이학석사학위논문

Large-sample test for joint signal extraction in multiblock data

다중원천 데이터의 공통 구조 도출을 위한 다표본 통계 검정

2020 년 2 월

서울대학교 대학원

통계학과

한 상 일

이학석사학위논문

Large-sample test for joint signal extraction in multiblock data

다중원천 데이터의 공통 구조 도출을 위한 다표본 통계 검정

2020 년 2 월

서울대학교 대학원

통계학과

한 상 일

Large-sample test for joint signal extraction in multiblock data

다중원천 데이터의 공통 구조 도출을 위한 다표본 통계 검정

지도교수 정 성 규

이 논문을 이학석사 학위논문으로 제출함

2019 년 12 월

서울대학교 대학원

통계학과

한 상 일

한상일의 이학석사 학위논문을 인준함

2019 년 12 월

위 원 장 오희석 (인)

부위원장 정성규 (인)

위 원 원중호 (인)

Abstract

This thesis is about a test of the existence of a joint structure in a joint-individual model, which is proposed by Lock, et al[Ann. Appl. Stat.,7(2013), 523]. in blocks of data obtained from multi-source data. For two data blocks, assume that each data blocks is a matrix of size $(p \times n), (q \times n)$ and written as $\mathbb{X} = vZ + E$ with a significant random row vector signal Z with rank 1 represented by a loading vector v, and a noise matrix, E. While Feng, et al.[J.Multivar.Anal.,166(2018),241-265] define the distance of signals as the principal angle, we define the measure of closeness of signals as the square of cosine principal angle. Assuming that each sample of a significant signals follows *iid* bivariate normal, a test of the joint signal rank selection is proposed using the fact that the closeness measure of the significant signals follows Beta($\frac{1}{2}$, $\frac{n-1}{2}$). In addition, we estimate unobservable v and significant signals Z by using the fact that from the form of the data block, v is the maximal eigenvector of the covariance matrix, $Var(X_1)$, where X_1 is the first column of $\mathbb{X}$. Replacing the unobservable Z by the predictions $\hat{Z}$, we develop a large-sample test procedure for the rank of the joint signal. In the simulation study, this asymptotic test is compared to the joint structure rank selection process of Feng, et al.[J.Multivar.Anal.,166(2018),241-265], called AJIVE. Simulation results show that this asymptotic test is similar to AJIVE's joint rank selection in performance. Under the assumption that significant signals follow the normal distribution, the asymptotic test is expected to be a good alternative to the AJIVE rank selection.

Keywords: Multisource data, principal angle, joint-individual structure

Student Number: 2018-20856

Contents

Chapter 1

Introduction

One of the major challenges of today is data integration. In this big data era, the volume of information is getting bigger and the variety and complexity is increasing accordingly. Numerous blocks of data have different types of different features and samples from different sources. Therefore, data integration is a field of research worthy of the big data era in that it can process these blocks of data and consider the information contained in them simultaneously. Currently, various methods that reasonably integrate different types of data blocks are proposed. Lock, et al. [1] proposed a joint-individual structure model, JIVE model, as one of the methods for data integration. In the data blocks with different features observed in the same subjects, the joint-individual model assumes that the concatenated data blocks consist of an joint structure shared by all blocks, the individual structures unique to each block, and noise. They proposed to estimate the rank of the joint-individual structures through permutation tests for the singular values of the concatenated data blocks and each of blocks. In addition, Feng, et al.[2] proposed AJIVE

model, a method of extracting significant signal (described above as 'structure') and noise from the concatenated blocks of data by using scree plots of singular values. In addition, they proposed the distance between the row spaces of significant signals as the principal angle and extract joint signal and individual signal using Wedin upper bound[3] of the principal angle and empirical quantile obtained by simulating principal angles in a random direction space. While JIVE and AJIVE models derived the joint-individual structure in an intuitive and reasonable process, the distance between the space of the two significant signals was not accurately measured in that they use the permutation test or the upper bound. In this thesis, assuming a proper probability distribution for row rank 1 significant signals from the two data blocks, we will investigate the exact and asymptotic distribution of the principal angle. In addition, we will propose a test that determine whether the significant signal of the two data blocks is a joint signal or not, and compare the test with AJIVE's joint signal selection process.

Chapter 2

Notation and abbreviation

Following notations and abbreviations are used in the present work.

$\mathbb{R}^{m,n}$	matrix with m rows, n columns
I_p	identity matrix with p rows, p columns
$N_p(\mu, \Sigma)$	p-multivariate normal distribution with mean μ, covariance matrix Σ
$\mathcal{L}(X)$	distribution of random vector or random matrix X
$O(n)$	orthogonal group, which is satisfying for $R \in \mathbb{R}^{n,n}$ if $R \in O(n)$ then $R^t R = I_n$
$Sp_n(\sigma^2)$	n-dimensional spherical distribution family with covariance matrix $\sigma^2 I_n$.
$\mathbb{E}(X), Var(X)$	expectation and covariance matrix of X
$\perp$	statiscally independent
$cor(z, w)$	correlation coefficient of random variable z and w
$Beta(\alpha, \beta)$	beta distribution with parameter α, β
S^{n-1}	unit sphere on $\mathbb{R}^n$
$Unif(S^{n-1})$	uniform distribution on vector space S^{n-1}
$\|\cdot\|$	2-norm
$\|\cdot\|$	determinant of matrix
$diag(a_1, \ldots, a_m)$	diagonal matrix with diagonal element $a_1, \ldots, a_m$
$O_p(a_n)$	big O notation..... with probability... matrix, vector every element
$\xrightarrow{p}$	conver...
$Gamma(\alpha, \beta)$	dd
mgf	moment generating function
iid	Independent and identically distributed

Chapter 3

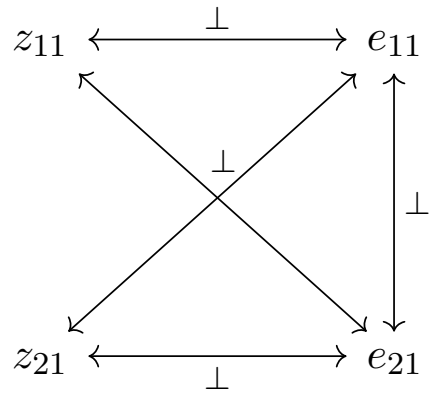
Model of data block

The two data blocks $\mathbb{X}_1, \mathbb{X}_2$ are matrices of size $(p \times n), (q \times n)$. The number of rows p, q of the block matrices is the number of variables each block has, and the number of columns n is the number of samples. We can understand that p and q features are measured from two different sources respectively for the same n subjects. The block data can be expressed as an additive model consisting of a significant signal with rank 1 and a noise as shown below.

