Validity of an efficient approach for principal component analysis in high dimensional data with mixed variable structure

Okada, Susumu
Institute of Statistical Science
3, Asagayakita,
Suginami-ku, Tokyo, 166-0001, Japan
E-mail: okada@statistics.co.jp

Sugiyama, Takakazu
Soka University
1, Tangimachi,
Hachiouji-shi, Tokyo, 192-8577, Japan
E-mail: stakakaz@gmail.com

1. Introduction

Let $\mathbf{X}$ be a $p \times N$ observation matrix $(p \le N)$ which is obtained by independently distributed p-dimensional variables $\mathbf{x} = \begin{pmatrix} x_1 & \cdots & x_p \end{pmatrix}'$ for N observations. Let $\mathbf{S}$ be a sample covariance matrix of $\mathbf{X}$. The covariance matrix $\mathbf{S}$ is expressed as

$$\mathbf{S} = \frac{1}{N-1} \sum_{\alpha=1}^{N} (\mathbf{x}_\alpha - \bar{\mathbf{x}})(\mathbf{x}_\alpha - \bar{\mathbf{x}})', \tag{1.1}$$

where $\bar{\mathbf{x}}$ is a sample mean vector of $\mathbf{x}$. Here $\mathbf{S}$ is positive semidefinite.

Suppose the covariance matrix $\mathbf{S}$ has a $p \times p$ real orthogonal matrix $\mathbf{B}$ which is formed by eigenvectors of p components, and is satisfies $\mathbf{B}'\mathbf{B} = \mathbf{I}$. Here $\mathbf{B}'$ means the transpose of $\mathbf{B}$, and $\mathbf{I}$ is the identity matrix. Let us now consider any orthogonal transformation

$$\mathbf{U} = \mathbf{B}'\mathbf{X} \tag{1.2}$$

as a multivariate linear model for $\mathbf{X}$, where $\mathbf{U}$ is a $p \times N$ transformed observation matrix which has p components. This is a normal approach for analyzing principal components.

The method of principal component analysis is often applied, when the number of variables under consideration is too large to treat. Principal components, which are obtained by principal component analysis, are use to reduce the dimension of a data set of original interrelated variables, where the principal components are constituted of uncorrelated linear combinations with large variance of these variables. However, the normal approach for analyzing principal components, which treat all variables simultaneously, requires many computing resources in high dimensional data with a large number of variables.

Then we propose an efficient approach for analyzing principal components as a proposal approach.

The proposal approach should be derived with a new strategy with a partitioned data model as an orthogonal transformation for a multivariate linear model. We also investigate the proposal approach through two concrete examples with an educational data set $(p=9, N=166)$ and a molecular genetics data set $(p=100, N=135)$. The validity of the proposal approach should be verified in high dimensional data with mixed variable structure.

2. An efficient approach for principal component analysis

In this section an efficient approach for analyzing principal components should be proposed. We shall now consider a new strategy with a partitioned data model to derive an orthogonal transformation

$$\mathbf{V}=\mathbf{H}'\mathbf{X} \tag{2.1}$$

as a multivariate linear model for $\mathbf{X}$, where $\mathbf{V}$ is a $p\times N$ transformed observation matrix which has p components and $\mathbf{H}$ is a $p\times p$ real orthogonal matrix such that $\mathbf{H}\mathbf{H}'=\mathbf{I}$. Here $\mathbf{H}'$ means the transpose of $\mathbf{H}$. Then the orthogonal matrix $\mathbf{H}$ should be derived with the new strategy with the partitioned data model as a proposal approach.

Let us partition the $p\times N$ observation matrix $\mathbf{X}$ into two sets of observation matrices $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ as follows:

$$\mathbf{X}=\begin{pmatrix}\mathbf{X}^{(1)}\\ \mathbf{X}^{(2)}\end{pmatrix}. \tag{2.2}$$

Let $\mathbf{X}^{(1)}$ be a $p_1\times N$ partitioned observation matrix formed by a p_1-dimensional vector $\mathbf{x}^{(1)}=\left(x_1 \ \cdots \ x_{p_1}\right)'$ for N subjects. Similarly, let $\mathbf{X}^{(2)}$ be a $p_2\times N$ partitioned observation matrix formed by a p_2-dimensional vector $\mathbf{x}^{(2)}=\left(x_{p_1+1} \ \cdots \ x_p\right)'$ for N subjects, where $p_2=p-p_1$.

