Open Journal of Statistics, 2015, 5, 165-172
Published Online April 2015 in SciRes. http://www.scirp.org/journal/ojs
http://dx.doi.org/10.4236/ojs.2015.53020

A New Algorithm for Generalized Least Squares Factor Analysis with a Majorization Technique

Kohei Adachi

Graduate School of Human Sciences, Osaka University, Osaka, Japan
Email: adachi@hus.osaka-u.ac.jp

Received 18 January 2015; accepted 22 April 2015; published 27 April 2015

Abstract

Factor analysis (FA) is a time-honored multivariate analysis procedure for exploring the factors underlying observed variables. In this paper, we propose a new algorithm for the generalized least squares (GLS) estimation in FA. In the algorithm, a majorization step and diagonal steps are alternately iterated until convergence is reached, where Kiers and ten Berge's (1992) majorization technique is used for the former step, and the latter ones are formulated as minimizing simple quadratic functions of diagonal matrices. This procedure is named a majorizing-diagonal (MD) algorithm. In contrast to the existing gradient approaches, differential calculus is not used and only elmentary matrix computations are required in the MD algorithm. A simuation study shows that the proposed MD algorithm recovers parameters better than the existing algorithms.

Keywords

Exploratory Factor Analysis, Generalized Least Squares Estimation, Matrix Computations, Majorization

1. Introduction

Using $\boldsymbol{y}$ for a $p\times 1$ observation vector whose expectation $E(\boldsymbol{y})$ equals the $p\times 1$ zero vector $\mathbf{0}_p$, the factor analysis (FA) model is expressed as

$$\boldsymbol{y} = \boldsymbol{\Lambda}\boldsymbol{x} + \boldsymbol{e} \tag{1}$$

with $\boldsymbol{\Lambda} = (\lambda_{ij})$ a p-variables × m-factors loading matrix, $\boldsymbol{x}$ an $m\times 1$ latent factor score vector, $\boldsymbol{e}$ a $p\times 1$ error vector, and $p > m$. The expectations for $\boldsymbol{x}$ and $\boldsymbol{e}$ are assumed to satisfy

How to cite this paper: Adachi, K. (2015) A New Algorithm for Generalized Least Squares Factor Analysis with a Majorization Technique. *Open Journal of Statistics*, **5**, 165-172. http://dx.doi.org/10.4236/ojs.2015.53020

$$E(\boldsymbol{x})=\mathbf{0}_m,\quad E(\boldsymbol{e})=\mathbf{0}_p,\quad E(\boldsymbol{xe}')=\boldsymbol{O},\quad E(\boldsymbol{xx}')=\boldsymbol{I}_m,\quad E(\boldsymbol{ee}')=\boldsymbol{\Psi} \tag{2}$$

Here, $\boldsymbol{O}$ is the $m\times p$ matrix of zeros, $\boldsymbol{I}_m$ is the $m\times m$ identity matrix, and $\boldsymbol{\Psi}$ is the $p\times p$ diagonal matrix whose diagonal elements are called unique variances. The FA model (1) with the assumptions in (2) imply that the covariance matrix $\boldsymbol{\Sigma}=E(\boldsymbol{yy}')$ is modeled as

$$\boldsymbol{\Sigma}=\boldsymbol{\Lambda}\boldsymbol{\Lambda}'+\boldsymbol{\Psi} \tag{3}$$

[1] [2]. A main purpose of FA is to estimate the parameter matrices $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ from the inter-variable sample covariance matrix $\boldsymbol{S}$ $(p\times p)$ corresponding to (3). Some authors classify FA as exploratory (EFA) or confirmatory (CFA) [2], where $\boldsymbol{\Lambda}$ is unconstrained in EFA, while some elements of $\boldsymbol{\Lambda}$ are constrained in CFA. In this paper, we refer to EFA simply as FA.

Three major approaches for the parameter estimation are least squares (LS), generalized least squares (GLS), and maximum likelihood (ML) procedures [3]. They differ in the definition of the loss function $f(\boldsymbol{\Theta})$ to be minimized over $\boldsymbol{\Theta}=[\boldsymbol{\Lambda},\boldsymbol{\Psi}]$. The functions for the LS and GLS estimation procedures are defined as $f(\boldsymbol{\Theta})=\mathrm{tr}(\boldsymbol{S}-\boldsymbol{\Sigma})^2$ and

$$f(\boldsymbol{\Theta})=\mathrm{tr}\left\{(\boldsymbol{S}-\boldsymbol{\Sigma})\boldsymbol{S}^{-1}\right\}^2=\mathrm{tr}\left\{\left[\boldsymbol{S}-(\boldsymbol{\Lambda}\boldsymbol{\Lambda}'+\boldsymbol{\Psi})\right]\boldsymbol{S}^{-1}\right\}^2 \tag{4}$$

respectively, while $f(\boldsymbol{\Theta})$ is defined as the negative of the log-likelihood derived under the normality assumption for $\boldsymbol{x}$ and $\boldsymbol{e}$ in the ML estimation [3] [4].