$$\begin{aligned}
\mathbb{X}_1 &= (X_{11}, \ldots, X_{1n}) = v_1 Z_1 + E_1 & \mathbb{X}_2 &= (X_{21}, \ldots, X_{2n}) = v_2 Z_2 + E_2 \\
&= v_1(z_{11}, \ldots, z_{1n}) + (e_{11}, \ldots, e_{1n}) & &= v_2(z_{21}, \ldots, z_{2n}) + (e_{21}, \ldots, e_{2n}) \\
Z_1^t &\sim N_n(0, \sigma_1^2 I_n), e_{1i} \overset{iid}{\sim} N_p(0, \tau_1^2 I_p) & Z_2^t &\sim N_n(0, \sigma_2^2 I_n), e_{2i} \overset{iid}{\sim} N_q(0, \tau_2^2 I_q) \\
\|v_1\| &= 1 & \|v_2\| &= 1
\end{aligned}$$

$X_{11}, \ldots, X_{1n}$ are the columns of $\mathbb{X}_1$ and can be understood as *iid* sampled n random vectors and so is $\mathbb{X}_2$. Z_1 and Z_2 are significant n dimensional row signals with rank 1 and respectively follow the normal distributions $N_n(0, \sigma_1^2 I_n)$, $N_n(0, \sigma_2^2 I_n)$. These significant signals Z_1 and Z_2 are respectively represented

by a loading vector $v_1 \in R^{p,1}$ and $v_2 \in R^{q,1}$. E_1 and E_2 are noise signals. Let every element of E_1 have an *iid* normal distribution $N_1(0, \tau_1^2)$. Similarly, every element of E_2 also have an *iid* normal distribution. The above model shows that n independent columns of $\mathbb{X}_1$ are represented with n independent signals $Z_1 = (z_{11}, \ldots, z_{1n})$ by the loading vector of v_1. Since all signals of $\mathbb{X}_1$ and $\mathbb{X}_2$ are column-wise *iid*, the independence of the significant signals and the noise signals comes from the independence of $(z_{11}, z_{21}, e_{11}, e_{21})$. The following figure is not exact but makes it easy to understand the independence of the signals.



The above independence symbol '$\perp$' does not mean pairwise independence. The noise signal is completely independent of other signals, and each significant signals are completely independent of the noise signal, i.e., $\big(e_{11}, e_{21}, (z_{11}, z_{21})\big)$ are independent,

Chapter 4

Joint and individual signals extraction

Our goal is to determine if two data blocks have a joint signal or not. More specifically, we need a test to determine how similar the significant signals Z_1 and Z_2 are. The following two scenarios are available.

(1) Z_1, Z_2 are close

(2) Z_1, Z_2 are not close

We need 'distance' between Z_1,Z_2. Since each data block comes from a different source, it is necessary to define a distance that is invariant to normalization. Feng, et al.[2] proposed the principal angle as distance, which is invariant to normalization. We will transform the principal angle to make it easier to handle.

4.1 Principal angle in n-dim vectors Z_1, Z_2

Let $Z_1, Z_2 \in \mathbb{R}^{1\times n}$. The principal angle θ between Z_1 and Z_2 means the angle between two lines spanned Z_1 and Z_2 respectively. Since we want to regard 0 and 180° as close, 90° as far, we want to consider the $cos^2\theta$ as the measure of closeness instead of angle θ. Define the test statistic T for the principal angles θ of Z_1 and Z_2 as follows.

$$T \overset{def}{=} cos^2\theta = \frac{(Z_1 Z_2^t)^2}{Z_1 {Z_1}^t \cdot Z_2 {Z_2}^t}$$

Since $(z_{1i}, z_{2i})_{i=1,\dots,n}$ samples are *iid*, T can be regarded as a consistent estimate of the square of the correlation coefficient of (z_{11}, z_{21}). Therefore, we can replace the above two scenarios with the null and alternative hypotheses about the correlation coefficient ρ of (z_{11}, z_{21}) as follows.

(H_0) $\rho^2 = 0$ (since Z_1, Z_2 are normal, $Z_1 \perp Z_2$)

(H_1) $\rho^2 > 0$ (since Z_1, Z_2 are normal, $Z_1 \not\perp Z_2$)

What is interesting is that the closeness between the two significant signals is not described as independence, but as correlation, which means linear relationships. However, since significant signals follow a normal distribution, the assumption that the significant signals are uncorrelated replaced with the assumption of independence of the significant signals, and the distribution of the test statistic T can be obtained as follows.

Theorem 1. *Let* $(z_{1i}, z_{2i})_{i=1,\dots,n}$ *be iid random samples from bivariate normal distribution,* $N_2(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix})$ *and* $Z_1 = (z_{11}, \dots, z_{1n}), Z_2 = (z_{21}, \dots, z_{2n})$. *If* $\rho = cor(z_{11}, z_{21}) = 0$, *then* $T = cos^2\theta$ *with principal angle* θ *between* Z_1, Z_2 *has the distribution Beta*$(\frac{1}{2}, \frac{n-1}{2})$

The following two lemmas are required to prove the Theorem 1.

Lemma 2. *[4] For $n \in \mathbb{N}, \sigma^2 > 0$, if $Z = (Z_1, \ldots, Z_n)^t \sim F_Z, F_Z \in Sp_n(\sigma^2)$ and $P(Z = 0) = 0$, then*

$$\frac{Z}{\sqrt{Z_1^2 + \ldots + Z_n^2}} = \frac{Z}{\|Z\|} \sim \ Unif(S^{n-1})$$

Proof. Since $\frac{Z}{\|Z\|}$ is almost surely random vector on S^{n-1}, we only need to show that for any $R \in O(n)$,

$$\mathcal{L}(\frac{Z}{\|Z\|}) = \mathcal{L}(R\frac{Z}{\|Z\|})$$

Note that since $F_Z \sim Sp_n(\sigma^2)$, so $\mathcal{L}(RZ) = \mathcal{L}(Z)$. Hence

$$\begin{aligned}\mathcal{L}(R\frac{Z}{\|Z\|}) = \mathcal{L}(\frac{RZ}{\sqrt{Z^tZ}}) &= \mathcal{L}(\frac{RZ}{\sqrt{Z^tR^tRZ}}) \\ &= \mathcal{L}(\frac{RZ}{\|RZ\|}) = \mathcal{L}(\frac{Z}{\|Z\|})\end{aligned}$$

□

This lemma tells us that the test statistic T is expressed as a product of two random vectors following the uniform distribution on unit sphere.

