \leq r \leq 4}{\mathrm{argmin}}\, L_{JPT}(\mathbf{T}, \mathbf{D}(\gamma_t)) \tag{4.37}$$

where $\mathbf{D}(\gamma_t)$ denotes the function that returns the diagonal matrix whose $i$-th diagonal element is $\gamma_t$ and the remaining elements are those of $\mathbf{D}$.

From the above results, the algorithm of the joint Procrustes transformation can be summarized as follows:

1. Initialize $\mathbf{D}$.
2. Update $\mathbf{T}$ by (4.35).
3. For all $i$, update $d_i$ by (4.37).
4. Repeat 2. and 3. until convergence.

# Chapter 5

# A Modified $K$-means clustering for obtaining a simple centroid matrix

**ABSTRACT**

In $K$-means clustering, between-clusters differences are not always clearly captured in the centroid matrix. The study addresses this problem and propose a modified $K$-means clustering procedure that produces a sparse centroid matrix. This allows easy interpretation of the matrix, since the elements equaling to zero can be ignored in the interpretation. An iterative algorithm for the constrained minimization is proposed, with a cardinality selection procedure for identifying the optimal cardinality. In addition, it is shown that a centroid matrix possesses a simple structure when some restrictions are imposed on the positions of nonzero elements. The performance of our proposed procedure were evaluated by simulation studies. Further, three real data examples are presented to exemplify that the proposed method works well.

## 5.1 Introduction

$k$-means clustering is one of the most popular procedures for classifying the rows of an observations × variables data matrix into a small number of clusters (Aggarwal, 2014). The $k$-means clustering procedure is widely used for classification purposes and recent advances in its development can be found in Steinley (2006). Various extensions and related procedures of the $k$-means clustering exist, fuzzy versions (Miyamoto et al., 2008), probabilistic models (Bock, 1996), and variable selection procedures (Brusco and Cradit, 2001), for example. Applications of $k$-means clustering can be found in various fields of sciences, such as biology (Jetti et al., 2014), environmental science (Dalton et al., 2016), agricultural science (Hyland et al., 2016), engineering (Peng

This chapter is based on the published article Yamashita and Adachi (2017) and Yamashita and Adachi (2019a) with some modifications and extensions.

et al., 2013), applied psychology (Cortina and Wasti, 2005; Kuerbis et al., 2014), and experimental psychology (Schloss et al., 2015; Alsius et al., 2016; Slobodenyuk et al., 2015).

For the $n$-observations $\times$ $p$-variable data matrix $\mathbf{X}$, $k$-means clustering is formulated as minimizing the least squares loss function

$$f(\mathbf{M}, \mathbf{Y}) = \sum_{i,l} m_{il} ||\mathbf{x}'_{(i)} - \mathbf{y}_l||^2 = ||\mathbf{X} - \mathbf{M}\mathbf{Y}'||^2 \tag{5.1}$$

over $\mathbf{M}$ and $\mathbf{Y}$. Here, $\mathbf{M} = \{m_{il}\}(i = 1, \cdots, n; l = 1, \cdots, k)$ is an $n$-observations $\times$ $k$-clusters binary membership matrix and $\mathbf{Y} = \{y_{jl}\}(j = 1, \cdots, p)$ is a $p$-variables $\times$ $k$-cluster centroid matrix. $\mathbf{x}_{(i)}$ and $\mathbf{y}_l$ denote the $i$-th row vector and $l$-th column vector of $\mathbf{X}$ and $\mathbf{Y}$, respectively. The operator $||\mathbf{X}||^2 = \mathrm{tr}(\mathbf{X}'\mathbf{X})$ denotes the sum of the squared elements in $\mathbf{X}$. The minimization of (5.1) can be attained by using the iterative algorithm proposed by MacQueen (1967), in which $\mathbf{M}$ and $\mathbf{Y}$ are alternately updated.

The resulting membership matrix $\mathbf{M}$ shows how observations are classified into clusters. In the interpretation of what variables characterize the clusters, the centroid matrix $\mathbf{Y}$ plays a key role. Let us illustrate such an interpretation process with a data matrix by Adachi (2006); 13 (jobs) $\times$ 12 (attributes) data matrix $\mathbf{X}$ called job impression data in this article, which describes the extent to which the jobs are described by the impression attributes. We applied $k$-means clustering to the column-centered version of the data matrix, with the number of clusters set to 4. Table 5.1 shows the resulting centroid matrix. Its $j$-th row equals the average vector of the observations classified into the $l$-th cluster and represents the variables that characterize the $l$-th cluster. For example, the third cluster (C3) is characterized by the impressions of "powerful,", "strong," and "fast," as the elements for them in C3 show larger values than those in the other clusters, where police, journalist, sailor, and athlete are classified into C3. The within-column and between-column contrasts in the centroids help to capture which variables feature clusters, although these contrasts are not always observed clearly in the centroid matrix.

A typical strategy to clarify the contrasts is to ignore the elements close to zeros (those less than 0.4 in absolute value, for example) in the centroid matrix and regard them as zeros. For example, we can consider that the first cluster (C1) is characterized by the "useful," "stubborn," and "busy" impression, by ignoring the elements having absolute values less than 0.4. This strategy is not desirable, however, because which elements can be ignored depends on the users' decisions. Such decisions are both subjective and potentially erroneous, because, in fact, they weaken the initial fit achieved by the centroid matrix in an intuitive manner.

In this article, considering the above problem of the interpretability of the resulting centroid matrix, we propose a new clustering procedure that produces an easily interpreted centroid matrix. This procedure is called *cardinality-constrained k-means*

TABLE 5.1: Estimated centroid matrix by $k$-means with four clusters to job impression data.

| | C1 | C2 | C3 | C4 |
|---|---|---|---|---|
| admirable | 0.338 | -0.379 | 0.121 | 0.005 |
| useful | 0.557 | -0.393 | 0.182 | -0.276 |
| good | 0.205 | -0.129 | 0.021 | -0.062 |
| large | -0.010 | -0.418 | 0.382 | 0.057 |
| powerful | -0.071 | -0.796 | 1.029 | -0.238 |
| strong | 0.031 | -0.711 | 1.014 | -0.436 |
| fast | -0.333 | 0.125 | 0.950 | -1.100 |
| noisy | -0.279 | 0.246 | 0.871 | -1.212 |
| young | 0.019 | 0.686 | 0.461 | -1.548 |
| honest | 0.443 | -0.332 | 0.168 | -0.224 |
| stubborn | 0.569 | -1.239 | 0.386 | 0.569 |
| busy | 0.590 | 0.282 | 0.207 | -1.243 |

*clustering (CCKM)*. In CCKM, a number of the elements in the centroid matrix $\mathbf{Y}$ are constrained to be zeros, where $Card(\mathbf{Y})$ for the cardinality of $\mathbf{Y}$, i.e., its number of nonzero elements. The constraint is expressed as

$$Card(\mathbf{Y}) = c \tag{5.2}$$

with $c$ a prespecified integer. That is, our proposed CCKM is formulated as minimizing (5.1) subject to (5.2). Here, it should be noted that it is unknown which elements in $\mathbf{Y}$ are zero/nonzero. They are also estimated optimally in CCKM. This provides the centroid matrix with $pk - c$ zero elements, which facilitates the interpretation of $\mathbf{Y}$, as we may ignore the zero elements in $\mathbf{Y}$. Here, it is noteworthy that no subjective decision is involved in what elements are ignored; as described above, which elements are to be zeros is estimated optimally.

The classic work Gordon (1973) firstly introduced constrained clustering procedure, in which a priori information as to clusters are incorporated to clustering. Such a priori information is thoroughly discussed in DeSarbo and Mahajan (1984); some pairs of objects are constrained to be in same or different cluster, for example. These constraints are helpful for obtaining interpretable and valid clusters. The related works can be found in Steinley and Hubert (2008) and Basu et al. (2008). These procedures requires clearly defined information as to cluster structures although is often unavailable before applying clustering in the case of exploratory data analysis. The proposed procedure in this paper does not require such external information with combination of cardinality selection procedure introduced in Section 5.3.

The proposed procedure is also related to variable selection technique in clustering (Brusco and Cradit, 2001). These procedures are designed to obtain appropriate cluster structure even if some musking variables Fowlkes and Mallows (1983) exist, which often spoils clustering result. In order to accomplish this, variable selection procedure specify the set of variables which manifest cluster structure by various strategies,

which are detailed and compared in Steinley and Brusco (2008). CCKM differs from the variable selection procedures in that it is aimed to improve interpretability of cluster centroids and therefore the set of variable corresponds to clusters are not always identical.

The remaining parts of this paper are organized as follows. In the following section, we present the algorithm for minimizing (5.1) under the cardinality constraint (5.2). In Section 5.3, a cardinality selection procedure is proposed, in which the best $c$ is chosen with an information criterion. In Section 5.4, a modified version of CCKM is presented, in which cardinality constraints are imposed row/column-wise, although (5.2) is imposed matrix-wise. Two numerical simulations and three real data examples are presented in Section 5.5 and 5.6, respectively, in order to assess and illustrate the performances of the proposed procedure. Section 5.7 is devoted to a general discussion.

### 5.1.1 Related Method: Factor Rotation

Interpretability of solutions is of importance also in the multivariate analysis procedures other than clustering. Factor rotation is a well-known technique to obtain interpretable solutions in factor analysis (FA), in which its initial solution is transformed into more interpretable one of simple structure (Browne, 2001). It is based on the rotational indeterminacy of the FA solution, which allows a factor loading matrix be transformed in a post-hoc manner. Varimax rotation (Kaiser, 1958), Oblimin rotation (Harman, 1976), and Promax rotation (Hendrickson and White, 1964) are known as common procedures for factor rotation. More generally, rotation can be applied to some other multivariate analysis techniques, not limited for factor analysis; Yamashita (2012a) and Satomura and Adachi (2013) proved that solutions can be rotated in canonical correlation analysis.

In clustering, however, such post-hoc transformations of solutions are not feasible. We therefore propose CCKM for improving the interpretability of solutions in clustering, and it is surely beneficial for practical users. Also, as a special case of CCKM, we propose RC-CCKM to produce the centroid matrix having a simple structure, by means of restricting the row/column cardinality of the centroid matrix, Such a structure is desired also in FA (Thurstone, 1947; Ullman, 2006).

## 5.2 Algorithm

The optimization algorithm for CCKM is outlined in Section 5.2.1. It is composed of two steps called the M- and Y-steps, which are iteratively alternated until convergence is reached. These steps are described in detail in Section 5.2.2.

### 5.2.1 Outline: Whole Algorithm

Our proposed CCKM is formulated as

$$minimize\ f(\mathbf{M},\mathbf{Y}) = ||\mathbf{X}-\mathbf{M}\mathbf{Y}'||^2\ over\ \mathbf{M}\ and\ \mathbf{Y} \tag{5.3}$$

subject to the cardinality constraint (5.2) and the membership constraint imposed on $\mathbf{M}$ such that

$$m_{il} \in \{0, 1\} \; and \; \sum_{l} m_{il} = 1. \tag{5.4}$$

Two parameter matrices are alternately updated in the M- and Y-steps, respectively, starting from multiple sets of initial values in order to avoid accepting a local minimum as a final solution. In each step, (5.1) is minimized over $\mathbf{M}$ or $\mathbf{Y}$ with the other parameter matrix kept fixed. The CCKM algorithm is thus summarized as follows.

Step 1. Set $t = 0$

Step 2. Set initial values for $\mathbf{M}$ and $\mathbf{Y}$.

Step 3. (M-step) Update $\mathbf{M}$ to that which minimizes $f(\mathbf{M}, \mathbf{Y})$ subject to (5.4) with $\mathbf{Y}$ fixed.

Step 4. (Y-step) Update $\mathbf{Y}$ to that which minimizes $f(\mathbf{M}, \mathbf{Y})$ subject to (5.2) with $\mathbf{M}$ fixed.

Step 5. If the current $\mathbf{M}$ has an empty column filled with zeros, return to Step 2.

Step 6. Increase $t$ by one and go to Step 7, if the decrease in the (5.1) value from the previous round is less than $1.0 \times 10^{-7}$; otherwise, return to Step 3.

Step 7. Update $\hat{\mathbf{M}}$ and $\hat{\mathbf{Y}}$ by the current $\mathbf{M}$ and $\mathbf{Y}$ if $f(\hat{\mathbf{M}}, \hat{\mathbf{Y}}) > f(\mathbf{M}, \mathbf{Y})$ or $t = 1$.

Step 8. If $t = t_{max}$, accept $\hat{\mathbf{M}}$ and $\hat{\mathbf{Y}}$ as the final solution; otherwise, return to Step 2.

The purpose of Step 5 is to avoid a solution with an empty cluster to which no observations belong. In order to avoid accepting a local minimum, Steps 2 to 6 are repeated $t_{max}$ times starting from different initial values. Among the resulting multiple solutions, that with the lowest (5.1) value is selected as the optimal solution. The update formulas used in the M- and Y-steps are presented in the following section.

### 5.2.2 Optimization in M-step and Y-step

For the M-step, the corresponding step in the $k$-means clustering (MacQueen, 1967) can be used: for fixed $\mathbf{Y}$, the optimal $\mathbf{M} = \{m_{il}\}$ minimizing (5.1) is given by

$$m_{il} = \begin{cases} 1 & (l = \underset{l}{\operatorname{argmin}} f(\mathbf{M}, \mathbf{Y})) \\ 0 & (otherwise) \end{cases}, \tag{5.5}$$

for $i = 1, \cdots, n$.

The problem in the Y-step cannot be solved straightforwardly and we therefore need a trick. As such a one, we use the fact that (5.1) can be decomposed as

$$f(\mathbf{M}, \mathbf{Y}) = ||\mathbf{X} - \mathbf{M}\mathbf{A}'||^2 + ||\mathbf{D}^{1/2}(\mathbf{A} - \mathbf{Y})||^2. \tag{5.6}$$

Here, $\mathbf{D} = \text{diag}\{d_{11}, \cdots, d_{ll}, \cdots, d_{kk}\}$ denotes the $k \times k$ diagonal matrix with $d_{ll}$ the number of the observations classified into the $l$-th cluster ($l = 1, \cdots, k$), while $\mathbf{A}$ is defined as

$$\mathbf{A} = \mathbf{X}'\mathbf{M}(\mathbf{M}'\mathbf{M})^{-1}, \tag{5.7}$$

the $l$-th column of which contains the averages of the observation in the $l$-th cluster. The identity in (5.6) can be proved as follows. (5.1) can be rewritten as

$$\begin{aligned} f(\mathbf{M}, \mathbf{Y}) &= ||\mathbf{X} - \mathbf{M}\mathbf{Y}'||^2 \\ &= ||\mathbf{X} - \mathbf{M}\mathbf{A}' + \mathbf{M}\mathbf{A}' - \mathbf{M}\mathbf{Y}'||^2 \\ &= ||\mathbf{X} - \mathbf{M}\mathbf{A}'||^2 + ||\mathbf{D}^{1/2}(\mathbf{A} - \mathbf{Y})||^2 - \text{tr}\{(\mathbf{X} - \mathbf{M}\mathbf{A}')'(\mathbf{M}\mathbf{A}' - \mathbf{M}\mathbf{Y}')\} \end{aligned} \tag{5.8}$$

with its last term vanishing as

$$\begin{aligned} & \text{tr}\{(\mathbf{X} - \mathbf{M}\mathbf{A}')'(\mathbf{M}\mathbf{A}' - \mathbf{M}\mathbf{Y}')\} \\ = \ & \text{tr}\{\mathbf{X}'\mathbf{M}(\mathbf{M}'\mathbf{M})^{-1}\mathbf{M}'\mathbf{X}\} - \text{tr}\{\mathbf{X}'\mathbf{M}(\mathbf{M}'\mathbf{M})^{-1}\mathbf{M}'\mathbf{X}\} - \text{tr}(\mathbf{X}'\mathbf{M}'\mathbf{M}) + \text{tr}(\mathbf{X}'\mathbf{M}'\mathbf{M}) \\ = \ & 0. \end{aligned} \tag{5.9}$$

In the right-hand side of (5.6), only $g(\mathbf{Y}) = ||\mathbf{D}^{1/2}(\mathbf{A} - \mathbf{Y})||^2$ is relevant to $\mathbf{Y}$. Thus, our task is to minimize $g(\mathbf{Y})$ subject to (5.2). The minimum can be found using the fact that $g(\mathbf{Y})$ satisfies the following identity and inequality:

$$g(\mathbf{Y}) = \sum_{(j,l)\in Z} d_{ll}^{1/2} a_{jl}^2 + \sum_{(j,l)\in Z^\perp} d_{ll}^{1/2}(a_{jl} - y_{jl})^2 \geq \sum_{(j,l)\in Z} d_{ll}^{1/2} a_{jl}^2, \tag{5.10}$$

where $Z$ denotes the set of $pk - c$ pairs $(j, l)$s for the elements $y_{jl}$s to be zero, while $Z^\perp$ is the complement of $Z$. The inequality in (5.10) shows that $g(\mathbf{Y})$ attains its lower limit $\sum_{(j,l)\in Z} d_{ll}^{1/2} a_{jl}^2$ when the elements $y_{jl}$s in the second term in (5.10) is equal to zero, that is, when the $y_{jl}$s with $(j, l) \in Z^\perp$ is set equal to zero. Further, the limit $\sum_{(j,l)\in Z} d_{ll}^{1/2} a_{jl}^2$ is minimal when $Z$ contains the indices for the $pk - c$ smallest $a_{jl}^2$s among all squared elements in $\mathbf{A}$. Therefore, the optimal $\mathbf{Y}$ is obtained as

$$y_{jl} = \begin{cases} 0 & (\textit{iff } a_{jl}^2 \leq a_{[pk-c]}^2) \\ a_{jl} & (\textit{otherwise}) \end{cases}. \tag{5.11}$$

for $l = 1, \cdots, k$ and $j = 1, \cdots, p$, where $a_{[pk-c]}^2$ denotes the $(pk - c)$-th smallest value among all $a_{jl}^2 s$.

Equations (5.5) and (5.11) are used for the updates in M- and Y-step, respectively. They guarantee the monotonic decrement in the the $f(\mathbf{M}, \mathbf{Y})$ value. In the following simulation studies and real data examples, we used 300 different initial values for $\mathbf{M}$ and $\mathbf{Y}$; i.e., $t_{max} = 300$.

## 5.3 Cardinality Selection based on Information Criteria

In the CCKM algorithm, the cardinality of centroid matrix $\mathbf{Y}$ has to be set to a positive integer $c$, as in (5.2). In this article, the minimum and maximum of $c$, $c_{min}$, $c_{max}$, are defined as

$$c_{min} = p, \quad c_{max} = pk. \tag{5.12}$$

It should be noted that $\mathbf{Y}$ has $p$ non-zero elements when $\mathbf{Y}$ has a perfect cluster structure; i.e., each variable is associated with only one cluster. The selection of the number of non-zero elements in $\mathbf{Y}$ can be viewed as a model selection problem, since the selection partially specifies the model part of CCKM, $\mathbf{MY}'$, fitted to the given data matrix $\mathbf{X}$. Thus, information criteria, such as Akaike information criterion (AIC) and Bayesian information criterion (BIC), for the model selection problem can be suitable for determining $c$, which directly constrains the cardinality of the model fitted to the dataset. In this study, we propose using AIC and BIC to select the "best" $c$ among the interval $[c_{min}, c_{max}]$.

Let $\mathbf{E} = \{e_{ij}\}$ be the matrix of errors defined as $\mathbf{E} = \mathbf{X} - \mathbf{MY}'$ and assume that data matrix $\mathbf{X}$ is modeled as $\mathbf{X} = \mathbf{MY}' + \mathbf{E}$ with $e_{ij}$ distributed independently and identically according to $N(0, \sigma^2)$ for all $i$ and $j$. Here, $N(0, \sigma^2)$ represents normal distribution with its mean zero and variance $\sigma^2$. Then, it can be shown that the least squares and maximum likelihood estimates for CCKM are equivalent. Under the above assumption, the log-likelihood function to be maximized in ML estimation is expressed as

$$l(\mathbf{M}, \mathbf{Y}) = -\frac{np}{2} \log ||\mathbf{X} - \mathbf{MY}'||^2. \tag{5.13}$$

Its maximization is equivalent to minimizing least squares function (5.1). For a positive integer $c$, the maximum of $l(\mathbf{M}, \mathbf{Y})$ is attained as

$$l(\mathbf{M}, \mathbf{Y}) \leq -\frac{np}{2} \log f_{min}(c). \tag{5.14}$$

where $f_{min}(c)$ denotes the attained function value of (5.1). Using (5.14), the information criteria $AIC(c)$ and $BIC(c)$ for a specific $c$ are defined as

$$AIC(c) = np \times \log f_{min}(c) + 2\nu(c), \tag{5.15}$$

$$BIC(c) = np \times \log f_{min}(c) + \log(np) \times \nu(c), \tag{5.16}$$

where $\nu(c)$ denotes the number of parameters to be estimated and equals the sum of the numbers of the memberships in $\mathbf{M}$, the nonzero elements in $\mathbf{Y}$, and error variance $\sigma^2$:

$$\nu(c) = n + c + 1. \tag{5.17}$$

Therefore, the best $c$ can be given by $c = \underset{c_{min} \leq c \leq c_{max}}{\operatorname{argmin}} AIC(c)$ or $BIC(c)$. AIC and BIC were originally proposed as model selection criteria and also used for cardinality selection in studies in the literature (Adachi and Trendafilov, 2016, 2018b). This

approach is not considered to be computationally efficient, however, because CCKM involves the optimization with respect to a binary membership matrix, as in the standard $k$-means, and therefore is sensitive to local minima. To select $c$ with $AIC(c)$ or $BIC(c)$ is not feasible because a heavy computational load is required for all runs, especially when the data matrix is large (i.e., $\mathbf{X}$ contains many objects and variables). In such a situation, the centroid matrix to be estimated is also large, although CCKM must facilitate the interpretation of such a large matrix.

In order to find a suitable $c$ value with a lower computational cost, we propose the following algorithm.

Step 1. Set $S_{initial}$ and $S_{decrease}$ to an integer within the range $[0, 1]$. Set $c_{current} = c_{min}$ and $S = c_{max} \times S_{initial}$.

Step 2. Repeat Steps 2 and 3 for $S > 1$.

Step 3. (*Forward search*) Repeat (a) to (c).

(a) Set $c = c_{current}$ and compute

$$\Delta AIC(c) = AIC(c+1) - AIC(c) \tag{5.18}$$

or

$$\Delta BIC(c) = BIC(c+1) - BIC(c) \tag{5.19}$$

(b) If $\Delta AIC(c)$ or $\Delta BIC(c)$ is smaller than 0, set $c_{current} = c_{current} + S$ and return to Step 2; otherwise, proceed to (c).