Let $\mathbf{S}_1$ and $\mathbf{S}_2$ be sample covariance matrices of $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$, respectively. The covariance matrices $\mathbf{S}_1$ and $\mathbf{S}_2$ are expressed as

$$\mathbf{S}_1=\frac{1}{N-1}\sum_{\alpha=1}^{N}\left(\mathbf{x}_\alpha^{(1)}-\overline{\mathbf{x}}^{(1)}\right)\left(\mathbf{x}_\alpha^{(1)}-\overline{\mathbf{x}}^{(1)}\right)', \tag{2.3}$$

$$\mathbf{S}_2=\frac{1}{N-1}\sum_{\alpha=1}^{N}\left(\mathbf{x}_\alpha^{(2)}-\overline{\mathbf{x}}^{(2)}\right)\left(\mathbf{x}_\alpha^{(2)}-\overline{\mathbf{x}}^{(2)}\right)', \tag{2.4}$$

where $\overline{\mathbf{x}}^{(1)}$ and $\overline{\mathbf{x}}^{(2)}$ are sample mean vectors of $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$, respectively. Here $\mathbf{S}_1$ and $\mathbf{S}_2$ are positive semidefinite.

Suppose the covariance matrix $\mathbf{S}_1$ has a $p_1\times p_1$ real orthogonal matrix $\mathbf{C}_1$ which is formed by eigenvectors of p_1 components, and is satisfies $\mathbf{C}_1'\mathbf{C}_1=\mathbf{I}$. Similarly, suppose the covariance matrix $\mathbf{S}_2$ has a $p_2\times p_2$ real orthogonal matrix $\mathbf{C}_2$ which is formed by eigenvectors of p_2 components, and is satisfies $\mathbf{C}_2'\mathbf{C}_2=\mathbf{I}$. Here $\mathbf{C}_1'$ and $\mathbf{C}_2'$ mean the transposes of $\mathbf{C}_1$ and $\mathbf{C}_2$, respectively.

Let $\mathbf{Y}^{(1)*}$ be a $k_1\times N$ transformed observation matrix which has k_1 components, and be

obtained by an orthogonal linear transformation

$$\mathbf{Y}^{(1)*} = \mathbf{C}_1^{*\prime}\mathbf{X}^{(1)}, \tag{2.5}$$

where $\mathbf{C}_1^*$ is a $p_1 \times k_1$ real orthogonal matrix obtained from $\mathbf{C}_1$ by selecting k_1 columns in descending order of the variance. Similarly, let $\mathbf{Y}^{(2)*}$ be a $k_2 \times N$ transformed observation matrix which has k_2 components, and be obtained by an orthogonal linear transformation

$$\mathbf{Y}^{(2)*} = \mathbf{C}_2^{*\prime}\mathbf{X}^{(2)}, \tag{2.6}$$

where $\mathbf{C}_2^*$ is a $p_2 \times k_2$ real orthogonal matrix obtained from $\mathbf{C}_2$ by selecting k_2 columns in descending order of the variance. Then let $\mathbf{Y}^*$ be a $k \times N$ transformed observation matrix formed by a k-component vector $\mathbf{y}^* = \begin{pmatrix} y_1 & \cdots & y_k \end{pmatrix}'$ for N subjects, and be obtained by connecting $\mathbf{Y}^{(1)*}$ and $\mathbf{Y}^{(2)*}$ as follows:

$$\mathbf{Y}^* = \begin{pmatrix} \mathbf{Y}^{(1)*} \\ \mathbf{Y}^{(2)*} \end{pmatrix}. \tag{2.7}$$

