修 士 論 文 の 和 文 要 旨

研究科・専攻	大学院 情報理工 学研究科 情報・ネットワーク工学 専攻 博士前期課程		
氏　　名	吉川　剛平	学籍番号	1931160
論 文 題 目	Tensor Common Component Analysis based on Kronecker Product Representation (クロネッカー積表現に基づくテンソル共通成分分析)		

要　旨

諸科学，産業界のあらゆる分野において複雑な現象を解明するために，大規模なデータの蓄積が盛んに行われている．近年では，蓄積されるデータの種類も多様になってきており，複数種類のデータ集合に内在する価値ある特性を発見する分析手法の開発は必要不可欠である．発見された特性を検討することは，関心がある現象を説明している因子の本質的な関係や因果関係を解明する一助となる．このような統計的な分析手法の一つに共通成分分析がある．共通成分分析は，複数種類のデータ集合の共分散構造に着目し，共通した特徴を抽出し，データの潜在的な線形構造を探索する手法である．

一方で，機械学習やデータマイニングの分野でテンソル構造を持ったデータに対する分析手法の研究が注目を集めている．データの持つテンソル構造を考慮しながら分析を行うことで，データに内在する高次の関係性を正確に抽出できることが多く報告されている．従来の統計的な分析手法をテンソルデータに適用する場合，テンソル構造を考慮しながら分析を行うことができないため，データに内在する特徴を正確に抽出できない問題が生じる．そのため，近年では多くの統計的な分析手法がテンソルデータに適用できるように拡張されている．

本論文では，共通成分分析をテンソルデータに対して拡張したテンソル共通成分分析を提案する．テンソル共通成分分析は，クロネッカー積に基づくテンソルデータの共分散構造を用いて，複数種類のテンソルデータの集合に共通した特徴を抽出する手法である．また，テンソルデータに内在する潜在構造を同定するために，新たな推定アルゴリズムを提案し，その収束性に関する理論解析も行う．本論文では，実データ解析を通して提案手法の有効性および推定アルゴリズムの収束性について検証する．検証の結果，従来の共通成分分析や類似手法に比べて提案手法の方が内在する特徴を抽出する性能が高いことが確認された．また，提案するアルゴリズムの有効性を確認できた．

令和二年度 修士論文

Tensor Common Component Analysis based on Kronecker Product Representation

電気通信大学 大学院情報理工学研究科

情報・ネットワーク工学専攻 情報数理工学プログラム

学籍番号 1931160

吉川 剛平

主任指導教員 川野 秀一 准教授

指導教員 植野 真臣 教授

提出年月日 2021年1月25日

Contents

1 Introduction **4**

2 Preliminaries for Tensors **7**

2.1 Tensors 7

2.2 Inner Product and Norm 9

2.3 Unfolding Tensors 10

2.4 Tensor Multiplication 11

3 Statistical Dimensionality Reduction Methodologies **13**

3.1 Principal Component Analysis 13

3.2 Multilinear Principal Component Analysis 14

3.3 Common Component Analysis 16

4 Multilinear Common Component Analysis **18**

5 Estimation **22**

5.1 Initialization 22

5.2 Iterative Update of Parameters 23

6 Theory **25**

6.1 Analysis of Upper and Lower Bounds 25

6.2 Convergence Analysis 31

6.3 Computational Analysis 34

7 Experiment **36**

7.1 Experimental Setting 36

7.2 Performance Assessment 38

7.3 Behavior of Contraction Ratio 42

7.4 Efficacy of Solving the Quadratic Programming Problem 43

8 Concluding Remarks **50**

1 Introduction

Various statistical methodologies for extracting useful information from a large amount of data have been studied over the decades since the appearance of big data. In the present era, it is important to discover a common structure of multiple datasets. In an early study, Flury (1984) focused on the structure of the covariance matrices of multiple datasets and discussed the heterogeneity of the structure. The author reported that population covariance matrices differ between multiple datasets in practical applications. Many methodologies have been developed for treating the heterogeneity between covariance matrices of multiple datasets (see, e.g., Flury (1986, 1988); Flury and Gautschi (1986); Pourahmadi et al. (2007); Wang et al. (2011); Park and Konishi (2018)).

Among such methodologies, common component analysis (CCA) (Wang et al., 2011) is an effective tool for statistics. The central idea of CCA is to reduce the number of dimensions of data while losing as little information of the multiple datasets as possible. To reduce the number of dimensions, CCA reconstructs the data with a few new variables which are linear combinations of the original variables. For considering the heterogeneity between covariance matrices of multiple datasets, CCA assumes that there is a different covariance matrix for each dataset. There have been many papers on various statistical methodologies using multiple covariance matrices: discriminant analysis (Bensmail and Celeux, 1996), spectral decomposition (Boik, 2002), and a likelihood ratio test for multiple covariance matrices (Manly and Rayner, 1987). It should be noted that principal component analysis (PCA) (Pearson, 1901; Jolliffe, 2002) is a technique similar to CCA. In fact, CCA is a generalization of PCA; PCA can only be applied to single dataset, whereas CCA can be applied to multiple datasets.

Meanwhile, in various fields of research, including machine learning and computer

vision, the main interest has been in tensor data, which has a multidimensional array structure. In order to apply the conventional statistical methodologies, such as PCA, to tensor data, a simple approach is to first transform the tensor data into vector data and then apply the methodology. However, such an approach causes the following problems:

1. In losing the tensor structure of the data, the approach ignores the higher-order inherent relationships of the original tensor data.

2. Transforming tensor data to vector data increases the number of features large. It also has a high computational cost.

To overcome these problems, statistical methodologies for tensor data analyses have been proposed which take the tensor structure of the data into consideration. Such methods enable us to accurately extract higher-order inherent relationships in a tensor dataset. In particular, many existing statistical methodologies have been extended for tensor data, for example, multilinear PCA (MPCA) (Lu et al., 2008) and sparse PCA for tensor data analysis (Allen, 2012; Wang et al., 2012; Lai et al., 2014), as well as others (see Carroll and Chang (1970), Harshman (1970), Kiers (2000), Badeau and Boyer (2008), and Kolda and Bader (2009)).

In this paper, we extend CCA to tensor data analysis, proposing *multilinear common component analysis* (MCCA). MCCA discovers the common structure of multiple datasets of tensor data while losing as little of the information of the datasets as possible. To identify the common structure, we estimate a common basis constructed as linear combinations of the original variables. For estimating the common basis, we develop a new estimation algorithm based on the idea of CCA. In developing the estimation algorithm, two issues must be addressed. One is the convergence properties of the algorithm. The other is its computational cost. To determine the convergence properties, we investigated first the relationship between the initial

values of the parameters and global optimal solution and then the monotonic convergence of the estimation algorithm. These analyses revealed that our proposed algorithm guarantees convergence of the mode-wise global optimal solution under some conditions. To analyze the computational efficacy, we calculate the computational cost of our proposed algorithm and compare it with the computational cost of MPCA.

The rest of the paper is organized as follows. In Section 2, we present the properties and the basic calculations of tensors. Next, we briefly review related researches in Section 3. In Section 4, we formulate the MCCA model by constructing the covariance matrices of tensor data, based on a Kronecker product representation. Then, we formulate the estimation algorithm for MCCA in Section 5. In Section 6, we present the theoretical properties for our proposed algorithm and analyze the computational cost. The efficacy of the MCCA is demonstrated through the results of numerical experiments in Section 7. Concluding remarks are presented in Section 8. Our implementation of MCCA is available at `https://github.com/yoshikawa-kohei/MCCA`.

2 Preliminaries for Tensors

Interest in tensor data analysis has been increasing in recent years because the analysis enables us to understand higher-order inherent relationships. In this section, we introduce notations of tensors and basic methods of calculation in tensor data analysis.

2.1 Tensors

The tensor is a multidimensional array. Mathematically, an M-th order tensor is defined by the element of a space which consists of the tensor product of M vector spaces. The *order* represents the dimensions of the tensor, also known as a *mode*. In particular, the 0-th, 1st, and 2nd order tensors are called scalars, vectors, and matrices, respectively. Thus, the tensor is often taken as an extension of these concepts. An example of real tensor data includes color images as 3rd order tensor data. The color images are composed of width, height, and color, and thus their order is three. Another example is electroencephalogram data consisting of x-axis, y-axis, z-axis, and time-axis as 4-th order tensor data.

Let $\mathcal{X}$ be an M-th order tensor with dimensions $P_1, P_2, \ldots, P_M$ corresponding to each order . Then, the element of the tensor $\mathcal{X}$ at the coordinate $(p_1, p_2, \ldots, p_M)$ is represented as $x_{p_1 p_2 \ldots p_M}$. For example, when $M = 3$, $\mathcal{X}$ becomes a 3rd order tensor with the dimensions P_1, P_2, and P_3, denoting the element of the tensor $x_{p_1 p_2 p_3}$, as shown in Figure 1.

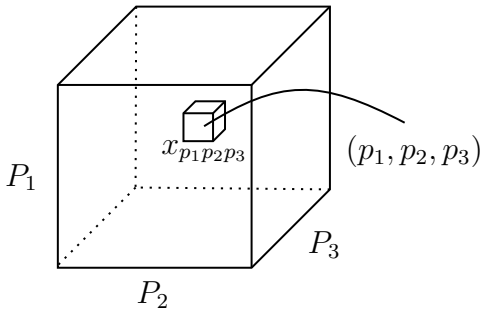
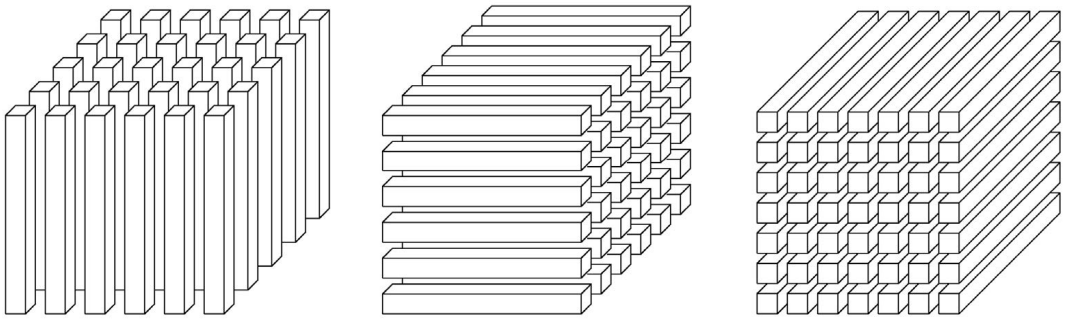


Figure 1: A 3rd order tensor

The concepts of rows and columns for matrices can be extended by considering a tensor subarray. The subarray is a tensor with a fixed subset of indices. When the index i corresponding to 1st mode of $\mathcal{X} \in \mathbb{R}^{P_1 \times P_2}$ is fixed, we obtain $\mathcal{X}_{i:} = [x_{i1}, x_{i2}, \ldots, x_{iP_2}]^\top$ as the subarray of $\mathcal{X}$. In such case, $\mathcal{X}_{i:}$ represents the i-th row vector of $\mathcal{X}$. Analogously, when we fix the index j corresponding to 2nd mode of $\mathcal{X}$, we can obtain the j-th column vector of $\mathcal{X}$ as $\mathcal{X}_{:j}$.

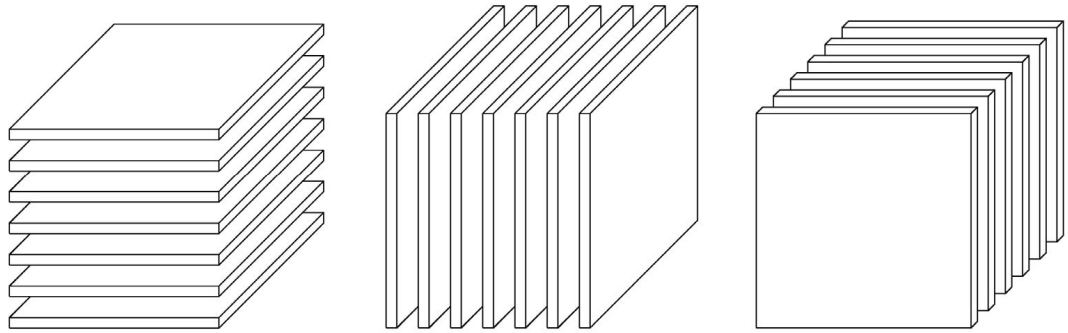
Now, for the 3rd order tensor $\mathcal{X} \in \mathbb{R}^{P_1 \times P_2 \times P_3}$, we consider $\mathcal{X}_{:jk}$, $\mathcal{X}_{i:k}$, and $\mathcal{X}_{ij:}$, denoting the subarrays of the tensor $\mathcal{X}$ with fixed all indices but one, respectively. These subarrays can be shown in Figure 2 and are called mode-i *fibers*. In particular, the mode-1 and mode-2 fibers correspond to the extension of column vectors and row vectors, respectively.

Similarly, $\mathcal{X}_{i::}$, $\mathcal{X}_{:j:}$, $\mathcal{X}_{::k}$ denote the subarrays of the tensor $\mathcal{X}$ with fixed all indices but two. These subarrays are illustrated as in Figure 3, and these are called *slices*.



(a) Mode-1 fibers : $\mathcal{X}_{:jk}$ (b) Mode-2 fibers : $\mathcal{X}_{i:k}$ (c) Mode-3 fibers : $\mathcal{X}_{ij:}$

Figure 2: Fibers of a 3rd order tensor (Kolda and Bader, 2009)



(a) Horizontal slices : $\mathcal{X}_{i::}$ (b) Lateral slices : $\mathcal{X}_{:j:}$ (c) Frontal slices : $\mathcal{X}_{::k}$

Figure 3: Slices of a 3rd order tensor (Kolda and Bader, 2009)

2.2 Inner Product and Norm

For two same size M-th order tensors $\mathcal{X}, \mathcal{Y} \in \mathbb{R}^{P_1 \times P_2 \times \cdots \times P_M}$, the inner product is defined by

$$\langle \mathcal{X}, \mathcal{Y} \rangle = \sum_{p_1=1}^{P_1} \sum_{p_2=1}^{P_2} \cdots \sum_{p_M=1}^{P_M} x_{p_1 p_2 \ldots p_M} y_{p_1 p_2 \ldots p_M}.$$

Similarly, a norm of tensor is the square root of the sum of the squares of all elements. For the M-th order tensor $\mathcal{X} \in \mathbb{R}^{P_1 \times P_2 \times \cdots \times P_M}$, the norm can be calculated as

$$\|\mathcal{X}\| = \sqrt{\sum_{p_1=1}^{P_1} \sum_{p_2=1}^{P_2} \cdots \sum_{p_M=1}^{P_M} x^2_{p_1 p_2 \ldots p_M}}.$$

By using the notation of inner product, we have $\langle \mathcal{X}, \mathcal{X} \rangle = \|\mathcal{X}\|^2$.