(c) Set $S = S \times S_{decrease}$ and proceed to the *backward search*.

Step 4. (*Backward search*) Repeat (a) to (c).

(a) Set $c = c_{current}$ and compute $\Delta AIC(c)$ or $\Delta BIC(c)$.

(b) If $\Delta AIC(c)$ or $\Delta BIC(c)$ is greater than 0, set $c_{current} = c_{current} - S$ and return to Step 4; otherwise proceed to (c).

(c) Set $S = S \times S_{decrease}$ and proceed to the *forward search*.

Step 5. If the previous step is *forward search*, repeat *backward search* with $S = 1$ until $\Delta AIC(c)$ or $\Delta BIC(c)$ becomes positive; otherwise, repeat *forward search* until $\Delta AIC(c)$ or $\Delta BIC(c)$ becomes negative.

The above algorithm seeks $c$ that minimizes $AIC(c)$ or $BIC(c)$ within the range $[m_{min}, m_{max}]$ by repeating the forward and backward searches and reducing the step size $S$ through iteration. The rate of decrement in the step size is controlled by $S_{decrease}$ and the initial step size is defined as $c_{max} \times S_{initial}$. The total computational cost is therefore dramatically reduced in comparison with that incurred by performing CCKM with computation of $AIC(c)$ or $BIC(c)$ for all $c$s. In the following simulation and real data examples, we set $S_{initial} = 0.9$ and $S_{decrease} = 0.7$, settings that were empirically confirmed to behave well.

## 5.4 CCKM with Row/Column-wise Cardinality Constraint (RC-CCKM)

While a matrix-wise cardinality is constrained by cardinality parameter $c$ in CCKM, we can also consider its modified version subject to row- and column-wise cardinality constraints. By combining row- and column-wise constraints, we can restrict both the cardinality of the centroid matrix and the positions of the nonzero elements, so that the resulting matrix approximates a simple structure. The utility of this approach can be illustrated with an example of a 5 variables $\times$ 3 clusters centroid matrix that has several zero elements, but is not easily interpretable:

$$\mathbf{Y} = \begin{bmatrix} & & * \\ * & * & * \\ & & * \\ * & * & * \\ & & * \end{bmatrix}. \tag{5.20}$$

Here, the cardinality is 10 and $*$ represents a nonzero element. Although this matrix has more zero elements than unconstrained ones, it is still difficult to interpret in that the centroid matrix has a column (the third column) filled with non-zero elements, which indicates that all the variables are associated with the third cluster. To interpret of such a cluster, abstraction and integration of all variables are required, and it is not always straightforward to name the cluster. Similarly, the second and fourth rows are filled with non-zeros. We therefore prefer the centroid matrix

$$\mathbf{Y} = \begin{bmatrix} * & * & \\ * & * & \\ & * & * \\ * & & * \\ & * & * \end{bmatrix} \tag{5.21}$$

because it does not contain any row/column vector filled with non-zeros in spite of its cardinality equaling that of (5.20). In order to obtain a centroid matrix as in (5.21), the positions of nonzero elements, in other words, the cardinality of rows and columns, have to be restricted. Such constraints are defined as

$$Card(\mathbf{y}_{(j)}) = r(j),\ Card(\mathbf{y}_l) = c(l) \tag{5.22}$$

where $r(j)$ and $c(l)$ denote the cardinality of the row and column of $\mathbf{Y}$, respectively.

To find the matrices $\mathbf{M}$ and $\mathbf{Y}$ that minimize (5.1) under the above constraint, the Y-step in Section 5.2.2 can be modified as follows. To minimize (5.1) subject to the constraint $Card(\mathbf{y}_{(j)}) = r(j)$, the set $Z$ is redefined as

$$Z = \{(j,l) | a_{jl}^2 \leq a_{j[r(j)]}^2\}, \tag{5.23}$$

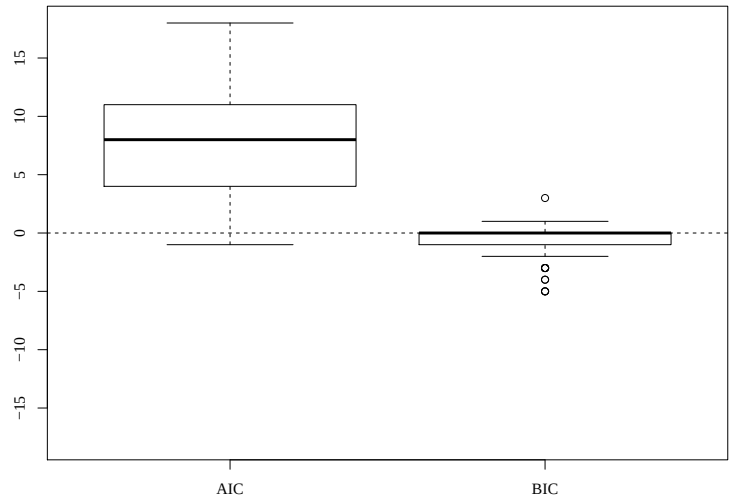



FIGURE 5.1: Boxplots of $SD(c_{AIC})$ and $SD(c_{BIC})$ as indices of discrepancy between true and estimated cardinality by AIC and BIC criterion.

where $\{a^2_{j[k-r(j)]}\}$ denotes the $\{k-r(j)\}$-th smallest element among $a^2_{j1}, \cdots, a^2_{jk}$ for $j = 1, \cdots, p$. Each row is therefore updated by using the above $Z$ and referencing the squared elements of $\mathbf{A}$ in (5.7). In a parallel manner, under $Card(\mathbf{y}_l) = c(l)$, $Z$ is redefined as

$$Z = \{(j,l) | a^2_{jl} \leq a^2_{[c(l)]l}\} \tag{5.24}$$

with $a^2_{[p-c(l)]l}$ the $\{p-c(l)\}$-th smallest element among $a^2_{1l}, \cdots, a^2_{pl}$ for $l = 1, \cdots, k$.

We refer to the above procedure as CCKM with row/column-cardinality constraint (RC-CCKM). The performance of the procedure is demonstrated later in one of the real data examples.

## 5.5 Simulation Studies

In order to assess the behaviors of the CCKM algorithm presented in Section 5.2 and the AIC/BIC-based cardinality selection procedure presented in Section 5.3, we performed two numerical simulation studies. The behaviors to be assessed are the following two. 1) The correctness of the identification of the true cardinality is identified by the cardinality selection procedure and 2) the recovery by CCKM of the true parameters from which artificial datasets are synthesized. Therefore, the purpose of the first simulation study was to assess the accuracy of the cardinality selection, and of the second to evaluate the performance of the parameter recovery.

### 5.5.1 Accuracy of Cardinality Selection

First, we examined the accuracy of true cardinality in the cardinality selection procedure. A hundred data matrices $\mathbf{X}$ of $n = 100$ by $p = 30$ were randomly generated

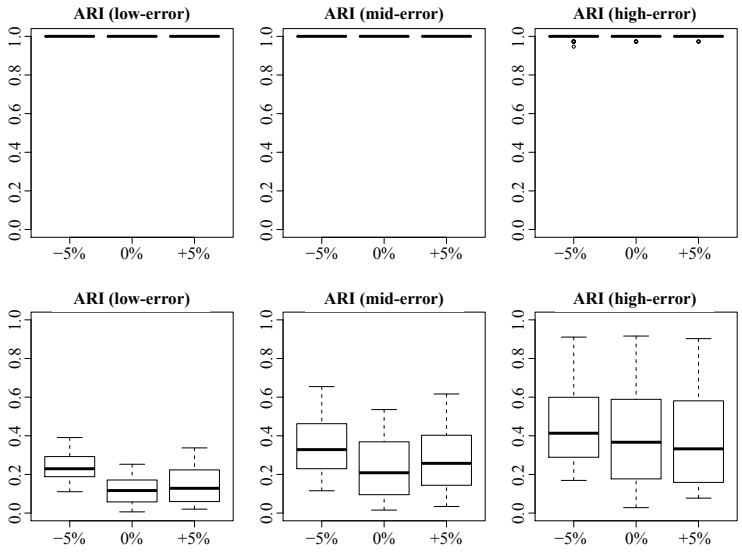



FIGURE 5.2: Boxplot of $ARI$ (adjusted Rand index) and $AAE(\hat{\mathbf{Y}}, \mathbf{Y})$ values in the case with $(n, p) = (100, 30)$.

with setting $k = 3$, as follows.

Step 1. A positive integer $c_T$ was randomly drawn from the interval $[0.1 \times pk, 0.9 \times pk]$ and used for the true cardinality of a centroid matrix.

Step 2. The $c_T$ nonzero elements in $\mathbf{Y}$ were drawn from the uniform distribution $U(1, 5)$, with their positions and signs randomly chosen.

Step 3. The true membership matrix $\mathbf{M}$ was formed by randomly assigning $n$ observations to $k$ clusters.

Step 4. The elements of $n \times p$ error matrix $\mathbf{E}$ were drawn from the standard normal distribution $N(0, 1)$.

Step 5. Data matrix $\mathbf{X}$ of $n$-observations $\times$ $p$-variables was synthesized with

$$\mathbf{X} = \mathbf{M}\mathbf{Y}' + \mu(\rho)\mathbf{E}. \tag{5.25}$$

where $\mu(\rho)$ is defined as

$$\mu(\rho) = \sqrt{\frac{1-\rho}{\rho} \times \frac{||\mathbf{M}\mathbf{Y}'||^2}{||\mathbf{E}||^2}} \tag{5.26}$$

with $\rho$ being the rate of the variance explained.

In Step 5, $\mu(\rho)$ was used for controlling the level of errors, so that $\rho$ approximated the proportion of the variances in $||\mathbf{X}||^2$ accounted for by the model part $\mathbf{M}\mathbf{Y}'$ (Adachi, 2009; Yamashita and Mayekawa, 2015). We considered the medium error level with $\rho = 0.70$, which represents the 70% of the variance of $\mathbf{X}$ explained by $\mathbf{M}\mathbf{Y}'$. For each

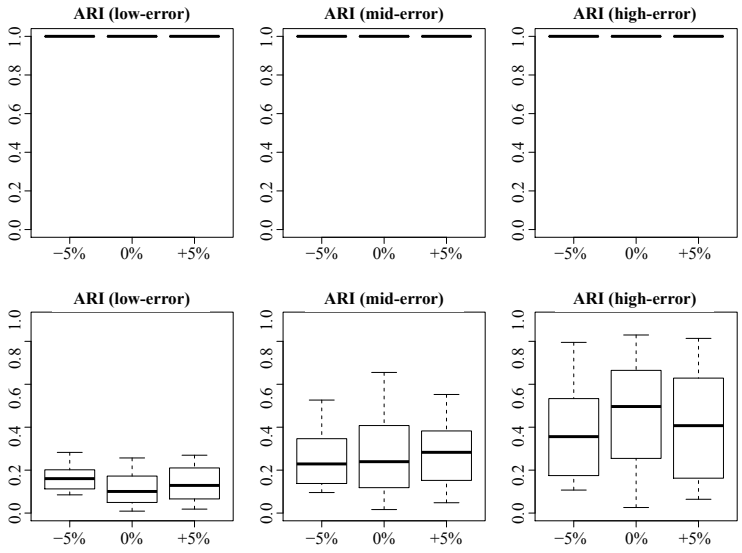



FIGURE 5.3: Boxplot of *ARI* (adjusted Rand index) and $AAE(\hat{\mathbf{Y}}, \mathbf{Y})$ values in the case with $(n, p) = (30, 100)$.

of 100 $\mathbf{X}$s, the best cardinality was selected with AIC and BIC: $c_{AIC}$ and $c_{BIC}$ were used for the selected cardinality, respectively. The bias of the estimated cardinality was evaluated with the standardized difference ($SD$) defined as

$$SD(m) = (pk)^{-1}(c_T - c). \tag{5.27}$$

Figure 5.1 shows the boxplots of the resulting $SD(c_{AIC})$ and $SD(c_{BIC})$ in the medium-error condition. The $SD$ values are found to approximate zero in the BIC-based cardinality selection, which indicates that the BIC-based procedure almost perfectly specifies the true cardinality, although it slightly overestimates it in a few cases. Conversely, the AIC-based selection tended to considerably underestimate the true cardinality, which is seen in the result that the discrepancy between the true and estimated cardinality is over 7% of all entries in $\mathbf{Y}$ at 50 percentile of $SD(c)$. These results show that the BIC-based is more precise than the AIC-based selection, suggesting that the former should be used.

### 5.5.2 Accuracy of Parameter Estimation

In order to assess how well the parameter matrices are recovered in the CCKM algorithm, we considered the cases where $(n, p) = (100, 30)$ and $(30, 100)$ and three error levels ($\rho = 0.90, 0.70, 0.50$). A hundred artificial datasets $\mathbf{X}$ were generated as described in the previous section. Next, CCKM was applied to $\mathbf{X}$ with its cardinality set as $c = c_T$ (identical to the true cardinality), $c = c_T - 0.05 \times pk$ (5% fewer than the true cardinality), and $c = c_T + 0.05 \times pk$ (5% more than the true cardinality) in order to assess the sensitivity to cardinality misidentification. For each of 2 (dimension of data matrix; $(n, p) = (100, 30), (30, 100)$) $\times$ 3 (error level; $\rho = 0.90, 0.70, 0.50$)

$\times$ 3 (cardinality setting; 5% fewer, identical to true cardinality, and 5% more) combinations, we generated 100 datasets in the manner described in the previous section. The resulting parameters $(\hat{\mathbf{M}}, \hat{\mathbf{Y}})$ were compared to their true counterparts and the accuracy of their recovery was evaluated by the following indices. For the membership matrix, we compared $\hat{\mathbf{M}}$ and $\mathbf{M}$ in terms of the Adjusted Rand Index ($ARI$) (Rand, 1971; Hubert and Arabie, 1985) of these matrices, ranged from 0 to 1 with $ARI = 1$ representing the perfect coherence of the two partitions shown by $\hat{\mathbf{M}}$ and $\mathbf{M}$. The proximity between $\hat{\mathbf{Y}} = \{\hat{y}_{jl}\}$ and its true counterpart $\mathbf{Y} = \{y_{jl}\}$ was evaluated with Averaged Absolute Error ($AAE$) defined as $AAE(\hat{\mathbf{Y}}, \mathbf{Y}) = (pk)^{-1} \sum_{j,l} |\hat{y}_{jl} - y_{jl}|$, which indicates the averaged discrepancy between the $pk$ elements in $\hat{\mathbf{Y}}$ and $\mathbf{Y}$. It should be noted that, before computing the four indices, we must choose the $k \times k$ permutation matrix that minimizes $||\hat{\mathbf{Y}}\mathbf{P} - \mathbf{Y}||^2$, to eliminate the freedom of row-permutation of $\hat{\mathbf{Y}}$ shown as $||\mathbf{X} - \mathbf{M}\mathbf{Y}'||^2 = ||\mathbf{X} - (\mathbf{M}\mathbf{P})(\mathbf{Y}\mathbf{P})'||$. Therefore, the permutation of $\mathbf{P}$ was chosen that minimizes $||\hat{\mathbf{Y}}\mathbf{P} - \mathbf{Y}||^2$ among all possible $p!$ permutation matrices. We use $\hat{\mathbf{M}}\mathbf{P}$ for $\hat{\mathbf{M}}$ and $\hat{\mathbf{Y}}\mathbf{P}$ for $\hat{\mathbf{Y}}$ with the chosen $\mathbf{P}$.

The CCKM algorithm is run starting from $t_{max}$ different initial values in order to avoid accepting a local minimum as the final solution. In this simulation, we ran the CCKM algorithm from 300 different initial values; i.e., $t_{max} = 300$.

Figures 5.2 and 5.3 show the parameter recovery results in the two cases of data sizes. Overall, the parameters were correctly recovered by CCKM, in that the $ARI$ indices attained their maximum 1 in almost all cases, and that the discrepancy between $\hat{\mathbf{Y}}$ and $\mathbf{Y}$ indicated by the $AAE$ values are sufficiently small. Even in the case where $\rho = 0.50$, where the error variance amounts to the half of the total variance of the dataset, the $AAE$ value is lower than 0.3 at 50 percentile. Further, even if the cardinality is misidentified, the levels of the $ARI$ and $AAE$ values are still satisfactory. Thus, it can be considered that the true cardinality can be identified fairly accurately in CCKM. These results allow us to conclude that the performances of CCKM is suitable for dealing with practical problems, in that the CCKM almost perfectly recovers the true parameters.

## 5.6 Real Data Examples

In this section, we illustrate CCKM with the cardinality selection procedure, using three real data examples. Further, the modified version of CCKM with row/column-cardinality constraints are finally illustrated with an additional example.

### 5.6.1 Example 1: Fisher's Iris Data

CCKM was applied to Iris data (Fisher, 1936), in which the 150 observations sampled from 3 species were measured with respect to 4 variables. Note that the data matrix was column-wise standardized beforehand. In order to find the optimal cardinality, the BIC-based cardinality selection procedure was used and the results suggested that

TABLE 5.2: Estimated centroid matrices by CCKM with $c = 10$ (left) and $k$-means (right); a blank cell shows exact-zero element.

| | CCKM | | | $k$-means | | |
|---|---|---|---|---|---|---|
| | C1 | C2 | C3 | C1 | C2 | C3 |
| Sepal.Length | 1.030 | | -0.999 | 1.132 | -0.050 | -1.011 |
| Sepal.Width | | -0.969 | 0.903 | 0.088 | -0.880 | 0.850 |
| Petal.Length | 0.940 | | -1.299 | 0.993 | 0.347 | -1.301 |
| Petal.Width | 0.969 | | -1.252 | 1.014 | 0.281 | -1.251 |

$c = 8$ is the best. We also applied $k$-means clustering to the dataset for comparison. The estimated centroid matrices are shown in Table 5.2.

In the centroid matrix estimated by CCKM in Table 5.2, we can find that the first cluster is contrasted to the second one for Sepal.Length and Sepal.Width. Further, the second cluster is different from the other clusters for Sepal.Width, with Versicolor in the second cluster characterized by narrow sepals. The contrast between clusters is clearer in the centroid matrix of CCKM than in that of the resulting $k$-means clustering, because the former has exactly zero elements. Further, as shown in Table 5.3, the three clusters are associated with the three species, Setosa, Versicolor, and Virginica, respectively. It can be seen that the estimated memberships correspond to the species, since $(49 + 37 + 42)/150 = 85.3\%$ of 150 observations are correctly classified into the true species, while $(50 + 39 + 36)/150 = 83.3\%$ in the $k$-means clustering.

The *ARI* (with its 95 % confidence interval by Steinley et al. (2016)) between the partition obtained by CCKM for the three species was 0.645 ($[0.626, 0.664]$), while $ARI = 0.620$ ($[0.601, 0.638]$) for $k$-means clustering. Also, *ARI* for the two partitions was 0.832 ($[0.813, 0.851]$), and it suggests that two partitions obtained by CCKM and $k$-means are similar, even if the cardinality of centroid matrix is restricted in CCKM.

These results demonstrate that CCKM yields an easily interpreted centroid matrix subject to the cardinality constraint, and the classification accuracy is improved.

TABLE 5.3: Contingency table for species vs. the estimated partitions by CCKM and $k$-means.

| | CCKM($c = 8$) | | | $k$-means | | |
|---|---|---|---|---|---|---|
| | C1 | C2 | C3 | C1 | C2 | C3 |
| Setosa | 49 | 0 | 0 | 50 | 0 | 0 |
| Versicolor | 1 | 37 | 8 | 0 | 39 | 14 |
| Virginica | 0 | 13 | 42 | 0 | 11 | 36 |

### 5.6.2 Example 2: Wine Data

The second example is that of the Wine data available at the UCI machine learning repository, in which the 178 types of wine are evaluated in terms of 13 chemical features. Note that the data matrix was column-wise standardized beforehand. CCKM

was performed for the dataset with $c = 35$ selected by the BIC-based procedure. Table 5.4 shows the estimated centroid matrix. It can be seen that each cluster is characterized by a small number of chemical features, which highlights the contrasts among clusters. For example, the contrast between the fourth and fifth clusters can be found in "Malic acid" and "Hue"; the observations classified into the fourth cluster are characterized by lower "Malic acid" and higher "Hue." Similarly, the other contrast between clusters can be found in several variables, such as "Ash," "Total phenols," and "Nonflavonoid phenols," which are clearly highlighted by the cardinality constrained estimation of nonzero and exact-zero elements.

Originally, the wines in the dataset are classified into three categories. In Table 5.5, we compare these categories and the clusters obtained by CCKM and $k$-means. Note that the number of clusters in $k$-means was set at 5 for comparison. It is interesting that clusters obtained by the two procedures are fairly similar, as indicated that $ARI = 0.715$ with its 95% interval [0.700, 0.731] for the two partitions. For both procedures, the first and fifth clusters correspond to Category 1 and 3, respectively. Further, the wines in Category 2 is splitted into three clusters (second, third, and fourth clusters). It suggests that even if CCKM reduces cardinality of centroid matrix, it produces similar cluster structure to $k$-means clustering.

TABLE 5.4: Centroid matrix obtained by CCKM with cardinality of $r = 35$. The exact-zero elements are shown as blank cells.

| | C1 | C2 | C4 | C4 | C5 |
|---|---|---|---|---|---|
| Alcohol | 0.930 | -0.613 | -0.993 | -0.737 | |
| Malic acid | | | | -0.843 | 0.930 |
| Ash | | 1.960 | | -1.225 | |
| Alcalinity of ash | -0.780 | 1.499 | 0.668 | -0.587 | 0.524 |
| Magnesium | 0.505 | 1.548 | -0.701 | | |
| Total phenols | 0.881 | | | | -1.004 |
| Flavonoids | 0.958 | 0.491 | | | -1.233 |
| Nonflavonoid phenols | -0.611 | | | | 0.743 |
| Proanthocyanidins | 0.599 | | | | -0.774 |
| Color intensity | | | -1.007 | -0.709 | 0.953 |
| Hue | | | | 0.975 | -1.180 |
| OD280 315 | 0.772 | | | | -1.275 |
| Proline | 1.204 | | -0.807 | -0.652 | |

TABLE 5.5: Contingency table for wine categories vs. the estimated partitions by CCKM and $k$-means.

| | CCKM($c = 35$) | | | | | $k$-means | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | C1 | C2 | C3 | C4 | C5 | C1 | C2 | C3 | C4 | C5 |
| Category 1 | 56 | 2 | 0 | 1 | 0 | 54 | 4 | 0 | 1 | 0 |
| Category 2 | 1 | 4 | 37 | 27 | 2 | 0 | 18 | 25 | 27 | 1 |
| Category 3 | 0 | 1 | 0 | 0 | 47 | 0 | 0 | 1 | 0 | 47 |

### 5.6.3 Example 3: Job Impression Data

Here, we illustrate RC-CCKM by using the job impression dataset that was also used in Section 5.1. Note that the data matrix was column-wise centered beforehand. Row- and column-wise cardinalities were set to $r(j) = k/2 = 2$ and $c(l) = p/2 = 6$ for $j = 1, \cdots, 12$, $l = 1, \cdots, 4$. This implies "Each row should contain at least one zero and each column should contain at least $k$ zeros," which is a property of an easily interpreted matrix (Thurstone 1947).