In all estimation procedures, iterative algorithms are needed for minimizing loss function $f(\boldsymbol{\Theta})$. They can be roughly classified into gradient and inequality-based algorithms. Here, the gradient ones refer to the algorithms using Newton and related methods [5], in which the partial differentiation of $f(\boldsymbol{\Theta})$ with respect to $\boldsymbol{\Theta}$ is used for updating it. On the other hand, the term "inequality-based algorithms" is not a popular one. We use the term for the algorithms, in which differentiation is not used and the inequality $f(\boldsymbol{\Theta})\geq f(\boldsymbol{\Theta}_{\mathrm{new}})$ underlies that which guarantees the weakly monotone decrease in the loss function value with updating $\boldsymbol{\Theta}$ to $\boldsymbol{\Theta}_{\mathrm{new}}$. Similar dichotomization of minimization methodology is also found in [6].

For all of the LS, GLS, and ML estimation, gradient algorithms have been developed: those with the Fletcher-Powell and Newton-Raphson methods have been proposed for the ML estimation [7] [8], while the algorithms using the Newton-Raphson and Gauss-Newton methods have been developed for GLS [9] [10] with the gradient algorithms for GLS also used for LS. On the other hand, inequality-based algorithms have been developed for the LS and ML estimation excluding GLS. Such an algorithm for LS is MINRES [11] in which $\boldsymbol{\Theta}$ is partitioned into the subsets of parameters with $\boldsymbol{\Theta}=\{\boldsymbol{\theta}_1,\cdots,\boldsymbol{\theta}_q\}$ and the minimization of $f(\boldsymbol{\Theta})$ over each subset $\boldsymbol{\theta}_i$ for $i=1,\cdots,q$ is iterated. The inequality-based one for the ML estimation is the EM algorithm for FA [12] in which $f(\boldsymbol{\Theta})$ decreases monotonically with the alternate iteration of so-called E- and M-steps [13]. A feature of MINRES and the EM algorithm is that only simple matrix computations such as the inversion of matrices are required and their computer-programs are easily formed. In contrast, the gradient algorithms require more complicated computations such as obtaining or numerically approximating the second derivatives of $f(\boldsymbol{\Theta})$.

As found in the above discussion, an inequality-based algorithm has not been developed for the GLS estimation in which (4) is minimized over $\boldsymbol{\Theta}$. To propose it is the purpose of this paper. The algorithm to be proposed is also computationally simple as in the existing inequality-based ones: only elementary matrix computations are required such as the inversion and singular value decomposition (SVD) of matrices. A feature of the proposed algorithm to be addressed is using majorization in one of steps. The majorization generally refers to a class of the techniques in which a *majorizing function* $h(\boldsymbol{\Phi},\boldsymbol{\Xi})$ is utilized for minimizing a function $g(\boldsymbol{\Phi})$ over $\boldsymbol{\Phi}$. Here, $h(\boldsymbol{\Phi},\boldsymbol{\Xi})$ satisfies $g(\boldsymbol{\Phi})=h(\boldsymbol{\Phi},\boldsymbol{\Phi})\geq h(\boldsymbol{\Phi}^*,\boldsymbol{\Phi})\geq g(\boldsymbol{\Phi}^*)$, with $\boldsymbol{\Phi}^*$ being the minimizer of the majorizing function $h(\boldsymbol{\Phi}^*,\boldsymbol{\Phi})$ for its latter argument matrix $\boldsymbol{\Phi}$ kept fixed [14]. It shows that $g(\boldsymbol{\Phi})$ decreases with the update of $\boldsymbol{\Phi}$ into $\boldsymbol{\Phi}^*$. As described in the next section, the step with a majorization technique and the steps for minimizing the functions of diagonal matrices form the algorithm to be presented. It is thus called majorizing-diagonal (MD) algorithm in this paper.

The MD algorithm is not the first one with majorization in FA. Indeed, the above EM algorithm [12] can be regarded as a majorization procedure with its majorizing function being the full log likelihood derived by supposing that latent factor scores in $\boldsymbol{x}$ were observed. [15] has also proposed an FA algorithm with a majorization technique. However, in that algorithm, the estimation of a new type [16] [17] is considered, which are different from the LS, GLS, and ML estimation treated as the major procedures in this paper: [15] is beyond the

scope of this paper.