2.3 Unfolding Tensors

Unfolding is an operation of rearranging the elements of a tensor into a matrix or a vector. In particular, the transformation to matrices and vectors are known as *matricization* and *vectorization*, respectively. The mode-k unfolding from an M-th order tensor $\mathcal{X} \in \mathbb{R}^{P_1 \times P_2 \times \cdots \times P_M}$ to a matrix $\mathbf{X}^{(k)} \in \mathbb{R}^{P_k \times (\prod_{j \neq k} P_j)}$ means mapping the element of tensor $x_{p_1 p_2 \cdots p_M}$ to the element of matrix at the coordinate (p_k, l), where

$$l = 1 + \sum_{\substack{t=1, \\ t \neq k}}^{M} (p_t - 1) L_t \quad \text{with} \quad L_t = \prod_{\substack{m=1, \\ m \neq k}}^{t-1} P_m.$$

In the example of Kolda and Bader (2009), let the frontal slices of $\mathcal{X} \in \mathbb{R}^{3 \times 4 \times 2}$ be

$$\mathcal{X}_{::1} = \begin{bmatrix} 1 & 4 & 7 & 10 \\ 2 & 5 & 8 & 11 \\ 3 & 6 & 9 & 12 \end{bmatrix}, \ \mathcal{X}_{::2} = \begin{bmatrix} 13 & 16 & 19 & 22 \\ 14 & 17 & 20 & 23 \\ 15 & 18 & 21 & 24 \end{bmatrix}.$$

Then, the mode-k unfolded matrices for $k = 1, 2, 3$ are

$$\mathbf{X}^{(1)} = \begin{bmatrix} 1 & 4 & 7 & 10 & 13 & 16 & 19 & 22 \\ 2 & 5 & 8 & 11 & 14 & 17 & 20 & 23 \\ 3 & 6 & 9 & 12 & 15 & 18 & 21 & 24 \end{bmatrix},$$

$$\mathbf{X}^{(2)} = \begin{bmatrix} 1 & 2 & 3 & 13 & 14 & 15 \\ 4 & 5 & 6 & 16 & 17 & 18 \\ 7 & 8 & 9 & 19 & 20 & 21 \\ 10 & 11 & 12 & 22 & 23 & 24 \end{bmatrix},$$

$$\mathbf{X}^{(3)} = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & \cdots & 9 & 10 & 11 & 12 \\ 13 & 14 & 15 & 16 & 17 & \cdots & 21 & 22 & 23 & 24 \end{bmatrix}.$$

In particular, the unfolding a tensor into a vector is represented as $\mathrm{vec}(\cdot)$. In the above example, the unfolding is performed as follows:

$$\mathrm{vec}(\mathcal{X}) = \begin{bmatrix} 1 \\ 2 \\ \vdots \\ 24 \end{bmatrix}.$$

2.4 Tensor Multiplication

For an M-th order tensor $\mathcal{X} \in \mathbb{R}^{P_1 \times P_2 \times \cdots \times P_M}$ and a matrix $\mathbf{A} \in \mathbb{R}^{Q \times P_k}$, the k-th mode product is the multiplication of the tensor and the matrix with respect to k-th mode. The k-th mode product is represented as $\mathcal{X} \times_k \mathbf{A}$ and its size is $P_1 \times P_2 \times \cdots \times Q \times P_{k+1} \times \cdots \times P_M$. The element of the tensor $(\mathcal{X} \times_k \mathbf{A})$ can be

calculated as follows:

$$(\mathcal{X} \times_k \mathbf{A})_{p_1 p_2 \dots q p_{k+1} \dots p_M} = \sum_{p_k=1}^{P_k} x_{p_1 p_2 \dots p_M} a_{q p_k}.$$

Let $\mathbf{A}^{(k)} \in \mathbb{R}^{Q_k \times P_k}$ be a matrix for all $k \in \{1, 2, \dots, M\}$. Then, the multiplications of the tensor $\mathcal{X}$ and matrices $\mathbf{A}^{(k)}$ for $k = 1, 2, \dots, M$ can be represented as

$$\mathcal{Y} = \mathcal{X} \times_1 \mathbf{A}^{(1)} \times_2 \mathbf{A}^{(2)} \cdots \times_M \mathbf{A}^{(M)}.$$

Moreover, we obtain the following lemma.

Lemma 1. *Let $\mathcal{X} \in \mathbb{R}^{P_1 \times P_2 \times \cdots \times P_M}$ and $\mathbf{A}^{(k)} \in \mathbb{R}^{Q_k \times P_k}$ be an M-th order tensor and matrices for $k = 1, 2, \dots, M$, respectively. For any k, we have*

$$\mathcal{Y} = \mathcal{X} \times_1 \mathbf{A}^{(1)} \times_2 \mathbf{A}^{(2)} \cdots \times_M \mathbf{A}^{(M)} \Longleftrightarrow$$
$$\mathbf{Y}^{(k)} = \mathbf{A}^{(k)} \mathbf{X}^{(k)} \Big(\mathbf{A}^{(M)} \otimes \cdots \otimes \mathbf{A}^{(k+1)} \otimes \mathbf{A}^{(k-1)} \otimes \cdots \otimes \mathbf{A}^{(1)} \Big)^\top,$$

where $\otimes$ denotes the Kronecker product operator.

This lemma reveals the relationships between the tensor multiplications and the unfolding. Hence, this lemma helps us to reduce the calculations on tensor spaces to matrix algebra. For the proof of this lemma, see Kolda (2006).

3 Statistical Dimensionality Reduction Methodologies

In this section, we introduce the mathematical formulations of PCA, MPCA, and CCA and estimation methods for the models.

3.1 Principal Component Analysis

PCA is a well-known traditional statistical methodology for the dimensionality reduction in multivariate analysis. The main purpose of PCA is to reduce the dimensions of the dataset while losing as little information of the dataset. For preserving most of the information, PCA reconstructs the dataset with a few new variables which are linear combinations of original variables so that the new variables maximize the variance of the dataset. In this section, we introduce the mathematical formulation of PCA.

Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^\top$ be an $N \times P$ data matrix, which consists of N independent observations of the P-dimensional vector $\mathbf{x}_i$.

The procedure of PCA consists of two step. First, we project the original variables from P-dimensional space to R-dimensional space as follows:

$$\mathbf{y} = \mathbf{W}^\top \mathbf{x},$$

where $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_R]$ is a $P \times R$ orthogonal projection matrix, in which $R \leq P$. By projecting the original variables $\mathbf{x}$ to $\mathbf{y}$, we obtain projected data matrix $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_N]^\top$, where $\mathbf{y}_i = \mathbf{W}^\top \mathbf{x}_i$ for $i = 1, \ldots, N$. Second, we identify the projection matrix $\mathbf{W}$ which maximizes the variance of projected data matrix $\mathbf{Y}$. The sample

covariance matrix for projected data $\mathbf{Y}$ can be calculated as follows:

$$\mathbf{\Lambda} = \frac{1}{n}\sum_{i=1}^{N}(\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})^\top,$$

where $\bar{\mathbf{y}} = \frac{1}{n}\sum_{i=1}^{N}\mathbf{y}_i$ is an R-dimensional mean vector. By using the formulation $\mathbf{y}_i = \mathbf{W}^\top\mathbf{x}_i$ for $i = 1, \ldots, N$, we have

$$\mathbf{\Lambda} = \mathbf{W}^\top\mathbf{S}\mathbf{W},$$

where $\mathbf{S}$ is the sample covariance matrix for the data matrix $\mathbf{X}$ calculated as $\mathbf{S} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top$ in which $\bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i$ is a P-dimensional mean vector. Thus, we can obtain the following maximization problem:

$$\max_{\mathbf{W}} \operatorname{tr}\left(\mathbf{W}^\top\mathbf{S}\mathbf{W}\right), \text{ s.t. } \mathbf{W}^\top\mathbf{W} = \mathbf{I}_R,$$

where $\operatorname{tr}(\cdot)$ denotes the trace of a matrix and $\mathbf{I}_R$ is an identity matrix of the size $R \times R$. By solving this maximization problem, we can identify the projection matrix $\mathbf{W}$ which maximizes the variance of projected data matrix $\mathbf{Y}$. The maximizer consists of the R eigenvectors, corresponding to the R largest eigenvalues, obtained by eigenvalue decomposition of $\mathbf{S}$.

3.2 Multilinear Principal Component Analysis

In this paper, we introduce MPCA (Lu et al., 2008), which has been proposed as the extension of PCA for tensor data. The purpose of MPCA is also to find the projection matrices for the dimensionality reduction.

Let $\left\{\mathcal{X}_i \in \mathbb{R}^{P_1 \times P_2 \times \cdots \times P_M}, i = 1, 2, \ldots, N\right\}$ be an independently obtained M-th or-

der tensor dataset. Then, we obtain the following projected tensors:

$$\mathcal{Y}_i = \mathcal{X}_i \times_1 \mathbf{W}^{(1)} \times_2 \mathbf{W}^{(2)} \cdots \times_M \mathbf{W}^{(M)},\ i = 1, 2, \ldots, N,$$

where $\mathbf{W}^{(k)} \in \mathbb{R}^{R_k \times P_k}$ for all $k \in \{1, 2, \ldots, M\}$ are the orthogonal projection matrices. MPCA captures the information of the tensor dataset by the variance like as PCA. In tensor data analysis, the sample covariance matrix is defined by unfolding the tensors. Let $\mathbf{Y}_i^{(k)} \in \mathbb{R}^{R_k \times (\prod_{j \neq k} R_j)}$ be the k-th mode unfolded matrix of $\mathcal{Y}_i$. Then, the sample covariance matrix for k-th mode is defined by

$$\mathbf{\Lambda}^{(k)} = \frac{1}{N} \sum_{i=1}^{N} \left(\mathbf{Y}_i^{(k)} - \bar{\mathbf{Y}}^{(k)}\right)\left(\mathbf{Y}_i^{(k)} - \bar{\mathbf{Y}}^{(k)}\right)^\top, \tag{3.1}$$

where $\bar{\mathbf{Y}}^{(k)}$ is the k-th mode unfolded matrix of the mean tensor $\bar{\mathcal{Y}}$ calculated as $\bar{\mathcal{Y}} = \frac{1}{N} \sum_{i=1}^{N} \mathcal{Y}_i$. By applying Lemma 1 to the equation (3.1), we have

$$\mathbf{\Lambda}^{(k)} = \mathbf{W}^{(k)} \left\{ \frac{1}{N} \sum_{i=1}^{N} \left(\mathbf{X}_i^{(k)} - \bar{\mathbf{X}}^{(k)}\right) \mathbf{W}^{(-k)^\top} \mathbf{W}^{(-k)} \left(\mathbf{X}_i^{(k)} - \bar{\mathbf{X}}^{(k)}\right)^\top \right\} \mathbf{W}^{(k)^\top},$$

where $\mathbf{X}_i^{(k)} \in \mathbb{R}^{P_k \times (\prod_{j \neq k} P_j)}$ and $\bar{\mathbf{X}}^{(k)}$ are the k-th mode unfolded matrix of $\mathcal{X}_i$ and the mean tensor $\bar{\mathcal{X}}$ calculated as $\bar{\mathcal{X}} = \frac{1}{N} \sum_{i=1}^{N} \mathcal{X}_i$, respectively. $\mathbf{W}^{(-k)}$ is represented as the Kronecker product of the projection matrices but $\mathbf{W}^{(k)}$:

$$\mathbf{W}^{(-k)} = \mathbf{W}^{(M)} \otimes \cdots \otimes \mathbf{W}^{(k+1)} \otimes \mathbf{W}^{(k-1)} \otimes \cdots \otimes \mathbf{W}^{(1)}.$$

To find the projection matrices which maximize the variance, we consider the following maximization problem.

$$\max_{\mathbf{W}^{(k)}} \operatorname{tr} \left[\mathbf{W}^{(k)} \left\{ \frac{1}{N} \sum_{i=1}^{N} \left(\mathbf{X}_i^{(k)} - \bar{\mathbf{X}}^{(k)}\right) \mathbf{W}^{(-k)^\top} \mathbf{W}^{(-k)} \left(\mathbf{X}_i^{(k)} - \bar{\mathbf{X}}^{(k)}\right)^\top \right\} \mathbf{W}^{(k)^\top} \right]$$
$$\text{s.t. } \mathbf{W}^{(k)^\top} \mathbf{W}^{(k)} = \mathbf{I}_{R_k}, k = 1, 2, \ldots, M, \tag{3.2}$$

where $\mathbf{I}_{R_k}$ is an identity matrix of the size $R_k \times R_k$. The estimator $\mathbf{W}^{(k)}$ can be obtained by solving this maximization problem for $k = 1, 2, \ldots, M$. However, since the estimator $\mathbf{W}^{(k)}$ depends on the other estimators $\mathbf{W}^{(1)}, \ldots, \mathbf{W}^{(k-1)}, \mathbf{W}^{(k+1)}, \ldots, \mathbf{W}^{(M)}$, we cannot explicitly solve the maximization problem. Thus, MPCA obtains the estimates by iteratively solving the maximization problem for $k = 1, 2, \ldots, M$.

3.3 Common Component Analysis

In this subsection, we briefly review the CCA, which is proposed as an extension of the PCA.

Suppose that we obtain data matrices $\mathbf{X}_{(g)} = [\mathbf{x}_{(g)1}, \ldots \mathbf{x}_{(g)N_g}]^\top \in \mathbb{R}^{N_g \times P}$ with N_g observations and P variables for $g = 1, \ldots G$, where $\mathbf{x}_{(g)i}$ is the P-dimensional vector corresponding to the i-th row of $\mathbf{X}_{(g)}$ and G is the number of datasets. Then, the sample covariance matrix in group g is

$$\mathbf{S}_{(g)} = \frac{1}{N_g} \sum_{i=1}^{N_g} \Big(\mathbf{x}_{(g)i} - \bar{\mathbf{x}}_{(g)}\Big)\Big(\mathbf{x}_{(g)i} - \bar{\mathbf{x}}_{(g)}\Big)^\top, \quad g = 1, \ldots, G,$$

where $\mathbf{S}_{(g)} \in \mathbb{S}_+^P$, in which $\mathbb{S}_+^P$ is a set of symmetric positive definite matrices of the size $P \times P$, and $\bar{\mathbf{x}}_{(g)} = \frac{1}{N_g} \sum_{i=1}^{N_g} \mathbf{x}_{(g)i}$ is a P-dimensional mean vector in group g.

The main idea of the CCA model is to find the common structure of multiple datasets by projecting the data onto a common lower-dimensional space with the same basis as the datasets. Wang et al. (2011) assumed that the covariance matrices $\mathbf{S}_{(g)}$ for $g = 1, \ldots, G$ can be decomposed to a product of latent covariance matrices and an orthogonal matrix for the linear transformation as follows:

$$\mathbf{S}_{(g)} = \mathbf{V}\mathbf{\Lambda}_{(g)}\mathbf{V}^\top + \mathbf{E}_{(g)}, \quad \text{s.t.} \quad \mathbf{V}^\top\mathbf{V} = \mathbf{I}_R, \tag{3.3}$$

where $\mathbf{\Lambda}_{(g)} \in \mathbb{S}_+^R$ is the latent covariance matrix in group g, $\mathbf{V} \in \mathbb{R}^{P \times R}$ is an

orthogonal matrix for the linear transformation, and $\mathbf{E}_{(g)} \in \mathbb{S}_+^P$ is the error matrix in group g. $\mathbf{E}_{(g)}$ consists of the sum of outer products for independent random vectors $\sum_{i=1}^{N_g} \mathbf{e}_{(g)i}\mathbf{e}_{(g)i}^\top$ with mean $\mathrm{E}\Big[\mathbf{e}_{(g)i}\Big] = \mathbf{0}$ and covariance matrix $\mathrm{Cov}\Big[\mathbf{e}_{(g)i}\Big]$ ($> \boldsymbol{O}$) ($i = 1, 2, \ldots, N_g$). $\mathbf{V}$ determines the R-dimensional common subspace of the multiple datasets. In particular, by assuming $R < P$, the CCA can discover the latent structures of the datasets. Wang et al. (2011) referred to the model (3.3) as *common component analysis* (CCA).

The parameters $\mathbf{V}$ and $\boldsymbol{\Lambda}_{(g)}$ ($g = 1, \ldots, G$) are estimated by solving the minimization problem

$$\min_{\substack{\mathbf{V}, \boldsymbol{\Lambda}_{(g)} \\ g=1,\ldots,G}} \sum_{g=1}^{G} \left\| \mathbf{S}_{(g)} - \mathbf{V}\boldsymbol{\Lambda}_{(g)}\mathbf{V}^\top \right\|_F^2, \quad \text{s.t.} \quad \mathbf{V}^\top\mathbf{V} = \mathbf{I}_R, \tag{3.4}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. The estimator of latent covariance matrices $\boldsymbol{\Lambda}_{(g)}$ for $g = 1, \ldots, G$ can be obtained by solving the minimization problem as $\hat{\boldsymbol{\Lambda}}_{(g)} = \mathbf{V}^\top\mathbf{S}_{(g)}\mathbf{V}$. By using the estimated value $\hat{\boldsymbol{\Lambda}}_{(g)}$, the minimization problem can be reformulated as the following maximization problem:

$$\max_{\mathbf{V}} \mathrm{tr} \left\{ \mathbf{V}^\top \sum_{g=1}^{G} \Big(\mathbf{S}_{(g)}\mathbf{V}\mathbf{V}^\top\mathbf{S}_{(g)}\Big)\mathbf{V} \right\}, \quad \text{s.t.} \quad \mathbf{V}^\top\mathbf{V} = \mathbf{I}_R, \tag{3.5}$$

A crucial issue for solving the maximization problem (3.5) is the non-convexity. Certainly, the maximization problem is non-convex since the problem is defined on a set of orthogonal matrices, which is a non-convex set. Generally speaking, it is difficult to find the global optimal solution in non-convex optimization problems, such as the problem (3.5). To overcome this drawback, Wang et al. (2011) proposed an estimation algorithm in which the estimated parameters are guaranteed to constitute the global optimal solution under some conditions.

4 Multilinear Common Component Analysis

In this section, we introduce a mathematical formulation of the MCCA, which is an extension of the CCA in terms of tensor data analysis. Moreover, we formulate an optimization problem of MCCA and investigate its convergence properties.