The estimated centroid matrix is presented in Table 5.6. The resulting cardinality of the centroid matrix is 19, which is $19/48 \approx 39.6\%$ of all entries. As in the previous two examples, we can see clear differences between clusters and the homogeneity in each cluster. As the row-wise cardinality was set at 2, each row has at least two zero elements and the remaining nonzero elements show the contrast of extreme positive and negative values, as found in the rows of "powerful" and "strong," for example. Furthermore, each cluster corresponds to fewer variables, as compared with the results of $k$-means in Table 5.1, which facilitates the interpretation of clusters. It should be noted that the proportion of the variance explained by the CCKM is 60.8%, which is not substantially lower than the proportion of 68.9% in the $k$-means solution. Further, the memberships estimated in CCKM perfectly correspond to those in the $k$-means clustering. These results demonstrate promising performances of RC-CCKM: it gives easily interpreted solutions while the goodness of fit to data is retained.

TABLE 5.6: Estimated centroid matrix by CCKM with row/column-cardinality constraint and four clusters for job impression data; exact-zero elements are shown as blank cells.

| | C1 | C2 | C3 | C4 |
|---|---|---|---|---|
| admirable | 0.338 | | | |
| useful | 0.557 | -0.393 | | |
| good | | | | |
| large | | -0.418 | | |
| powerful | | -0.796 | 1.029 | |
| strong | | -0.711 | 1.014 | |
| fast | | | 0.950 | -1.100 |
| noisy | | | 0.871 | -1.212 |
| young | | 0.686 | | -1.548 |
| honest | 0.443 | | | |
| stubborn | 0.569 | -1.239 | | |
| busy | 0.590 | | | -1.243 |

## 5.7 Concluding Remarks

In this paper, we addressed the difficulty in interpreting the centroid matrix resulting from the standard $k$-means clustering. We proposed a new procedure called CCKM in which the cardinality of the centroid matrix is directly constrained ti improve its

interpretability. CCKM produces a centroid matrix with reduced cardinality and its interpretation is easier than that of the standard $k$-means clustering, because between-cluster contrasts are highlighted by exact-zero elements. We also proposed a cardinality selection procedure and modified version of CCKM. The results of the simulation studies show that the BIC-based cardinality selection is more accurate than the AIC-based one, and the parameter estimation of CCKM is not sensitive to error contamination and misidentification of cardinality. Real data examples were presented to demonstrate the promising performance of CCKM and its modified version.

In clustering, the interpretability of solutions is of importance, as well as the classification accuracy. Thus, the cardinality constraint is considered to be useful for users in that the number of non-zero elements is directly associated with interpretability. Further, users can control the balance of low cardinality and model fit by tuning the cardinality parameter $c$ within a restricted range. Therefore, we can conclude that the proposed procedure is suitable for extracting interpretable clusters.

# Chapter 6

# Multivariate Analysis Procedures with Layered Simple Structure

**ABSTRACT**

The study proposes a new class of multivariate analysis procedures and it is called Layered Multivariate Analysis (LMA). In LMA, a solution matrix is expressed as a sum of several sparse matrices, which are called "layers". How sparse is the resulting solution matrix is controlled by the number of layers employed. As special cases, Layered Multivariate Regression (LMR) and Layered Principal Component Analysis (LPCA) are proposed and a regression coefficient matrix and a component loading matrix are constrained to have layered structure, respectively. A general framework of parameter estimation is proposed by means of majorization technique. It is shown that LMA's solution converges to unconstrained solution as the number of layers increases. Real data examples are presented to exemplify the effectiveness of the proposed procedures.

## 6.1 Introduction

One of the key roles of multivariate analysis is to provide insights of the relationships between variables (Izenman, 2008), and it is accomplished by interpreting of their solutions. For example, principal component analysis (Hotelling, 1936; Jolliffe, 2011) for an $N$-observations $\times$ $P$-variables data matrix $\mathbf{X}$ seeks an $N \times r(\leq P)$ component matrix $\mathbf{F}$ and a $P \times r$ loading matrix $\mathbf{A}$ that minimize

$$l_{PCA}(\mathbf{F}, \mathbf{A}) = ||\mathbf{X} - \mathbf{F}\mathbf{A}'||^2 \tag{6.1}$$

subject to the constraint that $n^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_r$ where $\mathbf{I}_r$ denotes the $r \times r$ identity matrix. The relationship between the $P$ variables in $\mathbf{X}$ is summarized by the $r$ principal components. A loading matrix $\mathbf{A}$ indicates the correspondence between $P$ variables in

This chapter is based on the published article Yamashita & Adachi(in press) with some modifications and extensions.

rows and $r$ principal components in columns. Namely, it serves to find some subsets of variables corresponding to principal components, and to understand what the components stand for. Interpreting $\mathbf{A}$ is, however, not always straightforward because the contrast between/within column(s) of $\mathbf{A}$ is often unclear, and this is mainly because interpretability is not considered in the traditional formulation of PCA.

Vichi and Saporta (2009) considered the above problem and proposed a modified PCA procedure called *clustering and disjoint principal component analysis* (CDPCA) where a loading matrix is constrained to have a perfect cluster structure (Harris and Kaiser, 1964; Kaiser, 1974; Bernaards and Jennrich, 2003). CDPCA is aimed to minimize the least squares loss function

$$l_{CDPCA}(\mathbf{U}, \tilde{\mathbf{F}}, \mathbf{A}) = ||\mathbf{X} - \mathbf{U}\tilde{\mathbf{F}}\mathbf{A}'||^2 \tag{6.2}$$

where $\mathbf{U}$ and $\tilde{\mathbf{F}}$ denote an $I$ (objects) $\times k$ (clusters) membership matrix and an $k \times r$ component score matrix that has $k(< I)$ component scores corresponding to the $k$ clusters for the $I$ objects. (6.2) is minimized subject to the constraint that $\mathbf{U}$ is a binary and row-wise stochastic matrix and $\mathbf{A}$ has a perfect cluster structure; each row of $\mathbf{A}$ has a single non-zero element and zeros elsewhere. The variables in rows are therefore classified into $r$ disjoint sets each of which corresponds to a column, and thus the components are self-evidently characterized by these sets. Note that original CDPCA disjointly classifies not only $I$ objects into $k$ classes but also $P$ variables into $r$ classes. In this paper, we focus on variable classification because it is of prominent importance for interpreting components. Importantly, Vichi (2017) proposed a procedure for estimating a loading matrix having such structure in factor analysis and the above approach can be therefore extended to other techniques of multivariate analysis. It is considered to be beneficial to extend the above procedure to multivariate regression (Izenman, 2008) and $k$-means clustering (MacQueen, 1967; Aggarwal, 2014), for example. On the other hand, the drawback of this approach is that the constraint imposed on the solution (i.e., perfect cluster structure) is too restrictive in many practical cases, and the resulting fit to a data matrix is generally poor. Therefore, a family of procedures that produces interpretable solutions with sufficient fitness to a data matrix is considered to be beneficial in practical use cases, especially for exploratory purpose of multivariate analysis.

The main contribution of this study is to propose a new approach of multivariate analysis that is called *layered multivariate analysis* (LMR). The concept of LMA is that a solution matrix is constrained to be a sum of matrices, which are called *layers*, having a certain simple structure. For example, *layered PCA* (LPCA) as a member of LMA family is formulated as the minimization of

$$f_{LPCA}(\mathbf{A}_1, \cdots, \mathbf{A}_L) = ||\mathbf{X} - \mathbf{F}\sum_{l=1}^{L}\mathbf{A}_l||^2 \;\; s.t.\; n^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_r. \tag{6.3}$$

In the formulation, notice that an unknown loading matrix $\mathbf{A}$ is expressed as a sum

of $L$ layers $\mathbf{A}_1, \cdots, \mathbf{A}_L$. Each layer is constrained to posses a certain structure that is helpful for the interpretation of $\mathbf{A}$; e.g., perfect cluster structure. The proposed procedure therefore considered as a generalization of CDPCA, which corresponds to CDPCA with $L = 1$. As a result, even if a layer-level structure is restrictive as in CDPCA, by superimposing multiple layers, it is possible to obtain a solution matrix that is easy to interpret with better fitness to a data matrix. Or equivalently, it allows users to control the balance between the interpretability and the goodness-of-fit of solutions by increasing or decreasing layers that are used in LMA. As an interesting property of the proposed method, an unconstrained solution is guaranteed to be recovered by a finite set of layers and the resulting goodness of fit is improved as the number of layers increases. Based on this property, one can safely choose a desired level of simplicity between the most restrictive but simplest case ($L = 1$) and the original solution that fits to data best. The study proposes a general framework for parameter estimation in LMA, followed by introducing *Layered Multivariate Regression* (LMR) and LPCA where a coefficient matrix and a component loading matrix are expressed as a sum of layers, respectively.

The remaining parts of this chapter are organized as follows. The next section presents a general algorithm for parameter estimation in LMA, and it is specialized into LPCA and LMR. Also an algorithm for determining how many layers should be employed is proposed in Section 6.2.3. In Section 6.2.4, an important property of the proposed procedures is introduced. Two real data examples are presented in Section 6.3, in order to exemplify the effectiveness of the proposed procedures. The final section is devoted to a concluding remark.

## 6.2 Proposed Method

In this section, a general framework of LMA is proposed. Two procedures, LPCA and LMR, are proposed as a special case of the formulation, together with optimization algorithms for the minimizations. As a starting point, we define the least squares criterion in LMA as

$$f(\mathcal{P}) = ||\mathbf{M} - \mathbf{NP}||^2 \tag{6.4}$$

that is minimized over the set of parameters $\mathcal{P}$ under some constraints, if necessary. Above, matrices $\mathbf{M}$, $\mathbf{N}$, and $\mathbf{P}$ are arbitrary matrices having suitable dimensions. Entries of $\mathcal{P}$ and three matrices noted as $\mathbf{M}, \mathbf{N}$, and $\mathbf{P}$ are determined according to the problem to be solved. For example, in LPCA, $\mathcal{P}$ is given by $\{\mathbf{F}, \mathbf{A}\}$ with $(\mathbf{M}, \mathbf{N}, \mathbf{P}) = (\mathbf{X}, \mathbf{F}, \mathbf{A}')$. The criterion is defined as $f(\mathcal{P}) = ||\mathbf{X} - \mathbf{FA}'||^2$. Here, consider the following structure on $\mathbf{P}$ called layered structure;

$$\mathbf{P} = \sum_{l=1}^{L} \mathbf{P}_l \tag{6.5}$$

where $\mathbf{P}_l$ is the $l$-th layer having a specific structure, e.g., a perfect cluster structure. This serves for obtaining an interpretable solution matrix $\mathbf{P}$ having a simple structure and controlling the degree of its simplicity against the fitness to a data matrix by increasing or decreasing how many layers are employed.

Here, $f(\mathcal{P})$ is rewritten as

$$f(\mathcal{P}) = ||(\mathbf{M} - \mathbf{N}\sum_{l' \neq l} \mathbf{P}_{l'}) - \mathbf{N}\mathbf{P}_l||^2 = ||\tilde{\mathbf{M}}_l - \mathbf{N}\mathbf{P}_l||^2 \tag{6.6}$$

where $\tilde{\mathbf{M}}_l = \mathbf{M} - \mathbf{N}\sum_{l' \neq l} \mathbf{P}_{l'}$ stands for the partial residual of other than $l$-th layer. Therefore, $\mathbf{P}_l$ is obtained as the one minimizes $f^{(l)}(\mathbf{P}_l) = ||\tilde{\mathbf{M}}_l - \mathbf{N}\mathbf{P}_l||^2$ given other parameter(s) and layers kept fixed. The overall algorithm is thus composed of the following four steps where $\mathcal{P}^\sharp$ stands for a set of parameters other than $\mathbf{P}$.

1. Initialize the entries of $\mathcal{P}$.
2. For $l$-th layer ($l = 1, \cdots, L$), update the current $\mathbf{P}_l$ by the one that minimizes $f^{(l)}(\mathbf{P}_l)$ satisfying constraint on its structure, keeping other layers and parameter(s) fixed.
3. If $\mathcal{S}^\sharp \neq \emptyset$, update $\mathcal{P}^\sharp$ by the one that minimizes $f(\mathcal{P}^\sharp|\mathbf{P})$; otherwise proceed to Step 4.
4. Finish if convergence is reached; otherwise go back to Step 2.

It is guaranteed that the function value monotonically decreases while repeating Step 2 to 3. In order to reduce the possibility to select a local minimum as a final solution, it is recommended to use 100 initial values for $\mathcal{P}$ and accept the solution as optimal one that gives the lowest function value of $f(\mathcal{P})$. In the following, a general algorithm to minimize $||\tilde{\mathbf{M}}_l - \mathbf{N}\mathbf{P}_l||^2$ over $l$-th layer subject to constraint on its structure is derived, while the optimization in Step 3 is dependent on the problem to be solved.

For the minimization of $f^{(l)}(\mathbf{P}_l)$, we use *majorization* function that is defined as follows:

**Definition 6.2.1.** *Let $g(\mathbf{\Theta})$ be an arbitrary matrix function of a matrix $\mathbf{\Theta}$. The function $m(\mathbf{\Theta})$ is said to be majorization function of $g(\mathbf{\Theta})$ if $m(\mathbf{\Theta})$ satisfies*

1. $g(\tilde{\mathbf{\Theta}}) = m(\tilde{\mathbf{\Theta}})$
2. $g(\mathbf{\Theta}) \geq m(\mathbf{\Theta})$ *for all possible* $\mathbf{\Theta}$*s.*

These two conditions state that the majorization function $m(\mathbf{\Theta})$ touches $g(\mathbf{\Theta})$ at the point $\mathbf{\Theta} = \tilde{\mathbf{\Theta}}$ and is never less than $g(\mathbf{\Theta})$. See Kiers (1990) for details. Following Kiers (2002), a majorization function of $f^{(l)}(\mathbf{P}_l)$ is given by

$$m^{(l)}(\mathbf{P}_l) = \alpha||\mathbf{Q}_l - \mathbf{P}_l||^2 \tag{6.7}$$

where

$$\mathbf{Q}_l = \mathbf{P}_l^c - \alpha^{-1}(\mathbf{N}'\tilde{\mathbf{M}}_l - \mathbf{N}'\mathbf{N}\mathbf{P}_l^c). \tag{6.8}$$

Above, $\alpha$ and $\mathbf{P}_l^c$ denote a positive integer larger than the largest eigen value of $\mathbf{N}'\mathbf{N}\otimes\mathbf{I}$, and the current estimate for $\mathbf{P}_l$, respectively, with $\mathbf{I}$ being an identity matrix. Considering the definition of the majorization function, minimization of $f^{(l)}(\mathbf{P}_l)$ is accomplished by minimizing $m^{(l)}(\mathbf{P}_l)$ and replacing $\mathbf{P}_l$ as the minimizer. Further, this minimization is simply accomplished as follows. Here, as an example, consider the minimization of $m^{(l)}(\mathbf{P}_l)$ under the constraint that $\mathbf{P}_l$ has a perfect cluster structure. Since $m^{(l)}(\mathbf{P}_l)$ is proportional to a sum of squared differences between $\mathbf{Q}_l$ and $\mathbf{P}_l$, the $(i,j)$-th element of the optimal $\mathbf{P}_l$ denoted as $p_{ij}^{(l)}$ is simply obtained as

$$p_{ij}^{(l)} = \begin{cases} q_{ij}^{(l)} \ iff \ |q_{ij}^{(l)}| > |q_{kj}^{(l)}| \ (k \neq i) \\ 0 \ \ otherwise \end{cases} \tag{6.9}$$

where $q_{ij}^{(l)}$ and $p_{ij}^{(l)}$ denote the $(i,j)$-th element of $\mathbf{Q}_l$ and $\mathbf{P}_l$, respectively. (6.9) indicates that the optimal $\mathbf{P}_l$ is set as $\mathbf{Q}_l$ and replace the elements other than the maximum element in absolute with zeros for each row. In other cases, if the cardinality (i.e., number of non-zero elements) of $\mathbf{P}_l$ is constrained to be lower than a positive integer $c$, $\mathbf{P}_l$ is obtained as $\mathbf{Q}_l$ and replace the elements other than the $c$-th largest elements in $\mathbf{Q}_l$ with zeros. As seen here, various types of constraint on the structure of $\mathbf{P}$ can be considered in LMA, since the optimization with respect to $\mathbf{P}_l$ is simply reduced to the minimization of $m(\mathbf{P}_l) = \alpha||\mathbf{F}_l - \mathbf{P}_l||^2$, which can be minimized easily.

In the following, as special cases of the above algorithm, the study proposes two procedures in which solution matrices are constrained to be layered for their interpretability.

### 6.2.1 Layered Multivariate Regression (LMR)

In LMR a regression coefficient matrix $\mathbf{W}$ is expressed as a sum of $L$ layers, and it is formally defined as the minimization of

$$f_{LMR}(\mathbf{W}) = ||\mathbf{Y} - \mathbf{X}\mathbf{W}||^2 = ||\mathbf{Y} - \mathbf{X}\sum_{l=1}^{L}\mathbf{W}_l||^2 \tag{6.10}$$

where $\mathbf{Y}(n \times q)$ denotes the matrix of $q$ dependent variables with respect to $n$ observations. The $q$ dependent variables are regressed on $p$ independent ones in an $n \times p$ matrix $\mathbf{X}$ with an unknown $p \times q$ matrix of regression coefficients $\mathbf{W}$. Considering that $\{\mathbf{M}, \mathbf{N}, \mathbf{P}\}$ in (6.10) corresponds to $\{\mathbf{Y}, \mathbf{X}, \mathbf{W}\}$ and $\mathcal{P}^\sharp = \emptyset$ in (6.10), the majorization function in (6.7) is re-expressed as following in LMR;

$$m_{LMR}^{(l)}(\mathbf{W}_l) = \alpha||\{\mathbf{W}_l^c - \alpha^{-1}(\mathbf{X}'\tilde{\mathbf{Y}}_l) - \mathbf{X}'\mathbf{X}\mathbf{W}_l^c\} - \mathbf{W}_l||^2 \tag{6.11}$$

that majorizes $f_{LMR}^{(l)} = ||\tilde{\mathbf{Y}} - \mathbf{X}\mathbf{W}_l||^2$ to be minimized for updating the $l$-th layer $\mathbf{W}_l$, where $\tilde{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\sum_{l' \neq l}\mathbf{W}_{l'}$ stands for the partial residual of other than the $l$-th layer. Above, $\alpha$ is a positive integer larger than the largest eigen value of $\mathbf{X}'\mathbf{X} \otimes \mathbf{I}_q$ where $\mathbf{I}_q$ denotes the $q$-dimensional identity matrix, and $\mathbf{W}_l^c$ is the current estimate for $\mathbf{W}_l$.

As noted in the previous subsection, $m_{LMR}^{(l)}(\mathbf{W}_l)$ is easily minimized under some constraint on $\mathbf{W}_1, \cdots, \mathbf{W}_L$. This study considers a perfect cluster structure for all layers, and such layers are obtain by (6.9) for $l$-th layer with $\mathbf{Q}_l = \mathbf{W}_l^c - \alpha^{-1}(\mathbf{X}'\tilde{\mathbf{Y}}_l) - \mathbf{X}'\mathbf{X}\mathbf{W}_l^c$ and $\mathbf{P}_l = \mathbf{W}_l$.

### 6.2.2 Layered Principal Component Analysis (LPCA)

As already mentioned, LPCA in which a loading matrix is expressed as a sum of $L$ layers is formally defined as the minimization of (6.3) given a data matrix $\mathbf{X}$. It is equivalent to minimize

$$f_{LPCA}(\mathbf{F}, \mathbf{A}) = ||\mathbf{X} - \mathbf{F}\mathbf{A}'||^2 = ||\mathbf{X} - \mathbf{F}\sum_{l}^{L}\mathbf{A}_l'||^2 \tag{6.12}$$

that is equivalent to (6.4) with $\{\mathbf{M}, \mathbf{N}, \mathbf{P}\} = \{\mathbf{X}, \mathbf{F}, \mathbf{A}'\}$ and $\mathcal{P} = \{\mathbf{F}, \mathbf{A}\}$. Further, the criterion to be minimized for the $l$-th layer of $\mathbf{A}_l$ is expressed as

$$f_{LPCA}^{(l)}(\mathbf{F}, \mathbf{A}) = ||\tilde{\mathbf{X}}_l - \mathbf{F}\mathbf{A}_l'||^2 \tag{6.13}$$

with $\tilde{\mathbf{X}}_l = \mathbf{X} - \mathbf{F}\sum_{l' \neq l}\mathbf{A}_{l'}$. Therefore, using $n^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_r$, we obtain a majorization function of (6.13) as

$$\begin{aligned} m_{LPCA}^{(l)}(\mathbf{A}_l) &= \alpha||\{\mathbf{A}_l^c - \alpha^{-1}(\tilde{\mathbf{X}}_l'\mathbf{F} - \mathbf{A}_l^c\mathbf{F}'\mathbf{F})\} - \mathbf{A}_l||^2 & (6.14)\\ &= ||n^{-1}\tilde{\mathbf{X}}_l'\mathbf{F} - \mathbf{A}_l||^2 & (6.15)\end{aligned}$$

where $\mathbf{A}_l^c$ is the current estimate of $\mathbf{A}_l$. Above $\alpha = n$ is used since $\alpha$ is define as a positive integer larger than the largest eigen value of $\mathbf{F}'\mathbf{F} = \mathbf{I}_p = \mathbf{I}_{n+p}$ that is equal to $1 < n$. Under the constraint that $\mathbf{A}_1, \cdots, \mathbf{A}_L$ have a perfect cluster structure, (6.15) is easily minimized by the same manner as in LMR, namely, by row-wise searching of the highest element in absolute of $n^{-1}\tilde{\mathbf{X}}_l'\mathbf{F}$ for all $l$s.

