Universal Journal of Educational Research 3(6): 402-406, 2015 http://www.hrpub.org
DOI: 10.13189/ujer.2015.030608

Inferences of Coordinates in Multidimensional Scaling by a Bootstrapping Procedure in R

Donghoh Kim[1,*], Se-Kang Kim[2], Soyeon Park[3]

[1]Department of Applied Mathematics, Sejong University, Korea
[2]Department of Psychology, Fordham University, United States
[3]Department of Child and Adolescent Development, San Francisco State University, United States

Abstract Recently, MDS has been utilized to identify and evaluate cognitive ability latent profiles in a population. However, dimension coordinates do not carry any statistical properties. To cope with statistical incompetence of MDS, we investigated the common aspects of various studies utilizing bootstrapping, and provided an R function for its implementation.

Keywords Cognitive Ability Test, Multidimensional Scaling (MDS), Bootstrap Empirical Distribution, Confidence Intervals, Latent Profile

1. Introduction

In recent decades, multidimensional scaling (MDS) has been utilized by several authors [1-3] to identify and evaluate cognitive ability latent profiles in a population. However, dimension coordinates do not carry any statistical properties (e.g., standard errors to test statistical significance of the coordinates). Therefore, interpretation of the dimension coordinates could be arbitrary and misleading. To cope with statistical incompetence for the MDS, several authors [2,4-6] proposed methods utilizing bootstrapping [7]. In this paper, we investigated the common aspects of various studies which are generation of the empirical distribution and its statistical inferential procedure based on the bootstrapping method. We also provided an R function available for public for its implementation in the R platform. It would contribute to enhancing the analytical methodology for educational and behavioral scientists who want to identify latent profile patterns utilizing the bootstrapping approach embedded in MDS.

Typically, scores of individuals in education studies and psychology consist of the several subtest scores. For personality assessment [4,6], cognitive ability [1-3], and educational measurements [8], transformed scores rather than the original subtest scores are utilized to identify distinct characteristics of individuals. The multidimensional scaling (MDS) model has been adapted to transform the original scale of scores through the MDS coordinates. These coordinates reflect distinct patterns of latent profiles which encapsulate all possible observed score profiles of individuals in a population. However, statistical inferences for such coordinates are not provided. Therefore, interpretation of observed scores of individuals could be arbitrary and misleading.

The bootstrap method produces the empirical distribution of the coordinates, which allows its inference such as estimation of standard errors and empirical confidence intervals for the MDS coordinates whereas conventional MDS does not provide any statistical inference for the coordinates. Consider an MDS model as follows. Let $\mathbf{X}_j=(X_{1j},...,X_{ij},...,X_{nj})'$ be the measurement on subtest j $(j=1,...,p)$ for n persons. MDS identifies the common dimension through dimension coordinates $\mathbf{C}_{p\times q}$ $(q<p)$. That is, the common dimension $\mathbf{Y}$ is

$$\mathbf{Y}=\mathbf{XC}$$

where $\mathbf{X}=(\mathbf{X}_1,...,\mathbf{X}_t,...,\mathbf{X}_p)$.

The bootstrapping procedure consists of three steps. In the first step, bootstrap samples are generated at random from given observations that construct a finite population. That is, we randomly select rows from the observation matrix $\mathbf{X}$. Let $\mathbf{X}^b$ $(b=1,...,B)$ be the b^th^ bootstrap sample. Each bootstrap sample inherits original characteristics of the population under study. In the second step, each bootstrap sample $\mathbf{X}^b$ is analyzed by a MDS to obtain the bootstrap MDS coordinates $\mathbf{C}^b$. In the third step, the MDS coordinates $\mathbf{C}^b$ $(b=1,...,B)$ are aggregated to form the empirical distribution of the coordinates. Based on the empirical distribution of the coordinates, a statistical inference about the coordinates can be made. We provide the R function “BootMDS” that implements the bootstrapping procedure to construct the empirical coordinate distributions.