Suppose that we independently obtain M-th order tensor data $\mathcal{X}_{(g)i} \in \mathbb{R}^{P_1 \times P_2 \times \cdots \times P_M}$ for $i = 1, \ldots N_g$. We set the datasets of the tensors $\mathcal{X}_{(g)} = [\mathcal{X}_{(g)1}, \mathcal{X}_{(g)2}, \ldots, \mathcal{X}_{(g)N_g}] \in \mathbb{R}^{P_1 \times P_2 \times \cdots \times P_M \times N_g}$ for $g = 1, \ldots, G$, where G is the number of datasets. Then, the sample covariance matrix in group g for the tensor dataset is defined by

$$\mathbf{S}^*_{(g)} := \mathbf{S}^{(1)}_{(g)} \otimes \mathbf{S}^{(2)}_{(g)} \otimes \cdots \otimes \mathbf{S}^{(M)}_{(g)}, \tag{4.1}$$

where $\mathbf{S}^*_{(g)} \in \mathbb{S}^P_+$, in which $P = \prod_{k=1}^M P_k$, and $\mathbf{S}^{(k)}_{(g)} \in \mathbb{S}^{P_k}_+$ is the sample covariance matrix for k-th mode in group g defined by

$$\mathbf{S}^{(k)}_{(g)} := \frac{1}{N_g \prod_{j \neq k} P_j} \sum_{i=1}^{N_g} \left(\mathbf{X}^{(k)}_{(g)i} - \bar{\mathbf{X}}^{(k)}_{(g)}\right)\left(\mathbf{X}^{(k)}_{(g)i} - \bar{\mathbf{X}}^{(k)}_{(g)}\right)^\top. \tag{4.2}$$

Here, $\mathbf{X}^{(k)}_{(g)i} \in \mathbb{R}^{P_k \times (\prod_{j \neq k} P_j)}$ is the k-th mode unfolded matrix of $\mathcal{X}_{(g)i}$ and $\bar{\mathbf{X}}^{(k)}_{(g)} \in \mathbb{R}^{P_k \times (\prod_{j \neq k} P_j)}$ is the k-th mode unfolded matrix of $\bar{\mathcal{X}}_{(g)} = \frac{1}{N_g} \sum_{i=1}^{N_g} \mathcal{X}_{(g)i}$. A representation of the tensor covariance matrix by Kronecker products is often used (Kermoal et al., 2002; Yu et al., 2004; Werner et al., 2008).

To formulate CCA in terms of tensor data analysis, we consider CCA for the k-th mode covariance matrix in group g as follows:

$$\mathbf{S}^{(k)}_{(g)} = \mathbf{V}^{(k)} \mathbf{\Lambda}^{(k)}_{(g)} {\mathbf{V}^{(k)}}^\top + \mathbf{E}^{(k)}_{(g)}, \quad \text{s.t.} \quad {\mathbf{V}^{(k)}}^\top \mathbf{V}^{(k)} = \mathbf{I}_{R_k}, \tag{4.3}$$

where $\mathbf{\Lambda}^{(k)}_{(g)} \in \mathbb{S}^{R_k}_+$ is the latent k-th mode covariance matrix in group g, $\mathbf{V}^{(k)} \in \mathbb{R}^{P_k \times R_k}$ is an orthogonal matrix for the linear transformation, and $\mathbf{E}^{(k)}_{(g)} \in \mathbb{S}^{P_k}_+$ is the error matrix in group g. $\mathbf{E}^{(k)}_{(g)}$ consists of the sum of outer products for indepen-

dent random vectors $\sum_{i=1}^{N_g} \mathbf{e}_{(g)i}^{(k)} \mathbf{e}_{(g)i}^{(k)\top}$ with mean $\mathrm{E}\left[\mathbf{e}_{(g)i}^{(k)}\right] = \mathbf{0}$ and covariance matrix $\mathrm{Cov}\left[\mathbf{e}_{(g)i}^{(k)}\right]$ $(> \boldsymbol{O})$ $(i = 1, 2, \ldots, N_g)$. Since $\mathbf{S}_{(g)}^{*}$ can be decomposed to a Kronecker product of $\mathbf{S}_{(g)}^{(k)}$ for $k = 1, \ldots, M$ in the formula (4.1), we obtain the following model:

$$\mathbf{S}_{(g)}^{*} = \mathbf{V}^{*} \mathbf{\Lambda}_{(g)}^{*} \mathbf{V}^{*\top} + \mathbf{E}_{(g)}^{*}, \quad \text{s.t.} \quad \mathbf{V}^{*\top} \mathbf{V}^{*} = \mathbf{I}_R, \tag{4.4}$$

where $R = \prod_{k=1}^{M} R_k$, $\mathbf{V}^{*} = \mathbf{V}^{(1)} \otimes \mathbf{V}^{(2)} \otimes \cdots \otimes \mathbf{V}^{(M)}$, $\mathbf{\Lambda}_{(g)}^{*} = \mathbf{\Lambda}_{(g)}^{(1)} \otimes \mathbf{\Lambda}_{(g)}^{(2)} \otimes \cdots \otimes \mathbf{\Lambda}_{(g)}^{(M)}$, and $\mathbf{E}_{(g)}^{*}$ is the error matrix in group g. We refer to this model as *multilinear common component analysis* (MCCA).

To find the R-dimensional common subspace between the multiple tensor datasets, MCCA determines $\mathbf{V}^{(1)}, \mathbf{V}^{(2)}, \ldots, \mathbf{V}^{(M)}$. As with CCA, we obtain the estimate of $\mathbf{\Lambda}_{(g)}^{*}$ for $g = 1, \ldots, G$ as $\hat{\mathbf{\Lambda}}_{(g)}^{*} = \mathbf{V}^{*\top} \mathbf{S}_{(g)}^{*} \mathbf{V}^{*}$. With respect to $\mathbf{V}^{*}$, we can obtain the estimate by solving the following maximization problem, which is similar to (3.5):

$$\max_{\mathbf{V}^{*}} \mathrm{tr} \left\{ \mathbf{V}^{*\top} \sum_{g=1}^{G} \left(\mathbf{S}_{(g)}^{*} \mathbf{V}^{*} \mathbf{V}^{*\top} \mathbf{S}_{(g)}^{*} \right) \mathbf{V}^{*} \right\}, \quad \text{s.t.} \quad \mathbf{V}^{*\top} \mathbf{V}^{*} = \mathbf{I}_R. \tag{4.5}$$

However, the number of parameters will be very large when we try to solve this problem directly. This large number of parameters result in a highly computational cost. Moreover, it may not be possible to discover the inherent relationships between the variables in each mode simply by solving the problem (4.5).

To solve the maximization problem efficiently and identify the inherent relationships, the maximization problem (4.5) can be decomposed into the mode-wise maximization problems represented in the following lemma.

Lemma 2. *An estimate of the parameters* $\mathbf{V}^{(k)}$ *for* $k = 1, 2, \ldots, M$ *in the maximization problem* (4.5) *can be obtained by solving the following maximization problem for*

each mode:

$$\max_{\substack{\mathbf{V}^{(k)} \\ k=1,2,\ldots,M}} \sum_{g=1}^{G} \prod_{k=1}^{M} \operatorname{tr}\left\{ {\mathbf{V}^{(k)}}^{\top} \mathbf{S}_{(g)}^{(k)} \mathbf{V}^{(k)} {\mathbf{V}^{(k)}}^{\top} \mathbf{S}_{(g)}^{(k)} \mathbf{V}^{(k)} \right\}, \quad \text{s.t.} \quad {\mathbf{V}^{(k)}}^{\top} \mathbf{V}^{(k)} = \mathbf{I}_{R_k}. \quad (4.6)$$

We provide two lemmas about Kronecker products before we prove Lemma 2.

Lemma 3. *For matrices* $\mathbf{A}, \mathbf{B}, \mathbf{C}$*, and* $\mathbf{D}$ *such that matrix products* $\mathbf{AC}$ *and* $\mathbf{BD}$ *can be calculated,*

$$(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = \mathbf{AC} \otimes \mathbf{BD}.$$

Lemma 4. *For square matrices* $\mathbf{A}$ *and* $\mathbf{B}$*,*

$$\operatorname{tr}(\mathbf{A} \otimes \mathbf{B}) = \operatorname{tr}(\mathbf{A}) \operatorname{tr}(\mathbf{B}).$$

These lemmas are known as the mixed-product property and the spectrum property, respectively; see Harville (1998) for detailed proofs. Here, we prove Lemma 2 by using the properties of Kronecker products.

***Proof of Lemma 2*:**

For the maximization problem (4.5), move the summation over index g out of the $\operatorname{tr}(\cdot)$ and replace $\mathbf{S}_{(g)}^{*}$ and $\mathbf{V}^{*}$ with $\mathbf{S}_{(g)}^{(1)} \otimes \mathbf{S}_{(g)}^{(2)} \otimes \cdots \otimes \mathbf{S}_{(g)}^{(M)}$ and $\mathbf{V}^{(1)} \otimes \mathbf{V}^{(2)} \otimes \cdots \otimes \mathbf{V}^{(M)}$, respectively. Then

$$\max_{\substack{\mathbf{V}^{(k)} \\ k=1,2,\ldots,M}} \sum_{g=1}^{G} \operatorname{tr}\Big\{ \left(\mathbf{V}^{(1)} \otimes \cdots \otimes \mathbf{V}^{(M)}\right)^{\top} \left(\mathbf{S}_{(g)}^{(1)} \otimes \cdots \otimes \mathbf{S}_{(g)}^{(M)}\right) \left(\mathbf{V}^{(1)} \otimes \cdots \otimes \mathbf{V}^{(M)}\right) \left(\mathbf{V}^{(1)} \otimes \cdots \otimes \mathbf{V}^{(M)}\right)^{\top} \left(\mathbf{S}_{(g)}^{(1)} \otimes \cdots \otimes \mathbf{S}_{(g)}^{(M)}\right) \left(\mathbf{V}^{(1)} \otimes \cdots \otimes \mathbf{V}^{(M)}\right) \Big\}, \quad \text{s.t.} \quad {\mathbf{V}^{(k)}}^{\top} \mathbf{V}^{(k)} = \mathbf{I}_{R_k}.$$

By Lemmas 3 and 4, we have

$$\max_{\substack{\mathbf{V}^{(k)\top}\mathbf{V}^{(k)}=\mathbf{I}_{R_k}\\ k=1,2,\ldots,M}} \sum_{g=1}^{G} \mathrm{tr}\left\{\left(\mathbf{V}^{(1)\top}\mathbf{S}_{(g)}^{(1)}\mathbf{V}^{(1)}\mathbf{V}^{(1)\top}\mathbf{S}_{(g)}^{(1)}\mathbf{V}^{(1)}\right)\cdots\left(\mathbf{V}^{(M)\top}\mathbf{S}_{(g)}^{(M)}\mathbf{V}^{(M)}\mathbf{V}^{(M)\top}\mathbf{S}_{(g)}^{(M)}\mathbf{V}^{(M)}\right)\right\}$$

$$= \max_{\substack{\mathbf{V}^{(k)\top}\mathbf{V}^{(k)}=\mathbf{I}_{R_k}\\ k=1,2,\ldots,M}} \sum_{g=1}^{G}\prod_{k=1}^{M} \mathrm{tr}\left\{\mathbf{V}^{(k)\top}\mathbf{S}_{(g)}^{(k)}\mathbf{V}^{(k)}\mathbf{V}^{(k)\top}\mathbf{S}_{(g)}^{(k)}\mathbf{V}^{(k)}\right\}.$$

This leads to the maximization problem in Lemma 2. □

However, we cannot simultaneously solve this maximization problem (4.6) for $\mathbf{V}^{(k)}, k = 1, 2, \ldots, M$. Thus, by summarizing the terms unrelated to $\mathbf{V}^{(k)}$ in the maximization problem (4.6), we can obtain the maximization problem for k-th mode:

$$\max_{\mathbf{V}^{(k)}} f_k(\mathbf{V}^{(k)}) = \max_{\mathbf{V}^{(k)}} \mathrm{tr}\left\{\mathbf{V}^{(k)\top}\mathbf{M}(\mathbf{V}^{(k)})\mathbf{V}^{(k)}\right\}, \quad \text{s.t.} \quad \mathbf{V}^{(k)\top}\mathbf{V}^{(k)} = \mathbf{I}_{R_k}, \tag{4.7}$$

where $\mathbf{M}(\mathbf{V}^{(k)}) = \sum_{g=1}^{G} w_{(g)}^{(-k)}\mathbf{S}_{(g)}^{(k)}\mathbf{V}^{(k)}\mathbf{V}^{(k)\top}\mathbf{S}_{(g)}^{(k)}$, in which $w_{(g)}^{(-k)}$ is given by

$$w_{(g)}^{(-k)} = \prod_{j\neq k} \mathrm{tr}\left\{\mathbf{V}^{(j)\top}\mathbf{S}_{(g)}^{(j)}\mathbf{V}^{(j)}\mathbf{V}^{(j)\top}\mathbf{S}_{(g)}^{(j)}\mathbf{V}^{(j)}\right\}.$$

Although an estimate of $\mathbf{V}^{(k)}$ can be obtained by solving the maximization problem (4.7), this problem is non-convex, since $\mathbf{V}^{(k)}$ is assumed to be an orthogonal matrix. Thus, the maximization problem has several local maxima. However, by choosing the initial values of parameters in the estimation near the global optimal solution, we can obtain the global optimal solution. In Section 5, we develop not only an estimation algorithm but also an initialization method for choosing the initial values of the parameters near the global optimal solution. The initialization method helps guarantee the convergence of our algorithm to the mode-wise global optimal solution.

5 Estimation

Our estimation algorithm consists of two steps: initializing the parameters and iteratively updating the parameters. The initialization step gives us the initial values of the parameters near the global optimal solution for each mode. Next, by iteratively updating the parameters, we can monotonically increase the value of the objective function (4.7) until convergence.

5.1 Initialization

The first step is to initialize the parameters $\mathbf{V}^{(k)}$ for each mode. We define an objective function $f'_k(\mathbf{V}^{(k)}) = \mathrm{tr}\left\{\mathbf{V}^{(k)\top}\mathbf{M}\left(\mathbf{I}^{(k)}\right)\mathbf{V}^{(k)}\right\}$ for $k = 1, \ldots, M$, where $\mathbf{M}\left(\mathbf{I}^{(k)}\right) = \sum_{g=1}^{G} w_{(g)}^{(-k)}\mathbf{S}_{(g)}^{(k)}\mathbf{S}_{(g)}^{(k)}$. Next, we adopt a maximizer of $f'_k(\mathbf{V}^{(k)})$ as initial values of the parameters $\mathbf{V}^{(k)}$. To obtain the maximizer, we need an initial value of $\boldsymbol{w}^{(k)} = \left[w_{(1)}^{(-k)}, w_{(2)}^{(-k)}, \ldots, w_{(G)}^{(-k)}\right]^\top$. The initial value for $\boldsymbol{w}^{(k)}$ is obtained by solving the quadratic programming problem

$$\min_{\boldsymbol{w}^{(k)}} \boldsymbol{w}^{(k)\top}\boldsymbol{\lambda}_0^{(k)}\boldsymbol{\lambda}_0^{(k)\top}\boldsymbol{w}^{(k)}, \quad \text{s.t.} \quad \boldsymbol{w}^{(k)} > \mathbf{0},\ \boldsymbol{w}^{(k)\top}\boldsymbol{\lambda}_1^{(k)}\boldsymbol{\lambda}_1^{(k)\top}\boldsymbol{w}^{(k)} = 1, \tag{5.1}$$

where

$$\begin{aligned}
\boldsymbol{\lambda}_0^{(k)} &= \left[\sum_{i=R_k+1}^{P_k}\lambda_{(1)i}^{(k)}, \sum_{i=R_k+1}^{P_k}\lambda_{(2)i}^{(k)}, \ldots, \sum_{i=R_k+1}^{P_k}\lambda_{(G)i}^{(k)}\right]^\top, \\
\boldsymbol{\lambda}_1^{(k)} &= \left[\sum_{i=1}^{P_k}\lambda_{(1)i}^{(k)}, \sum_{i=1}^{P_k}\lambda_{(2)i}^{(k)}, \ldots, \sum_{i=1}^{P_k}\lambda_{(G)i}^{(k)}\right]^\top,
\end{aligned} \tag{5.2}$$

in which $\lambda_{(g)i}^{(j)}$ is the i-th largest eigenvalue of $\mathbf{S}_{(g)}^{(j)}\mathbf{S}_{(g)}^{(j)}$.

