Consistency of the objective general index in high dimensional settings

Takuma Bando*, Tomonari Sei* and Kazuyoshi Yata†

Abstract

The objective general index is a weighting method for ranking of multivariate data. We show that the sample objective general index is a consistent estimator of the population counterpart in high-dimensional settings if the population is multivariate normal and the covariance matrix satisfies a condition. The proof is based on the large deviation theory. Numerical experiments and real data analysis are conducted.

1 Introduction

Rankings are often determined by multivariate data. For example, the world university ranking provided by [11] is based on five attributes of universities: teaching, research, citations, industry income and international outlook. For a happiness index of prefectures in Japan [10], 65 attributes are used to make a ranking of 47 prefectures. In heptathlon of athletics, the scores of seven events are unified into an overall score. These rankings are, after some transformations, based on a weighted sum of variables.

We focus on the weights. In [8], an objective weight is proposed via diagonal scaling of the sample covariance matrix. The resultant index called the objective general index (OGI) has positive correlation with all the variables and is invariant with respect to scale transformation of the data. A precise definition is given in Section 2.

In some applications like the happiness ranking mentioned above, the number of variables is often large and comparable with the sample size. In other words, we have to deal with high-dimensional data for ranking. If we use the objective general index for such a data, a reliable estimator will

*The University of Tokyo
†University of Tsukuba

be required. The aim of this paper is to study consistency of the weight determined from a random sample.

A relevant but different method of determining a weight vector is the principal component analysis (PCA), consistency of which under high-dimensional settings has been studied in a vast literature (e.g. [2, 3, 4, 6, 7, 13]). Roughly speaking, PCA finds a majority direction of variables with ignoring minor variables, whereas OGI puts large weights to the minor variables as fair as possible. We also point out that PCA is orthogonally invariant whereas OGI is (coordinate-wise) scale invariant.

This paper is organized as follows. In Section 2, we formulate an estimation problem of OGI and introduce a loss function. In Section 3, we state the main theorem. Simulation results and application to genomic data are given in Section 4. We omit the proof of the theorems.

2 Problem setting

The objective general index is one of possible general indices for multivariate data. We consider a weighted sum $\boldsymbol{w}^\top \boldsymbol{x}$ of an observation $\boldsymbol{x} \in \mathbb{R}^p$ as a general index, where $\boldsymbol{w} \in \mathbb{R}^p$ is a weight vector. Each variate x_i is assumed to have a meaning that "larger is better" without loss of generality. Then it is natural to suppose that every coordinate of $\boldsymbol{w}$ is positive.

To state the definition of the objective general index, we first recall the diagonal scaling theorem established by Marshall and Olkin [5]. A symmetric matrix $\boldsymbol{A}$ is called *strictly copositive* if $\boldsymbol{v}^\top \boldsymbol{A}\boldsymbol{v} > 0$ for any $\boldsymbol{v} \in [0,\infty)^p \setminus \{\boldsymbol{0}\}$.

Lemma 1 ([5])**.** *Let $\boldsymbol{A}$ be a symmetric positive semi-definite and strictly copositive matrix. Then there exists a unique positive definite diagonal matrix $\boldsymbol{D}$ such that all the row sums (and column sums) of $\boldsymbol{DAD}$ are unity.*

The lemma implies that the following equation with respect to $\boldsymbol{v} \in \mathbb{R}^p_{>0}$ has a unique solution:

$$\boldsymbol{A}\boldsymbol{v} = \frac{\boldsymbol{1}}{\boldsymbol{v}}, \tag{1}$$

where $\boldsymbol{1} = \boldsymbol{1}_p = (1,\dots,1)^\top$ is the all-one vector and $\boldsymbol{1}/\boldsymbol{v}$ denotes the element-wise division. The vector $\boldsymbol{v}$ in (1) is a unique minimizer of a convex

function

$$\psi(\boldsymbol{v}) = \sum_{i=1}^{p}(-\log v_i) + \frac{1}{2}\boldsymbol{v}^\top \boldsymbol{A}\boldsymbol{v}, \quad \boldsymbol{v} \in \mathbb{R}^p_{>0}.$$

Hence $\boldsymbol{v}$ is numerically obtained by generic optimization packages.