### 6.2.3 Determining number of layers $L$

Before applying LMA including three procedures introduced above, the number of layers $L$ has to be specified beforehand. As to be discussed and numerically exemplified in the next section, the function value of LMA in (6.4), namely $f_{LMR}(\mathbf{W})$ (LMR) or $f_{LPCA}(\mathbf{F}, \mathbf{A})$ (LPCA), keep on increasing and approach to the attained function values for unconstrained solutions without layer structure, as $L$ increases. For example, in LMR, $f_{LMR}(\hat{\mathbf{W}}_{OLS}) - f_{LMR}(\hat{\mathbf{W}})$ approaches to zero as the number of layers increases,

where $\hat{\mathbf{W}}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ denotes the OLS estimate without layer structure for $\mathbf{W}$ and $\hat{\mathbf{W}}$ is the estimate of LMR.

Using this property, the study recommend to use *scree method* in order to determine $L$ where the function values after convergence are plotted against a specific range of $L$s. $L$ can be determined at a scree point of the plot, which can be considered as an enough number of layers to approximate the unconstrained solution by the one with layer structure. It should be noted that $L$ specified by the scree method should be considered as a reference for the best number of layers, and it is recommended to investigate the resulting solutions for some candidates for $L$. This is because $L$ controls the balance between interpretability and fitness of the solution and the their balance considerably depends on what is aimed in the data analysis. For example, if a scree plot suggest that $L = 3$ is best, it is recommended to check the results with $L = 1, 3, 4$, in order to confirm the resulting interpretability of solutions.

### 6.2.4 Properties

Since the novelty of the proposed method is that a parameter matrix is expressed as a sum of layers, theoretical investigation of LMA's solutions is necessary. For the purpose, the subsection begins with the following theorem;

**Theorem 6.2.1.** *Let $\mathbf{P}_l (l = 1, \cdots, L)$ be the $l$-th layer of a certain matrix $\mathbf{P}$, and $\mathbf{P}^{(L)} = \sum_l^L \mathbf{P}_l$. For each of the attained minima of $f_{\mathbf{P}}(\mathbf{P}) = ||\mathbf{M}-\mathbf{NP}||^2$ with different $L$s, we have*

$$\min . f_{\mathbf{P}}(\mathbf{P}^{(L)}) \geq \min . f_{\mathbf{P}}(\mathbf{P}^{(L+1)}) \geq f_{\mathbf{P}}(\hat{\mathbf{P}}_{OLS}) \tag{6.16}$$

*where $\hat{\mathbf{P}}_{OLS}$ denotes the unconstrained solution for $f_{\mathbf{P}}(\mathbf{P})$.*

*Proof.* Let $\mathbf{P}^{(L)\sharp} = \sum_l^L \mathbf{P}_l^\sharp$ and $\mathbf{P}^{(L+1)\star} = \sum_l^{L+1} \mathbf{P}_l^\star$ denote the minimizer of $f_{\mathbf{P}}(\mathbf{P}^{(L)})$ and $f_{\mathbf{P}}(\mathbf{P}^{(L+1)})$, respectively. Note that $\mathbf{P}_l^\sharp$ is not always identical to $\mathbf{P}_l^\star$, because their estimates depend on the current value for the layers other than the $l$-th layer. Here, consider the minimization of

$$f_{\mathbf{P}}(\mathbf{P}_{L+1}|\mathbf{P}^{(L)\sharp}) = ||\mathbf{M} - \mathbf{N}(\mathbf{P}^{(L)\sharp} + \mathbf{P}_{L+1})||^2 \tag{6.17}$$

that is equivalent to obtain $(L + 1)$-th layer $\mathbf{P}_{L+1}$ subject to the $1, \cdots, L$-th layers kept fixed. For its attained minimum, we have

$$f_{\mathbf{P}}(\mathbf{P}^{(L)\sharp}) \geq \min . f_{\mathbf{P}}(\mathbf{P}_{L+1}|\mathbf{P}^{(L)\sharp}) \tag{6.18}$$

since the function on the right side is over-parametrized than the left side. Further, as to the minimum of $\min . f_{\mathbf{P}}(\mathbf{P}^{(L+1)})$, we also have

$$\min . f_{\mathbf{P}}(\mathbf{P}_{L+1}|\mathbf{P}^{(L)\sharp}) \geq f_{\mathbf{P}}(\mathbf{P}^{(L+1)\star}) \tag{6.19}$$

because the right side is minimized over a total of $L+1$ layers, while the right side is minimized over only a single layer $\mathbf{P}_{L+1}$. In addition, we have

$$f_{\mathbf{P}}(\mathbf{P}^{(L+1)\star}) \geq f_{\mathbf{P}}(\hat{\mathbf{P}}_{OLS}) \tag{6.20}$$

since it is guaranteed that $\hat{\mathbf{P}}_{OLS}$ minimizes $f_{\mathbf{P}}(\mathbf{P})$ the most within the feasible set of $\mathbf{P}$s. The above inequalities indicate that $f_{\mathbf{P}}(\mathbf{P}^{(L)\sharp}) \geq \min . f_{\mathbf{P}}(\mathbf{P}_{L+1}|\mathbf{P}^{(L)\sharp}) \geq f_{\mathbf{P}}(\mathbf{P}^{(L+1)\star}) \geq f_{\mathbf{P}}(\hat{\mathbf{P}}_{OLS})$ that leads (6.16). □

The theorem indicates that, in LMA, the attained minimum of function value in (6.4) decreases as the number of layers increases, which means that fitness to dataset is improved as $L$ increases. Further, if $L$ is set to a sufficiently large value, the obtained solution as a sum of constrained layers perfectly recovers the unconstrained solution.

The above properties is exemplified here by a real data example. LMR and ordinal multiple regression analysis are both applied to Tobacco dataset (Izenman, 2008) in which the total of six variables of chemical compounds Nitrogen (N), Chlorine (Cl), Potassium (K), Phosphorus (P), Calcium (Cl), and Magnesium (Mg), are regressed on the three characteristics (Burn rate, Sugar and Nicotine content) of tobacco leaves. The number of layers $L$ in LMR is set at the range $[1, \cdots, 8]$, and the resulting sum of squared errors (SSE), obtained as a function value of $f_{LMR}(\mathbf{W})$, is plotted against $L$s in Figure 6.1. In the figure, SSE of unconstrained solution obtained by $||\mathbf{Y}-\mathbf{X}\mathbf{W}_{OLS}||^2$ is shown in dotted line, and it can be seen that fitness to the dataset is improved as the layers increases and never exceed the one of unconstrained solution, which exemplifies Theorem 2. In addition, the fitness of LMR solution is almost identical to the one of unconstrained solution at $L = 5$ or later. The result suggests that we can safely choose the number of layers between the continuum between $L = 1$ (simplest but poor fitness) and unconstrained solution.

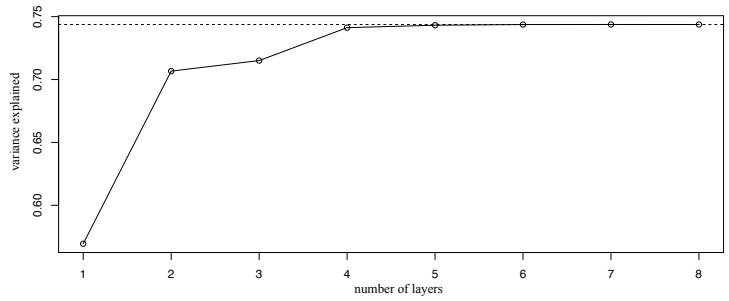



FIGURE 6.1: Scree plot of variance explained against the number of layers; variance explained of unconstrained solution is shown as dotted line

## 6.3 Real Data Examples

As demonstrations of the proposed procedures, the section presents two real data examples in which LMR and LPCA are applied.

### 6.3.1 Example of LMR: Tobacco Data

First, LMR is applied to Tobacco dataset that has been introduced above. Table 6.1 shows the results of LMR for $L = 1, \cdots, 4$ and the unconstrained solution, together with their proportion of explained variance as a measure of model-fitness. It can be seen that the number of the exact zero elements increases as $L$ increases, and the clear correspondences between the independent variables (rows) and the dependent variables (columns) are observed in the cases of fewer layers, although the goodness of fit is worse than the unconstrained solution. In other words, the result suggests trade-off of better fit and simplicity, because the value of variance explained gets higher but fewer exact zero elements as the number of layers increases. The scree plot in Figure 6.1 indicates that we have a scree point at $L = 3$, which can be chosen as an appropriate number of layers. Referring the solution with $L = 3$, one can see that concentration of Nicotine is mainly explained by chemical compounds other than Potassium and Phosphorus. 4 elements in coefficient matrix is estimated to be zero exactly, but the decrement of model fit is only $4\% = 74.4\% - 74.0\%$, which indicates that these elements are not necessary to explained the relation between independent/dependent variables.

TABLE 6.1: Estimated coefficient matrices by LMR with $L = 1, \cdots, 4$ and unconstrained solution with proportion of variance explained (Vexp.); element equaling to zero shows a blanc cell.

| | LMR ($L = 1$) | | | LMR ($L = 2$) | | | LMR ($L = 3$) | | | LMR ($L = 4$) | | | unconstrained | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Brate | Sug | Nic | Brate | Sug | Nic | Brate | Sug | Nic | Brate | Sug | Nic | Brate | Sug | Nic |
| N | | -0.69 | | | -0.65 | 0.51 | | -0.52 | 0.39 | | -0.58 | 0.30 | 0.10 | -0.58 | 0.29 |
| Cl | -0.61 | | | -0.60 | 0.38 | | -0.60 | 0.41 | -0.31 | -0.60 | 0.39 | -0.32 | -0.58 | 0.39 | -0.32 |
| K | 0.62 | | | 0.52 | 0.26 | | 0.52 | 0.14 | | 0.52 | 0.20 | 0.11 | 0.45 | 0.20 | 0.11 |
| P | | 0.17 | | -0.14 | 0.20 | | -0.14 | 0.22 | | -0.14 | 0.22 | | -0.13 | 0.22 | -0.05 |
| Ca | 0.33 | | | 0.47 | | | 0.47 | | 0.18 | 0.47 | 0.11 | 0.25 | 0.41 | 0.11 | 0.24 |
| Mg | | | 0.73 | -0.26 | | 0.43 | -0.26 | -0.21 | 0.40 | -0.26 | -0.21 | 0.47 | -0.32 | -0.22 | 0.48 |
| Vexp. | 0.569 | | | 0.701 | | | 0.740 | | | 0.742 | | | 0.744 | | |

### 6.3.2 Example of LPCA: Wine Data

In the second example, LPCA is applied to Wine dataset (Asuncion and Newman, 2007), in which 173 wines are described by 13 chemical features. The data matrix therefore consists of 173 rows and 13 columns, and standard PCA, CDPCA by Vichi and Saporta (2009), and LPCA were applied. Note that the dataset had been column-wise standardized and the number of principal components is set at 2. Before applying LPCA, how many layers should be employed for the dataset was specified via scree method. Figure 6.2 shows the scree plot of variance explained for LPCA solutions with $L = 1, \cdots, 5$. The scree plot suggests that the solution with $L = 3$ is best, while

the solution with $L = 2$ should be checked since the explained variance of the solution with $L = 3$ is almost equivalent to the one of standard PCA. The LPCA solution with $L = 2$ is therefore preferred because it is expected to be sparse and thus interpretable since it uses fewer layers.

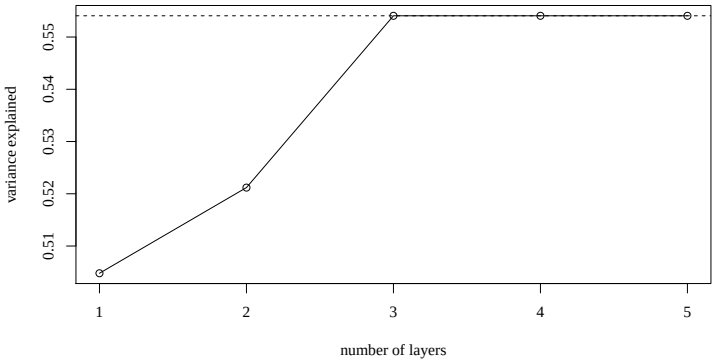



FIGURE 6.2: Scree plot of variance explained against the number of layers; variance explained of standard PCA is shown as dotted line

Figure 6.3 shows the obtained biplots of three PCA procedures, in which component scores and loadings are plotted as points and vectors, respectively. Although, the standard PCA solution fits to the dataset best, we can see that it is difficult to comprehend what each of the components stands for in the solution of standard PCA; correspondence between component (axis) and set of variables (vectors) is unclear. The solution obtained by LPCA fits slightly inferior to the one of standard PCA, but is considered to be easier to interpret, since each component is explained by the reduced number of variables because of exact zero elements in component loading matrix. For example, the first component (PC1) is manifested by the variables "Alc" and "Ash", while the second one (PC2) is characterized by "NFlaP", "Proan", and "OD". It is because, in the LPCA solution, some elements in loading matrix corresponding to the first or second component are estimated as zero. Also, in the solution of CDPCA, loading matrix is constrained have at least one zero element in each row, and it leads to the result that all variables are perfectly corresponds to the first or second component. Despite of the straightforwardness in interpretation, CDPCA fits poorer than LPCA, which explains 50.82% of total variation of the dataset. Further, the distribution of component scores is differ from the one of standard PCA, which indicates that the reduced space obtained by CDPCA is considerably differ from standard PCA that explains the dataset best. It is numerically shown in Figure6.4. The reduced space obtained by LPCA fairly approximates the one in standard PCA, while the one in CDPCA considerably differs. It can be therefore concluded that even if the CDPCA solution is easier to interpret than standard PCA and LPCA, its reduced space does not manifest the given data structure. On the other hand, LPCA successfully extract

the reduced space that is not only easy to interpret but also reflects the original data itself.

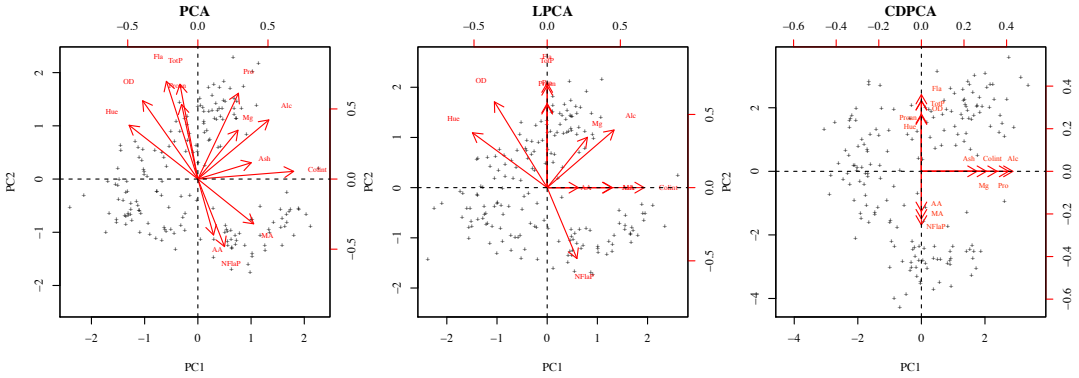



FIGURE 6.3: Principal component biplots of standard PCA, LPCA, and CDPCA

## 6.4 Concluding Remarks

The study considered to improve interpretability in multivariate analysis procedures, and proposed LMA, a novel framework of multivariate analysis procedure that has LMR and LMCA as special cases. In LMA, a coefficient matrix is expressed as a sum of layers having simple structure. For example, LMR allows to capture the relationship between independent and dependent variables more easier than unconstrained regression, because a reduced number of independent variables corresponds to each dependent variables. It is theoretically shown that LMR's solution approximates an unconstrained solution when sufficient number of layers are employed. Two numerical examples reveals that the proposed procedures works considerably well compared with the existing procedures.

Throughout this chapter, a perfect cluster structure is considered as a layer-level structure, in both LMR and LPCA. Another simple structure for layers should be considered in order to further improve interpretability. In addition, applications of the proposed method to another multivariate analysis procedures should be considered. For example, simple structure in canonical correlation analysis, multiple correspondence analysis, and $K$-means clustering would be of great help for the easier interpretation of those solutions.

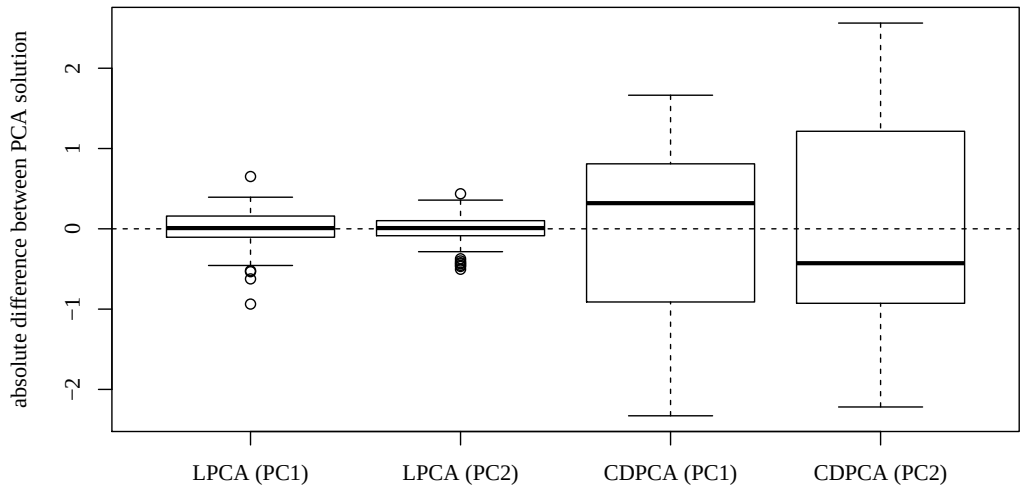



FIGURE 6.4: Differences between reduced space of LPCA/CDPCA and standard PCA

# Chapter 7

# Procrustes Penalty Function for Matching Matrices to Targets

**ABSTRACT**

Penalized estimation is widely used for obtaining sparse solutions, which facilitates easier interpretation compared with ordinal estimation procedures. In this study, as a generalized form of penalized estimation, a new penalty function is proposed. The proposed function shrinks solutions to a prespecified target matrix that possesses a certain simple structure. The resulting solution is therefore simple and easy to interpret, and its simplicity is controlled by some tuning parameters. Two applications of the proposed method are presented; sparse principal component analysis and three-way component analysis. The applications show that the proposed method surely produce sparse and interpretable solutions.

## 7.1 Introduction

Interpretability of solutions is one of the most important issues in modern multivariate analysis. Specifically, growing interest in machine learning technology makes the issue more important, in that it often involves less interpretable and black-box models (Rudin, 2019). Penalized estimation (Hastie et al., 2015) is widely used for obtaining sparse and interpretable solutions in various procedures of multivariate analysis, which uses penalized functions in order to shrink some elements toward zero. For example, Sparse Principal Component Analysis (SPCA) (Jolliffe et al., 2003; Zou et al., 2006) with an $L_1$ penalty aims to minimize

$$||\mathbf{Z} - \mathbf{F}\mathbf{A}'||^2 + \lambda||\mathbf{A}||_1 \tag{7.1}$$

This chapter is based on the published article Yamashita(in press) with some modifications and extensions.

over an $n$ (objects) $\times r$ (components) score matrix $\mathbf{F}$ and a $p$ (variables) $\times r$ loading matrix $\mathbf{A}$ given an $n \times p$ data matrix $\mathbf{Z}$, where $||\mathbf{A}||_1 = \sum |a_{jk}| (j = 1, \cdots, p;\ k = 1, \cdots, r)$ and $\lambda$ denote the $L_1$ norm of $\mathbf{A}$ and a tuning parameter, respectively. The second term in (7.1) is called a penalty function and serves to shrink the elements in $\mathbf{A}$ toward zero, and therefore the estimated $\mathbf{A}$ is of reduced cardinality.

The paper proposes a new penalty function that includes various penalty functions. The function is called a *Procrustes penalty function*, which is formally expressed as

$$P_{Pro}(\mathbf{A}|\mathbf{\Lambda}, \mathbf{T}) = ||(\mathbf{A} - \mathbf{T})\mathbf{\Lambda}||_1 \tag{7.2}$$

where $\mathbf{\Lambda}$ denotes the diagonal matrix of tuning parameters $\lambda_1, \cdots, \lambda_r (> 0)$. $\lambda_s$ controls the penalty strength on the $s$-th column of $\mathbf{A}$. $\mathbf{T}$ denotes a prespecified target matrix with the same dimension as $\mathbf{A}$. Here, consider to set $\mathbf{T}$ as a matrix with a certain simple structure. The elements in $\mathbf{A}$ shrink toward simple $\mathbf{T}$ so as to approximate its structure, and the resulting $\mathbf{A}$ is considered to be simple and easy interpret. The novel point of the proposed method is that a solution matrix directly approximates a simple structure, while the existing one shrinks all elements toward zeros without considering whether the resulting $\mathbf{A}$ has a simple structure or not. Procrustes penalty function therefore expected to yield a more simple and interpretable solution matrix compared with the existing methods. It should be noted that the article only consider the case with $L_1$-norm penalty, since the main goal of the study is to obtain sparse and interpretable solution matrices.

The remaining parts of the article are organized as follows. In the next section, an optimization algorithm for the proposed methods is presented, where a general case of regression model with Procrustes penalty is considered. There, a theorem of the minimizer for the regression problem is also presented. Further, the proposed procedure is applied to some machine learning problems, followed by their examples in Section 7.3. The final section is devoted to conclusions and future remarks.

## 7.2 Proposed Method

### 7.2.1 General Case: Multivariate Regression

Here, a multiple regression problem with a Procrustes penalty function as a general case is considered. It is formally expressed as the minimization of

$$f(\mathbf{W}) = ||\mathbf{Y} - \mathbf{X}\mathbf{W}||^2 + P_{Pro}(\mathbf{W}|\mathbf{\Lambda}, \mathbf{T}) \tag{7.3}$$

where $\mathbf{Y} = [\mathbf{y}_1, \cdots, \mathbf{y}_q]$ denotes the $n \times q$ matrix of the dependent variables, and it is regressed on the $p$ independent variables in an $n \times p$ data matrix $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_p]$. $\mathbf{W}$ is an unknown $p \times q$ matrix of regression coefficients. Above, the Procrustes penalty function applied to the coefficient matrix $\mathbf{W}$ serves to obtain an interpretable estimate of the parameter matrix $\mathbf{W}$. Here, as $P_{Pro}(\mathbf{W}|\mathbf{\Lambda}, \mathbf{T})$, the $L_1$-norm is considered;

$P_{Pro}(\mathbf{W}|\mathbf{\Lambda}, \mathbf{T}) = ||\mathbf{\Lambda}(\mathbf{A} - \mathbf{T})||_1$, because it is able to produce a $\mathbf{W}$ including some exact zero elements.

For the minimization of (7.3) over $\mathbf{W}$, we have the following theorem. Note that the theorem is valid only in the case with $L_1$-norm penalty.

