Journal of Multivariate Analysis 147 (2016) 247–264

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

An objective general index for multivariate ordered data

Tomonari Sei

Department of Mathematical Informatics, Graduate School of Information Science and Technology, The University of Tokyo,
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

ARTICLE INFO

Article history:
Received 19 July 2015
Available online 12 February 2016

AMS subject classifications:
15B48
62H20
62H25

Keywords:
Convex minimization
Correlation
General index
Ordered data
Quantification
Ranking

ABSTRACT

Multivariate quantitative data are often summarized into a general index as a weighted sum when the variates have a prescribed order. Although the sum of standardized scores is a sensible choice of index, it may have negative correlation with some of the variates. In this paper, a general index that has positive correlation with all the variates is constructed. The index is applied to study the fairness of decathlon scoring. Quantification of ordered categorical data is also discussed. The limit of quantification characterizes the Gaussian distribution.

1. Introduction

Consider a data matrix of n individuals with p variates. For example, the data may be scores on p academic subjects of n students in a school, stock prices of p companies at n time points, or decathlon data of n athletes about $p = 10$ events.

Our purpose is to construct a general index that combines the p variates into a univariate index in order to rank the n individuals. The task is unsupervised in the sense that no one knows the correct index or ranking. This is a fairly classical and fundamental problem. For example, in the study of animal breeding programs, index selection is used to combine several traits without economic weights (e.g., [1,7]). In sports data analysis, it is discussed in the context how to score combined events, such as decathlons and heptathlons (e.g., [5]). A number of university ranking systems are based on the weighted average of relevant measures (e.g., [6]).

A natural index is the sum of standardized scores (Z-scores). This is a sensible choice if all the variates are uncorrelated. For correlated data, the first principal component is sometimes used as a general index, since it maximizes the variance of the index under weight constraints. However, these indices can have negative correlation with some of the variates; this property is undesirable, since a general index should reflect all of the traits.

In this paper, we show that there is a general index that is always positively correlated with each of the variates. We will call this the *objective general index*. A mathematical property of positive definite matrices plays an important role in its construction. The weight is numerically obtained by simple convex programming. As an example, we study the fairness of the scoring rule for a decathlon.

Our index is extended to the case of ordered categorical variates. This is related to but distinct from the optimal scaling method for ordered categories (e.g., [2,18]). Remarkably, it is shown that the limit of the quantified data is a Gaussian random variable.

E-mail address: sei@mist.i.u-tokyo.ac.jp.

http://dx.doi.org/10.1016/j.jmva.2016.02.005

The paper is organized as follows. In Section 2, we introduce two conditions on general indices. Then, in Section 3, we recall a key property of positive definite matrices. The objective general index is defined in Section 4. Two examples are given in Section 5. We consider multicollinearity and subjective importance in Section 6 and ordered categorical data in Section 7. Section 8 is devoted to a functional version of the index, that characterizes the Gaussian distribution. Finally, we discuss our findings in Section 9.

2. Two conditions on general indices

Let $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p) \in \mathbb{R}^{n\times p}$ be a p-variate data matrix, where each $\boldsymbol{x}_i$ is a column vector representing scores of n individuals. Assume that each variate is centered:

$$\boldsymbol{1}_n^\top \boldsymbol{x}_i = 0,$$

where $\boldsymbol{1}_n$ is the vector $(1, \ldots, 1)^\top \in \mathbb{R}^n$, and the symbol $^\top$ denotes the vector/matrix transpose. We further assume that each variate has a prescribed order such that a larger value is associated with a better evaluation. For example, for decathlon data, the sign of the time for 100 m must be changed before analysis.

Let the covariance matrix of $\boldsymbol{X}$ be

$$\boldsymbol{S} = (S_{ij})_{i,j=1}^p, \qquad S_{ij} = \frac{1}{n}\boldsymbol{x}_i^\top \boldsymbol{x}_j.$$

Suppose that $\boldsymbol{S}$ is positive definite. This assumption will be relaxed in Section 6. We often, but not always, standardize the data in advance, such that $S_{ii} = 1$ for any i. For standardized data, $\boldsymbol{S}$ is equal to the correlation matrix.

A general index of $\boldsymbol{X}$ is a linear combination of p variates:

$$\boldsymbol{g} = \sum_{i=1}^p w_i \boldsymbol{x}_i = \boldsymbol{X}\boldsymbol{w},$$

where $\boldsymbol{w} = \boldsymbol{w}(\boldsymbol{S}) = (w_1, \ldots, w_p)^\top \in \mathbb{R}^p$ is a weight vector that depends on $\boldsymbol{S}$. The map $\boldsymbol{S} \mapsto \boldsymbol{w}(\boldsymbol{S})$ is called a *weight map*. A weight map determines a general index.

The most fundamental general index is the simple sum

$$\sum_{i=1}^p \boldsymbol{x}_i,$$

whose weight map is $\boldsymbol{w}(\boldsymbol{S}) = \boldsymbol{1}_p = (1, \ldots, 1)^\top$, independent of $\boldsymbol{S}$. If the columns of $\boldsymbol{X}$ have different units, then the sum of Z-scores

$$\sum_{i=1}^p \frac{\boldsymbol{x}_i}{\sqrt{S_{ii}}}$$

is more sensible.

Another example is the first principal component, where the weight map $\boldsymbol{w}$ is an eigenvector of the covariance matrix $\boldsymbol{S}$ with respect to the largest eigenvalue. This maximizes the variance of $\boldsymbol{X}\boldsymbol{w}$ under a given $\boldsymbol{w}^\top\boldsymbol{w}$.

Our purpose is to construct a general index that is as fair as possible. Consider the following two conditions. For a vector $\boldsymbol{a} = (a_i)_{i=1}^p$, $\boldsymbol{a} > \boldsymbol{0}$ denotes that $a_i > 0$ for every i.

Definition 1. A weight map $\boldsymbol{w} = \boldsymbol{w}(\boldsymbol{S})$ is said to be *consistent* if $\boldsymbol{w} > \boldsymbol{0}$ for any $\boldsymbol{S}$. It is said to be *covariance consistent* if the covariance between each variate and the general index is positive, or equivalently $\boldsymbol{S}\boldsymbol{w} > \boldsymbol{0}$, for any $\boldsymbol{S}$.

Consistency is a natural requirement for a general index: if an individual A is better than B in all the variates, then the general index of A should be better than that of B. In contrast, the meaning of covariance consistency is not trivial, but at least this implies that none of the variates has negative correlation with the index. These two conditions have a duality relation (see Appendix B).

The weight map $\boldsymbol{w} = \boldsymbol{1}_p$ is obviously consistent, but it does not have covariance consistency if $p \geq 3$. For example, consider a positive definite matrix

$$\boldsymbol{S} = \begin{pmatrix} 1 & -7/12 & -7/12 \\ -7/12 & 1 & 0 \\ -7/12 & 0 & 1 \end{pmatrix}, \tag{1}$$

whose eigenvalues are 1 and $1 \pm (7/12)\sqrt{2}$. In this case, covariance consistency fails:

$$\boldsymbol{S}\,\boldsymbol{1}_3 = \begin{pmatrix} -1/6 \\ 5/12 \\ 5/12 \end{pmatrix}.$$

The first principal component has neither consistency nor covariance consistency. For example, consider

$$\boldsymbol{S} = \begin{pmatrix} 1 & -1/2 \\ -1/2 & 1 \end{pmatrix}.$$