Let $n > p$ and consider a random sample

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_{(1)}^\top \\ \vdots \\ \boldsymbol{x}_{(n)}^\top \end{pmatrix} \in \mathbb{R}^{n\times p}$$

according to the multivariate normal distribution $N(\boldsymbol{0}, \boldsymbol{\Sigma})$ with covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$. Denote the sample covariance matrix by $\boldsymbol{S} = n^{-1}\sum_{t=1}^{n} \boldsymbol{x}_{(t)}\boldsymbol{x}_{(t)}^\top$.

Definition 1 ([8])**.** The *objective weight* $\boldsymbol{w}$ is defined by a solution of

$$\boldsymbol{\Sigma}\boldsymbol{w} = \frac{\boldsymbol{1}}{\boldsymbol{w}}, \quad \boldsymbol{w} \in \mathbb{R}^p_{>0}. \tag{2}$$

Similarly, *the sample objective weight* $\hat{\boldsymbol{w}}$ is defined by

$$\boldsymbol{S}\hat{\boldsymbol{w}} = \frac{\boldsymbol{1}}{\hat{\boldsymbol{w}}}, \quad \hat{\boldsymbol{w}} \in \mathbb{R}^p_{>0}. \tag{3}$$

The weighted sum $\boldsymbol{w}^\top\boldsymbol{x}$ of an observation $\boldsymbol{x} \in \mathbb{R}^p$ using the objective weight $\boldsymbol{w}$ is called *the objective general index* (OGI).

The OGI is determined by the objective weight $\boldsymbol{w}$. Therefore we sometimes call $\boldsymbol{w}$ itself the OGI in this paper.

Our problem is to find conditions on which $\hat{\boldsymbol{w}}$ is a consistent estimator of $\boldsymbol{w}$. The loss function we adopt is

$$l(\boldsymbol{\Sigma}, \hat{\boldsymbol{w}}) = \left\| \frac{\hat{\boldsymbol{w}}}{\boldsymbol{w}} - \boldsymbol{1} \right\|_2, \tag{4}$$

where $\boldsymbol{w}$ is determined by (2). The estimator $\hat{\boldsymbol{w}}$ is said to be consistent if $l(\boldsymbol{\Sigma}, \hat{\boldsymbol{w}}) \to 0$ in probability as $n \to \infty$.

At first glance, the loss function in (4) seems to be strange. But it is not unnatural from the viewpoint of invariance with respect to scale transformation. More specifically, consider a scale transformation $\boldsymbol{x}_{(t)} \mapsto \boldsymbol{D}_{\boldsymbol{a}}\boldsymbol{x}_{(t)}$ for each observation $\boldsymbol{x}_{(t)} \in \mathbb{R}^p$, where $\boldsymbol{a}$ is a positive vector and $\boldsymbol{D}_{\boldsymbol{a}}$ is the diagonal matrix with diagonal part $\boldsymbol{a}$. Then the population covariance matrix

$\boldsymbol{\Sigma}$ is transformed into $\boldsymbol{D_a \Sigma D_a}$. Likewise, $\boldsymbol{S}$ is transformed into $\boldsymbol{D_a S D_a}$. Then the weights $\boldsymbol{w}$ and $\hat{\boldsymbol{w}}$ determined by (2) and (3) are transformed into $\boldsymbol{w}/\boldsymbol{a}$ and $\hat{\boldsymbol{w}}/\boldsymbol{a}$, respectively. Hence the ratio $\hat{\boldsymbol{w}}/\boldsymbol{w}$ is scale invariant.