**Theorem 7.2.1.** *The minimizer of the penalized loss function* $f_1(\mathbf{W}) = ||\mathbf{Y}-\mathbf{X}\mathbf{W}||^2 + ||(\mathbf{W} - \mathbf{T})\mathbf{\Lambda}||_1$ *is given by*

$$w_{jl} = \begin{cases} \hat{w}_{jl} - \frac{\lambda_j}{2||\mathbf{x}_j||^2} & \left(\hat{w}_{jl} > t_{jl} + \frac{\lambda_j}{2||\mathbf{x}_j||^2}\right) \\ \hat{w}_{jl} + \frac{\lambda_j}{2||\mathbf{x}_j||^2} & \left(\hat{w}_{jl} < t_{jl} - \frac{\lambda_j}{2||\mathbf{x}_j||^2}\right) \\ t_{jl} & \left(|w_{jl}| \leq t_{jl} + \frac{\lambda_j}{2||\mathbf{x}_j||^2}\right) \end{cases} \tag{7.4}$$

*using* $\hat{\mathbf{W}} = \{\hat{w}_{kl}\} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$*, where* $t_{jl}$ *is the* $(j,l)$*-th element of* $\mathbf{T}$ *corresponding to* $w_{jl}$*, the* $(j,l)$*-th element of* $\mathbf{W}$*.*

*Proof.* $f_1(\mathbf{W})$ can be rewritten as

$$f_1(\mathbf{W}) = \sum_{l=1}^{q}\sum_{j=1}^{p}\left(||\mathbf{y}_l - w_{jl}\mathbf{x}_j||^2 + \lambda_k|w_{jl} - t_{jl}|\right). \tag{7.5}$$

For minimizing it, consider the following three cases; [1]$w_{jl} > t_{jl}$, [2]$w_{jl} < t_{jl}$, and [3]otherwise. In the case [1], $f_1(\mathbf{W})$ is reduced to

$$f_1(\mathbf{W}) = \sum_{l=1}^{q}\sum_{j=1}^{p}\left(||\mathbf{y}_l - w_{jl}\mathbf{x}_j||^2 + \lambda_j(w_{jl} - t_{jl})\right) \tag{7.6}$$

and therefore its minimizer in the case is obtained as follows;

$$\frac{\partial f_1(\mathbf{W})}{\partial w_{jl}} = -2\mathbf{y}_l'\mathbf{x}_k + 2w_{jl}||\mathbf{x}_j||^2 + \lambda_k = 0 \Leftrightarrow w_{jl} = \hat{w}_{jl} - \frac{\lambda_j}{2||\mathbf{x}_j||^2}. \tag{7.7}$$

Further, the condition can be rewritten as

$$\hat{w}_{jl} - \frac{\lambda_j}{2||\mathbf{x}_j||^2} > t_{jl} \Leftrightarrow \hat{w}_{jl} > t_{jl} + \frac{\lambda_j}{2||\mathbf{x}_j||^2}, \tag{7.8}$$

that leads the first case of (7.4). In the similar manner, the minimizer in the case [2] can be given by $w_{jl} = \hat{w}_{jl} - \frac{\lambda_j}{2||\mathbf{x}_j||^2}$ and the condition indicates that $\hat{w}_{jl} < t_{jl} - \frac{\lambda_j}{2||\mathbf{x}_j||^2}$. The case [3] indicates $w_{jl} = t_{kl}$, and its condition is $|w_{jl}| \leq t_{jl} + \frac{\lambda_j}{2||\mathbf{x}_j||^2}$, which is not covered by the first and second conditions. These results immediately lead (7.4). □ □

(7.4) can be simply expressed as below, with $(x)_+$ being $x$ if $x \geq 0$ and $-x$ otherwise;

$$w_{jl} = \text{sign}(\hat{w}_{jl} - t_{jl})\left(|\hat{w}_{jl} - t_{jl}| - \frac{\lambda_j}{2||\mathbf{x}_j||^2}\right)_+ + t_{jl}. \tag{7.9}$$

It should be noted that (7.4) is a generalization of LASSO solution in that both are equivalent when $\mathbf{T} = \mathbf{O}_{p\times q}$, where $\mathbf{O}_{p\times q}$ denotes the $p \times q$ matrix filled with 0s. This implies that a LASSO penalty function is aimed to shrink a solution matrix toward the zero matrix with the same dimension.

### 7.2.2 Applications in Machine Learning Problems

The above minimization procedure can be applied to SPCA, in which a component loading matrix $\mathbf{A}$ is sparsely estimated. It is formulated as the minimization of

$$f_{SPCA}(\mathbf{F}, \mathbf{A}) = ||\mathbf{Z} - \mathbf{F}\mathbf{A}'||^2 + ||(\mathbf{A} - \mathbf{T})\mathbf{\Lambda}||_1. \tag{7.10}$$

subject to $n^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_r$. Note that it is equivalent to SCoTLASS (Jolliffe et al., 2003) if one set $\mathbf{T} = \mathbf{O}$. The minimization of $f_{SPCA}(\mathbf{F}, \mathbf{A})$ is attained by repeating the following two steps until the decrement of the function value converges, starting from suitable initial values for $\mathbf{F}, \mathbf{A}$.

1. Update $\mathbf{F}$ by $\mathbf{F} = n^{1/2}\mathbf{K}_r$ where $\mathbf{K}_r$ $(n \times r)$ denotes the matrix of $r$ right-eigenvectors corresponding the $r$-largest eigenvalues of $\mathbf{X}$.

2. Update $\mathbf{A}$ by (7.9) with setting $\mathbf{W} \to \mathbf{A}'$, $\mathbf{X} \to \mathbf{F}$, and $\mathbf{Y} \to \mathbf{Z}$.

The example of the SPCA procedure is shown in the next section.

Also, the proposed procedure is also applicable for three-mode component analysis (Kroonenberg, 2008). Tucker3 (Tucker, 1966; Kroonenberg and De Leeuw, 1980) and CANDECOMP/PARAFAC (Carroll and Chang, 1970) models are common choices for the problem. It is known that the former is less restrictive but difficult to interpret its result, while the latter is easier to interpret but too restrictive. It is formally expressed as follows. Let $\bar{\mathbf{X}}$ be an $n \times p \times k$ data array. Tucker3 and PARAFAC models either compress the first, second, and third modes into $s$, $t$, and $u$ components, respectively, and an $s \times t \times u$ array $\bar{\mathbf{C}}$ called a core array is obtained. Note that $s = t = u$ in PARAFAC model. In PARAFAC, the core array is restricted to be super-diagonal array and easier to be interpreted in that it contains several zero elements; the number of linkages between the components is limited. On the other hand, no constraint is imposed on the array in Tucker3 and thus often difficult to interpret (Frølich et al., 2018). In terms of fitness to the data array, Tucker3 is better because it has more unknown parameters than the other.

Here, Procrustes penalty function is used to an intermediate solution between two models in order to balance interpretability and fitness to the data array. It is accomplished by minimizing

$$f_{TP}(\mathbf{G}, \mathbf{C}, \mathbf{H}, \mathbf{E}) = ||\mathbf{Z} - \mathbf{G}\mathbf{C}(\mathbf{H} \otimes \mathbf{E})'||^2 + ||(\mathbf{C} - \mathbf{T})\mathbf{\Lambda}||_1 \tag{7.11}$$

where the former is Tucker3's loss function. Above, $\mathbf{G}(n \times s)$, $\mathbf{H}(p \times t)$, and $\mathbf{E}(k \times u)$ are unknown component loading matrices of the first, second, third mode of $\bar{\mathbf{X}}$,

respectively. $\mathbf{Z} = \{\mathbf{X}_1, \cdots, \mathbf{X}_k\}$ is the $n \times pk$ matrix of the horizontal slices of $\bar{\mathbf{X}}$. Also, $\mathbf{C} = \{\mathbf{C}_1, \cdots, \mathbf{C}_k\}$ is the $s \times tu$ matrix composed of the horizontal slices of $\bar{\mathbf{C}}$, and it is matched to $\mathbf{T}$ having a suitable dimension. Here, consider to set $\mathbf{T}$ as the matrix of horizontal slices of the core array estimated by PARAFAC. The estimated core array by minimizing (7.11) therefore shrinks towards the one by PARAFAC, and thus the array is considered to be in the middle of PARAFAC and Tucker3. The relative strength of Procrustes penalty to the Tucker3's loss function is controlled by $\mathbf{\Lambda}$; $\mathbf{\Lambda}$ having high values for its diagonal elements leads to a more PARAFAC-like core array, and vice-versa.

The minimization of (7.11) is accomplished as follows. Nelder-Mead's numerical optimization is used for minimizing $f_{TP}(\mathbf{G}, \mathbf{C}, \mathbf{H}, \mathbf{E})$ over $\mathbf{G}$, $\mathbf{H}$, $\mathbf{E}$. The core array minimizing the loss function is obtained by extending Theorem 1 to Penrose regression in which

$$f_{Pen}(\mathbf{W}) = ||\mathbf{Y} - \mathbf{X}_1\mathbf{W}\mathbf{X}_2||^2 + ||\mathbf{\Lambda}(\mathbf{W} - \mathbf{T})||^2 \tag{7.12}$$

is minimized over $\mathbf{W}$.

**Corollary 7.2.1.** *The minimizer of the loss function in (7.12) over* $\mathbf{W}$ *is given by*

$$w_{jk} = \text{sign}(\tilde{w}_{ik} - t_{jk}) \left( |\tilde{w}_{jk} - t_{jk}| - \frac{\lambda_j}{2||\mathbf{x}_j^{(1)}||^2||\mathbf{x}_k^{(2)}||^2} \right)_+ + t_{jk} \tag{7.13}$$

*where* $\mathbf{x}_j^{(1)}$ *and* $\mathbf{x}_k^{(2)}$ *denote the* $j$*-th and* $k$*-th columns of* $\mathbf{X}_1$ *and* $\mathbf{X}_2$*, respectively.*

*Proof.* It can be easily verified in the same way as Theorem 7.2.1. □ □

Using the corollary, the optimal $\mathbf{C}$ is obtained by setting $\mathbf{X}_1 \to \mathbf{G}$ and $\mathbf{X}_2 \to \mathbf{H} \otimes \mathbf{E}$. The solution for the minimization of (7.11) is obtained by repeating the following steps; updating $\mathbf{C}$ by (7.13), and updating the other parameter matrices by numerical optimization of the loss function.

## 7.3 Example

The proposed procedures are illustrated in order to show how well they work in dealing with real datasets.

### SPCA to Wine Data

The SPCA procedure combined with a Procrustes penalty function is to be demonstrated. The dataset used here is Wine data (Dua and Graff, 2019) that consists of 178 samples (wines) and 13 variables (chemical ingredients), and it is available at UCI machine learning repository.

First, the target matrix $\mathbf{T}$ was specified by the following manner. An original principal component analysis with three components was applied to the dataset, and

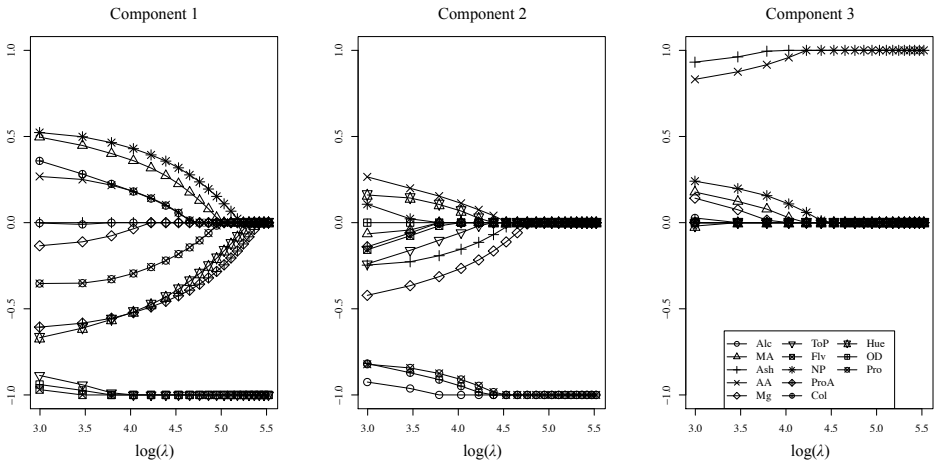



FIGURE 7.1: Solution paths for three components; names of the thirteen variables are shortened. The $y$-axes stand for value of elements. The legend is common for all the plots.

the score and loading matrices were obtained denoted as $\tilde{\mathbf{F}}$ and $\tilde{\mathbf{A}}$, respectively. Let $\tilde{\mathbf{A}}^{\sharp} = \{\tilde{a}^{\sharp}_{jk}\}$ be a Varimax-rotated $\tilde{\mathbf{A}}$ and $\mathbf{T}$ was set by

$$t_{jk} = \begin{cases} \text{sign}(a^{\sharp}_{jk}) & (|\tilde{a}^{\sharp}_{jk}| > \tau) \\ 0 & (otherwise) \end{cases} \tag{7.14}$$

where $\tau$ is a suitable threshold. The value of $\tau$ was set at 0.4 in this example. $\mathbf{T}$ is considered as a possible simple structure because it is based on Varimax-rotated loading matrix, and its simplicity was further emphasized by substituting all elements with $\pm 1$ or zeros.

The study restricted $\lambda_k = \lambda$ for $k = 1, \cdots, r$ for the simplicity of the example, and estimated $\mathbf{A}$s for $\lambda$s within the range of $\lambda \in [20, 250]$. Figure 7.1 shows solution paths of $\mathbf{A}$ where each of the estimated elements of $\mathbf{A}$ is plotted against $\lambda$s. It can be seen that some of the elements shrink toward zero as $\lambda$ increases, while the others shrinks toward 1 or -1. The figure indicates that most of the elements in $\mathbf{A}$ take 0, 1, or -1 when $\lambda$ is large, and therefore $\lambda$ controls the resulting simplicity of the solution. To illustrate this, the example picked up some $\lambda$s within the range of $\lambda$s and showed the estimated $\mathbf{A}s$ with the target matrix $\mathbf{T}$ as Table reftab:As . The correspondences between the variables and the components are clearly captured, because the estimated loading matrix contains many exact zero elements when $\lambda$ is sufficiently large, and also has simple structure as well as $\mathbf{T}$. For example, referring $\mathbf{A}$ with $\lambda = 50$, the second component is characterized by *Alcohol* (Alc) and *Color intensity* (Col), and *Proline* (Pro), but it is not clearly captured in the one with $\lambda = 20$.

TABLE 7.1: Estimated **A**s for $\lambda = 20, 50, 100$ and the target matrix. Blanc cells indicate exact zero elements.

| | $\lambda = 20$ | | | $\lambda = 50$ | | | $\lambda = 100$ | | | Target Matrix **T** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PC1 | PC2 | PC3 | PC1 | PC2 | PC3 | PC1 | PC2 | PC3 | PC1 | PC2 | PC3 |
| Alc | -0.003 | -0.925 | | | -1.000 | | | -1.000 | | | -1.000 | |
| MA | 0.497 | -0.066 | 0.177 | 0.381 | | 0.055 | 0.196 | | | | | |
| Ash | | -0.247 | 0.931 | | -0.174 | 1.000 | | | 1.000 | | | 1.000 |
| AA | 0.268 | 0.265 | 0.832 | 0.201 | 0.132 | 0.937 | 0.031 | | 1.000 | | | 1.000 |
| Mg | -0.134 | -0.422 | 0.142 | -0.058 | -0.290 | | | -0.079 | | | | |
| ToP | -0.887 | -0.239 | | -1.000 | -0.078 | | -1.000 | | | -1.000 | | |
| Flv | -0.969 | -0.157 | | -1.000 | | | -1.000 | | | -1.000 | | |
| NP | 0.523 | 0.107 | 0.241 | 0.448 | | 0.135 | 0.293 | | | | | |
| ProA | -0.606 | -0.140 | | -0.540 | | | -0.405 | | | | | |
| Col | 0.358 | -0.818 | 0.027 | 0.201 | -0.930 | | 0.030 | -1.000 | | | -1.000 | |
| Hue | -0.668 | 0.161 | -0.019 | -0.540 | 0.090 | | -0.358 | | | | | |
| OD | -0.939 | | | -1.000 | | | -1.000 | | | -1.000 | | |
| Pro | -0.354 | -0.822 | | -0.314 | -0.891 | | -0.159 | -1.000 | | | -1.000 | |

### 7.3.1 Three-Mode Component Analysis to Multiple Personality Data

The proposed method is also used for estimating an intermediate core array between Tucker3 (unconstrained) and PARAFAC (constrained to be super-diagonal). The three-mode component analysis with Procrustes penalty function was applied to Multiple Personality Data (Kroonenberg, 2008) in which fifteen concepts were evaluated by ten scales by six personalities. The data array $\bar{\mathbf{X}}$ is thus has a dimension of $13 \times 10 \times 6$, and see Osgood and Luria (1954) for details of the data. To set a target matrix **T**, PARAFAC with $s = t = u = 2$ was firstly applied to $\bar{\mathbf{X}}$, and the resulting $2 \times 2 \times 2$ core array $\bar{\mathbf{C}}$ is transformed into $2 \times 4$ a target matrix **T**. For simplicity, the isotropic penalty parameters were used; $\mathbf{\Lambda} = \lambda \mathbf{I}_2$. With the target, the proposed method was applied to the data, using the penalty parameters $\lambda = 1$, 10, and 100.

Table 7.2 shows frontal slices of the estimated core arrays together with Tucker3 ($\lambda = 0$) and PARAFAC ($\lambda = \infty$) solutions. There are two extreme cases, Tucker3 and PARAFAC, and components are fully connected in the former solution while they are sparsely connected in the latter. The proposed method allows to obtain intermediate solutions between them. As $\lambda$ gets increases, the solution gradually approximates PARAFAC solution that is the most restrictive but interpretable core array. Users freely choose appropriate value of $\lambda$, by checking resulting simplicity of the core array. If one considers $\lambda = 10$ is appropriate, connections between several components can be ignored while all connection have to be interpreted in Tucker3.

## 7.4 Conclusions

The study proposes a new penalty function for penalized optimization in multivariate analysis procedures. It is called Procrustes penalty function, and it shrinks a solution matrix to a prespecified target matrix. The target matrix possesses a certain desired stricture that the resulting solution should possess, such as a simple structure. If a target with a simple structure is employed, the resulting solution matrix is thus simple

TABLE 7.2: Estimated core arrays by Tucker3, PARAFAC, and the proposed method with three $\lambda$s. Blanc cells stand for exact zero elements.

| | | mode 3 (personalities) | | | |
|---|---|---|---|---|---|
| | | comp.1 | | comp.2 | |
| | mode 2 (scales) | comp.1 | comp.2 | comp.1 | comp.2 |
| | mode 1 (concepts) | | | | |
| Tucker3 | comp.1 | -15.803 | -5.723 | 3.437 | -3.363 |
| | comp.2 | 4.423 | -12.243 | -1.229 | -2.460 |
| $\lambda = 1$ | comp.1 | -5.185 | | -0.189 | -2.404 |
| | comp.2 | 1.295 | -0.934 | | -3.008 |
| $\lambda = 10$ | comp.1 | -0.277 | | 0.088 | 0.075 |
| | comp.2 | 0.022 | -0.305 | | 0.266 |
| $\lambda = 100$ | comp.1 | 0.582 | | | |
| | comp.2 | | | | 0.493 |
| PARAFAC | comp.1 | 0.582 | | | |
| | comp.2 | | | | 0.493 |

and interpreted, while the existing penalty functions hardly consider the matrix-wise simplicity of a solution matrix. The proposed method is applicable to various multivariate analysis procedures, in order to obtain a interpretable solution matrix or other purposes. The study used the Procrustes penalty in SPCA and three-mode component analysis, aiming to simplify the loading matrix in the former, and to obtain an intermediate solution between Tucker3 and PARAFAC.

So far, the author treated the target matrix $\mathbf{T}$ as a prespecified matrix need to be fixed. $\mathbf{T}$ is, however, possible to be estimated jointly as well as other parameter matrices. There exist some cases where it is difficult specify $\mathbf{T}$ in advance, and such extension of the proposed method serves to relax the hurdle for using the Procrustes penalty function.

# Chapter 8

# Discussions

**ABSTRACT**

To conclude the thesis, the procedures proposed in the preceding chapters are reviewed and classified based on their strategies for enhancing the interpretability of solutions. Their limitations and directions for future studies are also discussed. Finally, the superiority of the proposed procedures as compared to the existing ones are discussed in terms of their ability to simultaneously achieve sparseness and simple structures, which are essential for interpretability.

As discussed in Chapter 1, there is a growing demand for interpretability in modern multivariate data analysis and machine learning procedures, in order to drive scientific research processes and assure their social accountability. The preceding chapters proposed seven procedures to improve interpretability, which are reviewed in the following section.

## 8.1 Summary of Chapters 2 to 7

Chapters 2 and 3 proposed two procedures for the post-hoc transformation of solution matrices for their interpretability. Chapter 2 proposed the oblique Procrustes rotation procedure called Permutimin, which directly approximates a target matrix with a refinement of Thurston's simple structure. One of the major novelties of Permutimin is that it jointly estimates a target matrix and its optimal row permutations, together with an optimal rotation of a loading matrix, in order to approximate the target. Further, row permutations enable the rows of target and loading matrices to be matched, and therefore, Permutimin produces a simpler loading matrix compared to existing procedures, in which these rows are matched deterministically.

In Chapter 3, a new formulation of canonical correlation analysis (CANO) was proposed, and its equivalence to the existing formulation was proved. The new formulation is based on maximizing the sum of squared correlations of canonical variates, and it is proven that the individual rotation of structure matrices is allowed, unlike in

the existing formulation. Structure matrices are further simplified by individual rotation, whereas their simultaneous rotation was shown to be too restrictive to simplify the matrices.

Two novel biplot procedures were proposed in Chapter 4. The first called crisp cluster-wise biplot analysis (CCBA), and it yields a simple and interpretable biplot by joint classification of the objects and variables in a data matrix. The resulting biplot is interpretable in that it is composed of fewer points and vectors compared with the existing one, and it would be beneficial when there is a large number of objects and variables. Further, CCBA is formulated as the minimization of a single loss function, and thus differs from *tandem analysis* (Arabie and Hubert, 1994) which often produces an erroneous result. CCBA was further generalized to fuzzy clustering, which was proposed as fuzzy cluster-wise biplot analysis (FCBA). The simulation study and the real data example revealed that the proposed procedures worked fairly well.