Lemma 3. *Let X, Y be independent random vectors with uniform distribution on the unit sphere, then $(X^tY)^2$ has the cumulative distribution function F satisfying $F(t) = P((X^tY)^2 \leq t) = B(t)$ where $B(t)$ is cumulative distribution function of $Beta(\frac{1}{2}, \frac{n-1}{2})$.*

Proof. Since $P((X^tY)^2 > t) = \int P((x^tY)^2 > t|X = x)dP_X(x)$, enough to show $P(x^tY > t|X = x)$ does not depend x, and has cdf $B(t)$. By using **lemma 2**, $\exists U = (U_1, \ldots, U_n)^t \overset{\text{iid}}{\sim} N_n(0, I_n)$ s.t $\mathcal{L}(\frac{U}{\|U\|}) = \mathcal{L}(Y)$. For fixed

$x \in S^{n-1}$, $\exists R_x \in O(n)$ such that $R_x x = (1, 0, \ldots, 0)^t = e_1 \in S^{n-1}$. hence

$$\begin{aligned} x^t Y|_{X=x} &= x^t Y \qquad\qquad (\because X \perp Y) \\ &= x^t R_x^t R_x Y \\ &= e_1^t R_x Y \\ &\overset{d}{=} \frac{e_1^t R_x U}{\sqrt{U^t R_x^t R_x U}} \end{aligned}$$

Define $Z = (Z_1, \ldots, Z_n)^t = R_x U$, then Z follows the distribution $\sim N_n(0, I_n)$ and does not depend on x since the normal distribution $N_n(0, I_n)$ belongs to the $Sp_n(1)$. Hence,

$$P((x^t Y)^2 > t) = P((\frac{e_1^t Z}{\sqrt{Z^t Z}})^2 > t) = P(\frac{Z_1^2}{Z^t Z} > t)$$

Note that $Z_i^2 \overset{\text{iid}}{\sim} Gamma(1/2, 2)$. Therefore,

$$P((x^t Y)^2 > t) = P(\frac{Z_1^2}{Z^t Z} > t) = 1 - B(t)$$

□

Proof of Theorem 1 can be easily summarized by substituting Z_1 and Z_2 into Lemmas 2 and 3. First, since each Z_1, Z_2 follows the normal distribution, which belongs to the Sp_n and have zero measure on the origin of $\mathbb{R}^n$, by substituting Z_1, Z_2 for the Lemma 2, the test statistic T can be expressed as the product of two random vectors $\frac{Z_1}{\|Z_1\|}, \frac{Z_2}{\|Z_2\|}$ uniformly distributed on the unit sphere. In addition, they are independent because they are a function of the uncorrelated normal distribution, i.e. independent, Z_1, Z_2. Hence, by plugging $\frac{Z_1}{\|Z_1\|}, \frac{Z_2}{\|Z_2\|}$ into Lemma 3, The test statistic T follows the Beta($\frac{1}{2}$,$\frac{n-1}{2}$) distribution. Thus, if we know exactly two significant signals Z_1 and Z_2, we can test as follow.

Principal angle test for Z_1, Z_2

Let $(z_{1i}, z_{2i})_{i=1,\ldots,n}$ be iid random samples from bivariate normal distribution, $N_2(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix})$. For significant signals $Z_1 = (z_{11}, \ldots, z_{1n}), Z_2 = (z_{21}, \ldots, z_{2n})$ in data blocks, define $T = \dfrac{(Z_1 Z_2^t)^2}{Z_1 {Z_1}^t \cdot Z_2 {Z_2}^t}$, then for two hypothesis,

$$H_0 : \rho^2 = 0 \qquad\qquad H_1 : \rho^2 > 0$$

reject H_0 if $T > B_\alpha(\frac{1}{2}, \frac{n-1}{2})$ where $B_\alpha(\frac{1}{2}, \frac{n-1}{2})$ is the $1-\alpha$ quantile of the beta distribution Beta$(\frac{1}{2}, \frac{n-1}{2})$ with a significant level α.

Chapter 5

Significant signals extraction and test for $\hat{Z}_1, \hat{Z}_2$

Although we have the test for the principal angle between Z_1, Z_2, we can't not know the exact value of Z_1, Z_2 due to noise. In this section, it would be introduced estimation of significant signals Z_1, Z_2, noted $\hat{Z}_1, \hat{Z}_2$. Moreover, we will define a statistic called $\hat{T}$ with $\hat{Z}_1, \hat{Z}_2$ substituted forZ_1, Z_2 in statistics T, and investigate the asymptotic probability properties of $\hat{T}$.

5.1 Estimation of the significant signal

Let $\mathbb{X}_1 = (X_{11}, \ldots, X_{1n}) = v_1 Z_1 + E_1$ be a data block with $(p \times n)$-dimension in the above model structure. The data block $\mathbb{X}_1$ is formed by column-wise stacking *iid* n samples. Hence the role of v_1 is the same in n samples. Thus, we want to examine the role of v_1 in the first column X_{11} we can observe.