Let $\mathbf{S}_\mathbf{Y}^*$ be a sample covariance matrix of $\mathbf{Y}^*$. The covariance matrix $\mathbf{S}_\mathbf{Y}^*$ is expressed as

$$\mathbf{S}_\mathbf{Y}^* = \frac{1}{N-1}\sum_{\alpha=1}^{N}\left(\mathbf{y}_\alpha^* - \overline{\mathbf{y}}^*\right)\left(\mathbf{y}_\alpha^* - \overline{\mathbf{y}}^*\right)', \tag{2.8}$$

where $\overline{\mathbf{y}}^*$ is a sample mean vector of $\mathbf{y}^*$. Here $\mathbf{S}_\mathbf{Y}^*$ is positive semidefinite.

Suppose the covariance matrix $\mathbf{S}_\mathbf{Y}^*$ has a $k \times k$ real orthogonal matrix $\mathbf{D}^*$ which is formed by eigenvectors of k components, and is satisfies $\mathbf{D}^{*\prime}\mathbf{D}^* = \mathbf{I}$. Here $\mathbf{D}^{*\prime}$ means the transpose of $\mathbf{D}^*$. Let us partition the $k \times k$ real orthogonal matrix $\mathbf{D}^*$ into two sets of real orthogonal matrices $\mathbf{D}_1^*$ and $\mathbf{D}_2^*$ as follows:

$$\mathbf{D}^* = \begin{pmatrix} \mathbf{D}_1^* \\ \mathbf{D}_2^* \end{pmatrix}, \tag{2.9}$$

where $\mathbf{D}_1^*$ is a $k_1 \times k$ real orthogonal matrix, and $\mathbf{D}_2^*$ is a $k_2 \times k$ real orthogonal matrix.

Let $\mathbf{Z}^*$ be a $k \times N$ transformed observation matrix which has k components, and be obtained by an orthogonal linear transformation

$$\mathbf{Z}^* = \mathbf{D}^{*\prime}\mathbf{Y}^*. \tag{2.10}$$

Then $\mathbf{Z}^*$ is expressed by $\mathbf{D}_1^*$ and $\mathbf{D}_2^*$ as

$$\mathbf{Z}^* = \begin{pmatrix} \mathbf{D}_1^* \\ \mathbf{D}_2^* \end{pmatrix}' \begin{pmatrix} \mathbf{Y}^{(1)*} \\ \mathbf{Y}^{(2)*} \end{pmatrix} = \mathbf{D}_1^{*\prime}\mathbf{Y}^{(1)*} + \mathbf{D}_2^{*\prime}\mathbf{Y}^{(2)*} = \mathbf{D}_1^{*\prime}\mathbf{C}_1^{*\prime}\mathbf{X}^{(1)} + \mathbf{D}_2^{*\prime}\mathbf{C}_2^{*\prime}\mathbf{X}^{(2)}. \tag{2.11}$$

Let $\mathbf{H}^*$ be a $p \times k$ real orthogonal matrix, and let us partition $\mathbf{H}^*$ into two sets of real orthogonal matrices $\mathbf{H}_1^*$ and $\mathbf{H}_2^*$ as follows:

$$\mathbf{H}^* = \begin{pmatrix} \mathbf{H}_1^* \\ \mathbf{H}_2^* \end{pmatrix} = \begin{pmatrix} \mathbf{C}_1^*\mathbf{D}_1^* \\ \mathbf{C}_2^*\mathbf{D}_2^* \end{pmatrix}, \tag{2.12}$$

where $\mathbf{H}_1^*$ is a $p_1 \times k$ real orthogonal matrix, and $\mathbf{H}_2^*$ is a $p_2 \times k$ real orthogonal matrix.

Thus the proposal orthogonal linear transformation

$$\mathbf{V}^{*} = \mathbf{H}^{*\prime}\mathbf{X} \tag{2.13}$$

is derived with the new strategy with the partitioned data model, where $\mathbf{V}^{*}$ is a $k \times N$ transformed observation matrix which has k components. Here $\mathbf{V}^{*}$ means the principal components in $\mathbf{V}$. This is the proposal approach as the efficient approach for analyzing principal components.