Since the distribution of the loss function is scale invariant, we can assume $\boldsymbol{w} = \mathbf{1}$, or equivalently,

$$\boldsymbol{\Sigma}\mathbf{1} = \mathbf{1} \tag{5}$$

without loss of generality. This argument will be used in the next section. We refer to the equation (5) as the *equisum* property according to [1].

3 Main result

We consider a high-dimensional setting in that the dimension p grows with the sample size n and the covariance matrix $\boldsymbol{\Sigma}$ changes with n while keeping the equisum property (5). We first state a result on weak consistency.

Theorem 1. *Suppose that* $\boldsymbol{\Sigma}\mathbf{1} = \mathbf{1}$. *If*

$$\frac{p}{n} \operatorname{tr}(\boldsymbol{\Sigma}) \to 0 \tag{6}$$

as $n \to \infty$, *then* $\hat{\boldsymbol{w}}$ *is weakly consistent in the sense that* $\|\hat{\boldsymbol{w}} - \mathbf{1}\|_2$ *converges to* 0 *in probability.*

Now we state the main theorem. Denote the entries of $\boldsymbol{\Sigma}$ as $(\sigma_{ij})_{i,j=1}^{p}$.

Theorem 2. *Suppose that* $\boldsymbol{\Sigma}\mathbf{1} = \mathbf{1}$. *Then there exists a constant* $C > 0$ *such that*

$$\mathrm{P}(\|\hat{\boldsymbol{w}} - \mathbf{1}\| \geq \varepsilon) \leq 4p \exp\left(-\frac{nC\varepsilon^2}{(\max_i \sigma_{ii})p^2}\right) \tag{7}$$

for any $\varepsilon > 0$ *and any* $n \geq n_0$ *with some* $n_0 = n_0(\varepsilon)$. *In particular, if*

$$\max_i \sigma_{ii} = \mathrm{O}(1) \tag{8}$$

and

$$\frac{p^2 \log p}{n} = \mathrm{o}(1) \tag{9}$$

as $n \to \infty$, *then* $\hat{\boldsymbol{w}}$ *is strongly consistent in the sense that* $\|\hat{\boldsymbol{w}} - \mathbf{1}\|_2$ *converges to* 0 *almost surely.*

The condition (8) is satisfied for many cases. For example, if $\mathbf{\Sigma}$ is the identity matrix, then $\sigma_{ii} = 1$. On the other hand, if $\mathbf{\Sigma}$ is a degenerated matrix $\mathbf{\Sigma} = \mathbf{1}\mathbf{1}^\top/p$, then $\sigma_{ii} = 1/p$. In the latter example, the condition (9) can be further weakened to $(p \log p)/n = \mathrm{o}(1)$.

Note that (8) and (9) imply (6) from the fact that

$$\mathrm{tr}(\mathbf{\Sigma}) = \sum_{i=1}^{p} \sigma_{ii} \leq p \max_i \sigma_{ii}.$$

The factor $(\max_i \sigma_{ii})^{-1}$ on the right hand side of (7) is not arbitrarily large since the following lemma holds.

Lemma 2. *If $\mathbf{\Sigma}\mathbf{1} = \mathbf{1}$, then $\sigma_{ii} \geq 1/p$ for any i.*

Proof. Since $\mathbf{1}$ is an eigenvalue of $\mathbf{\Sigma}$, the spectral decomposition of $\mathbf{\Sigma}$ is

$$\mathbf{\Sigma} = \frac{\mathbf{1}\mathbf{1}^\top}{p} + \sum_{j=2}^{p} \lambda_j \boldsymbol{v}_j \boldsymbol{v}_j^\top,$$

where $\lambda_2 \geq \ldots \geq \lambda_p > 0$ are positive numbers (which are not necessarily less than one) and $\boldsymbol{v}_2, \ldots, \boldsymbol{v}_p$ are vectors orthogonal to $\mathbf{1}$. Denoting $\boldsymbol{v}_j = (v_{ij})_{i=1}^p$, we have

$$p\sigma_{ii} - 1 = \sum_{j=2}^{p} p\lambda_j v_{ij}^2 \geq \sum_{j=2}^{p} p\lambda_p v_{ij}^2 = (p-1)\lambda_p > 0,$$

where the last equality follows from $\sum_{j=2}^{p} \boldsymbol{v}_j \boldsymbol{v}_j^\top = \boldsymbol{I}_p - (\mathbf{1}\mathbf{1}^\top/p)$. □

The proof also shows that σ_{ii} can be arbitrarily large since λ_p can take any positive value.