Observation 1. v_1 is maximal eigenvector of $Var(X_{11})$

Let $\Sigma_1 = Var(X_{11})$ and $u_2, \ldots, u_p$ be orthonormal vectors where

$U = (v_1, u_2, \ldots, u_p)$ are basis for $\mathbb{R}^p$. Then $\Sigma_1 = \sigma_1^2 v_1 v_1^t + \tau_1^2 I_p = U\Lambda U^t$, where $\Lambda = diag(\sigma^2 + \tau^2, \tau^2, \ldots, \tau^2)$. Moreover,

$$\begin{aligned} |tI_p - \Sigma| = |U(tI - \Lambda)U^t| &= |U||tI - \Lambda||U^t| \\ &= |tI - \Lambda| \\ &= (t - (\sigma^2 + \tau^2))(t - \tau^2)^{p-1} \end{aligned}$$

Thus it is obvious that v is the maximal eigenvector with eigenvalue $\sigma^2 + \tau^2$

The key point of the estimation is that v_1 and v_2 become the maximum eigenvector of $Var(X_1)$ and $Var(X_2)$, respectively. Therefore, it is natural to estimate v_1 and v_2 as the maximal eigenvectors of sample covariance matrix $\hat{\Sigma}_1 = \mathbb{X}_1\mathbb{X}_1^t/n$ and $\hat{\Sigma}_2 = \mathbb{X}_2\mathbb{X}_2^t/n$, which converge to the maximal eigenvectors v_1, v_2. Under assumption that the influence of the noise signals is very small compared to the significant signals, if we know the the estimate $\hat{v}_1, \hat{v}_2$ close to true value v_1, v_2, the estimate for Z_1, Z_2 noted $\hat{Z}_1, \hat{Z}_2$ is naturally induced as follow,

$$\hat{Z}_1 = \hat{v}_1^t \mathbb{X}_1 = \hat{v}_1^t v_1 Z_1 + \hat{v}_1^t E_1 \approx v_1^t v_1 Z_1 + v_1^t E_1 \approx Z_1$$

Briefly, each $\hat{v}_1, \hat{v}_2$ is the maximal eigenvector of $\mathbb{X}_1\mathbb{X}_1^t, \mathbb{X}_2\mathbb{X}_2^t$, and $\hat{Z}_1 = \hat{v}_1^t\mathbb{X}_1, \hat{Z}_2 = \hat{v}_2^t\mathbb{X}_2$.

5.2 Test for principal angle $\hat{Z}_1, \hat{Z}_2$

For the data block $\mathbb{X}_1, \mathbb{X}_2$, we only can know $\hat{Z}_1, \hat{Z}_2$, not Z_1, Z_2. Let $\hat{T}$ be the square of the cosine of principal angle between $\hat{Z}_1, \hat{Z}_2$, i.e., $\hat{T} = \dfrac{(\hat{Z}_1\hat{Z}_2^t)^2}{\hat{Z}_1\hat{Z}_1^t \cdot \hat{Z}_2\hat{Z}_2^t}$. We need to find the distribution of $\hat{T}$, but unlike T, because of noise, it is difficult to calculate the distribution of T accurately. Therefore, in that $\hat{T}$ is an estimate of T, it is necessary to find out how asymptotically similar the two are under the null-hypothesis ($H_0 : \rho = 0$).

Theorem 4. *Let* $(z_{1i}, z_{2i})_{i=1,\dots,n}$ *be iid random samples from bivariate normal distribution,* $N_2(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix})$. *Define* $\hat{T} = \frac{(\hat{Z}_1\hat{Z}_2^t)^2}{\hat{Z}_1\hat{Z}_1^t \cdot \hat{Z}_2\hat{Z}_2^t}, T = \frac{(Z_1Z_2^t)^2}{Z_1{Z_1}^t \cdot Z_2{Z_2}^t}$, *then, Under the null hypothesis* $(H_0 : \rho = cor(z_{11}, z_{21}) = 0)$, *test statistic* $\hat{T} = T + O_p(1/n)$

Proof. Note that, for $k = 1, 2$, $\hat{Z}_k = \hat{v}_k^t\mathbb{X}_k = v_k^t\mathbb{X}_k + (\hat{v}_k - v_k)^t\mathbb{X}_k$. Since $(\hat{v}_k - v_k)$ is $O_p(1/\sqrt{n})$ and $\frac{\mathbb{X}_1\mathbb{X}_2^t}{n}$ is $O_p(1/\sqrt{n})$ under H_0,

$$\begin{aligned}\frac{\hat{Z}_1{\hat{Z}_2}^t}{n} =& \frac{(v_1^t\mathbb{X}_1)(v_2^t\mathbb{X}_2)^t}{n} + (\hat{v}_1 - v_1)^t\frac{\mathbb{X}_1\mathbb{X}_2^t}{n}(\hat{v}_2 - v_2)\\ &+ v_1^t\frac{\mathbb{X}_1\mathbb{X}_2^t}{n}(\hat{v}_2 - v_2) + (\hat{v}_1 - v_1)^t\frac{\mathbb{X}_1\mathbb{X}_2^t}{n}v_2\\ =& \frac{(v_1^t\mathbb{X}_1)(v_2^t\mathbb{X}_2)^t}{n} + O_p(1/n),\end{aligned}$$

and for $k = 1, 2$,

$$\begin{aligned}\frac{\hat{Z}_k{\hat{Z}_k}^t}{n} =& \frac{(v_k^t\mathbb{X}_k)(v_k^t\mathbb{X}_k)^t}{n} + (\hat{v}_k - v_k)^t\frac{\mathbb{X}_k\mathbb{X}_k^t}{n}(\hat{v}_k - v_k)\\ &+ v_1^t\frac{\mathbb{X}_k\mathbb{X}_k^t}{n}(\hat{v}_k - v_k) + (\hat{v}_k - v_k)^t\frac{\mathbb{X}_k\mathbb{X}_k^t}{n}v_k\\ =& \frac{(v_k^t\mathbb{X}_k)(v_k^t\mathbb{X}_k)^t}{n} + O_p(1/n).\end{aligned}$$

Therefore,

$$\begin{aligned}\frac{\hat{Z}_1{\hat{Z}_2}^t}{\sqrt{\hat{Z}_1{\hat{Z}_1}^t}\sqrt{\hat{Z}_2{\hat{Z}_2}^t}} =& \frac{\frac{(v_1^t\mathbb{X}_1)(v_2^t\mathbb{X}_2)^t}{n} + O_p(1/n)}{\sqrt{\frac{(v_1^t\mathbb{X}_1)(v_1^t\mathbb{X}_1)^t}{n} + O_p(1/n)}\sqrt{\frac{(v_2^t\mathbb{X}_2)(v_2^t\mathbb{X}_2)^t}{n} + O_p(1/n)}}\\ =& \frac{(v_1^t\mathbb{X}_1)(v_2^t\mathbb{X}_2)^t}{\sqrt{(v_1^t\mathbb{X}_1)(v_1^t\mathbb{X}_1)^t}\sqrt{(v_2^t\mathbb{X}_2)(v_2^t\mathbb{X}_2)^t}} + O_p(1/n).\end{aligned}$$

Since T is also $O_p(1/\sqrt{n})$,

$$n\hat{T} = n\frac{(\hat{Z}_1{\hat{Z}_2}^t)^2}{\hat{Z}_1{\hat{Z}_1}^t \cdot \hat{Z}_2{\hat{Z}_2}^t} = nT + O_p(1/\sqrt{n}).$$

□

Now we know that $n\hat{T} = nT + O_p(1/\sqrt{n})$. If we find an asymptotic distribution of nT, we can make a asymptotic test for $n\hat{T}$ by using the asymptotic distribution of nT. We have already shown that T follows a beta distribution. For the random variable T following the beta distribution, it can be easily shown that nT asymptotically follows the chi-square distribution using the convergence and uniqueness of the mgf.