Adapting the bootstrapping procedure with “BootMDS”, researchers can (1) generate of the empirical distribution of the MDS coordinates, (2) estimate various statistics based on the empirical distribution, (3) perform statistical inferential procedure of the coordinates, and (4) plot the various aspects of the empirical distribution of the coordinates.

2. Availability

The “BootMDS” is written with R. R is a statistical system freely available at CRAN (Comprehensive R Archive Network) from the website http://CRAN.r-project.org/, and works under Windows, Linux, and MacOS platforms. Note that the R package “smacof” should be preinstalled for the implementation of “BootMDS”. The source code of “BootMDS” function and cognitive ability test data “wj7.txt” are available for free from the author’s website http://dasan.sejong.ac.kr/~dhkim/BootMDS.html. For a step-by-step tutorial one may download sample code and data from the author’s website.

3. Empirical Analysis

We analyze “wj7” data and provide the potential applications of the proposed bootstrapping procedure through R implementation. The “wj7” data consists of the seven Woodcock-Johnson III cognitive ability tests [9]. The seven cognitive ability tests are Comprehension Knowledge (CK), Long-term Memory (LT), Visual-Spatial Thinking (VS), Auditory Processing (AP), Fluid Reasoning (FR), Processing Speed (PS), and Short-Term Memory (ST). Here, we load R functions for implementation, read the data file in R as the matrix of observation.

```
#### Load the R function.
source("BootMDS.R")
#### Reading data
testdata <- read.table(file="wj7.txt", header=TRUE)
# matrix of observation.
```

To run “BootMDS”, we need to specify seven input arguments as follows.

```
BootMDS(x, mds=c("smacof", "classical"),
distance=c("euclid", "sqeuclid"),
scale=TRUE, nBoot=2000, nprofile=3,
cl=0.95)
```

1. x : matrix of observation whose column is each test and row represents subject.
2. mds : specify the MDS method, either "smacof" or "classical". For "smacof", smacof package is required from R.
3. distance : specify the distance measure to be used. This must be one of "euclid" for Euclidean distance or "sqeuclid" for squared Euclidean distance.
4. scale : specify whether the observation is standardized or not.
5. nBoot : specify the number of bootstrap samples.
6. nprofile : specify the number of dimension or profiles.
7. cl : specify the empirical confidence level of the interval estimation of a profile.

The “BootMDS” function generates bootstrap samples at random from a given observations, analyzes each bootstrap sample by two MDS scaling methods (“smacof” or “classical” which is a metric scaling), and aggregates the MDS coordinates to form the empirical distribution of the coordinates. In this study, 2,000 bootstrap samples are generated from the original sample, and these samples are analyzed by the “smacof” procedure with squared Euclidean distance (and can also be analyzed by “classical”, although we did not include the results here). Then, three profiles are aggregated to form the empirical distribution of each profile. The following R code implements this procedure for obtaining the empirical distribution of each profile and 95% confidence interval for MDS coordinates, and saves its result to “empprofile” object. To duplicate its result, we set the random number seed fixed.

```
#### Generating empirical distribution of
#### profiles from observation by bootstrapping
library(smacof) # load smacof package
set.seed(1) # To duplicate the result set the seed as 1.
empprofile <- BootMDS(x=testdata,
mds="smacof", distance="sqeuclid",
scale=FALSE, nBoot=2000, nprofile=3,
cl=0.95)
```

The “BootMDS” function produces the empirical distribution of the profiles as well as a statistical summary of each profile. The output arguments of the “BootMDS” function are as follows:

1. stress : provide stress value of MDS.
2. profile : list of bootstrap samples of each dimension profile. The i^{th} list is the matrix of the i^{th} profile. The row represents bootstrap samples of i^{th} profile.
3. summary : list of summary statistics of bootstrap profiles. The i^{th} list is the dataframe of the summary statistics of the i^{th} dimension profile including original profile (Ori), standard error (SE), mean (Mean), lower bound and upper bound of confidence interval (Lower and Upper) and width of confidence interval (WD).
4. testname : the coordinate names.