Using the initial value of $\boldsymbol{w}^{(k)}$, we can obtain the initial value of the parameter $\mathbf{V}_0^{(k)}$ by maximizing $f'_k(\mathbf{V}^{(k)})$ for each mode. The maximizer consists of R_k eigenvectors, corresponding to the R_k largest eigenvalues, obtained by eigenvalue decomposition

of $\mathbf{M}\left(\mathbf{I}^{(k)}\right)$. The theoretical justification for this initialization will be discussed in Section 6.

5.2 Iterative Update of Parameters

The second step is to update parameters $\mathbf{V}^{(k)}$ for each mode. We update parameters such that the objective function $f_k(\mathbf{V}^{(k)})$ is maximized. Let $\mathbf{V}_s^{(k)}$ be the value of $\mathbf{V}^{(k)}$ at step s. Then, we solve the surrogate maximization problem

$$\max_{\mathbf{V}_{s+1}^{(k)}} \operatorname{tr}\left\{ {\mathbf{V}_{s+1}^{(k)}}^{\top} \mathbf{M}(\mathbf{V}_s^{(k)}) \mathbf{V}_{s+1}^{(k)} \right\}, \quad \text{s.t.} \quad {\mathbf{V}_{s+1}^{(k)}}^{\top} \mathbf{V}_{s+1}^{(k)} = \mathbf{I}_{R_k}. \tag{5.3}$$

The solution of (5.3) consists of R_k eigenvectors, corresponding to the R_k largest eigenvalues, obtained by eigenvalue decomposition of $\mathbf{M}(\mathbf{V}_s^{(k)})$. By iteratively updating the parameters, the objective function $f_k(\mathbf{V}^{(k)})$ is monotonically increased, which allows it to be maximized. The monotonically increasing property will be discussed in Section 6.

Our estimation procedure comprises the above estimation steps. The procedure is summarized as Algorithm 1.

Algorithm 1 Iteratively updating algorithm via eigenvalue decomposition

Input: M-th order tensor dataset $\left\{\mathcal{X}_{(g)} \in \mathbb{R}^{P_1 \times P_2 \times \cdots \times P_M \times N_g}, g = 1, 2, \ldots, G\right\}$.

Calculate covariance matrix for tensors: $\mathbf{S}^*_{(g)}$ via (4.1) and (4.2).

Step 1 Initialization:

$\boldsymbol{w}^{(k)} \leftarrow$ the solution of quadratic programming problem (5.1), $k = 1, 2, \ldots, M$.

$\mathbf{V}_0^{(k)} \leftarrow R_k$ eigenvectors obtained by the eigenvalue decomposition of $\mathbf{M}\left(\mathbf{I}^{(k)}\right)$, $k = 1, 2, \ldots, M$.

$\mathbf{\Lambda}_{(g)}^{(k)} \leftarrow \mathbf{V}^{(k)\top}\mathbf{S}_{(g)}^{(k)}\mathbf{V}^{(k)}, \quad k = 1, 2, \ldots, M; g = 1, 2, \ldots, G.$

Step 2 Updating parameters:

for $s = 1, 2, \ldots$ **do**

  Update $\mathbf{V}^{(k)}$**:** $\mathbf{V}_{s+1}^{(k)} \leftarrow R_k$ eigenvectors obtained by eigenvalue decomposition of $\mathbf{M}\left(\mathbf{V}_s^{(k)}\right), \quad k = 1, 2, \ldots, M.$

  Update $\mathbf{\Lambda}_{(g)}^{(k)}$**:** $\mathbf{\Lambda}_{(g)}^{(k)} \leftarrow \mathbf{V}_{s+1}^{(k)\top}\mathbf{S}_{(g)}^{(k)}\mathbf{V}_{s+1}^{(k)}, \quad k = 1, 2, \ldots, M; g = 1, 2, \ldots, G.$

return $\mathbf{V}^{(k)} \in \mathbb{R}^{P_k \times R_k}, \mathbf{\Lambda}_{(g)}^{(k)} \in \mathbb{S}_+^{R_k}, \quad k = 1, 2, \ldots, M; g = 1, 2, \ldots, G.$

6 Theory

This section presents the theoretical and computational analyses for Algorithm 1. Theoretical analyses consist of two steps. First, we prove that the initial values of parameters obtained in Section 5.1 are relatively close to the global optimal solution. If the initial values are close to the global maximum, then we can obtain the global optimal solution even if the maximization problem is non-convex. Second, we prove that the iterative updates of the parameters in Section 5.2 monotonically increase the value of objective function (4.7) by solving the surrogate problem (5.3). From the monotonically increasing property, the estimated parameters always converge at a stationary point. The combination of these two results enables us to obtain the mode-wise global optimal solution. In the computational analysis, we calculate computational cost for MCCA and then compare the cost with conventional methods. By comparing the costs, we investigate the computational efficacy of MCCA.

6.1 Analysis of Upper and Lower Bounds

The aim of this subsection is to provide the upper and lower bounds of the maximization problem (4.7). From the bounds, we find that the initial values in Section 5.1 are relatively close to the global optimal solution. As shown in the following lemma, the objective function for initialization $f_k'(\mathbf{V}^{(k)})$ gives the upper and lower bounds for $f_k(\mathbf{V}^{(k)})$.

Lemma 5. *Consider the maximization problem*

$$\max_{\mathbf{V}^{(k)}} f_k'(\mathbf{V}^{(k)}) = \max_{\mathbf{V}^{(k)}} \mathrm{tr} \left\{ \mathbf{V}^{(k)\top} \left(\sum_{g=1}^{G} w_{(g)}^{(-k)} \mathbf{S}_{(g)}^{(k)} \mathbf{S}_{(g)}^{(k)} \right) \mathbf{V}^{(k)} \right\}.$$

Let $M^{(k)} = \mathrm{tr}\left\{\sum_{g=1}^{G} w_{(g)}^{(-k)} \mathbf{S}_{(g)}^{(k)} \mathbf{S}_{(g)}^{(k)}\right\}$. *Then*

$$\frac{f_k'(\mathbf{V}^{(k)})^2}{M^{(k)}} \leq f_k(\mathbf{V}^{(k)}) \leq f_k'(\mathbf{V}^{(k)}).$$

***Proof of Lemma 5*:**

First, we prove $f_k(\mathbf{V}^{(k)}) \leq f_k'(\mathbf{V}^{(k)})$. For any orthogonal matrix $\mathbf{V}^{(k)} \in \mathbb{R}^{P_k \times R_k}$, we can always find an orthogonal matrix $\mathbf{V}_\perp^{(k)} \in \mathbb{R}^{P_k \times (P_k - R_k)}$ that satisfies $\mathbf{V}^{(k)\top}\mathbf{V}_\perp^{(k)} = \boldsymbol{O}$. Then the equation $\mathbf{V}^{(k)}{\mathbf{V}^{(k)}}^\top + \mathbf{V}_\perp^{(k)}{\mathbf{V}_\perp^{(k)}}^\top = \mathbf{I}_{P_k}$ holds. By definition,

$$\begin{aligned}
f_k(\mathbf{V}^{(k)}) &= \mathrm{tr}\left\{ {\mathbf{V}^{(k)}}^\top \left(\sum_{g=1}^{G} w_{(g)}^{(-k)} \mathbf{S}_{(g)}^{(k)} \mathbf{V}^{(k)} {\mathbf{V}^{(k)}}^\top \mathbf{S}_{(g)}^{(k)} \right) \mathbf{V}^{(k)} \right\} \\
&\leq \mathrm{tr}\left\{ {\mathbf{V}^{(k)}}^\top \left(\sum_{g=1}^{G} w_{(g)}^{(-k)} \mathbf{S}_{(g)}^{(k)} \left(\mathbf{V}^{(k)} {\mathbf{V}^{(k)}}^\top + \mathbf{V}_\perp^{(k)} {\mathbf{V}_\perp^{(k)}}^\top \right) \mathbf{S}_{(g)}^{(k)} \right) \mathbf{V}^{(k)} \right\} \\
&= \mathrm{tr}\left\{ {\mathbf{V}^{(k)}}^\top \left(\sum_{g=1}^{G} w_{(g)}^{(-k)} \mathbf{S}_{(g)}^{(k)} \mathbf{S}_{(g)}^{(k)} \right) \mathbf{V}^{(k)} \right\} \\
&= f_k'(\mathbf{V}^{(k)}).
\end{aligned}$$

Thus, we have obtained $f_k(\mathbf{V}^{(k)}) \leq f_k'(\mathbf{V}^{(k)})$.

Next, we prove $\frac{f_k'(\mathbf{V}^{(k)})^2}{M^{(k)}} \leq f_k(\mathbf{V}^{(k)})$. We define the following block matrices:

$$\begin{aligned}
\mathbf{A} &= \left[\sqrt{w_{(1)}^{(-k)}} \mathbf{S}_{(1)}^{(k)\frac{1}{2}} \mathbf{V}^{(k)} {\mathbf{V}^{(k)}}^\top \mathbf{S}_{(1)}^{(k)\frac{1}{2}}, \ldots, \sqrt{w_{(G)}^{(-k)}} \mathbf{S}_{(G)}^{(k)\frac{1}{2}} \mathbf{V}^{(k)} {\mathbf{V}^{(k)}}^\top \mathbf{S}_{(G)}^{(k)\frac{1}{2}} \right], \\
\mathbf{B} &= \left[\sqrt{w_{(1)}^{(-k)}} \mathbf{S}_{(1)}^{(k)}, \ldots, \sqrt{w_{(G)}^{(-k)}} \mathbf{S}_{(G)}^{(k)} \right]
\end{aligned}$$

Note that since $\mathbf{S}_{(g)}^{(k)}$ is a symmetric positive definite matrix, $\mathbf{S}_{(g)}^{(k)}$ can be decomposed

to $\mathbf{S}_{(g)}^{(k)\frac{1}{2}}\mathbf{S}_{(g)}^{(k)\frac{1}{2}}$. We calculate the traces of $\mathbf{AA}$, $\mathbf{AB}$, and $\mathbf{BB}$, respectively:

$$\begin{aligned}
\mathrm{tr}\,(\mathbf{AA}) &= \sum_{g=1}^{G} w_{(g)}^{(-k)}\,\mathrm{tr}\left(\mathbf{S}_{(g)}^{(k)\frac{1}{2}}\mathbf{V}^{(k)}\mathbf{V}^{(k)\top}\mathbf{S}_{(g)}^{(k)\frac{1}{2}}\mathbf{S}_{(g)}^{(k)\frac{1}{2}}\mathbf{V}^{(k)}\mathbf{V}^{(k)\top}\mathbf{S}_{(g)}^{(k)\frac{1}{2}}\right)\\
&= \sum_{g=1}^{G} w_{(g)}^{(-k)}\,\mathrm{tr}\left\{\mathbf{V}^{(k)\top}\mathbf{S}_{(g)}^{(k)}\mathbf{V}^{(k)}\mathbf{V}^{(k)\top}\mathbf{S}_{(g)}^{(k)}\mathbf{V}^{(k)}\right\}\\
&= \mathrm{tr}\left\{\mathbf{V}^{(k)\top}\left(\sum_{g=1}^{G} w_{(g)}^{(-k)}\mathbf{S}_{(g)}^{(k)}\mathbf{V}^{(k)}\mathbf{V}^{(k)\top}\mathbf{S}_{(g)}^{(k)}\right)\mathbf{V}^{(k)}\right\}\\
&= f_k(\mathbf{V}^{(k)}),
\end{aligned}$$

$$\begin{aligned}
\mathrm{tr}\,(\mathbf{AB}) &= \sum_{g=1}^{G} w_{(g)}^{(-k)}\,\mathrm{tr}\left(\mathbf{S}_{(g)}^{(k)\frac{1}{2}}\mathbf{V}^{(k)}\mathbf{V}^{(k)\top}\mathbf{S}_{(g)}^{(k)\frac{1}{2}}\mathbf{S}_{(g)}^{(k)}\right)\\
&= \sum_{g=1}^{G} w_{(g)}^{(-k)}\,\mathrm{tr}\left(\mathbf{S}_{(g)}^{(k)\frac{1}{2}}\mathbf{V}^{(k)}\mathbf{V}^{(k)\top}\mathbf{S}_{(g)}^{(k)\frac{1}{2}}\mathbf{S}_{(g)}^{(k)\frac{1}{2}}\mathbf{S}_{(g)}^{(k)\frac{1}{2}}\right)\\
&= \sum_{g=1}^{G} w_{(g)}^{(-k)}\,\mathrm{tr}\left\{\mathbf{V}^{(k)\top}\mathbf{S}_{(g)}^{(k)}\mathbf{S}_{(g)}^{(k)}\mathbf{V}^{(k)}\right\}\\
&= \mathrm{tr}\left\{\mathbf{V}^{(k)\top}\left(\sum_{g=1}^{G} w_{(g)}^{(-k)}\mathbf{S}_{(g)}^{(k)}\mathbf{S}_{(g)}^{(k)}\right)\mathbf{V}^{(k)}\right\}\\
&= f_k'(\mathbf{V}^{(k)}),
\end{aligned}$$

$$\mathrm{tr}\,(\mathbf{BB}) = \mathrm{tr}\left(\sum_{g=1}^{G} w_{(g)}^{(-k)}\mathbf{S}_{(g)}^{(k)}\mathbf{S}_{(g)}^{(k)}\right) = M^{(k)}.$$

From the Cauchy–Schwarz inequality, we have

$$f_k(\mathbf{V}^{(k)})M^{(k)} = \mathrm{tr}\,(\mathbf{AA})\,\mathrm{tr}\,(\mathbf{BB}) \geq \{\mathrm{tr}\,(\mathbf{AB})\}^2 = f_k'(\mathbf{V}^{(k)})^2.$$

By dividing both sides of the inequality by $M^{(k)}$, we obtain $\frac{f_k'(\mathbf{V}^{(k)})^2}{M^{(k)}} \leq f_k(\mathbf{V}^{(k)})$. This completes the proof. □

By using Lemma 5, we can obtain the bounds for the global maximum in the problem (4.7). Before providing the bounds, we define a contraction ratio.

Definition 1. *Let* $f_k'^{\max}$ *be the global maximum of* $f_k'(\mathbf{V}^{(k)})$ *and* $M^{(k)} = \mathrm{tr}\left\{\mathbf{M}\left(\mathbf{I}^{(k)}\right)\right\}$. *Then a contraction ratio of data for* k*-th mode is defined by*

$$\alpha^{(k)} = \frac{f_k'^{\max}}{M^{(k)}} = \frac{\mathrm{tr}\left\{\mathbf{V}_0^{(k)^\top}\mathbf{M}\left(\mathbf{I}^{(k)}\right)\mathbf{V}_0^{(k)}\right\}}{\mathrm{tr}\left\{\mathbf{M}\left(\mathbf{I}^{(k)}\right)\right\}}. \tag{6.1}$$

Note that a contraction ratio $\alpha^{(k)}$ satisfies $0 \leq \alpha^{(k)} \leq 1$ and $\alpha^{(k)} = 1$ if and only if $R_k = P_k$.

Using $f_k'^{\max}$ and the contraction ratio $\alpha^{(k)}$, we have the following theorem that reveals the upper and lower bounds of the global maximum in the problem (4.7).

Theorem 1. *Let* $f_k^{\max}$ *be the global maximum of* $f_k(\mathbf{V}^{(k)})$. *Then*

$$\alpha^{(k)} f_k'^{\max} \leq f_k^{\max} \leq f_k'^{\max},$$

where $\alpha^{(k)}$ *is the contraction ratio defined in (6.1) and* $f_k'^{\max}$ *is the global maximum of* $f_k'(\mathbf{V}^{(k)})$.