Theorem 5. *Suppose T is a random variable that follows $Beta(\frac{1}{2})$, then nT converges in distribution to $\chi^2(1)$.*

Proof. In this proof, we want to use the uniqueness of mgf. It is sufficient to show that the mgf of the random variable nT converges to the mgf of a random variable following the chi-square distribution as n increases. Specifically,

$$\begin{aligned} mgf_{nT}(s) = \mathbb{E}(e^{snT}) &= 1 + \sum_{k=1}^{\infty} \Big(\prod_{r=0}^{k-1} \frac{\frac{1}{2} + r}{\frac{1}{2} + \frac{n-1}{2} + r} \Big) \frac{(ns)^k}{k!} \\ &= 1 + \sum_{k=1}^{\infty} \Big(\prod_{r=0}^{k-1} \frac{1+2r}{n+2r} \Big) \frac{(ns)^k}{k!}. \end{aligned}$$

Then,

$$\begin{aligned} \lim_{n\to\infty} mgf_{nT}(s) &= 1 + \lim_{n\to\infty} \sum_{k=1}^{\infty} \Big(\prod_{r=0}^{k-1} \frac{1+2r}{n+2r} \Big) \frac{(ns)^k}{k!} \\ &= 1 + \sum_{k=1}^{\infty} \lim_{n\to\infty} \Big(\prod_{r=0}^{k-1} \frac{1+2r}{n+2r} \Big) \frac{(ns)^k}{k!} \\ &= 1 + \sum_{k=1}^{\infty} \Big(\prod_{r=0}^{k-1} (1+2r) \Big) \frac{(s)^k}{k!} \\ &= (1-2t)^{-\frac{1}{2}}. \end{aligned}$$

The first equality uses the fact that it could change the limitation order of positive sequence. The last equality uses taylor series expansion. □

The above theorem indicates that nT asymptotically follows the $\chi^2(1)$ distribution. Since $n\hat{T}$ is $nT + O_p(\frac{1}{\sqrt{n}})$, We present the following asymptotic test using $n\hat{T}$.

Asymptotic principal angle test for $\hat{Z}_1, \hat{Z}_2$

Let $(z_{1i}, z_{2i})_{i=1,\dots,n}$ be iid random samples from bivariate normal distribution, $N_2(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix})$. For $\hat{Z}_1 = \hat{v}_1^t \mathbb{X}_1, \hat{Z}_2 = v_2^t \mathbb{X}_2$, define $\hat{T} = \dfrac{(\hat{Z}_1 \hat{Z}_2^t)^2}{\hat{Z}_1 \hat{Z}_1^t \cdot \hat{Z}_2 \hat{Z}_2^t}$, then for two hypothesis,

$$H_0 : \rho^2 = 0 \qquad\qquad H_1 : \rho^2 > 0$$

reject H_0 if $n\hat{T} > \chi^2_\alpha(1)$ where $\chi^2_\alpha(1)$ is the $1-\alpha$ quantile of the chi square distribution $\chi^2(1)$ with a significant level α.

Next theorem shows that the probability of rejecting the null hypothesis H_0 under the alternative hypothesis H_1, i.e., $\rho = cor(z_{11}, z_{21}) \neq 0$, converges to 1. The rejection region of this test with a significant level α is $(n\hat{T} > \chi^2_\alpha(1))$. It could be changed to $(\hat{T} > \dfrac{\chi^2_\alpha(1)}{n})$. It would be helpful to know the convergence of $\hat{T}$ under the alternative hypothesis.

Theorem 6. *Let ρ be the correlation coefficient of (z_{11}, z_{21}). For $\hat{Z}_1 = \hat{v}_1^t \mathbb{X}_1, \hat{Z}_2 = \hat{v}_2^t \mathbb{X}_2$ in the data blocks with $\hat{v}_1, \hat{v}_2$, maximal eigenvector of $\mathbb{X}_1\mathbb{X}_1^t, \mathbb{X}_2\mathbb{X}_2^t$ respectively.*

$$\frac{\hat{Z}_1 \hat{Z}_2}{\left\|\hat{Z}_2\right\| \left\|\hat{Z}_2\right\|} \xrightarrow{p} \frac{\rho}{\sqrt{1+\frac{\tau_1^2}{\sigma_1^2}}\sqrt{1+\frac{\tau_2^2}{\sigma_2^2}}} \text{ as } n \to \infty$$

Proof. Let each $\hat{\lambda_1}, \hat{\lambda_2}$ be the maximal eigenvalue of $\frac{1}{n}\mathbb{X}_1\mathbb{X}_1^t, \frac{1}{n}\mathbb{X}_2\mathbb{X}_2^t$ and $\hat{v}_1, \hat{v}_2$

be the corresponding eigenvector. Then,

$$\frac{\hat{Z_1}\hat{Z_2}^t}{\left\|\hat{Z_2}\right\|\left\|\hat{Z_2}\right\|} = \frac{\frac{1}{n}\hat{Z_1}\hat{Z_2}^t}{\sqrt{\hat{v_1}^t\frac{\mathbb{X}_1\mathbb{X}_1^t}{n}\hat{v_1}}\sqrt{\hat{v_2}^t\frac{\mathbb{X}_2\mathbb{X}_2^t}{n}\hat{v_2}}} = \frac{\frac{1}{n}\hat{Z_1}\hat{Z_2}^t}{\sqrt{\hat{\lambda_1}}\sqrt{\hat{\lambda_2}}},$$

where

$$\begin{aligned}\frac{1}{n}\hat{Z_1}\hat{Z_2}^t &= \frac{1}{n}(\hat{v_1}^t v_1 Z_1 + \hat{v_1}^t E_1)(\hat{v_2}^t v_2 Z_2 + \hat{v_2}^t E_2)^t \\ &= (\hat{v_1}^t v_1)(\frac{Z_1 Z_2^t}{n})(v_2^t\hat{v_2}) + (\hat{v_1}^t)(\frac{E_1 Z_2^t}{n})(v_2^t\hat{v_2}) \\ &\quad + (\hat{v_1}^t v_1)(\frac{Z_1 E_2^t}{n})(\hat{v_2}) \quad + \ (\hat{v_1}^t)(\frac{E_1 E_2^t}{n})(\hat{v_2}).\end{aligned}$$