The remaining parts of this paper are organized as follows: the MD algorithm is detailed in the next section, and it is illustrated with a real data set in Section 3. A simulation study for assessing the algorithm is reported in Section 4, which is followed by discussions.

2. Proposed Algorithm

We propose the MD algorithm for minimizing the GLS loss function (4) over the loadings in $\boldsymbol{\Lambda}$ $(p \times m)$ and the unique variances in the diagonal matrix $\boldsymbol{\Psi}$ $(p \times p)$. Here, it is supposed that the sample covariance matrix $\boldsymbol{S}$ is positive-definite and $\boldsymbol{\Lambda}$ is of full-column rank, *i.e.*, its rank is m with $p > m$. This supposition and the covariance matrix being modeled as (3) imply that, without loss of generality, we can reparameterize $\boldsymbol{\Lambda}$ as

$$\boldsymbol{\Lambda} = \boldsymbol{L}\boldsymbol{\Delta}^{1/2} \tag{5}$$

where $\boldsymbol{L}$ is a $p \times m$ matrix satisfying

$$\boldsymbol{L}'\boldsymbol{L} = \boldsymbol{I}_p \tag{6}$$

and $\boldsymbol{\Delta}$ is an $m \times m$ positive-definite diagonal matrix. By substituting (5) into the GLS loss function (4), it is rewritten as

$$\begin{aligned} f\left(\boldsymbol{L},\boldsymbol{\Delta},\boldsymbol{\Psi}\right) &= \mathrm{tr}\left\{\left[\boldsymbol{S}-\left(\boldsymbol{L}\boldsymbol{\Delta}\boldsymbol{L}'+\boldsymbol{\Psi}\right)\right]\boldsymbol{S}^{-1}\right\}^2 = \mathrm{tr}\left\{\boldsymbol{I}_p-\left(\boldsymbol{L}\boldsymbol{\Delta}\boldsymbol{L}'+\boldsymbol{\Psi}\right)\boldsymbol{S}^{-1}\right\}^2 \\ &= p-2\mathrm{tr}\boldsymbol{S}^{-1}\left(\boldsymbol{L}\boldsymbol{\Delta}\boldsymbol{L}'+\boldsymbol{\Psi}\right)+\mathrm{tr}\boldsymbol{S}^{-2}\left(\boldsymbol{L}\boldsymbol{\Delta}\boldsymbol{L}'+\boldsymbol{\Psi}\right)^2 \\ &= p-2\mathrm{tr}\boldsymbol{S}^{-1}\boldsymbol{L}\boldsymbol{\Delta}\boldsymbol{L}'-2\mathrm{tr}\boldsymbol{S}^{-1}\boldsymbol{\Psi}+\mathrm{tr}\boldsymbol{S}^{-2}\boldsymbol{L}\boldsymbol{\Delta}^2\boldsymbol{L}'+2\mathrm{tr}\boldsymbol{\Psi}\boldsymbol{S}^{-2}\boldsymbol{L}\boldsymbol{\Delta}\boldsymbol{L}'+\mathrm{tr}\boldsymbol{S}^{-2}\boldsymbol{\Psi}^2 \end{aligned} \tag{7}$$

This function is minimized over $\boldsymbol{L}$, $\boldsymbol{\Delta}$, and $\boldsymbol{\Psi}$ subject to (6) and the latter two matrices being diagonal ones, by alternately iterating the majorizing and diagonal steps described in the next subsections.

2.1. Majorization Step

Let us consider minimizing (7) over $\boldsymbol{L}$ subject to (6) while $\boldsymbol{\Delta}$ and $\boldsymbol{\Psi}$ are kept fixed. Summarizing the parts irrelevant to $\boldsymbol{L}$ in (7) into $c = p-2\mathrm{tr}\boldsymbol{S}^{-1}\boldsymbol{\Psi}+\mathrm{tr}\boldsymbol{S}^{-2}\boldsymbol{\Psi}^2$, the loss function (7) is rewritten as

$$f\left(\boldsymbol{L}\middle|\boldsymbol{\Delta},\boldsymbol{\Psi}\right) = 2\mathrm{tr}\left(\boldsymbol{\Psi}\boldsymbol{S}^{-2}-\boldsymbol{S}^{-1}\right)\boldsymbol{L}\boldsymbol{\Delta}\boldsymbol{L}'+\mathrm{tr}\boldsymbol{S}^{-2}\boldsymbol{L}\boldsymbol{\Delta}^2\boldsymbol{L}'+c\,. \tag{8}$$