Chapter 5 proposed a cardinality-constrained $K$-means clustering (CCKM) as a new clustering procedure that produces a centroid matrix with higher interpretability. It is well known that post-hoc transformation of solution matrices is not allowed in $K$-means clustering, while it is allowed in factor analysis (FA) and CANO. As a result, an estimated centroid matrix is often difficult to interpret. In CCKM, a centroid matrix $\mathbf{C}$ is estimated under a cardinality constraint; the matrix has at least $r$ zero elements equaling to zero. Thus, CCKM controls the resulting sparseness by the number of zero elements that must be contained in $\mathbf{C}$. The tuning parameter in CCKM therefore has an intuitive meaning, different from the existing procedures that use $L_1$ or $L_2$ penalty functions. In this respect, CCKM is considered more efficient for the purpose of obtaining interpretable solutions. Importantly, the study also proposed a variant of CCKM that approximates a simple structure by specifying the column-wise cardinality rather than the entire matrix.

Chapter 6 and 7 both proposed emerging techniques named layered multivariate analysis (LMA) and penalized estimation by Procrustes penalty function, respectively. Their approaches are considered to be different from those proposed above. LMA, including a layered version of multivariate analysis and PCA as special cases, constrains a solution matrix as a sum of matrices called layers, and they are required to possess a certain simple structure. Procrustes penalty aims to approximate a prespecified target matrix having a simple structure by a solution matrix. Note that these approaches can easily be extended to other multivariate procedures introduced in this research, because each of them is proposed as a general framework.

## 8.2 Four Approaches towards Enhanced Simplicity

This section classifies the seven procedures proposed herein into four classes. Their limitations and possible future directions are also discussed.

### 8.2.1 Rotation of Solution Matrix

This class of approach attempts to develop a new post-transformation method of a solution matrix, exploiting the indeterminacy of orthogonal or oblique rotation in some of the multivariate analysis procedures. Permutimin, proposed in Chapter 2, is classified into this class, in which a simple target matrix is approximated by obliquely rotating a factor loading matrix. In Psychometrics, a variety of rotation procedures have been proposed, as reviewed in Browne (2001). Here, the existing rotation procedures can be classified into three families according to their strategy of rotation, that is, how the simplicity (or complexity) of a matrix is defined, and how to maximize the simplicity. The first family is the *simplicity-maximization* family, which defines the simplicity of a rotated loading matrix as a function of a rotation matrix and maximizes it over the rotation matrix. Varimax rotation can be classified into this family. Recently, Jennrich (2004, 2006) proposed a component loss function as a novel measure of matrix simplicity, and found that maximizing the function successfully yielded a simple loading matrix.

The second family, called the *complexity-minimization* family, aims to minimize the complexity of a matrix, contrary to the first family. For instance, Kaiser (1958) defined a matrix complexity as sum of inter-column covariances and found a minimizer of the complexity function. Note that the first and second family are equivalent in that the maximization of simplicity is equivalent to the minimization of complexity.

Now, the author introduces the third family the *target rotation* family, which is considerably different from the other families. In general, it is formulated as the minimization of

$$L_{PR}(\mathbf{S}) = \|\mathbf{T} - \mathbf{\Lambda S}\|^2 \tag{8.1}$$

over a rotation matrix $\mathbf{S}$ with a suitable dimensionality. Rotation procedure based on this type of minimization is called a Procrustes rotation, and Promax (Hendrickson and White, 1964) rotation is known as a special case of Procrustes rotation. A target matrix is specified, before the rotation, as a matrix with a property desired for interpretation (e.g., simple structure). Permutimin, classified into the target rotation family, is novel because it does not require the specification but estimates an optimal target matrix, which is similar to Simplimax rotation (Kiers, 1994).

These rotation procedures are useful in that they can simplify solution matrices a posteriori, under the condition that rotational indeterminacy exists. It would be beneficial to extend the area in which rotation is allowed by discovering new indeterminacy for other multivariate analysis procedures. As proposed in Section 3, it has been proven that solution matrices in CANO have rotational indeterminacy, and thus their individual and orthogonal rotation is allowed by reformulating CANO as maximization of sum of squared canonical correlations. Followed by the study, Satomura and Adachi (2013) proposed another new formulation of CANO, in which oblique rotation of solution matrices is allowed.

The most critical drawback of this approach is that rotation is not applicable

in all of the multivariate analysis procedures. For example, to the best of the author's knowledge, it appears impossible to reformulate $K$-means clustering such that a centroid matrix can be rotated. This issue may be solved by a technological breakthrough in matrix-based multivariate statistical analysis, which is unlikely to happen at present.

Further, as shown in Table 3.1, a canonical correlation matrix is no longer diagonal after rotation, and it indicates that rotated canonical variates have several correlation structures, which might produce another problem regarding their interpretation. A similar problem also occurs in FA; an inter-factor correlation matrix is often difficult to interpret after oblique rotation, because rotated factors are not orthogonal. As seen here, although the optimality of a solution itself does not change after its rotation, some properties of the solution may change, and thus, its interpretation may be difficult.

One possible approach in FA is to consider the simplicity of an inter-factor correlation matrix as well as a factor loading matrix, thus simultaneously simplifying both matrices. This can be accomplished by maximizing

$$S(\mathbf{\Lambda S}) - ||\mathbf{SS}'||^2 \tag{8.2}$$

over a rotation matrix $\mathbf{S}$, where $\dot{S()}$ is a certain simplicity index of a matrix. Above, the second term represents the squared sum of elements in an inter-factor correlation matrix after rotation and it can be interpreted as a complexity of the correlation matrix. The second term serves as a penalty function that shrinks the off-diagonal elements in the correlation matrix toward zero.

### 8.2.2 Combination with Clustering

CCBA and FCBA, proposed in Chapter 4, both provide simple biplots that are easily interpreted by joint classification of the object and variables in a data matrix. Generally, the combination of a multivariate analysis procedure with clustering considerably reduces the cognitive load required for interpretation. This approach is particularly effective not only for biplots but also for multivariate data analysis procedures that involve visualization, such as multidimensional scaling and multiple correspondence analysis. In fact, some procedures in which clustering is incorporated with other multivariate analysis procedures have been proposed. Hwang et al. (2010) proposed the combined use of multiple correspondence analysis (MCA) (Gifi, 1990) and fuzzy $K$-means clustering in order to extract heterogeneous subgroups of objects, called fuzzy cluster MCA (FMCA). In addition, as a generalization of linear cluster-wise regression analysis (Späth, 1979), several procedures for fuzzy linear cluster-wise regression have been proposed (Bezdek et al., 1981a,b; Wedel and Steenkamp, 1989; Hathaway and Bezdek, 1993; D'Urso et al., 2010) Likewise, some methods combining clustering and another method of multivariate analysis in a single framework have been developed

(e.g., Rocci and Vichi (2005), Hwang and Dillon (2010), and Hwang et al. (2007)). A good review of such procedures can be found in Bezdek et al. (1999).

These these works indicate that incorporating clustering with certain multivariate techniques is sufficient to handle real-world problems. An insightful suggestion for such an approach can be found in Coppi et al. (2006).

As represented in the algorithm for $K$-means clustering in Chapter 3, updating $\mathbf{M}$ requires computations of $||\mathbf{x}_{(i)} - \mathbf{c}_k||^2$ for all possible $k$s, assigning the $i$-th object to the cluster minimizing the function value. It indicates that, even when combined with other multivariate analysis procedures, $\mathbf{M}$ can be updated in the same manner because the form of the loss function is irrelevant to the update process. This useful property makes it easier to integrate $K$-means clustering with other procedures.

A shortcoming of the combination is that it degrades computational efficiency, because updating $\mathbf{M}$ requires a type of *brute-force* computation for all clusters and objects; namely the local minimum frequently occurs in the computation. This issue becomes critical when attempting to handle a high dimensional data matrix that contains many objects and variables. FCBA partly solves the issue by means of fuzzy clustering, as shown in the simulation study in Section 4.3, which serves as a useful suggestion for the future study of the approach.

### 8.2.3 Sparse Estimation with Cardinality Constraint

Sparse estimation with penalized estimation by an $L_1$ penalty function for example is applicable to various types of loss functions, including the least squares loss function and maximum likelihood function. Existing sparse estimation methods may sometimes yield inconveniences when used to improve interpretability. For instance, a penalty parameter that balances sparseness and fitness is difficult to specify because it ranges between $[0, \inf]$, and its value does not have any clear interpretation; in the example in Table 1.2, $\lambda = 12$ only indicates the relative importance of the LASSO penalty function to the loss function.

Sparse estimation with the cardinality constraint treated in Chapter 5 is comparable to the aforementioned combination with clustering approach in that both require lesser cognitive load for interpretation owing to the fewer elements that need to be interpreted. Further, it achieves a sparse solution matrix via a cardinality constraint instead of an $L_1$ penalty function. There, the resulting number of non-zero elements in a solution matrix functions as a tuning parameter, and it has a clear interpretation: the number of non-zero elements that should be possessed in the solution matrix. For example, a solution matrix with 30% elements zero equaling to zero can be obtained. In this respect, cardinality constraint is considered to be more suitable for treating the issue of interpretability as compared with other penalized estimation methods.

In the same chapter, it was shown that column-wise and row-wise cardinality constraints yielded a solution with sparseness and, importantly, Thurstone's simple structure. Note that a sparse solution matrix with a certain simple structure is the one that the rotation of the solution attempts to approximate but always fails. The

successful result shown in the chapter indicates that not only sparseness but also positions of zero elements can be considered in the cardinality constraint, unlike in rotation and existing penalized estimation procedures.

### 8.2.4 Some Emerging Techniques

The procedures proposed in Chapters 6 and 7 are considered to novel in that both the procedures are based on unique computational techniques which have not been proposed yet. Namely, multivariate layered analysis in Chapter 6 assumes a layered simple structure for a solution matrix, while Procrustes penalty function in Chapter 7 shrinks a solution matrix to a prespecified target matrix. They are applicable to various multivariate analysis procedures including PCA and CANO, for example. Further, these procedures also achieve a solution matrix with sparsity and a simple structure. Further research is desired for these emerging techniques, as they demonstrate promising performance in achieving interpretable solutions.

## 8.3 Hierarchical Relationships of the Proposed Methods

This thesis considers two sub-concepts of interpretability: sparseness and simple structure. A sparse solution matrix is easy to interpret, because it contains several zero elements, and therefore, its rows and columns are clearly matched. The simple structure aims to emphasize these correspondences, and Thurstone's simple structure, presented in Chapter 2, can be assumed to be the the most ideal one. Considering Thurstone's rules, shown in Table 2.1, it can be observed that all the rules refer to zero elements, and each rule seems to require that a solution matrix contain some zeros. For example, [T1] and [T2] both require rows and columns to contain some zero elements, and [T3] to [T5] respectively require that the arrangement of zero and nonzero elements satisfy certain conditions. Therefore, a simple structure is considered to be a sparseness plus a constraint on the position of zero elements. In other words, sparseness is a necessary condition for a simple structure, and in certain cases, a solution matrix may be sparse but may not have a simple structure represented by the one in (5.20); although the matrix is sparse, the position of the zero element is inappropriate, and thus [T3] is not satisfied.

Based on the above discussions, the thesis concludes that the the proposed and existing procedures possess the hierarchical relationship shown in Figure 8.1.

The figure shows that the majority of the procedures proposed in the thesis can simultaneously achieve sparsity and simple structure, while $L_1$ penalization only achieves sparseness, which cannot achieved by the rest of the procedures. The figure also reveals another critical drawback of rotation: its result is merely an approximation of an ideal structure which is sparse and has a simple structure. Therefore, a sparse solution cannot be obtained by rotation, and Thurstone's rules can not be precisely satisfied. In practical cases, elements whose absolute values are below a prespecified threshold would be regarded as zeros, which is called *thresholding*, in order

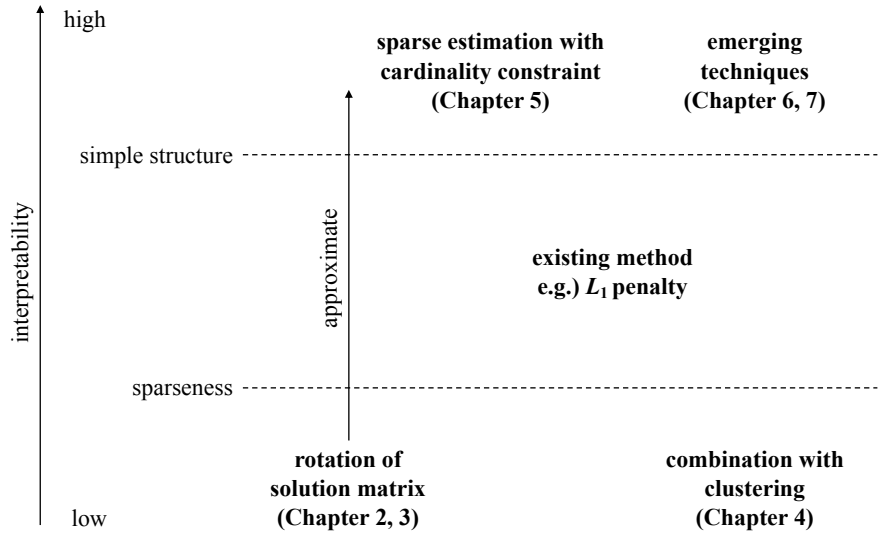



FIGURE 8.1: Hierarchical relationship between the proposed and existing procedures.

to transform a non-sparse loading matrix into a sparse one. However, thresholding should be avoided because 1) a threshold must be subjectively specified, 2) it often yields misleading results (Trendafilov, 2014; Trendafilov et al., 2017), and 3) the transformed loading matrix does not satisfy any optimality. To make matters worse, even with thresholding is used in the interpretation process, the thresholding value is rarely reported.

Cardinality-constrained estimation and some other emerging techniques can solve the above issue by directly approximating a sparse and simple structure. Existing methods, including $L_1$ penalization, do not have this property, partly because they were originally developed and used for the purpose of variable selection and building robust models and were thus not always suitable for improving the interpretability of solution. This thesis has focused on this weakness of the existing methods, developed a series of methods that accomplished both sparseness and simple structure, and their effective functioning has been demonstrated by means of simulation studies and real data examples.

# Appendix A

# Short Paper (In Japanese)

## 研究の背景と目的

心理学，社会学などの人間科学を含む経験科学では，その研究の途上において，質的もしくは量的，構造的もしくは非構造的を問わず様々な形で，何らかのデータが収集される．さらに，そのデータや過去の研究知見から演繹的に研究仮説を構築し，その真偽判定およびより確からしい仮説の構築を繰り返すことで，研究が対象とする現象の理解，ひいては将来の現象の予測を目指す．例えば心理学では，人間の心理や認知に関する基本的なメカニズムを理解することを目的として，先行研究による知見や経験的事実に基づいた研究仮説を立案し，それを検証するための心理実験を計画，実施する．さらに，その実験により得られたデータをもとに，研究仮説の確からしさを検証する．このように，今日の科学的研究において収集されるデータは，仮説の検証と修正のために中心的な役割を果たす．

さて本研究が中心に扱う，多変量データ解析法，特に，探索的データ解析法は，仮説の立案にあたって，研究の対象とする複雑な事象のふるまいを規定する，少数の根本原因に関する手がかりを抽出するために，しばしば用いられる．例えば主成分分析（Principal Component Analysis; PCA）は，多変量データを少数の次元に圧縮することで，冗長な情報を除外した，データの本質的な成分を抽出することができる．特に近年では，探索的データ解析は，データマイニングとも呼ばれる．データマイニングは，大量のデータを要約したり，頑健な予測モデルを構築するために，予測に有用な少数の変数を大量の変数群から特定することにも用いられる．このように探索的データ解析は，仮説構築という科学的研究の出発点における重要なツールとしてだけではなく，データの大規模化，高次元化といった今日のデータを取り巻くトレンドの中で，その重要性を増しつつある．

探索的データ解析では，その結果得られた解が，どの程度容易に解釈できるかどうか，すなわち解の解釈可能性（interpretability）が，非常に重要な性質であると考えられる．例えば，もしある多変量解析の解に対して，明快な解釈を与えることができなければ，研究の出発点となる仮説を立案することがそもそも困難となるであろう．このことは，表A.1に示す，日常の行動の傾向を評定したデータに対する因子分析の適用例（仮想数値例）を参照することによって明瞭に示される．

表には，変数（行動の傾向）と因子の対応を示す因子負荷量行列を示しているが，初期解を参照しても，行に並ぶ変数と列に並ぶ主成分の対応関係が曖昧であり，その因子が何を意味するか解釈しづらい．したがって，例えばこのデータから，研究者が，

TABLE A.1: 性格データに関する因子分析の結果．太字は絶対値が 0.5 以上の要素．

| | 初期解 | | Varimax 回転後 | |
|---|---|---|---|---|
| | 因子 1 | 因子 2 | 因子 1 | 因子 2 |
| 話し好き | **0.79** | 0.43 | **0.87** | 0.25 |
| 無愛想 | **-0.74** | -0.40 | **-0.80** | -0.24 |
| 陽気 | **0.61** | **0.50** | **0.78** | 0.07 |
| 人気 | **0.70** | 0.18 | **0.63** | -0.37 |
| やる気 | **0.76** | -0.44 | 0.23 | **0.85** |
| 積極性 | **0.77** | -0.38 | 0.28 | **0.81** |
| 躊躇 | **-0.63** | 0.46 | -0.12 | **-0.77** |
| 先導性 | **0.67** | -0.36 | 0.22 | **0.73** |

行動傾向を要約する因子に対する仮説を得ようとしても，因子分析によって得られた因子を明快に解釈できないため，仮説の立案は困難である．一方，初期解に Varimax 回転 (Kaiser, 1958) を施した結果を参照すると，変数と因子の対応関係が明瞭に把握できる．この点で，回転後の因子負荷量行列は解釈可能性が高い．この結果に基づけば，第 1 因子は社交性，第 2 因子は活動性といった解釈を与えることができ，この結果に基づいて「行動の活動性はやる気，積極性，先導性という 3 つの下位尺度から構成される」といった仮説をたてることが可能であろう．この例では，たかだか変数の数が 8 つのケースを取り上げたが，変数の数が非常に多い場合には，ある現象を規定する少数の要因を特定するという意味において，単純構造によって達成される解釈可能性の重要性はより大きくなる．

また，解釈可能性は，予測という文脈においては，予測モデルの説明可能性（accountability）のために重要な性質である．予測のための数理モデルが学術のみならず社会でも広く用いられる今日において，ある予測モデルがどのようにその予測を導くかを明快に説明できること，すなわちモデルの説明可能性が高いことは，予測精度を維持するのみならず，その社会的な信用を担保するためにも重要である．

ただし，一般に，多変量データ解析法で得られる解は，データに基づく観測値と，モデルから導かれる予測値の不一致度が最も小さくなるようにして推定される．このプロセスでは，得られる解の解釈が可能かどうか，容易に行えるかどうかは，全く考慮されていない．

そこで本論文では，解釈可能性が研究の出発点となる仮説を導き出すことに必要となる性質であることを鑑みて，様々な多変量解析法の解釈可能性を向上させるための方法を提案する．具体的には，解釈可能性を構成する 2 つの要素として，[1] スパース性（解行列が多くのゼロを含むこと），[2] 単純構造，を想定し，それらの性質を解が有するように，多変量解析法のアルゴリズムを修正する．

## 本研究で提案する手法の要旨

本研究で提案する 6 つの手法は，以下の 4 つのカテゴリに分類することができる．

1 つ目は，解の回転およびその適用範囲の拡張である．ここでは，第 1 節で例示したバリマックス回転のような，解の事後的な回転法に関して，既存法がもつ問題点を解

決する，新たな回転のアルゴリズムを提案する．さらに，因子分析や主成分分析など，これまで解の回転が許容されていた多変量データ解析法以外の分析法でも，同様に回転が許容されることを数学的に証明し．解の回転の適用範囲に関する理論的基礎を拡張する．第 2 章，第 3 章で提案する方法は，このカテゴリに該当する．

2 つ目は，クラスタリングとの融合であり，第 4 章で提案する方法が該当するカテゴリである．ここでは，ある多変量解析法をクラスタリングと融合することで，データ中に含まれる多くの個体や変数を少数のクラスタに分類し，より直感的な解釈を行える解を出力する方法を提案する．

続く 5 章は，解のスパース推定と呼ばれるカテゴリであり，解行列をスパースに推定する，すなわち，多くのゼロを含むように推定する方法を提案する．解行列をスパースに推定することによって，解釈を必要とする要素が非ゼロ要素に限定され，解釈に伴う認知的な負荷を大きく軽減することが期待できる．また，表A.1の解釈において，例えば絶対値が 0.3 以下の要素を無視するといった，主観的かつ adhoc な解釈法を取り入れる必要がなくなる点でも望ましい．

4 つ目のカテゴリは，新たな計算原理に基づく方法であり，第 6，7 章では，これまで述べたカテゴリとは異なる，新しい計算原理を導入した 2 つの方法を提案する．上で述べたカテゴリは，これまでの先行研究を拡張するものであった一方で，ここで提案する方法は，全く新しい研究アプローチを提供するものであり，解釈可能性に関する今後の研究において，大きな示唆を与えるであろう．

これらの 4 カテゴリと，解釈可能性を構成するスパース性，単純構造との関係は，図A.1に示される．本論文で提案する方法の最も特徴的な点は，その多くが，スパースかつ単純構造を有する解が得られる点である．元来機械学習の分野で提案され，解釈可能性向上にしばしば用いられる $L_1$ 正則化などに基づく既存手法では，これらの 2 つの性質を有する解を得ることは困難であった．本論文では，$L_1$ 正則化とは異なるアプローチで解釈可能性の問題に取り組むことにより，スパースかつ単純構造を有する解を得る方法を数多く考案しており，この点が多変量データ解析に関する研究に対する，本論文の最も大きな貢献である．

以下では，上の 6 つの手法の概要を紹介する．

### Permutimin: Factor Rotation to Simple Structure with Permutation of Variables(第 2 章)

ターゲット回転と呼ばれる回転法では，解釈のために望ましい構造をもつターゲット行列を用意して，それとの差異がなるべく小さくなるように解行列を回転することによって，解行列の単純化を試みる．そのような方法として，Promax 回転 (Hendrickson and White, 1964) が知られている．この手法の問題点として，[1] ターゲット行列を明示的に与えられない場合が多いこと，[2] ターゲット行列と回転対象の解行列の行が必ずしも一致しない事があることが知られている．特に 2 つ目の問題点は，達成可能な単純構造を制約してしまう点で大きな問題であったものの，著者の知る限り，これを解決可能な回転法はこれまで存在しなかった．

本研究では，上の 2 つの問題に対して，[1] ターゲット行列を固定するのではなく，推定および更新しつつ回転すること，[2][1] と同時にターゲット行列の最適な行置換を行うこと，の 2 点を新たに有する回転法として，Permutimin 回転を提案した．Cureton and

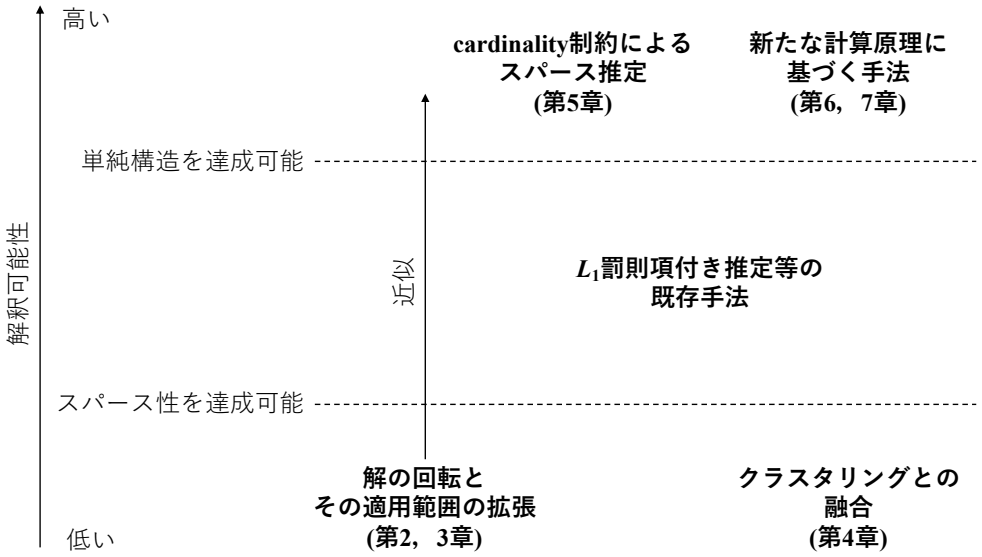



FIGURE A.1: 提案手法，既存手法とスパース性/単純構造の関係.