Note that $\hat{v_i}^t v_i \xrightarrow{p} \|v_i\|^2 = 1$ since $\hat{v_i} \xrightarrow{p} v_i$,

$$\begin{aligned}&\frac{Z_1 Z_2^t}{n} \xrightarrow{p} Cov(z_1, z_2) = \rho\sigma_1\sigma_2, \\ &\frac{E_1 Z_2^t}{n} \xrightarrow{p} \mathbb{E}(z_2\epsilon_1) = \mathbb{E}(z_2)\mathbb{E}(\epsilon_1) = 0, \text{ similarly } \frac{Z_1 E_2^t}{n} \xrightarrow{p} 0, \\ &\frac{E_1 E_2^t}{n} \xrightarrow{p} Cov(\epsilon_1, \epsilon_2) = 0, \hat{\lambda}_1 \xrightarrow{p} \sigma_1^2 + \tau_1^2, \hat{\lambda}_2 \xrightarrow{p} \sigma_2^2 + \tau_2^2.\end{aligned}$$

Therefore,

$$\therefore \frac{\hat{Z_1}\hat{Z_2}^t}{\left\|\hat{Z_2}\right\|\left\|\hat{Z_2}\right\|} \xrightarrow{p} \frac{\rho\sigma_1\sigma_2}{\sqrt{\sigma_1^2+\tau_1^2}\sqrt{\sigma_2^2+\tau_2^2}} = \frac{\rho}{\sqrt{1+\frac{\tau_1^2}{\sigma_1^2}}\sqrt{1+\frac{\tau_2^2}{\sigma_2^2}}} \neq 0.$$

□

Corollary 7. *In the asymptotic principal angle test for $(\hat{Z}_1, \hat{Z}_2)$, $P_{H_1}(\cdot)$ denotes the probability law under the alternative hypothesis $H_1 : \rho^2 > 0$ at any fixed ρ. Then $P_{H_1}(n\hat{T} > \chi^2_\alpha(1)) \to 1$ as $n \to \infty$ with a significant level α.*

Proof. The reject region of the test with a significant level α is $(n\hat{T} > \chi^2_\alpha(1))$. Under $H_1 : \rho^2 > 0$,

$$P_{H_1}\left(n\hat{T} > \chi^2_\alpha(1)\right) = P_{H_1}\left(\frac{(\hat{Z_1}\hat{Z_2}^t)^2}{\hat{Z_1}\hat{Z_1}^t \cdot \hat{Z_2}\hat{Z_2}^t} > \chi^2_\alpha(1)/n\right)$$

Since $\frac{(\hat{Z}_1\hat{Z}_2{}^t)^2}{\hat{Z}_1\hat{Z}_1{}^t\cdot\hat{Z}_2\hat{Z}_2{}^t} \xrightarrow{p} \frac{\rho}{\sqrt{1+\frac{\tau_1^2}{\sigma_1^2}}\sqrt{1+\frac{\tau_2^2}{\sigma_2^2}}}, \chi^2_\alpha(1)/n \xrightarrow{p} 0,$ and

$\frac{(\hat{Z}_1\hat{Z}_2{}^t)^2}{\hat{Z}_1\hat{Z}_1{}^t\cdot\hat{Z}_2\hat{Z}_2{}^t} - \chi^2_\alpha(1)/n \xrightarrow{p} \frac{\rho}{\sqrt{1+\frac{\tau_1^2}{\sigma_1^2}}\sqrt{1+\frac{\tau_2^2}{\sigma_2^2}}}$, for some small $\epsilon > 0$,

$$
\begin{aligned}
P_{H_1}\left(n\hat{T} > \chi^2_\alpha(1)\right) &= P_{H_1}\left(\frac{(\hat{Z}_1\hat{Z}_2{}^t)^2}{\hat{Z}_1\hat{Z}_1{}^t\cdot\hat{Z}_2\hat{Z}_2{}^t} > \chi^2_\alpha(1)/n\right) \\
&= P_{H_1}\left(\frac{(\hat{Z}_1\hat{Z}_2{}^t)^2}{\hat{Z}_1\hat{Z}_1{}^t\cdot\hat{Z}_2\hat{Z}_2{}^t} - \chi^2_\alpha(1)/n > 0\right) \\
&\geq P_{H_1}\left(\left|\frac{(\hat{Z}_1\hat{Z}_2{}^t)^2}{\hat{Z}_1\hat{Z}_1{}^t\cdot\hat{Z}_2\hat{Z}_2{}^t} - \chi^2_\alpha(1)/n - \frac{\rho}{\sqrt{1+\frac{\tau_1^2}{\sigma_1^2}}\sqrt{1+\frac{\tau_2^2}{\sigma_2^2}}}\right| \leq \epsilon\right) \\
&\xrightarrow{n\to\infty} 1
\end{aligned}
$$

□

The above Corollary 7 ensures that the probability of rejecting null hypothesis in the asymptotic principal angle test for$\hat{Z}_1, \hat{Z}_2$ converges to 1 under the alternative hypothesis. In AJIVE, the rank of joint signal is determined by using the empirical distribution of the principal angle between random direction in $\mathbb{R}^n$ and wedin bound. However, under the large n, this asymptotic principal angle test is expected to make it relatively simple to determine if the rank of joint signals is 1 or 0.

Chapter 6

Simulation study

In this section, we compare the performance of the rank selection in AJIVE and the rank selection in the asymptotic test. The two data block were generated randomly in the programming language R. The detailed process is as follows.