In Section 4, we conduct numerical experiments to support the main result.

4 Numerical results

4.1 Simulation

We numerically demonstrate that the OGI is estimated with enough accuracy if $(p^2 \log p)/n$ is small. Figure 1 shows that the mean L_2 norm of

$\hat{\boldsymbol{w}} - \mathbf{1}$ against $n/(p^2 \log p)$, where the true covariance matrix is the first-order autoregressive model

$$(\sigma_{ij}) = \rho^{|i-j|}.$$

The covariance matrix is not equisum in the sense of Eq. (5). Therefore, in advance of the experiments, we applied a diagonal transformation to $\boldsymbol{\Sigma}$ in order that Eq. (5) is satisfied. The examined pairs of (p, n) are

$$p \in \{2^2, 2^3, \ldots, 2^7\} \quad \text{and} \quad n \in \{p, 2p, \ldots, 2^{10}\}.$$

The number of experiments for each pair (p, n) is 10^3.

From the figure, the mean L_2 norm decreases as $n/(p^2 \log p)$ becomes large. We also observe that the norm is smaller if ρ is closer to 1. This is consistent with the factor $\max_i \sigma_{ii}$ in Theorem 2. Indeed, the equisum matrix $\tilde{\boldsymbol{\Sigma}} = (\tilde{\sigma}_{ij})$ corresponding to $\boldsymbol{\Sigma}$ is approximately

$$\tilde{\sigma}_{ij} \asymp (1 - \rho)\rho^{|i-j|}$$

as $p \to \infty$ due to $\sum_{k=0}^{\infty} \rho^k = 1/(1 - \rho)$, and hence $\max_i \tilde{\sigma}_{ii}$ is reduced if ρ tends to 1.

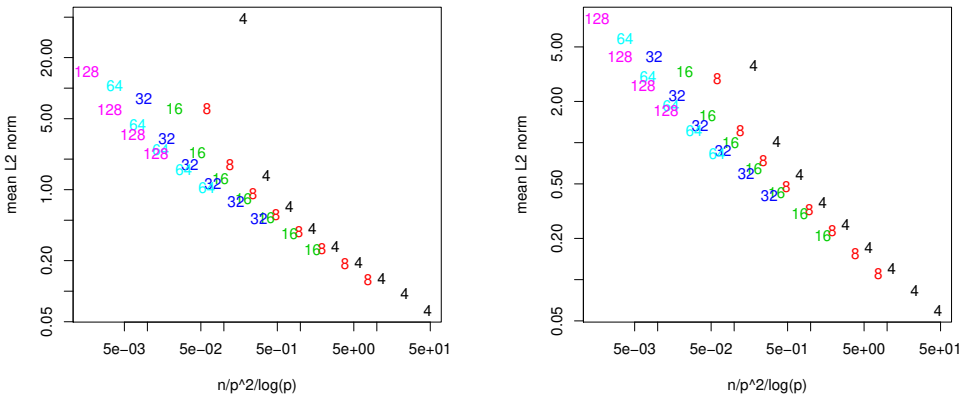
4.2 Real data

We computed the objective weight of a microarray data provided by [12]. The result suggests a variable selection method for unsupervised data. Details will be provided in the presentation.