Mulaik (1975) による Box データに対して，Permutimin 回転，Oblimin 回転 (Harman, 1976)，および Promax 回転を適用した結果を表A.2に示した．ほか 2 手法と比べて，Permutimin 回転では，明瞭に単純構造が得られており，3 つの因子がそれぞれ，変数中に含まれる $x$, $y$, $z$ に対応することがはっきりと見て取れる．

なお，以上の内容はYamashita and Adachi (2019b) の内容を一部発展させたものである．

TABLE A.2: Box データに対する Permutimin 回転，Oblimin 回転，Promax 回転の結果.

| | Permutimin | | | Oblimin | | | Promax | | |
|---|---|---|---|---|---|---|---|---|---|
| | F1 | F2 | F3 | F1 | F2 | F3 | F1 | F2 | F3 |
| $x$ | **0.98** | 0.03 | 0.03 | **0.44** | **0.39** | **0.73** | **-0.52** | **0.48** | **0.88** |
| $y$ | 0.02 | **0.96** | 0.03 | -0.17 | **0.98** | **-0.33** | **0.46** | **0.82** | **-0.37** |
| $z$ | 0.03 | 0.04 | **0.96** | **0.97** | -0.16 | **-0.34** | **0.84** | **-0.32** | **0.43** |
| $xy$ | **0.61** | **0.67** | 0.00 | 0.11 | **0.91** | 0.23 | -0.04 | **0.86** | 0.26 |
| $xy^2$ | **0.35** | **0.84** | 0.03 | 0.00 | **0.98** | -0.03 | 0.22 | **0.87** | -0.03 |
| $2x + 2y$ | **0.54** | **0.75** | -0.04 | 0.02 | **0.98** | 0.17 | 0.01 | **0.91** | 0.15 |
| $(x^2 + y^2)^{1/2}$ | **0.51** | **0.74** | -0.02 | 0.03 | **0.95** | 0.15 | 0.03 | **0.88** | 0.14 |
| $xz$ | **0.61** | 0.01 | **0.67** | **0.93** | 0.09 | 0.21 | 0.24 | 0.04 | **0.83** |
| $xz^2$ | **0.42** | -0.03 | **0.82** | **1.02** | -0.05 | 0.03 | **0.46** | -0.13 | **0.75** |
| $2x + 2z$ | **0.58** | -0.02 | **0.71** | **0.97** | 0.04 | 0.19 | 0.27 | -0.01 | **0.84** |
| $(x^2 + z^2)^{1/2}$ | **0.55** | -0.02 | **0.71** | **0.95** | 0.03 | 0.17 | **0.30** | -0.02 | **0.81** |
| $yz$ | -0.03 | **0.60** | **0.63** | **0.50** | **0.46** | **-0.46** | **0.85** | 0.26 | 0.00 |
| $yz^2$ | -0.01 | **0.43** | **0.77** | **0.68** | 0.26 | **-0.44** | **0.88** | 0.07 | 0.14 |
| $2y + 2z$ | -0.01 | **0.61** | **0.63** | **0.49** | **0.48** | **-0.45** | **0.84** | 0.28 | 0.00 |
| $(y^2 + z^2)^{1/2}$ | 0.01 | **0.61** | **0.59** | **0.47** | **0.50** | **-0.42** | **0.79** | **0.30** | 0.01 |

## Canonical Correlation Analysis Formulated as Maximizing Sum of Squared Correlations and Rotation of Structure Matrices（第 3 章）

正準相関分析は，2 つのデータ行列のそれぞれに対して，最も相関が高くなるような合成変数を，観測変数の線形結合により構成する分析法である．2 つのデータ行列を取

TABLE A.3: 体力診断テスト/運動能力テストに対する正準相関分析の結果と提案手法による回転の結果.

| | 初期解 | | | 提案手法 | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 1 | 2 | 3 |
| 体力診断テスト | | | | | | |
| 反復横跳び | **-0.670** | 0.278 | -0.187 | **-0.697** | 0.117 | 0.246 |
| 背筋力 | **-0.744** | -0.167 | -0.333 | **-0.733** | 0.359 | -0.160 |
| 垂直跳び | **-0.670** | **-0.605** | 0.307 | -0.202 | **0.926** | -0.105 |
| 握力 | **-0.251** | 0.006 | **0.365** | 0.013 | **0.334** | 0.29 |
| 踏み台昇降 | -0.172 | **0.773** | **0.526** | -0.015 | -0.103 | **0.944** |
| 運動能力テスト | | | | | | |
| 長距離走 | **0.667** | -0.117 | 0.441 | **0.784** | 0.149 | -0.129 |
| ボール投げ | **-0.775** | -0.168 | -0.294 | **-0.688** | **-0.473** | 0.193 |
| 走り幅跳び | **-0.789** | **-0.497** | 0.038 | -0.387 | **-0.827** | 0.136 |
| 懸垂 | **-0.649** | -0.214 | **0.565** | 0.020 | **-0.640** | **0.614** |
| 50m 走 | **0.676** | **-0.362** | **-0.497** | 0.174 | 0.163 | **-0.882** |

り扱うことができる点で，他の多変量データ解析法と比べユニークなこの方法は，その結果出力される解行列の解釈が困難であることが知られている．ただし，因子分析や主成分分析と異なり，正準相関分析では，解の回転の不定性が存在しないとされてきており，そのため，上でのべた Permutimin 回転のような方法を利用して，解を単純化することが許容されなかった．

そこで本研究では，まず，正準相関分析の新たな定式化を提案する．この新たな定式化は，正準相関係数の平方和の最大化に基づくものであり，この新しい定式化が，既存の定式化と同等であり，なおかつ同じ解を与えることを証明する．さらに，この新たな定式化のもとでは，これまで許容されなかった，解の回転を行うことができる．すなわち，これまで提案されている様々な回転法を自由に適用して，正準相関分析の解を単純化することが，理論的に保証される．

表A.3に，体力診断テストと運動能力テストそれぞれに対する成績データに，正準相関分析を適用し，得られた解と，それを提案手法によって回転した結果を示す．初期解は，ほとんどすべての変数が第 1 正準変数に大きく関係しており，いわば「一般的運動能力」として解釈可能な因子が得られてしまっている．もし，分析者の関心が「一般運動能力」を構成する下位の概念を抽出することにあるならば，初期解は無意味である．一方で，提案手法による結果は，その下位尺度とみなせるような正準変数が確かに得られており，かつその解釈が容易であることがはっきりと確認できる．

なお，以上の内容はYamashita (2012a) の内容を一部発展させたものである．

## A New Biplot Procedure with Joint Classification Objects and Variables by Crisp and Fuzzy $K$-means Clustering（第 4 章）

主成分分析の結果の優れた視覚化法であるバイプロットは，データ行列に含まれる個体と変数の関係性，等質性の直感的な把握を目的として，広く用いられている．ただし，データ行列が非常に多くの個体と変数を含む場合，バイプロットを構成する点（各個体に対応する）とベクトル（各変数に対応する）の数が非常に多くなるため，直感的

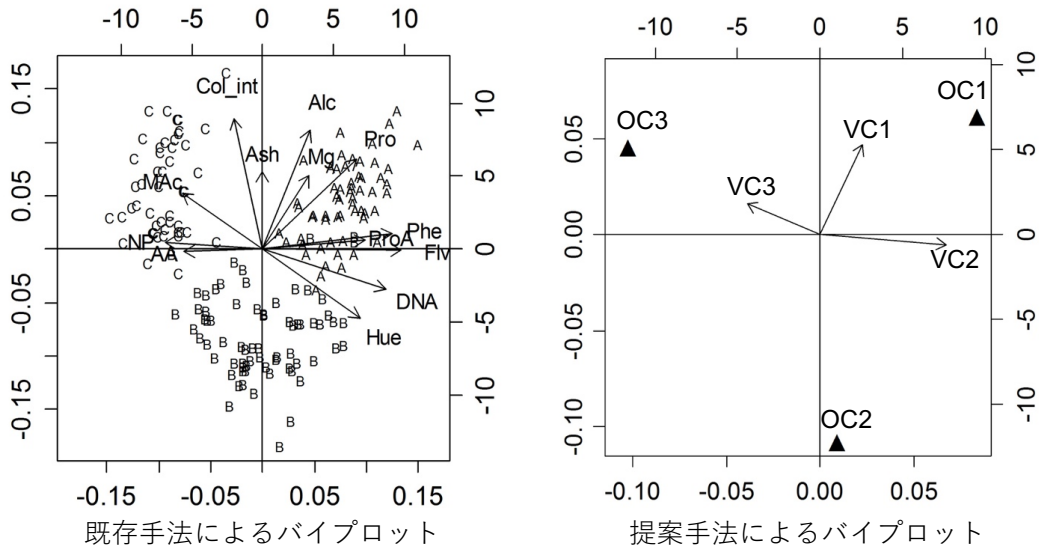



FIGURE A.2: 既存手法/提案手法によるバイプロットの比較.

な結果の把握が妨げられることが懸念される．本研究では，そのような大規模なデータであっても，解釈が容易なバイプロットを得るための方法として，Crisp Clusterwise Biplot Analysis（CCBA）を提案する．CCBA では，個体と変数のクラスタリングと，それによって得られた，個体クラスタおよび変数クラスタのバイプロットを同時に行う．このことにより，バイプロットはクラスタ数分の点およびベクトルから構成されるため，データ行列が巨大であったとしても，そのバイプロットの解釈は容易になると考えられる．

178 個の個体と，13 個の変数からなるデータ行列に対して，個体と変数のクラスタ数がそれぞれ 3 の提案手法を適用した結果を，図A.2に示す．既存手法によるバイプロットは，データが多数の個体と変数から構成されるがゆえに，煩雑な見た目となり，データの構造が直感的に把握しづらい．一方提案手法では，個体と変数がそれぞれ 3 つのクラスタに分類されることで，解釈が容易なバイプロットが得られている．

本研究では上の結果に加えて，個体/変数のクラスタリングを，所属クラスタの重複を認める，ファジィクラスタリングに拡張した，Fuzzy Clusterwise Biplot Analsys（FCBA）も提案する．なお，以上の結果，Yamashita and Mayekawa (2012) およびYamashita and Mayekawa (2015) を一部発展させたものである．

## A Modified $K$-means clustering procedure for obtaining a cardinality-constrained centroid matrix（第 5 章）

個体の分類に広く用いられる $K$-means クラスタリングでは，得られたクラスタがどのような変数によって特徴づけられるかを解釈する上で，セントロイド行列と呼ばれる解行列が参照される．ただし，$K$-means クラスタリングでは，正準相関分析と同様に，解の回転が許容されていない．また，第 3 章と同じアプローチにより，回転が許容されるように $K$-means クラスタリングを再定式化することは，計算上の困難が伴う．

TABLE A.4: $K$-means と RSKM により得られたセントロイド行列．空白のセルはゼロに等しい要素を表す．

| | $K$-means | | | | RSKM | | | |
|---|---|---|---|---|---|---|---|---|
| | C1 | C2 | C3 | C4 | C1 | C2 | C3 | C4 |
| 立派な | 0.34 | -0.38 | 0.12 | 0.00 | 0.34 | | | |
| 役に立つ | 0.56 | -0.39 | 0.18 | -0.28 | 0.56 | -0.39 | | |
| 良い | 0.20 | -0.13 | 0.02 | -0.06 | | | | |
| 大きい | -0.01 | -0.42 | 0.38 | 0.06 | | -0.42 | | |
| 力強い | -0.07 | -0.80 | 1.03 | -0.24 | | -0.80 | 1.03 | |
| 強力な | 0.03 | -0.71 | 1.01 | -0.44 | | -0.71 | 1.01 | |
| 速い | -0.33 | 0.13 | 0.95 | -1.10 | | | 0.95 | -1.10 |
| 騒がしい | -0.28 | 0.25 | 0.87 | -1.21 | | | 0.87 | -1.21 |
| 若い | 0.02 | 0.69 | 0.46 | -1.55 | | 0.69 | | -1.55 |
| 真面目な | 0.44 | -0.33 | 0.17 | -0.22 | 0.44 | | | |
| 頑固な | 0.57 | -1.24 | 0.39 | 0.57 | 0.57 | -1.24 | | |
| 忙しい | 0.59 | 0.28 | 0.21 | -1.24 | 0.59 | | | -1.24 |

そこで本研究では，セントロイド行列中の非ゼロ要素の数（cardinality）を制約することにより，スパースで解釈が容易なセントロイド行列を得る方法を，Reduced Structure $K$-Means（RSKM）として提案する．RSKM では，事前に設定した個数以上のゼロ要素を含むように，セントロイド行列が推定されるため，分析者は，その結果の解釈において，一部のゼロ要素にのみ注目すればよく，解釈に伴う認知的コストを大きく削減できる．また，ゼロ要素と非ゼロ要素が混在することで，行に並ぶ変数と列に並ぶ変数の対応関係がより明瞭に把握できるようになる．さらに，提案手法の変法として，行ごと，列ごとの cardinality を制約して，解の回転が目指す，単純構造を直接得られる方法も併せて提案する．すなわち，スパース性と単純構造を両方有したセントロイド行列を得ることができる．

表A.4には，通常の $K$-means と，提案手法の RSKM を，13 の職業を 12 の印象語で評定した，職業評定データに対して適用した例を示す．得られた 4 つのクラスタがそれぞれ何を意味するかに関して，値がゼロの要素を含まない $K$-means の結果を参照しても，直感的に理解することは難しい．一方 RSKM は，単純構造に近い構造を取り，クラスタと印象語の対応関係がはっきりと理解できる．

なお，以上の内容はYamashita and Adachi (2019a) およびYamashita and Adachi (2017) の結果を一部発展させたものである．

### Layered Multivariate Analysis(第 6 章)

解の解釈可能性向上に関する新たな研究アプローチとして，Vichi (2017) は，完全クラスタ構造をもつ主成分負荷量行列，つまり，各行にただひとつの非ゼロ要素しか存在しない解を出力する主成分分析を提案している．ただし，完全クラスタ行列のように制約の強い構造が，実データに対して良好に適合することは稀である．そこで本研究では，上の方法から着想を得て，完全クラスタ行列の和として解行列を表現する，Layered Multivariate Analysis，および，その特別版としての Layered Multibariate Regression（LMR）および Layered PCA（LPCA）を提案する．提案手法では，解行列が，layer と呼ばれる，完全クラスタ構造をもつ行列の和として表現される．そのため，Layer の

個数を増減させることで，データに対する適合性と，解行列の単純性のバランスを調整することが可能である．LMR では回帰係数行列が，LPCA では主成分負荷量行列が，それぞれ layer の和で表現される．また，layer の個数を大きくしていくと，無制約時の解に収束することが，数学的に保証される．

LMR を，タバコ葉の科学的組成に関するデータに適用した例を表A.5に示す．表には，レイヤ数（$L$）が 1 から 4 までの場合と，無制約時の解（最小二乗解）を示しているが，layer の数を増やしていくにつれて，ゼロに等しい要素の数が減り，分散説明率が上昇，すなわち，データに対する適合度が上昇していることが見て取れる．例えば，この例では，分散説明率の変化が小さくなる，$L = 3$ 程度が適当な layer 数だと思われるが，タバコ葉の燃焼率，および含まれる糖類と Nic（ニコチン）の量に影響を与える科学的組成が明瞭に特定できる．

以上の内容は，Yamashita and Adachi (ress) を一部拡張したものである．

TABLE A.5: Layered Multivariate Regression の適用例．空白のセルはゼロに等しい要素を表す．

| | LMR ($L=1$) | | | LMR ($L=2$) | | | LMR ($L=3$) | | | LMR ($L=4$) | | | 最小二乗解 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 燃焼率 | 糖類 | Nic | 燃焼率 | 糖類 | Nic | 燃焼率 | 糖類 | Nic | 燃焼率 | 糖類 | Nic | 燃焼率 | 糖類 | Nic |
| N | | -0.69 | | | -0.65 | 0.51 | | -0.52 | 0.39 | | -0.58 | 0.30 | 0.10 | -0.58 | 0.29 |
| Cl | -0.61 | | | -0.60 | 0.38 | | -0.60 | 0.41 | -0.31 | -0.60 | 0.39 | -0.32 | -0.58 | 0.39 | -0.32 |
| K | 0.62 | | | 0.52 | 0.26 | | 0.52 | 0.14 | | 0.52 | 0.20 | 0.11 | 0.45 | 0.20 | 0.11 |
| P | | 0.17 | | -0.14 | 0.20 | | -0.14 | 0.22 | | -0.14 | 0.22 | | -0.13 | 0.22 | -0.05 |
| Ca | 0.33 | | | 0.47 | | | 0.47 | | 0.18 | 0.47 | 0.11 | 0.25 | 0.41 | 0.11 | 0.24 |
| Mg | | | 0.73 | -0.26 | | 0.43 | -0.26 | -0.21 | 0.40 | -0.26 | -0.21 | 0.47 | -0.32 | -0.22 | 0.48 |
| 分散説明率 | 0.569 | | | 0.701 | | | 0.74 | | | 0.742 | | | 0.744 | | |

## Procrustes Penalty Function for Matching Matrices to Targets with its Applications(第 7 章)

スパースな解行列を得るための方法として，解行列の要素絶対値和（$L_1$ ノルム）を罰則項として用いた，罰則項付き推定が用いられることが多い．この罰則項付き推定では，解行列中のすべての要素を 0 に向けて縮退させることで，スパースな解行列を推定することができる．しかしながら，解の解釈可能性を目標とすることを考えると，この推定法では不十分であると考えられる．なぜならば，これまで提案した方法で見たように，解釈可能性のためには，値のスパース性だけではなく，単純構造も同じく重要であり，それを達成する上で，必ずしもすべての要素を 0 へ向けて縮退させることが効果的であるとは限らないためである．

そこで本研究では，解行列の要素ごとに 0 へ縮退させるのではなく，解行列全体を，あるターゲット行列へ向けて縮退させることにより，スパース性と単純構造の両立を目指す罰則項，Procrustes penalty を提案する．提案手法を用いた罰則項付き推定では，行列全体をターゲット行列へ近づけるため，すべての要素を 0 へ向けて縮退するとは限らず，一部の要素は非ゼロの値に近づいていく．この事によって，ターゲットが単純構造を取っていれば，推定される解行列もまた単純構造に近づくことが期待できる．

提案手法を PCA と組み合わせて，スパース PCA を適用した結果を表A.6に示す．罰則項の相対的な強さを調整する $\lambda$ の値を大きくするに従って，表中に示したターゲット行列に解行列が近づき，主成分と変数との対応関係がより明確に表現されているこ

TABLE A.6: 提案手法によるスパース PCA の結果と，用いたターゲット行列．空白はゼロに等しい要素を表す．

| | $\lambda = 20$ | | | $\lambda = 50$ | | | $\lambda = 100$ | | | Target Matrix $\mathbf{T}$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PC1 | PC2 | PC3 | PC1 | PC2 | PC3 | PC1 | PC2 | PC3 | PC1 | PC2 | PC3 |
| Alc | -0.003 | -0.925 | | | -1.000 | | | -1.000 | | | -1.000 | |
| MA | 0.497 | -0.066 | 0.177 | 0.381 | | 0.055 | 0.196 | | | | | |
| Ash | | -0.247 | 0.931 | | -0.174 | 1.000 | | | 1.000 | | | 1.000 |
| AA | 0.268 | 0.265 | 0.832 | 0.201 | 0.132 | 0.937 | 0.031 | | 1.000 | | | 1.000 |
| Mg | -0.134 | -0.422 | 0.142 | -0.058 | -0.290 | | | -0.079 | | | | |
| ToP | -0.887 | -0.239 | | -1.000 | -0.078 | | -1.000 | | | -1.000 | | |
| Flv | -0.969 | -0.157 | | -1.000 | | | -1.000 | | | -1.000 | | |
| NP | 0.523 | 0.107 | 0.241 | 0.448 | | 0.135 | 0.293 | | | | | |
| ProA | -0.606 | -0.140 | | -0.540 | | | -0.405 | | | | | |
| Col | 0.358 | -0.818 | 0.027 | 0.201 | -0.930 | | 0.030 | -1.000 | | | -1.000 | |
| Hur | -0.668 | 0.161 | -0.019 | -0.540 | 0.090 | | -0.358 | | | | | |
| OD | -0.939 | | | -1.000 | | | -1.000 | | | -1.000 | | |
| Pro | -0.354 | -0.822 | | -0.314 | -0.891 | | -0.159 | -1.000 | | | -1.000 | |

とが見て取れる．また，Procrustes panelty は，PCA だけではなく，多変量回帰分析や $K$-means クラスタリングなど，様々な多変量データ解析法にも適用可能である．

以上の内容はYamashita (ress) を拡張したものである．

## 総合論議

解の回転は，何らかの基準を最適化することによって，単純構造を有するように解行列を変換することを企図したものである．ただし，ここで重要なことは，その変換が，直交回転もしくは斜交回転に限定されているため，得られる解はあくまで単純構造を近似したものであり，回転後の解は，一般にスパースではなく，それゆえ単純構造を厳密な意味で有するものではない．言い換えれば，スパース性と単純構造をどちらも達成できない．

一方，解釈の容易なセントロイド行列を出力する RSKM では，セントロイド行列の cardinality を制約することによって，セントロイド行列のスパース性を達成が達成される．さらに，行ごと，列ごとの cardinality を制約することで，ゼロ要素の位置を制約することができる．これらの 2 つの特徴により，RSKM は，スパース性の達成に加えて，単純構造の実現において不可欠である，ゼロ要素の位置まで制約することが可能であり，RSKM により得られた解は，スパース性と単純構造をどちらも有することが期待できる．一方で，機械学習におけるスパース推定で頻繁に用いられる $L_1$ 罰則項による罰則項付き推定では，すべての要素を一斉に 0 に向けて shrink させるため，ゼロ要素の位置の flexible な調整は，一般には不可能である．よって，結果として得られる解は，スパースであるが，単純構造を取り得ない．

さらに，新たな計算原理に基づく手法として，Layered Multivariate Analysis と，Procrustes penalty をそれぞれ提案したが，これらも RSKM と同様に，ゼロ要素の位置を制約することで，スパース性だけではなく，単純構造を有する解を出力可能である．

ここで見たように，本研究で提案した手法のほとんどは，スパース性と単純構造を同時に有する解を出力できる．$L_1$ 正則化などのこれまで提案されている方法とは異なる，本研究の提案する方法の大きなアドバンテージである．元来，$L_1$ 正則化は，解の

解釈可能性向上よりもむしろ，変数選択や予測モデルのロバスト化のために提案され，用いられてきた方法であった．そのような経緯のために，罰則項付き推定による解の単純化には限界があるといえるだろう．第 1 章で議論したとおり，昨今，解の解釈可能性やそれがもたらすモデルの accountability が重要視される中で，多変量データ解析法の解釈可能性を考慮することは，必須であるといえる．そのような状況の中，本研究では，スパース性を超えた解釈可能性の重要性に注目し，既存の方法には困難であった，スパース性と単純構造の両立が可能な方法を，数多く開発し，その有用性を示した．

# Bibliography

Adachi, K. (2006). *Multivariate data analysis*. Nakanishiya Shuppan, Tokyo (in Japanese).