***Proof of Theorem 1*:**

Let $f_k'^{\max}$ be the global maximum of $f_k'(\mathbf{V}^{(k)})$ and $\mathbf{V}_0^{(k)} = \arg\max_{\mathbf{V}^{(k)}} f_k'(\mathbf{V}^{(k)})$. From Lemma 5 and the definition of $\alpha^{(k)}$, we have

$$\alpha^{(k)} f_k'^{\max} = \frac{f_k'(\mathbf{V}_0^{(k)})^2}{M^{(k)}} \leq f_k(\mathbf{V}_0^{(k)}).$$

Let $f_k^{\max}$ be the global maximum of $f_k(\mathbf{V}^{(k)})$. It then holds that $f_k(\mathbf{V}_0^{(k)}) \leq f_k^{\max}$. Thus

$$\alpha^{(k)} f_k'^{\max} \leq f_k^{\max}.$$

Let $\mathbf{V}_{0^*}^{(k)} = \arg\max_{\mathbf{V}^{(k)}} f_k(\mathbf{V}^{(k)})$. From Lemma 5, we have

$$f_k^{\max} = f_k(\mathbf{V}_{0^*}^{(k)}) \leq f_k'(\mathbf{V}_{0^*}^{(k)}).$$

Since $f_k'(\mathbf{V}_{0^*}^{(k)}) \leq f_k'^{\max}$, we have

$$f_k^{\max} \leq f_k'^{\max}.$$

Hence, we have obtained $\alpha^{(k)} f_k'^{\max} \leq f_k^{\max} \leq f_k'^{\max}$. □

This theorem indicates that $f_k'^{\max} \to f_k^{\max}$ when $\alpha^{(k)} \to 1$. Thus, it is important to obtain an $\alpha^{(k)}$ that is as close as possible to one. Since $\alpha^{(k)}$ depends on $\mathbf{V}_0^{(k)}$ and $\boldsymbol{w}^{(k)}$, $\mathbf{V}_0^{(k)}$ depends on $\boldsymbol{w}^{(k)}$. From this dependency, if we could set the initial value of $\boldsymbol{w}^{(k)}$ such that $\alpha^{(k)}$ is as large as possible, then we could obtain an initial value of $\mathbf{V}_0^{(k)}$ that attains a value near $f_k^{\max}$. The following theorem shows that we can compute the initial value of $\boldsymbol{w}^{(k)}$ such that $\alpha^{(k)}$ is maximized.

Theorem 2. *Let $\boldsymbol{\lambda}_0^{(k)}$ and $\boldsymbol{\lambda}_1^{(k)}$ be the vectors consisting of eigenvalues defined in (5.2). For $\boldsymbol{w}^{(k)} = \left[w_{(1)}^{(-k)}, w_{(2)}^{(-k)}, \ldots, w_{(G)}^{(-k)}\right]$ $(k = 1, 2, \ldots, M)$, suppose that the estimate $\hat{\boldsymbol{w}}^{(k)}$ is obtained by solving (5.1) for $k = 1, 2, \ldots, M$. Then $\hat{\boldsymbol{w}}^{(k)}$ maximizes $\alpha^{(k)}$.*

***Proof of Theorem 2*:**

By definition, we obtain

$$\alpha^{(k)} = \frac{f_k'^{\max}}{M^{(k)}} = \frac{\mathrm{tr}\left\{\mathbf{V}_0^{(k)\top}\left(\sum_{g=1}^G w_{(g)}^{(-k)}\mathbf{S}_{(g)}^{(k)}\mathbf{S}_{(g)}^{(k)}\right)\mathbf{V}_0^{(k)}\right\}}{\mathrm{tr}\left\{\sum_{g=1}^G w_{(g)}^{(-k)}\mathbf{S}_{(g)}^{(k)}\mathbf{S}_{(g)}^{(k)}\right\}}.$$

By using the eigenvalue representation, we can rewrite the numerator of $\alpha^{(k)}$ as

follows:

$$f_k'^{\max} = \sum_{g=1}^{G} w_{(g)}^{(-k)} \sum_{i=1}^{R_k} \lambda_{(g)i}^{(k)}.$$

On the other hand, the denominator of $\alpha^{(k)}$ can be represented as the sum of eigenvalues as follows:

$$M^{(k)} = \sum_{g=1}^{G} w_{(g)}^{(-k)} \sum_{i=1}^{P_k} \lambda_{(g)i}^{(k)}.$$

Thus, we can transform $\alpha^{(k)}$ as follows:

$$\alpha^{(k)} = \frac{\sum_{g=1}^{G} w_{(g)}^{(-k)} \sum_{i=1}^{R_k} \lambda_{(g)i}^{(k)}}{\sum_{g=1}^{G} w_{(g)}^{(-k)} \sum_{i=1}^{P_k} \lambda_{(g)i}^{(k)}}.$$

When we set

$$\begin{aligned}
\boldsymbol{\lambda}_0^{(k)} &= \left[\sum_{i=R_k+1}^{P_k} \lambda_{(1)i}^{(k)}, \sum_{i=R_k+1}^{P_k} \lambda_{(2)i}^{(k)}, \ldots, \sum_{i=R_k+1}^{P_k} \lambda_{(G)i}^{(k)} \right]^\top, \\
\boldsymbol{\lambda}_1^{(k)} &= \left[\sum_{i=1}^{P_k} \lambda_{(1)i}^{(k)}, \sum_{i=1}^{P_k} \lambda_{(2)i}^{(k)}, \ldots, \sum_{i=1}^{P_k} \lambda_{(G)i}^{(k)} \right]^\top, \\
\boldsymbol{w}^{(k)} &= \left[w_{(1)}^{(-k)}, w_{(2)}^{(-k)}, \ldots, w_{(G)}^{(-k)} \right]^\top,
\end{aligned}$$

we can reformulate $\alpha^{(k)}$ as

$$\alpha^{(k)} = \frac{\left(\boldsymbol{\lambda}_1^{(k)} - \boldsymbol{\lambda}_0^{(k)} \right)^\top \boldsymbol{w}^{(k)}}{\boldsymbol{\lambda}_1^{(k)\top} \boldsymbol{w}^{(k)}}.$$

Thus, we obtain the following maximization problem:

$$\max_{\boldsymbol{w}^{(k)}} \frac{\left(\boldsymbol{\lambda}_1^{(k)} - \boldsymbol{\lambda}_0^{(k)} \right)^\top \boldsymbol{w}^{(k)}}{\boldsymbol{\lambda}_1^{(k)\top} \boldsymbol{w}^{(k)}}, \quad \text{s.t.} \quad \boldsymbol{w}^{(k)} > \mathbf{0}.$$

Note that the constraints can be obtained by the definition of $\boldsymbol{w}^{(k)}$. In addition,

this maximization problem can be reformulated as

$$\max_{\boldsymbol{w}^{(k)}} \frac{\left(\boldsymbol{\lambda}_1^{(k)} - \boldsymbol{\lambda}_0^{(k)}\right)^\top \boldsymbol{w}^{(k)}}{\boldsymbol{\lambda}_1^{(k)\top} \boldsymbol{w}^{(k)}} = \max_{\boldsymbol{w}^{(k)}} 1 - \frac{\boldsymbol{\lambda}_0^{(k)\top} \boldsymbol{w}^{(k)}}{\boldsymbol{\lambda}_1^{(k)\top} \boldsymbol{w}^{(k)}} = \min_{\boldsymbol{w}^{(k)}} \frac{\boldsymbol{\lambda}_0^{(k)\top} \boldsymbol{w}^{(k)}}{\boldsymbol{\lambda}_1^{(k)\top} \boldsymbol{w}^{(k)}}.$$

Since $\boldsymbol{\lambda}_0^{(k)\top} \boldsymbol{w}^{(k)} / \boldsymbol{\lambda}_1^{(k)\top} \boldsymbol{w}^{(k)}$ is non-negative, solving the optimization problem for the squared function of the objective function maintains generality. Thus, we can consider the following minimization problem:

$$\min_{\boldsymbol{w}^{(k)}} \frac{\boldsymbol{w}^{(k)\top} \boldsymbol{\lambda}_0^{(k)} \boldsymbol{\lambda}_0^{(k)\top} \boldsymbol{w}^{(k)}}{\boldsymbol{w}^{(k)\top} \boldsymbol{\lambda}_1^{(k)} \boldsymbol{\lambda}_1^{(k)\top} \boldsymbol{w}^{(k)}}, \quad \text{s.t.} \quad \boldsymbol{w}^{(k)} > \mathbf{0}.$$

Additionally, from the invariance under multiplication of $\boldsymbol{w}^{(k)}$ by a constant, we obtain the following objective function of the quadratic programming problem.

$$\min_{\boldsymbol{w}^{(k)}} \boldsymbol{w}^{(k)\top} \boldsymbol{\lambda}_0^{(k)} \boldsymbol{\lambda}_0^{(k)\top} \boldsymbol{w}^{(k)}, \quad \text{s.t.} \quad \boldsymbol{w}^{(k)} > \mathbf{0},\ \boldsymbol{w}^{(k)\top} \boldsymbol{\lambda}_1^{(k)} \boldsymbol{\lambda}_1^{(k)\top} \boldsymbol{w}^{(k)} = 1.$$

The proof is complete. □

In fact, $\alpha^{(k)}$ is very close to one with the initial values given in Theorem 2 even if R_k is small. This resembles the cumulative contribution ratio in PCA.

6.2 Convergence Analysis

We next verify that our proposed procedure for iteratively updating parameters maximizes the optimization problem (4.7). In Algorithm 1, the parameter $\mathbf{V}_{s+1}^{(k)}$ can be obtained by solving the surrogate maximization problem (5.3). The following Theorem 3 shows that we can monotonically increase the value of the function $f_k(\mathbf{V}^{(k)})$ in (4.7) by Algorithm 1.

Theorem 3. *Let* $\mathbf{V}_{s+1}^{(k)}$ *be* R_k *eigenvectors, corresponding to the* R_k *largest eigenvalues, obtained by eigenvalue decomposition of* $\mathbf{M}(\mathbf{V}_s^{(k)})$*. Then*

$$f_k(\mathbf{V}_s^{(k)}) \leq f_k(\mathbf{V}_{s+1}^{(k)}).$$

***Proof of Theorem 3*:**

We define the following block matrices:

$$\mathbf{A}_s = \left[\sqrt{w_{(1)}^{(-k)}}\mathbf{S}_{(1)}^{(k)\,\frac{1}{2}}\mathbf{V}_s^{(k)}\mathbf{V}_s^{(k)\top}\mathbf{S}_{(1)}^{(k)\,\frac{1}{2}}, \ldots, \sqrt{w_{(G)}^{(-k)}}\mathbf{S}_{(G)}^{(k)\,\frac{1}{2}}\mathbf{V}_s^{(k)}\mathbf{V}_s^{(k)\top}\mathbf{S}_{(G)}^{(k)\,\frac{1}{2}}\right].$$

Here, we calculate the traces of $\mathbf{A}_s\mathbf{A}_s$, $\mathbf{A}_s\mathbf{A}_{s+1}$, and $\mathbf{A}_{s+1}\mathbf{A}_{s+1}$. The calculations of $\mathrm{tr}\,(\mathbf{A}_s\mathbf{A}_s)$ and $\mathrm{tr}\,(\mathbf{A}_{s+1}\mathbf{A}_{s+1})$ are the same as that of $\mathrm{tr}\,(\mathbf{A}\mathbf{A})$ by replacing $\mathbf{V}^{(k)}$ with $\mathbf{V}_s^{(k)}$ and $\mathbf{V}^{(k)}$ with $\mathbf{V}_{s+1}^{(k)}$, respectively, in Lemma 5. Thus, we obtain

$$\mathrm{tr}\,(\mathbf{A}_s\mathbf{A}_s) = f_k(\mathbf{V}_s^{(k)}),$$

$$\begin{aligned}
\mathrm{tr}\,(\mathbf{A}_s\mathbf{A}_{s+1}) &= \sum_{g=1}^{G} w_{(g)}^{(-k)}\,\mathrm{tr}\left(\mathbf{S}_{(g)}^{(k)\,\frac{1}{2}}\mathbf{V}_s^{(k)}\mathbf{V}_s^{(k)\top}\mathbf{S}_{(g)}^{(k)\,\frac{1}{2}}\mathbf{S}_{(g)}^{(k)\,\frac{1}{2}}\mathbf{V}_{s+1}^{(k)}\mathbf{V}_{s+1}^{(k)\top}\mathbf{S}_{(g)}^{(k)\,\frac{1}{2}}\right) \\
&= \sum_{g=1}^{G} w_{(g)}^{(-k)}\,\mathrm{tr}\left\{\mathbf{V}_{s+1}^{(k)\top}\mathbf{S}_{(g)}^{(k)}\mathbf{V}_s^{(k)}\mathbf{V}_s^{(k)\top}\mathbf{S}_{(g)}^{(k)}\mathbf{V}_{s+1}^{(k)}\right\} \\
&= \mathrm{tr}\left\{\mathbf{V}_{s+1}^{(k)\top}\left(\sum_{g=1}^{G} w_{(g)}^{(-k)}\mathbf{S}_{(g)}^{(k)}\mathbf{V}_s^{(k)}\mathbf{V}_s^{(k)\top}\mathbf{S}_{(g)}^{(k)}\right)\mathbf{V}_{s+1}^{(k)}\right\},
\end{aligned}$$

$$\mathrm{tr}\,(\mathbf{A}_{s+1}\mathbf{A}_{s+1}) = f_k(\mathbf{V}_{s+1}^{(k)}).$$

Since $\mathbf{V}_{s+1}^{(k)} = \arg\max_{\mathbf{V}^{(k)}} \mathrm{tr}\left\{\mathbf{V}^{(k)\top}\left(\sum_{g=1}^{G} w_{(g)}^{(-k)}\mathbf{S}_{(g)}^{(k)}\mathbf{V}_s^{(k)}\mathbf{V}_s^{(k)\top}\mathbf{S}_{(g)}^{(k)}\right)\mathbf{V}^{(k)}\right\}$, we have

$$\begin{aligned} f_k(\mathbf{V}_s^{(k)}) &= \mathrm{tr}\left\{\mathbf{V}_s^{(k)\top}\left(\sum_{g=1}^{G} w_{(g)}^{(-k)}\mathbf{S}_{(g)}^{(k)}\mathbf{V}_s^{(k)}\mathbf{V}_s^{(k)\top}\mathbf{S}_{(g)}^{(k)}\right)\mathbf{V}_s^{(k)}\right\} \\ &\leq \mathrm{tr}\left\{\mathbf{V}_{s+1}^{(k)\top}\left(\sum_{g=1}^{G} w_{(g)}^{(-k)}\mathbf{S}_{(g)}^{(k)}\mathbf{V}_s^{(k)}\mathbf{V}_s^{(k)\top}\mathbf{S}_{(g)}^{(k)}\right)\mathbf{V}_{s+1}^{(k)}\right\} \\ &= \mathrm{tr}\left(\mathbf{A}_s\mathbf{A}_{s+1}\right). \end{aligned}$$

From the positivity of both sides of the inequality, it holds that

$$f_k(\mathbf{V}_s^{(k)})^2 \leq \left[\mathrm{tr}\left(\mathbf{A}_s\mathbf{A}_{s+1}\right)\right]^2.$$

In addition, from the Cauchy–Schwarz inequality, we have

$$\begin{aligned} f_k(\mathbf{V}_s^{(k)})f_k(\mathbf{V}_{s+1}^{(k)}) &= \mathrm{tr}\left(\mathbf{A}_s\mathbf{A}_s\right)\mathrm{tr}\left(\mathbf{A}_{s+1}\mathbf{A}_{s+1}\right) \\ &\geq \left[\mathrm{tr}\left(\mathbf{A}_s\mathbf{A}_{s+1}\right)\right]^2. \end{aligned}$$

Thus,

$$f_k(\mathbf{V}_s^{(k)})f_k(\mathbf{V}_{s+1}^{(k)}) \geq \left[\mathrm{tr}\left(\mathbf{A}_s\mathbf{A}_{s+1}\right)\right]^2 \geq f_k(\mathbf{V}_s^{(k)})^2.$$

Thus, we have obtained $f_k(\mathbf{V}_s^{(k)})^2 \leq f_k(\mathbf{V}_s^{(k)})f_k(\mathbf{V}_{s+1}^{(k)})$. By dividing both sides of the inequality by $f_k(\mathbf{V}_s^{(k)})$, we obtain the relation $f_k(\mathbf{V}_s^{(k)}) \leq f_k(\mathbf{V}_{s+1}^{(k)})$. □

From Theorem 1, we obtain initial values of the parameters that are near the global optimal solution. By combining Theorem 1 and Theorem 3, the solution from Algorithm 1 can be characterized by the following corollary.

Corollary 1. *Consider the maximization problem (4.7). Suppose that the initial value of the parameter is obtained by* $\mathbf{V}_0^{(k)} = \arg\max_{\mathbf{V}^{(k)}} \widetilde{f_k}'(\mathbf{V}^{(k)})$ *and the parameter* $\mathbf{V}_s^{(k)}$ *is repeatedly updated by Algorithm 1. Then the mode-wise global maximum for*

the maximization problem (4.7) is achieved when all the contraction ratios $\alpha^{(k)}$ *for* $k = 1, 2, \ldots, M$ *go to one.*

Algorithm 1 does not guarantee the global solution, due to the fundamental problem of non-convexity, but it is enough for pragmatic purposes. We will investigate the issue of convergence to global solution through numerical studies in Section 7.3.