3. Numerical studies

In this section the validity of the proposal approach should be verified in high dimensional data with mixed variable structure. We shall now investigate the proposal approach through two concrete examples. These examples are an educational data set of scholastic ability in Japan $(p = 9, N = 166)$, and a molecular genetics data set extracted from a data set on the International HapMap Project $(p = 100, N = 135)$.

The educational data set, which is obtained from the results of an examination to junior high school students, is constituted of ordinary low dimensional large sample size data. And the molecular genetics data set, which is obtained from two ethnic groups as Japanese $(N = 45)$ and European $(N = 90)$, is constituted of high dimensional data with mixed variable structure with two sets of different types of observations.

In each example, we present eigenvalues with the normal approach, eigenvalues with the proposal approach, and inner products of eigenvectors as diagonal elements of $\mathbf{B}'\mathbf{H}$. The results with the proposal approach are corresponding to the results with the normal approach well in principal components, though the computational cost with the proposal approach is relatively small.

3.1. Examples with an educational data set of scholastic ability in Japan $(p = 9, N = 166)$

Example 1: $k = 3$ $(k_1 = 1, k_2 = 2)$

1. Eigenvalues with the normal approach:	3211.94 706.30 271.07
2. Eigenvalues with the proposal approach:	3211.12 670.47 114.00
3. Inner products of the eigenvectors:	1.00 0.97 0.09

Example 2: $k = 5$ $(k_1 = 2, k_2 = 3)$

1. Eigenvalues with the normal approach:	3211.94 706.30 271.07 211.39 125.30
2. Eigenvalues with the proposal approach:	3211.67 699.84 255.33 122.19 111.55
3. Inner products of the eigenvectors:	1.00 0.99 0.95 0.41 0.51

The set of the explanatory variables in $p_1 (= 4)$ is constituted of Japanese language, social studies, mathematics, and science. And the set of the explanatory variables in $p_2 (= 5)$ is constituted of music, art, physical education, technical arts and home economics, and English language.

3.2. Examples with a molecular genetics data set extracted from a data set on the International HapMap Project ($p = 100, N = 135$)

Example 1: $k = 10$ $(k_1 = 5, k_2 = 5)$

1. Eigenvalues with the normal approach: 3.95 1.41 1.30 1.23 1.21 1.14 1.09 1.01 0.98 0.92

2. Eigenvalues with the proposal approach: 3.92 1.25 1.17 1.01 0.89 0.82 0.74 0.72 0.60 0.35

3. Inner products of the eigenvectors: 1.00 0.90 0.85 0.72 0.53 0.15 0.30 0.29 0.20 0.08

Example 2: $k = 20$ $(k_1 = 10, k_2 = 10)$

1. Eigenvalues with the normal approach: 3.95 1.41 1.30 1.23 1.21 1.14 1.09 1.01 0.98 0.92 0.91 0.89 0.83 0.83 0.80 0.77 0.75 0.69 0.66 0.64

2. Eigenvalues with the proposal approach: 3.93 1.30 1.24 1.12 1.08 1.01 0.92 0.82 0.78 0.76 0.72 0.69 0.64 0.61 0.54 0.52 0.50 0.43 0.38 0.33

3. Inner products of the eigenvectors: 1.00 0.94 0.95 0.53 0.49 0.78 0.61 0.15 0.13 0.32 0.43 0.09 0.01 0.30 0.11 0.44 0.18 0.06 0.10 0.02