Adachi, K. (2009). Joint procrustes analysis for simultaneous nonsingular transformation of component score and loading matrices. *Psychometrika*, 74(4):667.

Adachi, K. (2011). Constrained principal component analysis of standardized data for biplots with unit-length variable vectors. *Advances in Data Analysis and Classification*, 5(1):23–36.

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

Adachi, K. and Trendafilov, N. T. (2016). Sparse principal component analysis subject to prespecified cardinality of loadings. *Computational Statistics*, 31(4):1403–1427.

Adachi, K. and Trendafilov, N. T. (2018a). Some mathematical properties of the matrix decomposition solution in factor analysis. *Psychometrika*, 83(2):407–424.

Adachi, K. and Trendafilov, N. T. (2018b). Sparsest factor analysis for clustering variables: a matrix decomposition approach. *Advances in Data Analysis and Classification*, 12(3):559–585.

Aggarwal, C. C. (2014). *Data classification: algorithms and applications*. CRC press.

Alsius, A., Wayne, R. V., Paré, M., and Munhall, K. G. (2016). High visual resolution matters in audiovisual speech perception, but only for some. *Attention, Perception, & Psychophysics*, 78(5):1472–1487.

Anderson, T. W. and Rubin, H. (1956). Statistical inference in factor analysis. In *Proceedings of the third Berkeley symposium on mathematical statistics and probability*, volume 5, pages 111–150.

Arabie, P. and Hubert, L. (1994). Cluster analysis in marketing research. In Bagozzi, R., editor, *Advanced Methods of Marketing Research*, pages 160–189. Blackwell, Oxford.

Asuncion, A. and Newman, D. (2007). Uci machine learning repository.

Bartholomew, D. J., Knott, M., and Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach*, volume 904. John Wiley & Sons.

Bartlett, M. S. (1950). Tests of significance in factor analysis. *British Journal of statistical psychology*, 3(2):77–85.

Basu, S., Davidson, I., and Wagstaff, K. (2008). *Constrained clustering: Advances in algorithms, theory, and applications*. CRC Press.

Bernaards, C. A. and Jennrich, R. I. (2003). Orthomax rotation and perfect simple structure. *Psychometrika*, 68(4):585–588.

Bezdek, J. C. (2013). *Pattern recognition with fuzzy objective function algorithms*. Springer Science & Business Media.

Bezdek, J. C., Coray, C., Gunderson, R., and Watson, J. (1981a). Detection and characterization of cluster substructure i. linear structure: Fuzzy c-lines. *SIAM Journal on Applied Mathematics*, 40(2):339–357.

Bezdek, J. C., Coray, C., Gunderson, R., and Watson, J. (1981b). Detection and characterization of cluster substructure ii. fuzzy c-varieties and convex combinations thereof. *SIAM Journal on Applied Mathematics*, 40(2):358–372.

Bezdek, J. C., Keller, J., Krisnapuram, R., and Pal, N. (1999). *Fuzzy models and algorithms for pattern recognition and image processing*, volume 4. Springer Science & Business Media.

Bock, H. H. (1996). Probabilistic models in cluster analysis. *Computational Statistics & Data Analysis*, 23(1):5–28.

Browne, M. W. (1972a). Oblique rotation to a partially specified target. *British Journal of Mathematical and Statistical Psychology*, 25:207–212.

Browne, M. W. (1972b). Orthogonal rotation to a partially specified target. *British Journal of Mathematical and Statistical Psychology*, 25:115–120.

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, 36:111–150.

Brusco, M. J. and Cradit, J. D. (2001). A variable-selection heuristic for k-means clustering. *Psychometrika*, 66(2):249–270.

Cadima, J. and Jolliffe, I. T. (1995). Loading and correlations in the interpretation of principle compenents. *Journal of Applied Statistics*, 22(2):203–214.

Carroll, J. B. (1953). An analytic rotation for approximating simple structure in factor analysis. *Pssychometrika*, 18:23–38.

Carroll, J. D. and Chang, J.-J. (1970). Analysis of individual differences in multi-dimensional scaling via an n-way generalization of "eckart-young" decomposition. *Psychometrika*, 35(3):283–319.

Cliff, N. and Krus, D. J. (1976). Interpretation of canonical analysis: Rotated vs. unrotated solutions. *Psychometrika*, 41(1):35–42.

Coppi, R., Gil, M. A., and Kiers, H. A. (2006). The fuzzy approach to statistical analysis. *Computational statistics & data analysis*, 51(1):1–14.

Cortina, L. M. and Wasti, S. A. (2005). Profiles in coping: Responses to sexual harassment across persons, organizations, and cultures. *Journal of Applied Psychology*, 90(1):182.

Cureton, E. and Mulaik, S. A. (1975). The weighted varimax rotation and the promax rotation. *Psychometrika*, 40:183–195.

Dalton, C., Jennings, E., O'Dwyer, B., and Taylor, D. (2016). Integrating observed, inferred and simulated data to illuminate environmental change: a limnological case study. In *Biology and Environment: Proceedings of the Royal Irish Academy*, volume 116, pages 279–294. JSTOR.

De Leeuw, J. (2004). Least squares optimal scaling of partially observed linear systems. In *Recent Developments on Structural Equation Models*, pages 121–134. Springer.

DeSarbo, W. S. and Mahajan, V. (1984). Constrained classification: the use of a priori information in cluster analysis. *Psychometrika*, 49(2):187–215.

Dua, D. and Graff, C. (2019). UCI machine learning repository [http://archive.ics.uci.edu/ml]. university of california, school of information and computer science.

D'Urso, P., Massari, R., and Santoro, A. (2010). A class of fuzzy clusterwise regression models. *Information Sciences*, 180(24):4737–4762.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188.

Fowlkes, E. B. and Mallows, C. L. (1983). A method for comparing two hierarchical clusterings. *Journal of the American statistical association*, 78(383):553–569.

Frølich, L., Andersen, T. S., and Mørup, M. (2018). Rigorous optimisation of multilinear discriminant analysis with tucker and parafac structures. *BMC bioinformatics*, 19(1):197.

Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58(3):453–467.

Gan, G., Ma, C., and Wu, J. (2007). *Data clustering: theory, algorithms, and applications*, volume 20. Siam.

Gifi, A. (1990). *Nonlinear multivariate analysis*. Wiley.

Gower, J. C. and Dijksterhuis, G. B. (2004). *Procrustes problems*. Oxford University Press, Oxford.

Gower, J. C. and Hand, D. J. (1995). *Biplots*, volume 54 of *Monographs on Statistics and Applied Probability*. CRC Press.

Gower, J. C., Lubbe, S. G., and Le Roux, N. J. (2011). *Understanding biplots*. John Wiley & Sons.

Green, B. F. (1952). The orthogonal approximation of an oblique structure in factor analysis. *Psychometrika*, 17:429–440.

Greenacre, M. J. (2010). *Biplots in practice*. Fundacion BBVA.

Harman, H. H. (1976). *Modern factor analysis*. University of Chicago press.

Harris, C. W. and Kaiser, H. F. (1964). Oblique factor analytic solutions by orthogonal transformations. *Psychometrika*, 29(4):347–362.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC.

Hathaway, R. J. and Bezdek, J. C. (1993). Switching regression models and fuzzy clustering. *IEEE Transactions on fuzzy systems*, 1(3):195–204.

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.

Heumann, C., Schomaker, M., et al. (2016). *Introduction to statistics and data analysis*. Springer.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Hogg, R. V., McKean, J., and Craig, A. T. (2005). *Introduction to mathematical statistics*. Pearson Education.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, 28(3/4):321–377.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification*, 2(1):193–218.

Huley, J. and Cattell, R. B. (1962). The procrustes program: Producing direct rotation to test a hypothesized factor structure. *Behavioral Science*, 7:258–262.

Hwang, H., DeSarbo, W. S., and Takane, Y. (2007). Fuzzy clusterwise generalized structured component analysis. *Psychometrika*, 72(2):181.

Hwang, H. and Dillon, W. R. (2010). Simultaneous two-way clustering of multiple correspondence analysis. *Multivariate behavioral research*, 45(1):186–208.

Hwang, H., Dillon, W. R., and Takane, Y. (2010). Fuzzy cluster multiple correspondence analysis. *Behaviormetrika*, 37(2):111–133.

Hyland, J. J., Jones, D. L., Parkhill, K. A., Barnes, A. P., and Williams, A. P. (2016). Farmers' perceptions of climate change: identifying types. *Agriculture and Human Values*, 33(2):323–339.

Izenman, A. J. (2008). *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. Springer, New York.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

Jennrich, R. I. (2002). A simple general method for oblique rotation. *Psychometrika*, 67:7–20.

Jennrich, R. I. (2004). Rotation to simple loadings using component loss functions: The orthogonal case. *Psychometrika*, 69:257–273.

Jennrich, R. I. (2006). Rotation to simple loadings using component loss functions: The oblique case. *Psychometrika*, 71:173–191.

Jetti, S. K., Vendrell-Llopis, N., and Yaksi, E. (2014). Spontaneous activity governs olfactory representations in spatially organized habenular microcircuits. *Current biology*, 24(4):434–439.

Jolliffe, I. (2011). *Principal component analysis*. Springer.

Jolliffe, I. T., Trendafilov, N. T., and Uddin, M. (2003). A modified principal component technique based on the lasso. *Journal of Computational and Graphical Statistics*, 12(3):531–547.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23:187–240.

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, 39(1):31–36.

Kiers, H. A. (1990). Majorization as a tool for optimizing a class of matrix functions. *Psychometrika*, 55(3):417–428.

Kiers, H. A. (1998). Joint orthomax rotation of the core and component matrices resulting from three-mode principal components analysis. *Journal of Classification*, 15(2):245–263.

Kiers, H. A. (2002). Setting up alternating least squares and iterative majorization algorithms for solving various matrix optimization problems. *Computational statistics & data analysis*, 41(1):157–170.

Kiers, H. A. L. (1994). Simplimax: Oblique rotation to an optimal target with simple structure. *Psychometrika*, 59:567–579.

Kristof, W. (1970). A theorem on the trace of certain matrix products and some applications. *Journal of Mathematical Psychology*, 7(3):515–530.

Kroonenberg, P. M. (2008). *Applied multiway data analysis*, volume 702. John Wiley & Sons.

Kroonenberg, P. M. and De Leeuw, J. (1980). Principal component analysis of three-mode data by means of alternating least squares algorithms. *Psychometrika*, 45(1):69–97.

Kuerbis, A., Armeli, S., Muench, F., and Morgenstern, J. (2014). Profiles of confidence and commitment to change as predictors of moderated drinking: A person-centered approach. *Psychology of Addictive Behaviors*, 28(4):1065.

Levine, M. S. (1977). *Canonical analysis and factor comparison.* Number 6. Sage Publications.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

Magnus, J. R. and Neudecker, H. (2019). *Matrix differential calculus with applications in statistics and econometrics.* John Wiley & Sons.

McDonald, R. P. (1981). Constrained least squares estimators of oblique common factors. *Psychometrika*, 46(3):337–341.

Miyamoto, S., Ichihashi, H., Honda, K., and Ichihashi, H. (2008). *Algorithms for fuzzy clustering.* Springer.

Mosier, C. I. (1939). Determining a simple structure when loadings for certain tests are known. *Psychometrika*, 4:149–162.

Mulaik, S. A. (2009). *Foundations of factor analysis.* Chapman and Hall/CRC.

Mulaik, S. A. (2010). *Foundations of factor analysis. Second Edition.* Boca Raton: CRC Press.

Neudecker, H. (1981). On the matrix formulation of kaiser's varimax criterion. *Psychometrika*, 46(3):343–345.

Osgood, C. E. and Luria, Z. (1954). A blind analysis of a case of multiple personality using the semantic differential. *The Journal of Abnormal and Social Psychology*, 49(4p1):579.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572.

Peng, X., Zhou, C., Hepburn, D. M., Judd, M. D., and Siew, W. H. (2013). Application of k-means method to pattern recognition in on-line cable partial discharge monitoring. *IEEE Transactions on Dielectrics and Electrical Insulation*, 20(3):754–761.

Popper, K. (1962). Conjectures and refutations: The growth of scientific knowledge.

Qi, X., Luo, R., and Zhao, H. (2013). Sparse principal component analysis by choice of norm. *Journal of multivariate analysis*, 114:127–160.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850.

Reynolds, T. J. and Jackosfsky, E. F. (1981). Interpreting canonical analysis: The use of orthogonal transformations. *Educational and Psychological Measurement*, 41(3):661–671.

Rocci, R. and Vichi, M. (2005). Three-mode component analysis with crisp or fuzzy partition of units. *Psychometrika*, 70(4):715–736.

Roubens, M. (1982). Fuzzy clustering algorithms and their cluster validity. *European journal of operational research*, 10(3):294–301.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206.

Satomura, H. and Adachi, K. (2013). Oblique rotaton in canonical correlation analysis reformulated as maximizing the generalized coefficient of determination. *Psychometrika*, 78(3):526–537.

Schloss, K. B., Hawthorne-Madell, D., and Palmer, S. E. (2015). Ecological influences on individual differences in color preference. *Attention, Perception, & Psychophysics*, 77(8):2803–2816.

Sherin, R. J. (1966). A matrix formulation of kaiser's varimax criterion. *Psychometrika*, 31(4):535–538.

Slobodenyuk, N., Jraissati, Y., Kanso, A., Ghanem, L., and Elhajj, I. (2015). Cross-modal associations between color and haptics. *Attention, Perception, & Psychophysics*, 77(4):1379–1395.

Späth, H. (1979). Algorithm 39 clusterwise linear regression. *Computing*, 22(4):367–373.

Steinley, D. (2006). K-means clustering: a half-century synthesis. *British Journal of Mathematical and Statistical Psychology*, 59(1):1–34.

Steinley, D. and Brusco, M. J. (2008). Selection of variables in cluster analysis: An empirical comparison of eight procedures. *Psychometrika*, 73(1):125.

Steinley, D. and Hubert, L. (2008). Order-constrained solutions in k-means clustering: even better than being globally optimal. *Psychometrika*, 73(4):647.

Suk, H. W. and Hwang, H. (2010). Regularized fuzzy clusterwise ridge regression. *Advances in data analysis and classification*, 4(1):35–51.

Takane, Y. (1978). Multi dimensional scaling. *Seminars in Nuclear Medicine*, 8:283–298.

Takane, Y. (1980). *Multi Dimensional Scaling*. Todai Syuppan, Tokyo (in Japanese).

Tanaka, Y., Tarumi, T., and Wakimoto, K. (1984). *Pasokon Toukei Kaiseki Handbook II Tahenryo Kaiseki Hen (Handbook of Computerized Statistical Analysis II, Multivariate Analysis)*. Kyoritsu Syuppan (in Japanese).

Ten Berge, J. M. (1983). A generalization of kristof's theorem on the trace of certain matrix products. *Psychometrika*, 48(4):519–523.

ten Berge, J. M. (1993). *Least squares optimization in multivariate analysis*. DSWO Press, Leiden University Leiden.

ten Berge, J. M., de Leeuw, J., and Kroonenberg, P. M. (1987). Some additional results on principal components analysis of three-mode data by means of alternating least squares algorithms. *Psychometrika*, 52(2):183–191.

Thurstone, L. L. (1947). Multiple-factor analysis; a development and expansion of the vectors of mind.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Trendafilov, N. T. (2014). From simple structure to sparse components: a review. *Computational Statistics*, 29(3-4):431–454.

Trendafilov, N. T., Fontanella, S., and Adachi, K. (2017). Sparse exploratory factor analysis. *psychometrika*, 82(3):778–794.

Tucker, L. R. (1951). A method for synthesis of factor analysis studies (personal research section report no. 984). Technical report, U.S. Department of the Army, Washington, D.C.

Tucker, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311.

Tukey, J. W. (1970). *Exploratory Data Analysis. Preliminary edition.* Addison-Wesley.

Ullman, J. B. (2006). Structural equation modeling: Reviewing the basics and moving forward. *Journal of personality assessment*, 87(1):35–50.

Unkel, S. and Trendafilov, N. T. (2010). Simultaneous parameter estimation in exploratory factor analysis: An expository review. *International Statistical Review*, 78(3):363–382.

Vichi, M. (2017). Disjoint factor analysis with cross-loadings. *Advances in Data Analysis and Classification*, 11(3):563–591.

Vichi, M. and Kiers, H. A. (2001). Factorial k-means analysis for two-way data. *Computational Statistics & Data Analysis*, 37(1):49–64.

Vichi, M., Rocci, R., and Kiers, H. A. (2007). Simultaneous component and clustering models for three-way data: within and between approaches. *Journal of Classification*, 24(1):71–98.

Vichi, M. and Saporta, G. (2009). Clustering and disjoint principal component analysis. *Computational Statistics & Data Analysis*, 53(8):3194–3208.

Ward, J. H. J. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.

Wedel, M. and Steenkamp, J.-B. E. (1989). A fuzzy clusterwise regression approach to benefit segmentation. *International Journal of Research in Marketing*, 6(4):241–258.

Weisstein, E. W. (2002). *CRC concise encyclopedia of mathematics.* Chapman and Hall/CRC.

Yamashita, N. (2012a). Canonical correlation analysis formulated as maximizing sum of squared correlation and rotation of structure matrix. *The Japanese Journal of Behaviometrics*, 39(1):1–12.

Yamashita, N. (2012b). Sparse multivariate multiple regression analysis and its application. In *The 40th Conference of the Behaviormetric Society of Japan.*

Yamashita, N. (in press). Procrustes penalty function for matching matrices to targets with its applications. In *The Annual Meeting of the Psychometric Society.* Springer.

Yamashita, N. and Adachi, K. (2017). Spark: A new clustering algorithm for obtaining sparse and interpretable centroids. In *The Annual Meeting of the Psychometric Society*, pages 407–416. Springer.

Yamashita, N. and Adachi, K. (2019a). A modified k-means clustering procedure for obtaining a cardinality-constrained centroid matrix. *Journal of Classification*, pages 1–17.

Yamashita, N. and Adachi, K. (2019b). Permutimin: Factor rotation to simple structure with permutation of variables. *Multivariate behavioral research*, pages 1–13.

Yamashita, N. and Adachi, K. (in press). Layered multivariate regression and its applications. In *Advanced Researches in Classification and Data Science*. Springer.

Yamashita, N. and Mayekawa, S. (2012). A new biplot procedure for joint classification of objects and variables with $k$-means clustering. *Bulletin of Data Analysis of Japanese Classification Society*, 2(1):33–51.

Yamashita, N. and Mayekawa, S. (2015). A new biplot procedure with joint classification of objects and variables by fuzzy c-means clustering. *Advances in Data Analysis and Classification*, 9(3):243–266.

Yanai, H. and Ichikawa, M. (2006). Factor analysis. In Rao, C. and Sinharay, S., editors, *Psychometrics*, volume 26 of *Handbook of Statistics*, pages 257 – 296. Elsevier.

Yanai, H., Kashiwagi, S., and Kokusho, R. (1987). Construction of new personality inventory by means of factor analysis based on promax rotation. *The Japanese Journal of Psychology*, 58:158–165. (in Japanese).

Yanai, H., Shigemasu, K., Mayekawa, S.-I., and Ichikawa, M. (1990). *Inshi Bunseki - Sono Riron To Houhou (Factor Analysis -Theory and Method)*. Asakura Syuppan, Tokyo (in Japanese).

Yates, A. (1987). *Multivariate exploratory data analysis: A perspective on exploratory factor analysis*. Albany: State University of New York Press.

Zadeh, L. A. (1965). Fuzzy sets. *Information and control*, 8(3):338–353.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320.

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286.

# Acknowledgements

Firstly, I would like to express my sincere gratitude to my prime supervisor Prof. Kohei Adachi for the continuous support of my Ph.D research and related publications, strict but kind comments on my (stupid) ideas, exhaustive comments on my manuscripts, and countless suggestions. His guidance and devotion on our research kept motivated me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D study. I also show my thank to Prof. Yasoshima and Prof. Nittono, vise-supervisors of my Ph.D research.

Besides my supervisors, I would like to thank to Prof. Shin-ichi Mayekawa (my supervisor at Tokyo Institute of Technology), Assoc. Prof. Satoshi Usami (University of Tokyo), and Assoc. Prof. Kensuke Okada (University of Tokyo) for their insightful comments and encouragement which incented me to widen my research from various perspectives.

Also, I would like to show gratitude to my previous colleagues at Japan Tobacco Inc. and current colleagues at Yahoo Japan Inc.

Last but not the least, I would like to thank my family: my wife Shizuka, my son Akihito, my parents Takeshi and Shigeko who devoted themselves to working hard to send me Osaka University. Their heart-warming supports throughout writing this thesis and my life in general encouraged me in all time.