6.1 Data block setting

Each data blocks $\mathbb{X}_1$and $\mathbb{X}_2$ is $(p \times n)$, $(q \times n)$ size matrix in the same way as described in the model structure. The number of variable in $\mathbb{X}_1, p$ is set to 100,and the number of variable in $\mathbb{X}_2, q$ is set to 150. The number of sample,n is 100 and 500. The data structure is written as $\mathbb{X}_1 = v_1 Z_1 + E_1, \mathbb{X}_2 = v_2 Z_2 + E_2$, and each element is constructed by the following process. First, Z_1, Z_2 are generated by sampling n (z_{1i}, z_{2i}) independently and repeatedly from the bivariate normal distribution with mean of 0 vectors, and the covariance matrix, $\begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$. Set $\sigma_1^2 = \sigma_2^2 = 5$, and ρ varies from 0 to 1 by 0.1. Second, v_1 and v_2 are fixed vectors in p-dimensional and q-dimensional vector

space, respectively. However, since we cannot observe them, v_1 was determined by generating random vector from the p-dimensional standard normal distribution and then normalizing this random vector. v_2,q-dimensional vector, was determined in a similar way as v_1. Third, E_1, E_2 is a random matrices that means noise. E_1, the size of $(p \times n)$, was generated by n randomly sampling the p-dimensional normal distribution with mean of 0 and a covariance matrix,$(\tau_1 I_p)$. Similarly, E_2, the size of $(q \times n)$, was generated by n randomly sampling the q-dimensional normal distribution with mean of 0 and a covariance matrix,$(\tau_2 I_q)$.

6.2 Joint signal rank selection

In AJIVE, the rank selection of the joint signal is determined by using the wedin bound[3] and the empirical distribution of the principal angle between the random direction. AJIVE's rank selection used the 'joint_rank_estimate' of the ajive function in 'ajive' package provided in R. For those who want to reproduce this simulation, note that 'joint_rank' value of the ajive function is supposed to be only 1 value. This is because the 'joint_rank' function is supposed to be presented as 1 in case there is no joint signal. Therefore, we should check 'joint_rank_estimate' instead of 'joint_rank'. For $\rho = cor(z_{11}, z_{21}) = 0, 0.1, \ldots, 0.9, 1$ and $n = 100, 500$, the joint signal rank selection in the above data block setting was repeated 100 times, and the Table 1 shows the number of times the joint rank was determined to be 1 among 100 times with a significant level $\alpha = 0.05$ in each ρ and n.

Overall, the asymptotic test shows similar results to AJIVE's rank selection processing. There are three interesting things in the table. The first is that as n increases from 100 to 500, the number of times to detect ρ which is not zero among 100 simulations is significantly increased. The second is that even when

Table 6.1 The number of rank 1 selection out of 100 simulations

ρ	$n = 100$		$n = 500$	
	AJIVE	Asymptotic test	AJIVE	Asymptotic test
$\rho = 0.0$	10	5	6	4
$\rho = 0.1$	12	10	48	45
$\rho = 0.2$	35	30	92	94
$\rho = 0.3$	70	69	100	100
$\rho = 0.4$	90	89	100	100
$\rho = 0.5$	100	100	100	100
$\rho = 0.6$	100	100	100	100
$\rho = 0.7$	100	100	100	100
$\rho = 0.8$	100	100	100	100
$\rho = 0.9$	100	100	100	100
$\rho = 1.0$	100	100	100	100

n is 100, if $\rho \geq 0.5$, the number of times the joint signal rank determined to be 1 is 100 times out of 100 simulations. The last is that AJIVE processing tends to determine that a joint signal exists rather than the asymptotic test. In particular, if $\rho = 0$ and $n = 100$, then 5 is the ideal number of signal rank selection number with the significant level 5%. However AJIVE processing determined that 10 out of 100 simulations had a joint signal. Therefore, it appears that the asymptotic test is performing a joint rank selection with a similar performance as the AJIVE process. When the joint rank selection through random direction sampling is difficult to do, using the asymptotic test could be a good alternative if there is sufficient evidence that the two significant signals have normality.

Chapter 7

Discussion

In this thesis, we propose to estimate the rank of the joint signal from the joint individual structure [1]. In the two data blocks from different sources and same sample subjects, each data blocks consists of the joint signals, the individual signals and noise. We assume a model in which each of the two data blocks is composed of rank 1 significant signals Z_1, Z_2 and noise E_1, E_2 and written in $\mathbb{X}_1 = v_1 Z_1 + E_1, \mathbb{X}_2 = v_2 Z_2 + E_2$ with loading vector v_1, v_2 representing Z_1, Z_2 to data blocks $\mathbb{X}_1, \mathbb{X}_2$ respectively. Based on the principal angle between the two significant signals proposed by Feng, et al. [2], the measure of closeness between the two significant signals is defined as the square of the cosine of the principal angle and denoted as T. Assuming that the two significant signals are samples of n independent bivariate normal distribution, it has been confirmed that T follows the distribution of Beta$(\frac{1}{2}, \frac{n-1}{2})$. Next, we estimate (v_1, v_2) and (Z_1, Z_2) using that fact that v_1, v_2 are the maximal eigenvector of the covariance matrices of the first column of data blocks. Then we calculated the difference between $n\hat{T}$ and nT is approximately $O_p(\frac{1}{\sqrt{n}})$, where $\hat{T}$ is

the closeness measure of estimate value of Z_1, Z_2 and T is the exact closeness measure of Z_1, Z_2. Finally, using the fact that nT follows an asymptotically chi square distribution and nT and $n\hat{T}$ are approximately similar, the asymptotic test for principal angle which determine the rank of the joint signal was created. In the simulation study, the asymptotic test was observed to find the joint signals rank 1 better as the sample n increase or the correlation coefficient ρ is futher away from 0. In addition, we observed that the asymptotic test was similar in performance compared to the AJIVE's joint rank selection process in the ajive package provided in the programming language R. Thus, under the assumption that the significant signals follow normal distribution, the asymptotic test seems to be a good alternative.