The set of the explanatory variables in $p_1 (= 50)$ is constituted of rs1000000, rs10000010, rs10000023, rs10000030, rs10000041, rs1000007, rs10000081, rs10000092, rs10000121, rs1000014, rs10000141, rs1000016, rs10000169, rs10000185, rs10000201, rs1000022, rs10000226, rs1000025, rs10000282, rs10000300, rs1000031, rs1000032, rs10000388, rs1000040, rs1000041, rs10000435, rs10000438, rs10000456, rs10000471, rs10000487, rs1000050, rs10000502, rs10000538, rs10000543, rs1000055, rs10000595, rs1000061, rs1000068, rs10000697, rs10000708, rs1000071, rs10000719, rs10000726, rs1000073, rs10000770, rs1000078, rs10000785, rs1000079, rs1000083, and rs10000856. And the set of the explanatory variables in $p_2 (= 50)$ is constituted of rs10000869, rs10000901, rs10000918, rs10000929, rs1000094, rs10000959, rs10000969, rs1000104, rs10001138, rs10001148, rs1000115, rs10001154, rs10001198, rs1000121, rs10001214, rs1000122, rs10001225, rs10001236, rs10001241, rs10001297, rs1000131, rs10001340, rs10001348, rs1000137, rs10001378, rs1000140, rs1000141, rs10001415, rs1000147, rs10001480, rs10001483, rs10001495, rs1000152, rs10001539, rs1000154, rs1000156, rs10001565, rs10001577, rs10001580, rs10001582, rs100016, rs10001608, rs10001613, rs10001638, rs10001657, rs10001661, rs10001689, rs10001694, rs10001725, and rs1000173.

4. Acknowledgements

The authors are very grateful to Takahiro Nakamura for helpful suggestions.

REFERENCES

[1] Daisuke Watanabe, Susumu Okada, Yasunori Fujikoshi and Takakazu Sugiyama. (2008). Large sample approximations for the LR statistic for equality of the smallest eigenvalues of a covariance matrix under elliptical population. Computational Statistics & Data Analysis. 52, 2714-2724.
[2] Yasunori Fujikoshi, Takayuki Yamada, Daisuke Watanabe and Takakazu Sugiyama. (2007). Asymptotic distribution of the LR statistic for equality of the smallest eigenvalues in high-dimensional principal component analysis. Journal of Multivariate Analysis. 98, 2002-2008.
[3] Yumi Yamaguchi-Kabata, Kazuyuki Nakazono, Atsushi Takahashi, Susumu Saito, Naoya Hosono, Michiaki Kubo, Yusuke Nakamura, and Naoyuki Kamatani. (2008). Japanese Population Structure, Based on SNP Genotypes from 7003 Individuals Compared to Other Ethnic Groups: Effects on Population-Based Association Studies. The American Journal of Human Genetics. 83, 445-456
[4] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. (2006). Nature Genetics. 38, 904-909.

ABSTRACT

An efficient approach for analyzing principal components (PCs) is proposed. Then the validity of the proposal approach with a new strategy with a partitioned data model is verified in high dimensional data with mixed variable structure. The method of principal component analysis (PCA) is often applied, when the number of variables under consideration is too large to treat. PCs, which are obtained by PCA, are use to reduce the dimension of a data set of original interrelated variables, where the PCs are constituted of uncorrelated linear combinations with large variance of these variables. However, a normal approach for analyzing PCs, which treat all variables simultaneously, requires many computing resources in high dimensional data with a large number of variables. Then we propose an efficient approach for analyzing PCs as a proposal approach with a new strategy with a partitioned data model, and we verify the validity of the proposal approach in high dimensional data with mixed variable structure. The novel approach for PCA based on our idea is to partition all variables into several blocks, and is to execute PCA to the sets of PCs of the every block. We also investigate the proposal approach through two concrete examples. These examples are an educational data set of scholastic ability in Japan, and a molecular genetics data set extracted from a data set on the International HapMap Project. The educational data set is constituted of ordinary low dimensional large sample size data. And the molecular genetics data set is constituted of high dimensional data with mixed variable structure with two sets of different types of observations. The two approaches for PCA with the normal approach and the proposal approach are performed to the two concrete examples, and the results are compared in terms of eigenvalues and eigenvectors. Then the results with the proposal approach are corresponding to the results with the normal approach well in PCs, though the computational cost with the proposal approach is relatively small.