Based on the empirical distribution of each profile, a statistical inference can be made. We can calculate standard errors to test statistical significance of the coordinates and produce confidence interval of the coordinates. For example, the “empprofile” object by the “BootMDS” function includes the bootstrap samples of the dimension profiles and its summary statistics. The summary statistics of the i^{th} dimension profile can be deduced from summary[[i]] of the "empprofile" object. For example, Table 1 of the summary statistics for the first dimension profile can be deduced from the "empprofile" object by the following R command.

```
#### The summary statistics for each
#### dimension profile
i <- 1    # for the first dimension profile
#i <- 2 # for the second dimension profile
#i <- 3 # for the third dimension profile
empprofile$summary[[i]]
```

Table 1. Summary statistics for the first dimension profile.

	Ori	SE	Mean	Lower	Upper	WD
CK	0.0056	0.2266	-0.0085	-0.4663	0.4679	0.9341
LT	0.1714	0.0573	0.1453	0.0113	0.2489	0.2377
VS	0.8090	0.1874	0.7165	0.1225	0.8677	0.7452
AP	-0.5782	0.1429	-0.5041	-0.6645	-0.0629	0.6016
FR	0.1925	0.1245	0.1629	-0.1565	0.3649	0.5215
PS	-0.2032	0.3768	-0.1671	-0.8471	0.6704	1.5175
ST	-0.3971	0.1280	-0.3450	-0.5371	0.0106	0.5477

Note: CK = Comprehension Knowledge; LT = Long-term Memory; VS = Visual-Spatial Thinking; AP = Auditory Processing; FR = Fluid Reasoning; PS = Processing Speed; ST = Short-Term Memory

Note that "Ori" is the coordinates estimated from the original sample, "SE" is the bootstrap standard errors of the coordinates, "Mean" is mean coordinates from 2000 bootstrap replicates, "Upper" and "Lower" is lower and upper bound of 95% empirical confidence interval, and "WD" is the width of the confidence interval. Based on the bootstrap empirical confidence interval, we can infer that the LT, VS, and AP coordinates are significant for constructing the first dimension profile, since they did not include zeros in their empirical confidence intervals.

From the empirical distribution of each profile, we can obtain graphical summary of the bootstrap empirical confidence interval of each dimension profile or each coordinate. These procedures can be implemented by the function "figureMDS" with four input arguments as follows.

```
figureMDS(result, type=c("ci", "hist"), dimension,
coordinates)
```

1. result : the object by the "BootMDS" function.
2. type : specify the figure type, either "ci" for bootstrap empirical confidence interval of dimension profile or "hist" for histogram and Q-Q plot of coordinates.
3. dimension : specify the dimension of profile.
4. coordinates : specify the coordinates.

Figure 1 illustrates the bootstrap empirical confidence interval of the first dimension profile. The solid line is the original dimension profile, the dotted line is the mean profile of the empirical distribution by bootstrapping, and the gray band is the 95% confidence interval. The codes for Figure 1 are:

```
#### Figure of empirical confidence
#### interval for Profile
i <- 1 # the first dimension profile
figureMDS(empprofile, type="ci", dimension=i)
```