6.3 Computational Analysis

First, we analyze the computational cost. To simplify the analysis, we assume $P = \arg\max_j P_j$ for $j = 1, 2, \ldots, M$. This implies that P is the upper bound of R_j for all j. We then calculate the upper bound of the computational complexity.

The expensive computations of the each iteration in Algorithm 1 consist of three parts: the formulation of $\mathbf{M}(\mathbf{V}_s^{(k)})$, the eigenvalue decomposition of $\mathbf{M}(\mathbf{V}_s^{(k)})$, and updating latent covariance matrices $\mathbf{\Lambda}_g^{(k)}$. These steps are $O(GM^2P^3)$, $O(P^3)$, and $O(GMP^3)$, respectively. The total computational complexity per iteration is then $O(GM^2P^3)$. This indicates that the MCCA algorithm is not limited by the sample size. In contrast, the MPCA algorithm is affecred by the sample size (Lu et al., 2008).

Next, we analyze the memory requirement of Algorithm 1. MCCA represents the original tensor data with fewer parameters by projecting the data onto a lower-dimensional space. This requires the $P_k \times R_k$ projection matrices $\mathbf{V}^{(k)}$ for $k = 1, 2, \ldots, M$. MCCA projects the data a the size of $N\left(\prod_{k=1}^{M} P_k\right)$ to $N\left(\prod_{k=1}^{M} R_k\right)$, where $N = \sum_{g=1}^{G} N_g$. Thus, the required size for the parameters is $\sum_{k=1}^{M} P_k R_k + N\left(\prod_{k=1}^{M} R_k\right)$. MPCA requires the same amount of memory as MCCA. Meanwhile, CCA and PCA need a projection matrix, which is size $R\left(\prod_{k=1}^{M} P_k\right)$. The required size for the parameters is then $R\left(\prod_{k=1}^{M} P_k\right) + NR$. It should be noted that MCCA and MPCA require a large amount of memory when the number of modes in a

dataset is large, but their memory requirements are much smaller than those of CCA and PCA.

7 Experiment

To demonstrate the efficacy of MCCA, we applied MCCA, PCA, CCA, and MPCA to image compression tasks.

7.1 Experimental Setting

For the experiments, we prepared the following three image datasets:

MNIST dataset consists of data of hand written digits $0, 1, \ldots, 9$ at image sizes of 28×28 pixels. The dataset includes a training dataset of 60,000 images and a test dataset of 10,000 images. We used the first 10 training images of the dataset for each group. The MNIST dataset (Lecun et al., 1998) is available at `http://yann.lecun.com/exdb/mnist/`.

AT&T (ORL) face dataset contains gray-scale facial images of 40 people. The dataset has 10 images sized 92×112 pixels for each person. We used images resized by a factor of 0.5 in order to improve the efficiency of the experiment. The AT&T face dataset is available at `https://git-disl.github.io/GTDLBench/datasets/att_face_dataset/`.

Cropped AR database has color facial images of 100 people. These images are cropped around the face. The size of images is $120 \times 165 \times 3$ pixels. The dataset contains 26 images in each group, 12 of which are images of people wearing sunglasses or scarves. We used the cropped facial images of 50 males which were not wearing sunglasses or scarves. Due to memory limitations, we resized these images by a factor of 0.25. The AR database (Martinez and Benavente., 1998; Martinez and Kak, 2001) is available at `http://www2.ece.ohio-state.edu/~aleix/ARdatabase.html`.

The dataset characteristics are summarized in Table 1.

Table 1: Summary of the datasets.

Dataset	Group size	Sample size (/group)	Number of dimensions	Number of groups
MNIST	Small	10	$28 \times 28 = 784$	10
AT&T(ORL)	Small	10	$46 \times 56 = 2576$	10
	Medium			20
	Large			40
Cropped AR	Small	14	$30 \times 41 \times 3 = 7380$	10
	Medium			25
	Large			50

To compress these images, we performed dimensionality reductions by MCCA, PCA, CCA, and MPCA, as follows. We vectorized the tensor dataset before performing PCA and CCA. In MCCA, the images were compressed and reconstructed according to the following steps.

1. Prepare the multiple image datasets $\mathcal{X}_{(g)} \in \mathbb{R}^{P_1 \times P_2 \times \cdots \times P_M \times N_g}$ for $g = 1, 2, \ldots, G$.
2. Compute the covariance matrix of $\mathcal{X}_{(g)}$ for $g = 1, 2, \ldots, G$.
3. From these covariance matrices, compute the linear transformation matrices $\mathbf{V}^{(k)} \in \mathbb{R}^{P_k \times R_k}$ for $k = 1, 2, \ldots, M$ for mapping to the $(R_1, R_2, \ldots, R_M)$-dimensional latent space.
4. Map the i-th sample $\mathcal{X}_{(g)i}$ to $\mathcal{X}_{(g)i} \times_1 {\mathbf{V}^{(1)}}^\top \times_2 {\mathbf{V}^{(2)}}^\top \cdots \times_M {\mathbf{V}^{(M)}}^\top \in \mathbb{R}^{R_1 \times R_2 \times \cdots \times R_M}$.
5. Reconstruct i-th sample $\tilde{\mathcal{X}}_{(g)i} = \mathcal{X}_{(g)i} \times_1 {\mathbf{V}^{(1)}}^\top \mathbf{V}^{(1)} \times_2 {\mathbf{V}^{(2)}}^\top \mathbf{V}^{(2)} \cdots \times_M {\mathbf{V}^{(M)}}^\top \mathbf{V}^{(M)}$.

Meanwhile, PCA and MPCA each require a single dataset. Thus, we aggregated the datasets as $\mathcal{X} = [\mathcal{X}_{(1)}, \mathcal{X}_{(2)}, \ldots, \mathcal{X}_{(G)}] \in \mathbb{R}^{P_1 \times P_2 \times \cdots \times P_M \times \sum_{g=1}^{G} Ng}$ and performed PCA and MPCA for the dataset $\mathcal{X}$.

7.2 Performance Assessment

For MCCA and MPCA, the reduced dimensions R_1 and R_2 were chosen as the same number, and then we fixed R_3 as two. All computations were performed by the software `R` (ver. 3.6) (R Core Team, 2019). In the initialization of MCCA, solving the quadratic programming problem was carried out using the function `ipop` in the package `kernlab`. MPCA was implemented as the function `mpca` in the package `rTensor`. The implementations of MCCA, PCA, and CCA are available at `https://github.com/yoshikawa-kohei/MCCA`.

To assess their performances, we calculated the reconstruction error rate (RER) under the same compression ratio (CR). RER is defined by

$$\mathrm{RER} = \frac{\left\|\mathcal{X} - \widetilde{\mathcal{X}}\right\|^2}{\|\mathcal{X}\|^2}, \tag{7.1}$$

where $\widetilde{\mathcal{X}} = [\widetilde{\mathcal{X}}_{(1)}, \widetilde{\mathcal{X}}_{(2)}, \ldots, \widetilde{\mathcal{X}}_{(G)}]$ is the aggregated dataset of reconstructed tensors $\widetilde{\mathcal{X}}_{(g)} = [\tilde{\mathcal{X}}_{(g)1}, \tilde{\mathcal{X}}_{(g)2}, \ldots, \tilde{\mathcal{X}}_{(g)N_g}]$ for $g = 1, 2, \ldots, G$. In addition, we defined CR as

$$\mathrm{CR} = \frac{\#\left\{\text{The number of required parameters}\right\}}{N \cdot \prod_{k=1}^{M} P_k}. \tag{7.2}$$

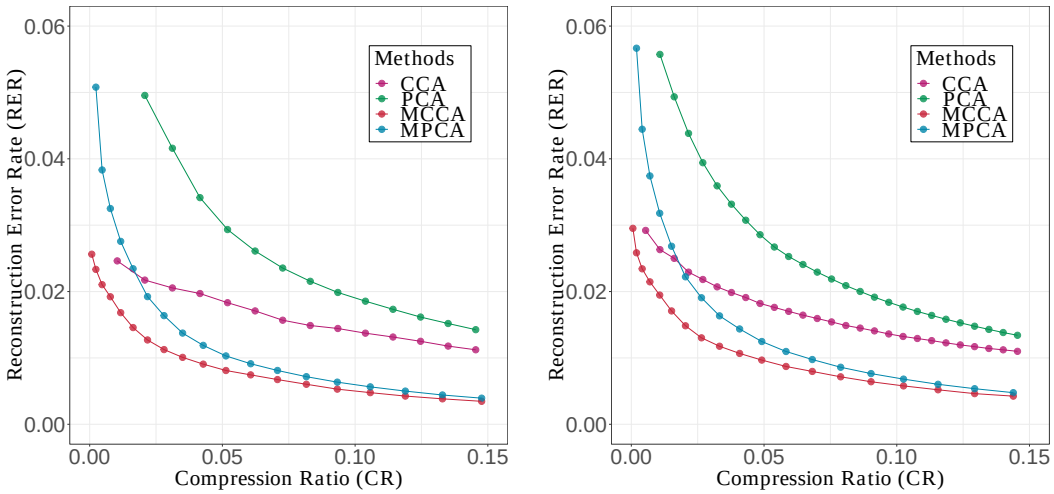
The number of parameters required for MCCA and MPCA is $\sum_{k=1}^{M} P_k R_k + N\left(\prod_{k=1}^{M} R_k\right)$, whereas that for CCA and PCA is $R\left(\prod_{k=1}^{M} P_k\right) + NR$.

Figures 4, 5, and 6 plot RER obtained by estimating various reduced dimensions for the dataset, AT&T(ORL), Cropped AR, and MNIST dataset with each group size, respectively. Since the trends in Figures 4, 5, and 6 are almost the same, we will only mention Figure 4.

From Figure 4, we observe that the RER of MCCA is the smallest for any value of CR. This indicates that the MCCA performs better than the other methods. In addition, note that CCA performs better than MPCA only for fairly small values

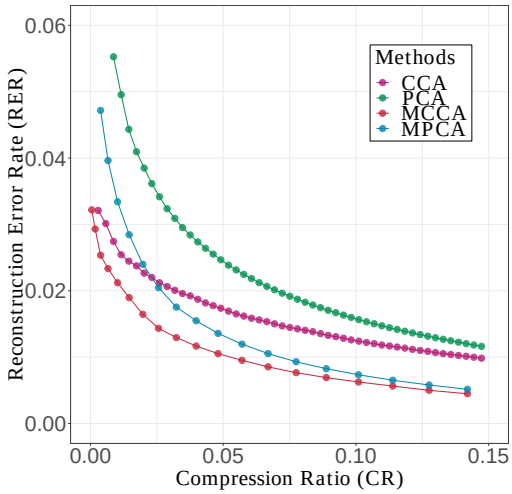
of CR, even though it is a method for vector data, whereas MPCA performs better for larger values of CR. This implies the limitations of CCA for vector data.

Next we cobsider group size by comparing (a), (b), and (c) in Figure 4. The value of CR at the intersection of CCA and MPCA increases with increasing the group size. This indicates that MPCA has more trouble extracting an appropriate latent space as the group size increases. Since MPCA does not consider the group structure, it is not possible to properly estimate the covariance structure when the group size is large.



(a)

(b)



(c)

Figure 4: Plots of RER versus CR for the AT&T(ORL) dataset of various group sizes: (a) Small; (b) Medium; and (c) Large.

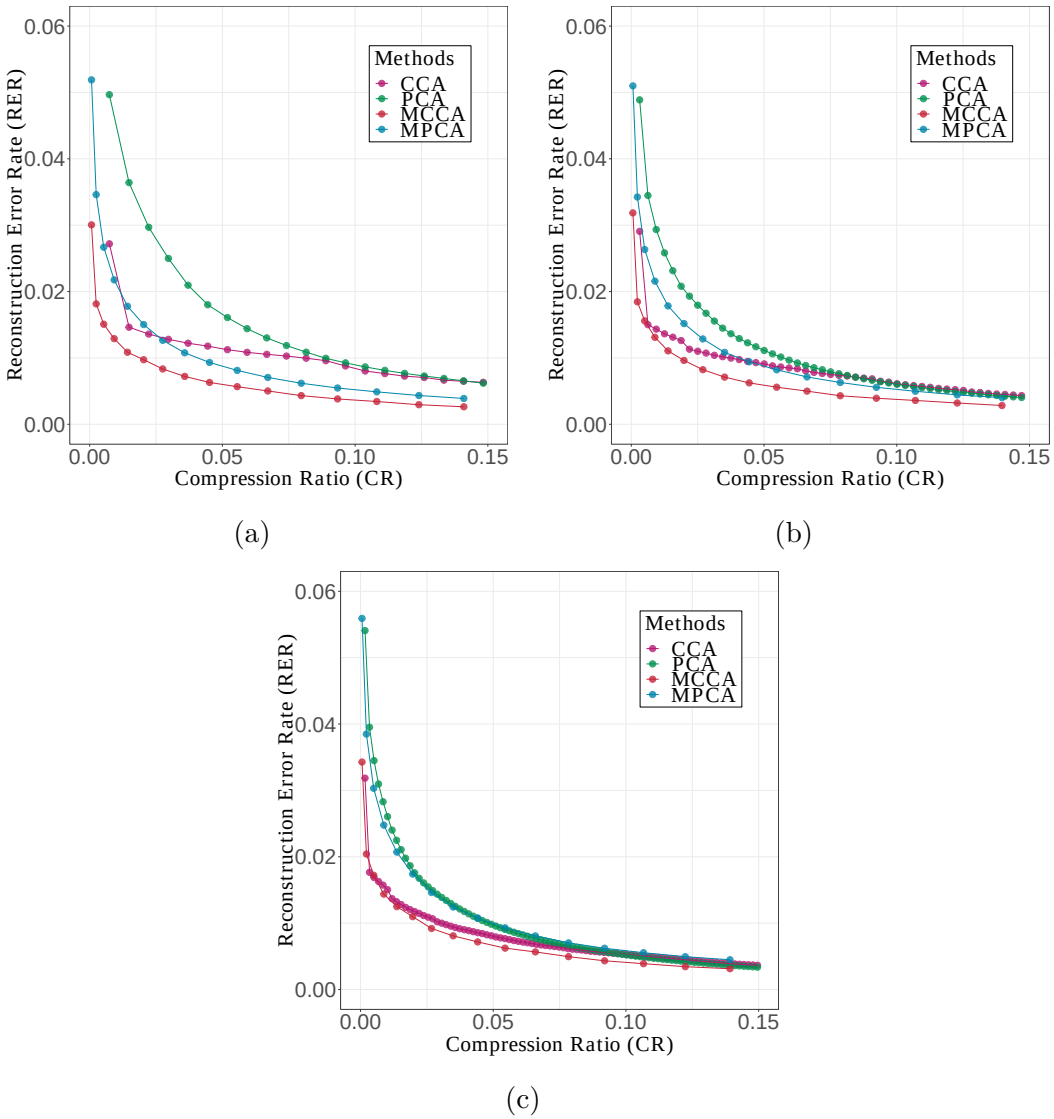


Figure 5: Plots of RER for the Cropped AR dataset for various group sizes: (a) Small; (b) Medium; and (c) Large.

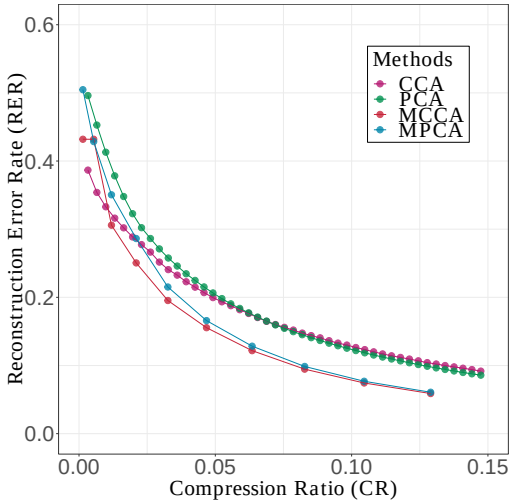


Figure 6: Plots of RER for the MNIST dataset.

7.3 Behavior of Contraction Ratio

We examined the behavior of contraction ratio $\alpha^{(k)}$. We performed MCCA on the AT&T(ORL) dataset with the medium group size and computed $\alpha^{(1)}$ and $\alpha^{(2)}$ with the various pairs of reduced dimensions $(R_1, R_2) \in \{1, 2, \ldots, 25\} \times \{1, 2, \ldots, 25\}$.