5 Discussion

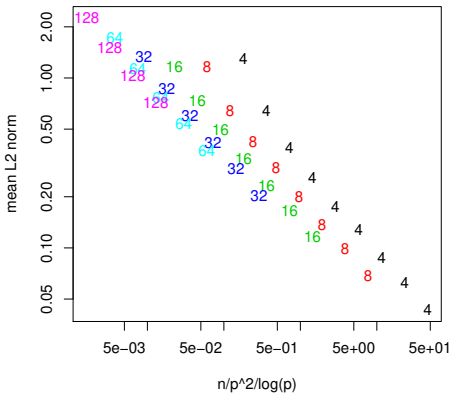
In this research, we obtained a consistency result of objective general index when a condition $p^2 \log p/n \to 0$ is satisfied.

A future work is to confirm whether the rate is the best possible or not. In [9], inconsistency is numerically observed if p/n tends to a positive number. In that paper, a limiting form of $\hat{\boldsymbol{w}}$ is conjectured via the replica method developed in statistical physics.

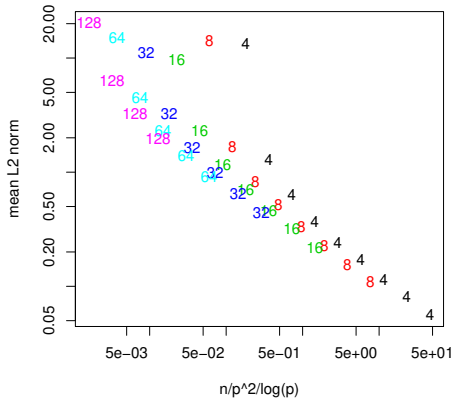


(a) $\rho = 0$

(b) $\rho = 0.5$



(c) $\rho = 0.9$



(d) $\rho = -0.9$

Figure 1: Average of $\|\hat{\boldsymbol{w}} - \mathbf{1}\|$ in 10^3 experiments for AR(1) models with the autoregression parameter ρ. The horizontal axis denotes $n/(p^2 \log p)$. The number and color on each point indicates p.

References

[1] Davis, P. J. and Najfeld, I. (2000). Equisum matrices and their permanence, *Quart. Appl. Math.*, **58** (1), 151–169.

[2] Johnstone, I. M. and Lu, A. Y. (2004). Sparse principal components analysis, Technical Report, Stanford University, Dept. of Statistics, arxiv:0901.4392.

[3] Johnstone, I. M. and Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions, *J. Amer. Statist. Assoc.*, **104**, 682–703.

[4] Jung, S. and Marron, J. S. (2009). PCA consistency in high-dimension, low sample size context, *Ann. Statist.*, **37**, 4104–4130.

[5] Marshall, A. W. and Olkin, I. (1968). Scaling of matrices to achieve specified row and column sums. *Numer. Math.*, **12**, 83–90.

[6] Nadler, B. (2008). Finite sample approximation results for principal component analysis: a matrix perturbation approach, *Ann. Statist.*, **36** (6), 2791–2817.

[7] Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model, *Statistica Sinica*, **17**, 1617–1642.

[8] Sei, T. (2016). An objective general index for multivariate ordered data, *J. Multivariate Anal.*, **147**, 247–264.

[9] Sei, T. (2018). Inconsistency of diagonal scaling under high-dimensional limit: a replica approach, Preprint, arXiv:1808.05781.

[10] Terashima, J., Japan Research Institute / Nihon Unisys Ltd. (eds.) (2016) *Happiness ranking of all the 47 prefectures in Japan, 2016* (in Japanese), Toyo Keizai.

[11] Times Higher Education, World University Rankings. https://www.timeshighereducation.com/

[12] Wille, A., Zimmermann, P., and Vranová, et al. (2004). Sparse graphical Gaussian modeling of the isoprenoid gene network in Arabidopsis thaliana, *Genome Biol.*, **5** (11), R92.

[13] Yata, K. and Aoshima, M. (2009). PCA consistency for non-Gaussian data in high-dimension, low sample size context, *Comm. Statist. Theory Methods*, **38**, 2634–2652.