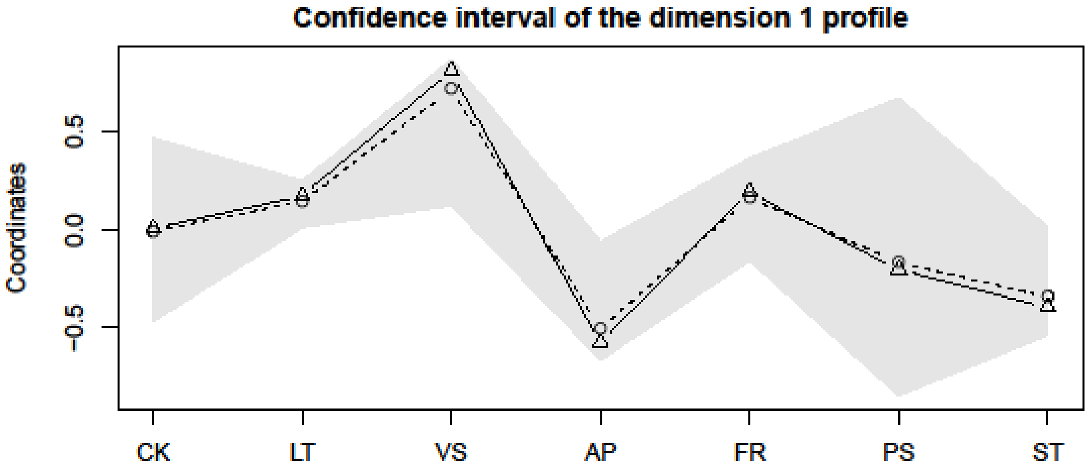

Figure 1. The confidence interval of the first dimension profile

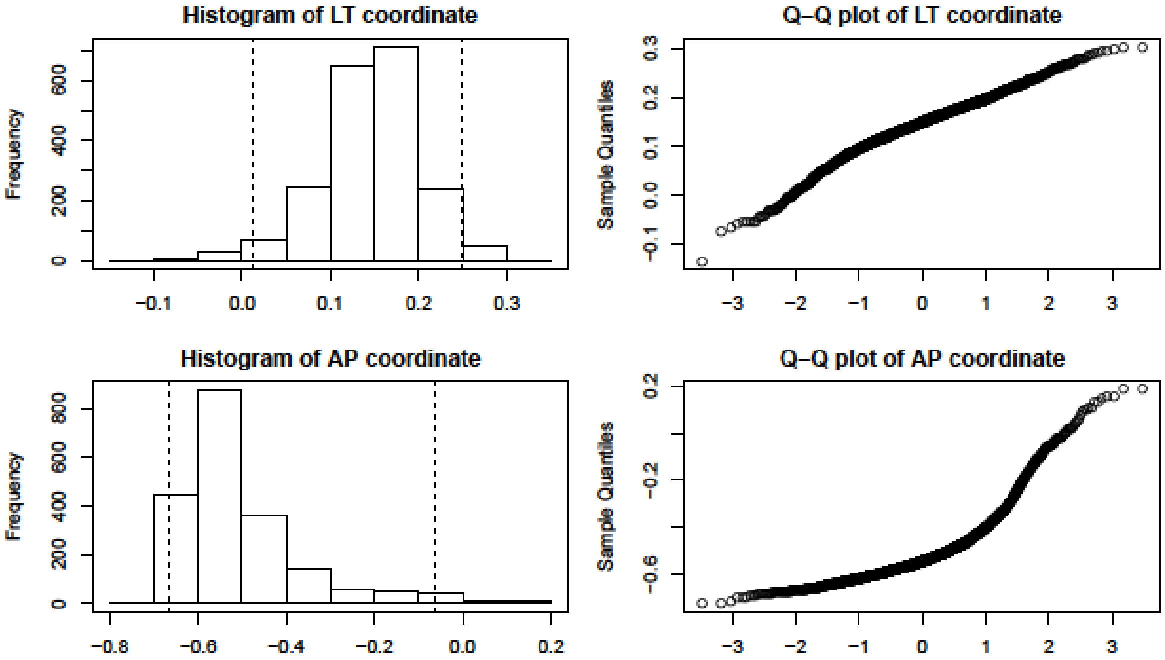

Figure 2. The histogram and Q-Q plot of LT and AP coordinate of the first dimension profile