Figure 7 shows the values of $\alpha^{(1)}$ and $\alpha^{(2)}$ for all pairs of R_1 and R_2. As shown, $\alpha^{(1)}$ and $\alpha^{(2)}$ were invariant to variations in R_2 and R_1, respectively. Therefore, to facilitate visualization of changes in $\alpha^{(k)}$, Figure 8 shows $\alpha^{(1)}$ and $\alpha^{(2)}$ for, respectively, $R_2 = 1$ and $R_1 = 1$. From these, we observe that when both R_1 and R_2 are greater than 8, both $\alpha^{(1)}$ and $\alpha^{(2)}$ are close to one.

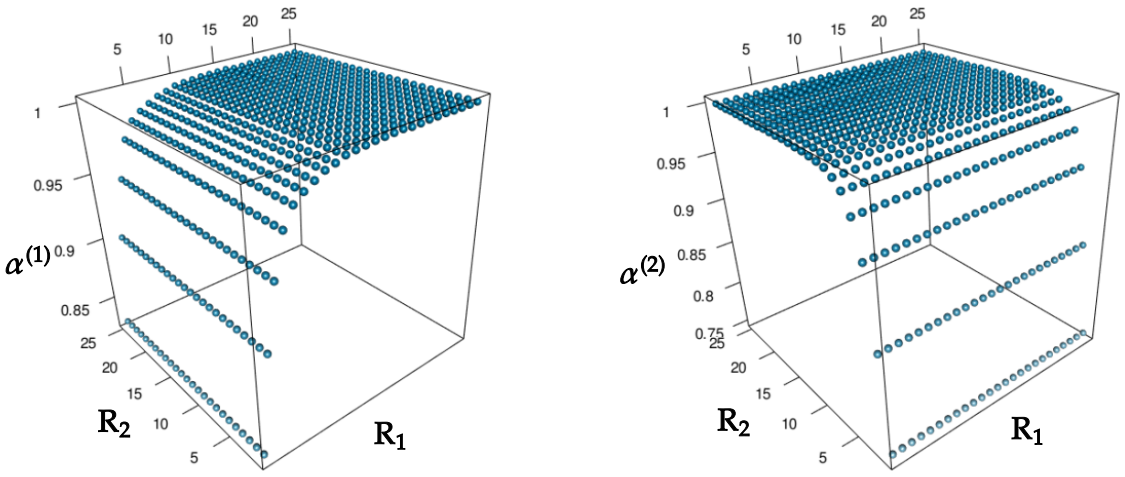


Figure 7: $\alpha^{(1)}$ and $\alpha^{(2)}$ versus pairs of reduced dimensions (R_1, R_2).

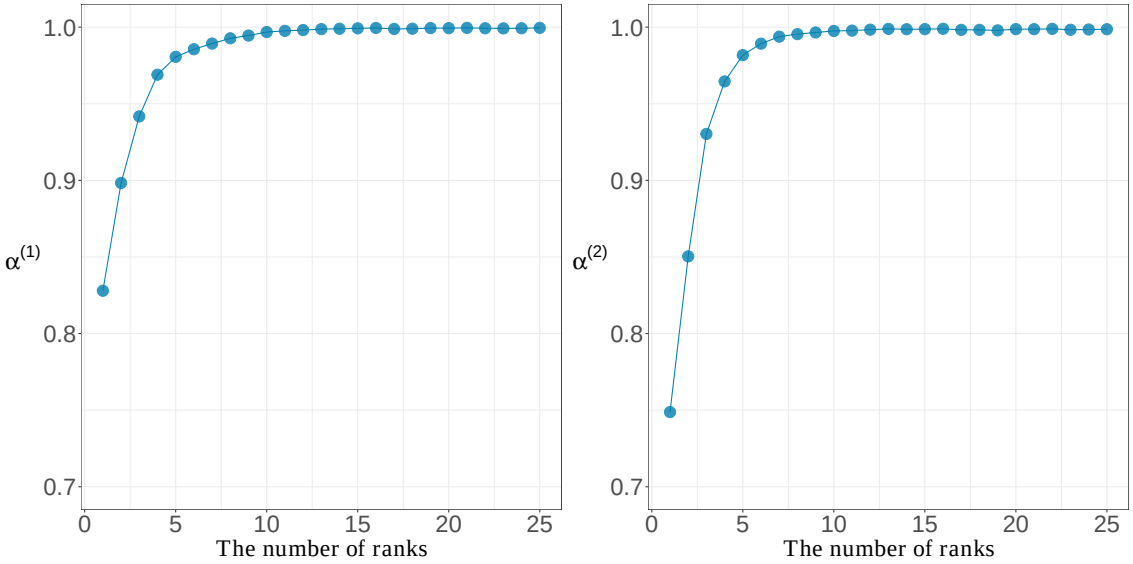


Figure 8: $\alpha^{(1)}$ and $\alpha^{(2)}$ versus R_1 and R_2, respectively.

7.4 Efficacy of Solving the Quadratic Programming Problem

We investigated the usefulness of determining the initial value of $\boldsymbol{w}^{(k)}$ by solving the quadratic programming problem (5.1). We applied MCCA to the AT&T(ORL) dataset with the small, medium, and large number of groups. In addition, we

also used the smaller group size of three. For determining the initial value of $\boldsymbol{w}^{(k)}$, we consider three methods: solving the quadratic programming problem (5.1) (MCCA:QP), setting all values of $\boldsymbol{w}^{(k)}$ to one (MCCA:FIX), and setting the values by random sampling according to the uniform distribution $U(0,1)$ (MCCA:RANDOM). We computed the $\alpha^{(k)}$ with the reduced dimensions $R_1 = R_2$ ($\in \{1, 2, \ldots, 10\}$) for each of these methods.

To evaluate the performance of these methods, we compared the values of $\alpha^{(k)}$ and the number of iterations in the estimation. The number of iterations in the estimation is the number of repetitions of lines 7 to 9 in Algorithm 1. For MCCA(RANDOM), we performed 50 trials and calculated averages of each of these indices.

Figure 9 shows the comparisons of $\alpha^{(1)}$ and $\alpha^{(2)}$ when the initialization was performed by MCCA:QP, MCCA:FIX, and MCCA:RANDOM for AT&T(ORL) dataset with a group size of 3. It was confirmed that MCCA:QP provides the largest values of $\alpha^{(1)}$ and $\alpha^{(2)}$. Figure 10 shows that the number of iterations. MCCA:QP gives the smallest number of iterations for almost all values of the reduced dimensions. This result indicates that MCCA:QP converges to a solution faster than the other initialization methods. However, when the reduced dimension is greater than or equal to 8, the other methods are competitive with MCCA:QP. A lack of difference in the number of iterations could result from the closeness of the initial values and the global optimal solution. Note that when the R_1 and R_2 are greater than or equal to 8, $\alpha^{(1)}$ and $\alpha^{(2)}$ are sufficiently close to one, based on Figure 9. This indicates that the initial values are close to the global optimal solution obtained from Theorem 1. Hence, the result shows almost the same numbers of iterations for the three methods.

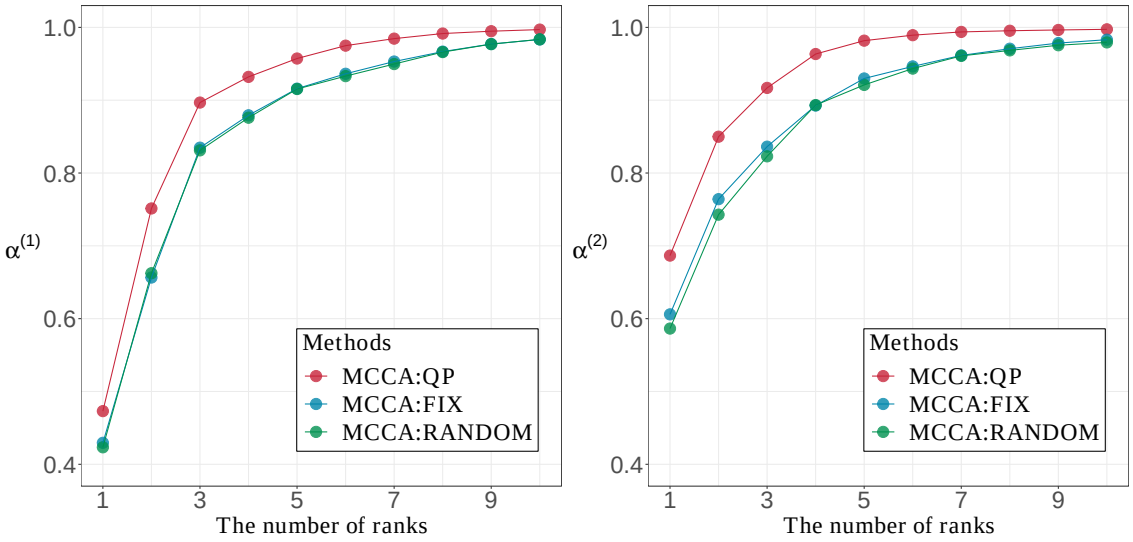


Figure 9: Comparisons of $\alpha^{(1)}$ and $\alpha^{(2)}$ computed by using the initial values obtained from the initializations MCCA:QP, MCCA:FIX, and MCCA:RANDOM with the AT&T(ORL) dataset for a group size of 3.

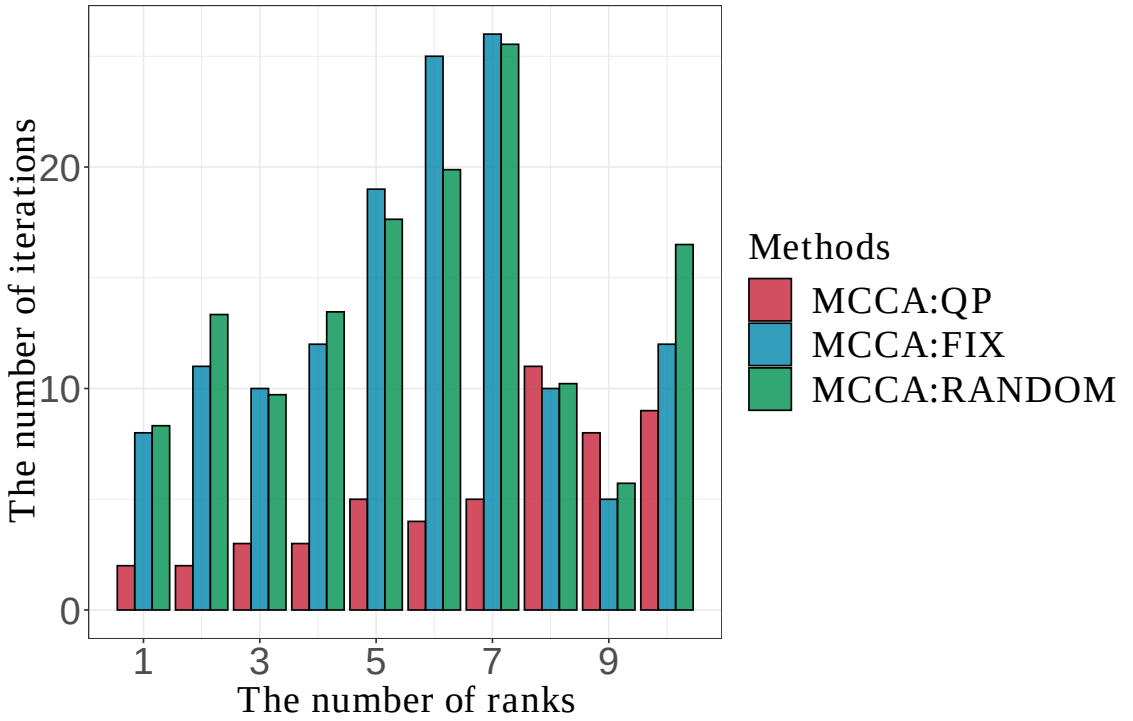


Figure 10: Comparison of the number of iterations when the initialization was performed by MCCA:QP, MCCA:FIX, and MCCA:RANDOM with the AT&T(ORL) dataset for a group size of 3.

Figures 11 to 16 show comparisons for the AT&T(ORL) dataset with the other group size. Figure 11, 13, and 15 show results similar those in Figure 9, whereas Figure 12, 14, and 16 show competitive performances for all reduced dimensions.

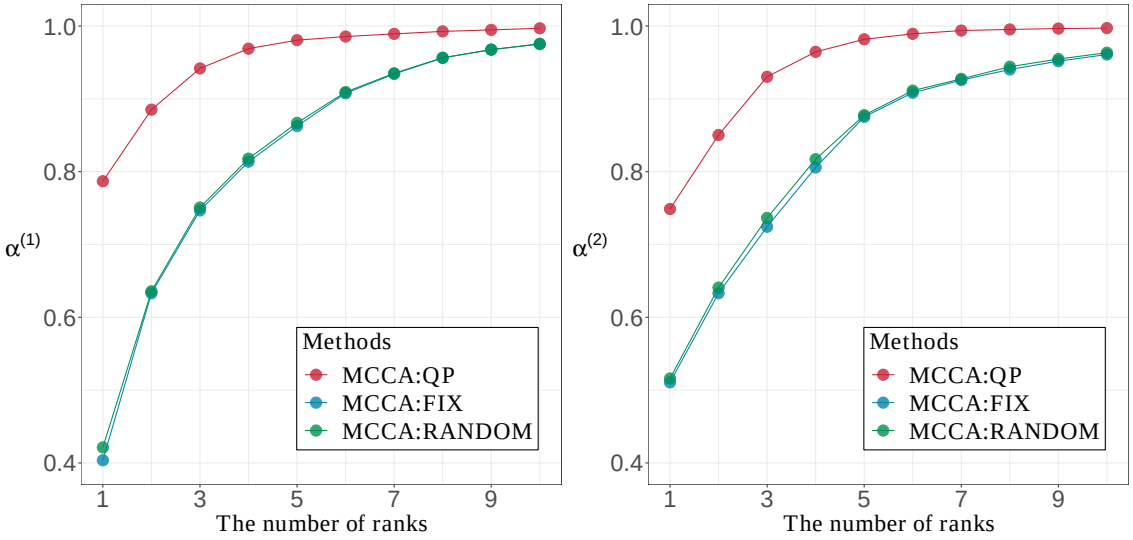


Figure 11: Comparisons of $\alpha^{(1)}$ and $\alpha^{(2)}$ computed using the initial values obtained from the initializations MCCA:QP, MCCA:FIX, and MCCA:RANDOM with the AT&T(ORL) dataset for the small group size.

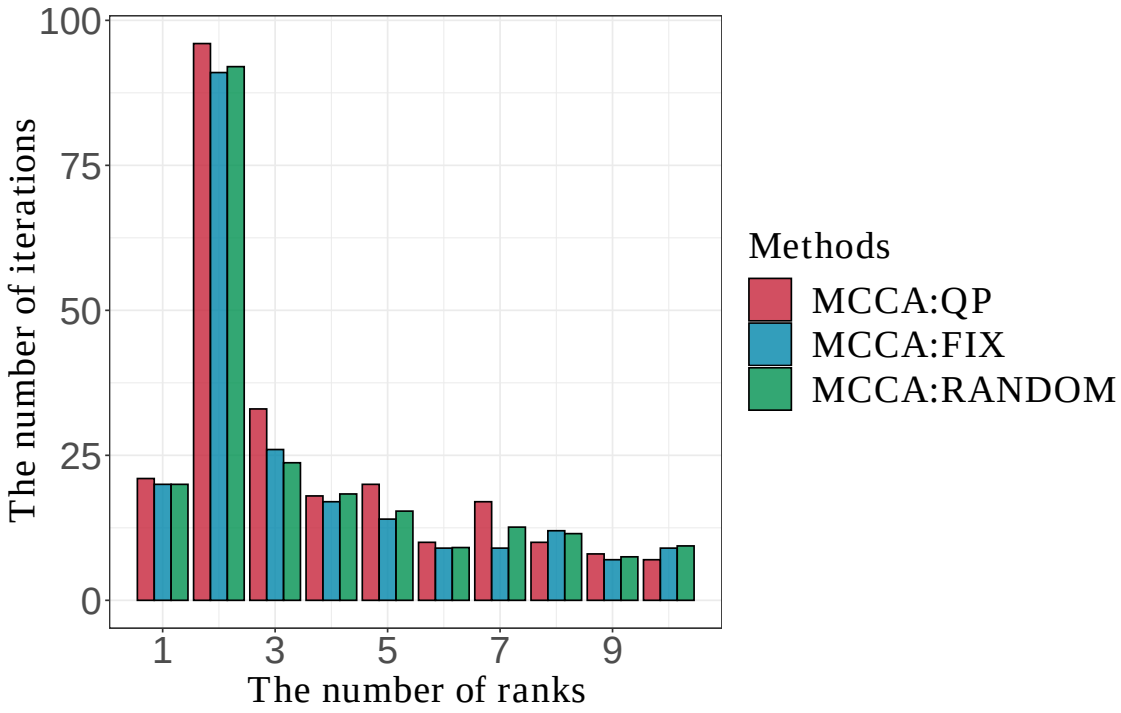


Figure 12: Comparison of the number of iterations when the initializations were performed by MCCA:QP, MCCA:FIX, and MCCA:RANDOM with the AT&T(ORL) dataset for the small group size.