The model structure includes rather strong assumptions, such as normality and independence. A strong property of normality is the fact that uncorrelatedness means independence and the fact that it is distribution which is invariant to rotation. it seems likely that the normality assumption will be weakened by the spherical distribution. In addition, the presented model structure assumes that the rank of the significant signals is 1. The reason why we regard the principal angle, not correlation coefficient is that the principal angle can be a more reasonable distance in a multi-dimensional significant signal matrix than the correlation coefficient. It seems necessary to investigate the probability property of the principal angle for the multi-dimensional significant signal matrices.

Chapter 8

Appendix : R code

the Simulation R code is as follow

```
library('MultiRNG') ;library('cowplot')
library('mvtnorm') ; library('ajive')
### normal dist
p <- 100 ; q <- 150
sig1 <- sqrt(5); sig2 <- sqrt(5)
tau1 <- 1; tau2 <- 1
rholist <- seq(0,1,by=0.1) ;nlist <- c(100,500)
result <- matrix(0.1,length(nlist),length(rholist))
result <- `colnames<-`(result,paste("rho=",as.character(rholist)))
result <- `rownames<-`(result,c('n=100','n=500'))
res.ajive <- result
res.test <- result
set.seed(1)
for (k in 1:length(rholist)) {
  for (j in 1:length(nlist)){
    rho <- rholist[k]
```

```
    n <- nlist[j]
    count1 <- 0 ; count2 <- 0
    for(i in 1:100){
      varX <- matrix(c(sig1^2,rho*sig1*sig2,rho*sig1*sig2,sig2^2),2,2)
      ### data generation
      z <- rmvnorm(n,c(0,0),varX)
      z1 <- z[,1] ; z2 <- z[,2]
      v1 <- matrix(rnorm(p,0,5),p,1) ; v1 <- v1/norm(v1,"2")
      v2 <- matrix(rnorm(q,0,5),q,1) ; v2 <- v2/norm(v2,"2")
      e1 <- t(rmvnorm(n,rep(0,p),tau1*diag(p)))
      e2 <- t(rmvnorm(n,rep(0,q),tau2*diag(q)))
      x1 <- t(v1%*%z1 + e1) ; x2 <- t(v2%*%z2 + e2)
      x <- list(x1,x2)
      ## ajive package
      initial_signal_ranks <- c(1, 1) # set by looking at scree plots
      jive_results <- ajive(x, initial_signal_ranks,
                 n_wedin_samples = 100, n_rand_dir_samples = 100)
      count1 <- count1+jive_results$joint_rank_sel$joint_rank_estimate
      ## chi-square test
      x1 <- t(x1) ; x2 <- t(x2)
      eigen1 <- eigen(x1%*%t(x1))  ;  eigen2 <- eigen(x2%*%t(x2))
      v1_hat <- eigen1$vectors[,1] ;  v2_hat <- eigen2$vectors[,1]
      z1_hat <- v1_hat%*%x1 ;  z2_hat <- v2_hat%*%x2
      count2 <- count2 + as.numeric(n*((z1_hat/norm(z1_hat,"2"))%*%
      t((z2_hat/norm(z2_hat,"2"))))^2 > qchisq(0.95,1))
    }
    res.ajive[j,k] <- count1 ; res.test[j,k] <- count2
  }
}
res.norm.ajive <- res.ajive
res.norm.test <- res.test
```

Bibliography

[1] E. F. Lock, K. A. Hoadley, J. S. Marron, and A. B. Nobel, "Joint and individual variation explained (jive) for integrated analysis of multiple data types," *The annals of applied statistics*, vol. 7, no. 1, p. 523, 2013.

[2] Q. Feng, M. Jiang, J. Hannig, and J. Marron, "Angle-based joint and individual variation explained," *Journal of multivariate analysis*, vol. 166, pp. 241–265, 2018.

[3] P.-Å. Wedin, "Perturbation bounds in connection with singular value decomposition," *BIT Numerical Mathematics*, vol. 12, no. 1, pp. 99–111, 1972.

[4] R. J. Muirhead, *Aspects of multivariate statistical theory*, vol. 197. John Wiley & Sons, 2009.

국문초록

본 연구는 다중원천 데이터에서 얻은 데이터 블럭에서 Lock, et al.[Ann. Appl. Stat.,7(2013),
523]가 제시한 공통-개별구조에서 공통구조의 존재성에 대한 검정을 다룬다. 두 개의 데이터 블럭에 대해, 각 데이터 블럭은 (p×n)와 (q × n) 사이즈인 행렬이며 각 데이터 블럭 $\mathbb{X}$는 랭크가 1인 유의미한 랜덤 행벡터 시그널 Z가 로딩 벡터 v로 표현되며 잡음 E가 더해지는 $\mathbb{X} = vZ + E$의 형태임을 가정한다. Feng, et al[J.Multivar.Anal.,166(2018),241-265]은 공통-개별구조에서 유의미한 시그널들의 거리를 principal angle로 정의한 반면, 본 연구에서는 유의미한 시그널들의 유사함의 측도를 principal angle의 코사인 함수 제곱값으로 정의한다. 유의미한 시그널의 각 표본이 iid 이변량 정규분포를 따른다고 가정할 때 유의미한 시그널의 유사함 측도가 Beta$(\frac{1}{2}, \frac{n-1}{2})$분포를 따름을 이용하여 유의미한 시그널의 공통구조 추출에 대한 검정을 제안한다. 더하여 데이터 블럭의 형태로 부터 v가 공분산행렬의 maxmal eigenvector라는 사실을 이용하면 관측 할 수 없는 v와 유의미한 시그널 Z를 추정할 수 있다. 관측 불가능한 Z의 값을 예측값 $\hat{Z}$로 대체하여, 우리는 공통 신호의 랭크에 대한 다표본 검정을 개발한다. 시뮬레이션 연구에서는 이 점근적 검정과 AJIVE로 불리는 Feng, et al.[J.Multivar.Anal.,166(2018),241-265]의 공통구조 랭크 판단을 비교하였다. 시뮬레이션 결과, 이 점근적 검정이 AJIVE의 기존 방법을 대체할만큼 AJIVE의 랭크 판단과 전반적으로 비슷한 결과를 보여줌을 확인하였다. 따라서 유의미한 신호들이 정규분포를 따른다는 가정하에서 이 점근적 테스트가 AJIVE 랭크 판단과정의 좋은 대안이 될 것으로 기대된다

주요어: 다중원천데이터,principal angle, 공통-개별구조
학번: 2018-20856