From the bootstrap samples, we can draw the empirical distribution of each dimension profile, and conduct any statistical procedure. For example, the bootstrap sample of the jth coordinate of the ith dimension profile can be deduced from profile[[i]][,j] of the "empprofile" object. Figure 2 describes the histogram and Q-Q plot for LT and AP coordinate of the first profile, which implies that the distribution of AP coordinate is right-skewed. The dotted line of histogram in Figure 2 indicates the 95% empirical confidence interval of each coordinate. The LT and AP coordinates are significant for the first profile since the 95% bootstrap interval does not include zero. In this way, the distributional characteristics of the MDS coordinates can be discovered and statistical inferences can be made. The following codes produce Figure 2.

```
#### Histogram and qqplot of the bootstrap
#### sample of the j-th coordinates of
#### the i-th profile
i <- 1    # specify the i-th profile.
j1 <- 2 # specify the j1-th coordinate of the i-th profile,
           #LT in this example.
j2 <- 4 # specify the j2-th coordinate of the i-th profile,
           #AP in this example.
figureMDS(empprofile, type="hist", dimension=i,
coordinates=c(j1, j2))
```

4. Concluding Remarks

We utilized the bootstrap method to produce the empirical distribution of the coordinates for a statistical inference. Based on the bootstrap empirical distribution, we conducted statistical inference of empirical confidence intervals of the MDS coordinates for cognitive ability test data. As the analysis results show, three cognitive ability clusters, LT, VS, and AP were statistically significant. Researchers and practitioners can adapt, modify, and extend the bootstrap procedure for their own purpose. Using the R function "BootMDS" for the current study, one can be ready for applying bootstrapping method to real data. It can be easily accessible from the author's website. We hope that the implementation of the bootstrap procedure for constructing the empirical distribution of the MDS coordinates helps researchers conduct a statistical inference of the dimension coordinates and interpret the coordinates for their real applications.

Acknowledgements

This research of D. Kim was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(2012R1A1A2039968).

REFERENCES

[1] M. L. Davison, M. Gasser, S. Ding. Identifying major profile patterns in a population: An exploratory study of WAIS and GATB pattern, Psychological Assessment, Vol. 8, No. 1, 26–31, 1996.

[2] S.-K. Kim, C. L. Frisby, M. L. Davison. Estimating cognitive profiles using profile analysis via multidimensional scaling (PAMS), Multivariate Behavioral Research, Vol. 39, No. 4, 595–624, 2004.

[3] C. L. Frisby, S.-K. Kim. Using profile analysis via multidimensional scaling (PAMS) to identify core profiles from the WMS-III, Psychological Assessment, Vol. 20, No. 1, 1–9, 2008.

[4] C. S. Ding. Determining the significance of scale values from multidimensional scaling profile analysis using a re-sampling method, Behavior Research Methods, Instruments, & Computers, Vol. 37, No. 1, 37–47, 2005.

[5] S.-K. Kim. Evaluating the invariance of cognitive profile patterns derived from profile analysis via multidimensional scaling (PAMS): A bootstrapping approach, Journal of International Testing, Vol. 10, No. 1, 33–46, 2010.

[6] S.-K. Kim, M. L. Davison, C. L. Frisby. Confirmatory factor analysis and profile analysis via multidimensional scaling, Multivariate Behavioral Research, Vol. 42, No. 1, 1–32, 2007.

[7] B. Efron, R. J. Tibshirani. An Introduction to the Bootstrap, New York, Chapman and Hall, 1993.

[8] C. S. Ding, M. L. Davison, A. C. Petersen. Multidimensional scaling analysis of growth and change, Journal of Educational Measurement, Vol. 42, No. 2, 171–191, 2005.

[9] R. W. Woodcock, K. S. McGrew, N. Mather. Woodcock-Johnson III Tests of Cognitive Abilities, Itasca, Riverside Publishing, 2001