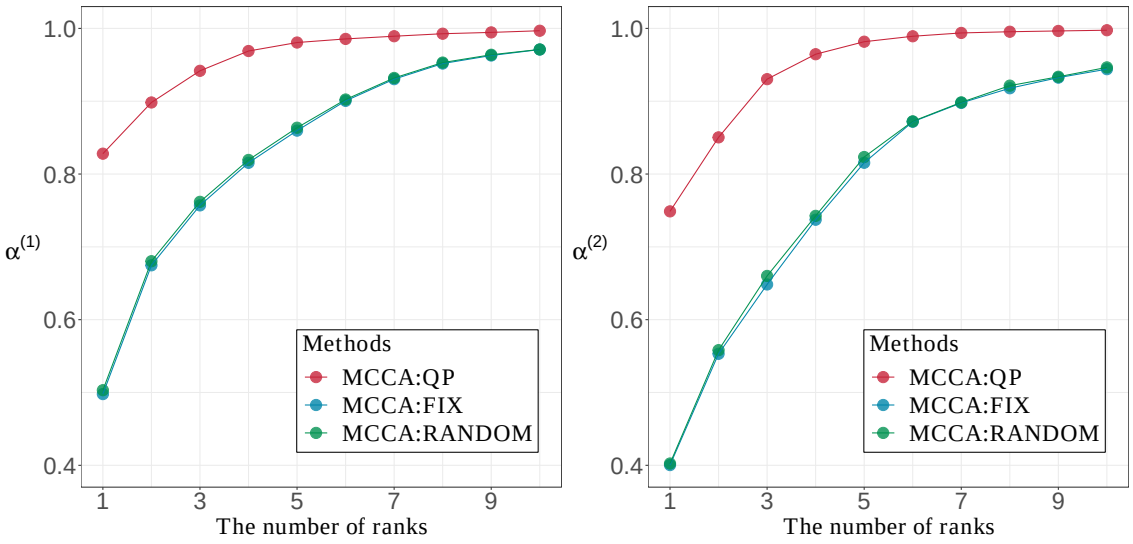


Figure 13: Comparisons of $\alpha^{(1)}$ and $\alpha^{(2)}$ computed using the initial values obtained from the initialization of MCCA:QP, MCCA:FIX, and MCCA:RANDOM with the AT&T(ORL) dataset and the medium group size.

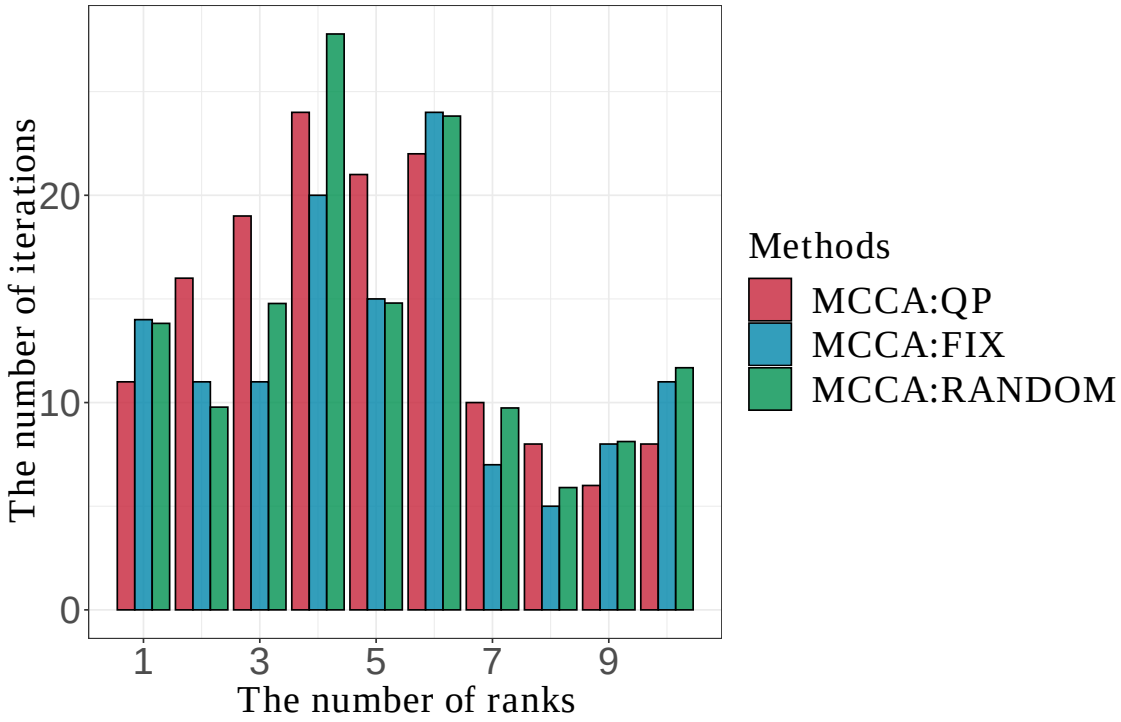


Figure 14: Comparison of the number of iterations when the initialization was perfomed by MCCA:QP, MCCA:FIX, and MCCA:RANDOM with the AT&T(ORL) dataset and the medium group size.

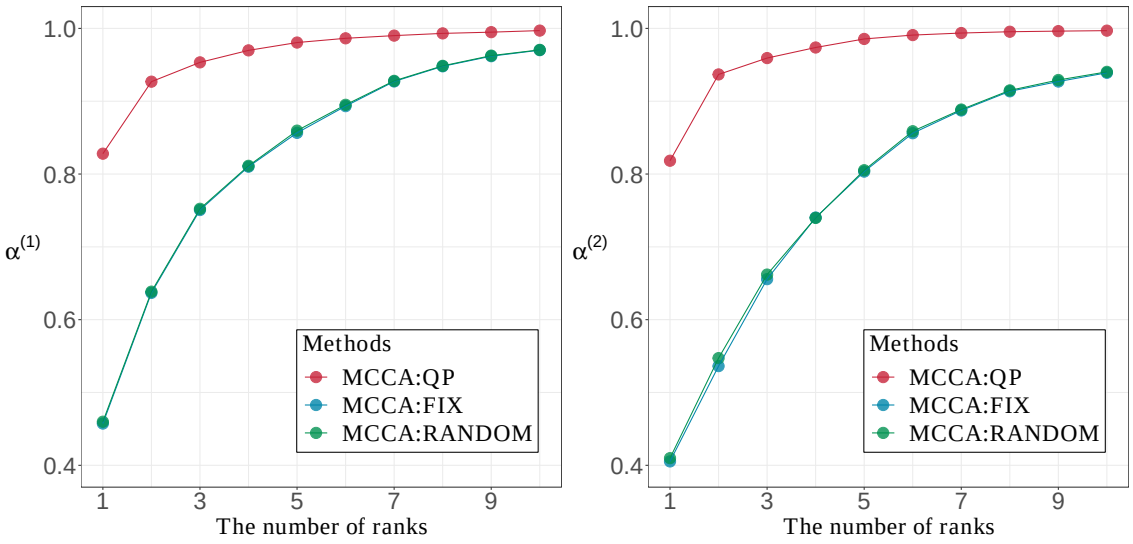


Figure 15: Comparisons of $\alpha^{(1)}$ and $\alpha^{(2)}$ computed using the initial values obtained from the initializations MCCA:QP, MCCA:FIX, and MCCA:RANDOM with the AT&T(ORL) dataset for the large group size.

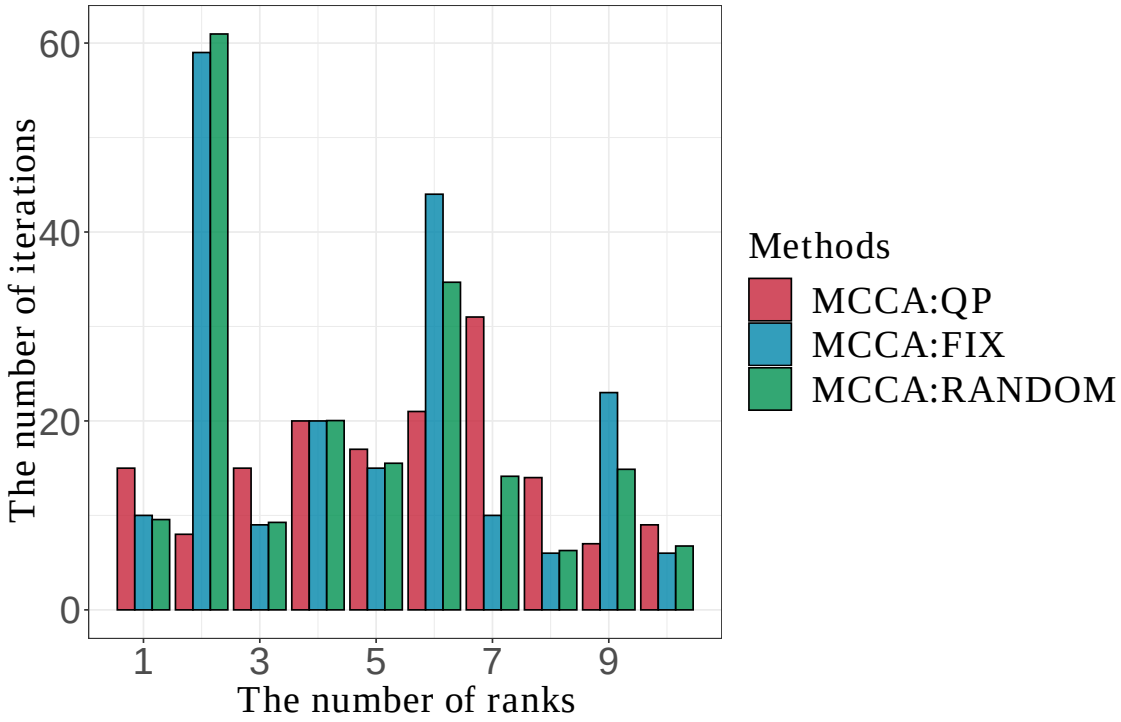


Figure 16: Comparison of the number of iterations when the initializations were performed by MCCA:QP, MCCA:FIX, and MCCA:RANDOM with the AT&T(ORL) dataset for the large group size.

8 Concluding Remarks

We have developed the multilinear common components analysis (MCCA) by introducing a covariance structure based on the Kronecker product. To efficiently solve the non-convex optimization problem for MCCA, we have proposed an iteratively updating algorithm. The proposed algorithm exhibits some superior theoretical convergence properties. Numerical experiments showed the usefulness of MCCA.

Specifically, MCCA was shown to be competitive among the initialization methods in terms of number of iterations. As the number of groups increases, the overall number of samples increases. This may be why all methods required almost the same number of iterations for small, medium, and large number of groups.

Note that, in this study, we used the Kronecker product representation to estimate the covariance matrix for tensor datasets. Greenewald et al. (2019) used the Kronecker sum representation for estimating the covariance matrix, and it would be interesting to extend the MCCA to this and other covariance representations.

Acknowledgments

I would like to express my gratitude to my advisor, Associate Professor Shuichi Kawano, for his careful and enthusiastic education. For three years, he has taught me not only the way of conducting research, but the way of making presentations and writing papers, and other knowledge necessary as a researcher. He also gave me the opportunities to present my research at many conferences and conduct collaborative research. I am grateful for the rich research experience through all the opportunities. Finally, I would like to thank all the members of Ueno, Kawano, Nishiyama, and Uto laboratories for their meaningful discussions and advice.

References

Allen, G. (2012). Sparse higher-order principal components analysis. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, 27–36.

Badeau, R. and Boyer, R. (2008). Fast multilinear singular value decomposition for structured tensors. *SIAM Journal on Matrix Analysis and Applications*, **30**(3), 1008–1021.

Bensmail, H. and Celeux, G. (1996). Regularized gaussian discriminant analysis through eigenvalue decomposition. *Journal of the American Statistical Association*, **91**(436), 1743–1748.

Boik, R. J. (2002). Spectral models for covariance matrices. *Biometrika*, **89**(1), 159–182.

Carroll, J. D. and Chang, J.-J. (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition. *Psychometrika*, **35**(3), 283–319.

Flury, B. N. (1984). Common principal components in k groups. *Journal of the American Statistical Association*, **79**(388), 892–898.

Flury, B. N. (1986). Asymptotic theory for common principal component analysis. *The Annals of Statistics*, **14**(2), 418–430.

Flury, B. N. (1988). *Common principal components & related multivariate models*. John Wiley & Sons, Inc.

Flury, B. N. and Gautschi, W. (1986). An algorithm for simultaneous orthogonal transformation of several positive definite symmetric matrices to nearly diagonal form. *SIAM Journal on Scientific and Statistical Computing*, **7**(1), 169–184.

Greenewald, K., Zhou, S., and Hero III, A. (2019). Tensor graphical lasso (Teralasso). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **81**(5), 901–931.

Harshman, R. A. (1970). Foundations of the PARAFAC procedure : Models and conditions for an "explanatory" multimodal factor analysis. *UCLA Working Papers in Phonetics*, **16**(1), 84.

Harville, D. A. (1998). *Matrix Algebra From a Statistician's Perspective*. Springer-Verlag, New York.

Jolliffe, I. (2002). *Principal Component Analysis*. Springer-Verlag, New York.

Kermoal, J. P., Schumacher, L., Pedersen, K. I., Mogensen, P. E., and Frederiksen, F. (2002). A stochastic mimo radio channel model with experimental validation. *IEEE Journal on Selected Areas in Communications*, **20**(6), 1211–1226.

Kiers, H. A. (2000). Towards a standardized notation and terminology in multiway analysis. *Journal of Chemometrics: A Journal of the Chemometrics Society*, **14** (3), 105–122.

Kolda, T. G. (2006). Multilinear operators for higher-order decompositions (No. SAND 2006-2081). *Sandia National Laboratories*.

Kolda, T. G. and Bader, B. W. (2009). Tensor decompositions and applications. *SIAM review*, **51**(3), 455–500.

Lai, Z., Xu, Y., Chen, Q., Yang, J., and Zhang, D.Oct . (2014). Multilinear sparse principal component analysis. *IEEE Transactions on Neural Networks and Learning Systems*, **25**(10), 1942–1950.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, **86**(11), 2278–2324.

Lu, H., Plataniotis, K. N., and Venetsanopoulos, A. N. (2008). MPCA: Multilinear principal component analysis of tensor objects. *IEEE transactions on Neural Networks*, **19**(1), 18–39.

Manly, B. F. J. and Rayner, J. C. W. (1987). The comparison of sample covariance matrices using likelihood ratio tests. *Biometrika*, **74**(4), 841–847.

Martinez, A. and Benavente., R. (1998). The AR face database. *CVC Technical Report*, **24**.

Martinez, A. M. and Kak, A. C. (2001). PCA versus LDA. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **23**(2), 228–233.

Park, H. and Konishi, S. (2018). Sparse common component analysis for multiple high-dimensional datasets via noncentered principal component analysis. *Statistical Papers*, **61**(6), 2283–2311.

Pearson, K. (1901). LIII. 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.

Pourahmadi, M., Daniels, M. J., and Park, T. (2007). Simultaneous modelling of the cholesky decomposition of several covariance matrices. *Journal of Multivariate Analysis*, **98**(3), 568–587.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, (2019).

Wang, H., Banerjee, A., and Boley, D. (2011). Common component analysis for multiple covariance matrices. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 956–964.

Wang, S., Sun, M., Chen, Y., Pang, E., and Zhou, C.Nov . (2012). STPCA: Sparse tensor Principal Component Analysis for feature extraction. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, 2278–2281.

Werner, K., Jansson, M., and Stoica, P. (2008). On estimation of covariance matrices with kronecker product structure. *IEEE Transactions on Signal Processing*, **56** (2), 478–491.

Yu, K., Bengtsson, M., Ottersten, B., McNamara, D., Karlsson, P., and Beach, M. (2004). Modeling of wide-band mimo radio channels based on nlos indoor measurements. *IEEE Transactions on Vehicular Technology*, **53**(3), 655–665.