Though the optimal $\boldsymbol{L}$ that minimizes (8) under (6) is not given explicitly, the solution can be obtained using Kiers and ten Berge's [18] majorization technique, whose earlier version is also found in [19]. This technique purposes to minimize a function expressed as the form $\phi\left(\boldsymbol{L}\right) = \mathrm{tr}\boldsymbol{A}\boldsymbol{L}+\sum_{j=1}^{q}\mathrm{tr}\boldsymbol{B}_j\boldsymbol{L}\boldsymbol{C}_j\boldsymbol{L}'+\text{constant}$. Comparing this with (8), we can find (8) to be a special case of the above $\phi\left(\boldsymbol{L}\right)$ with $\boldsymbol{A}$ being the zero matrix, $\boldsymbol{B}_1 = 2\left(\boldsymbol{\Psi}\boldsymbol{S}^{-2}-\boldsymbol{S}^{-1}\right)$, $\boldsymbol{B}_2 = \boldsymbol{S}^{-2}$, $\boldsymbol{C}_1 = \boldsymbol{\Delta}$, $\boldsymbol{C}_2 = \boldsymbol{\Delta}^2$, and $q = 2$. Therefore, the update formula in [18] (pp. 374-375) can be straightforwardly used for (8).

According to the formula, the update of $\boldsymbol{L}$ by

$$\boldsymbol{L} = \boldsymbol{P}\boldsymbol{Q}' \tag{9}$$

decreases the value of (8) with $f\left(\boldsymbol{L}\middle|\boldsymbol{\Delta},\boldsymbol{\Psi}\right) \le f\left(\boldsymbol{L}_{\mathrm{OLD}}\middle|\boldsymbol{\Delta},\boldsymbol{\Psi}\right)$. Here, $\boldsymbol{L}_{\mathrm{OLD}}$ stands for the matrix $\boldsymbol{L}$ before the update; $\boldsymbol{P}$ and $\boldsymbol{Q}$ are the column-orthonormal matrices that are obtained from the SVD defined as

$$2\left\{\left(\lambda_a-\lambda_b\right)\boldsymbol{I}_m-\boldsymbol{S}^{-2}\boldsymbol{\Psi}-\boldsymbol{\Psi}\boldsymbol{S}^{-2}+2\boldsymbol{S}^{-1}\right\}\boldsymbol{L}_{\mathrm{OLD}}\boldsymbol{\Delta}+2\left(\lambda_c-\boldsymbol{I}_m-\boldsymbol{S}^{-2}\right)\boldsymbol{L}_{\mathrm{OLD}}\boldsymbol{\Delta}^2 = \boldsymbol{P}\boldsymbol{\Theta}\boldsymbol{Q}' \tag{10}$$

with $\boldsymbol{\Theta}$ the diagonal matrix including the singular values of the matrix in the left-hand side, and λ_a, λ_b, and λ_c the largest eigenvalues of $\boldsymbol{\Psi}\boldsymbol{S}^{-2}+\boldsymbol{S}^{-2}\boldsymbol{\Psi}$, $2\boldsymbol{S}^{-1}$, and $\boldsymbol{S}^{-2}$, respectively.

2.2. Diagonal Steps

In this section, we describe updating each of diagonal matrices $\boldsymbol{\Delta}$ and $\boldsymbol{\Psi}$. First, let us consider minimizing the loss function (7) over $\boldsymbol{\Delta}$ with keeping $\boldsymbol{L}$ and $\boldsymbol{\Psi}$ fixed. Since the terms relevant to $\boldsymbol{\Delta}$ in the loss function (7) are the same as those relevant to $\boldsymbol{L}$, the expression (8) into which (7) is rewritten is to be noted again. By taking account of the fact that $\boldsymbol{\Delta}$ is a diagonal matrix, (8) can be rewritten as

$$f\left(\boldsymbol{\Delta}\middle|\boldsymbol{L},\boldsymbol{\Psi}\right)=\mathrm{tr}\left(\boldsymbol{L}'\boldsymbol{S}^{-2}\boldsymbol{L}\right)\boldsymbol{\Delta}^2-2\mathrm{tr}\boldsymbol{L}'\left(\boldsymbol{S}^{-1}-\boldsymbol{\Psi}\boldsymbol{S}^{-2}\right)\boldsymbol{L}\boldsymbol{\Delta}+c=\mathrm{tr}\boldsymbol{D}_1\boldsymbol{\Delta}^2-2\mathrm{tr}\boldsymbol{D}_2\boldsymbol{\Delta}+c \tag{11}$$

Here, $\boldsymbol{D}_1=\mathrm{diag}\left(\boldsymbol{L}'\boldsymbol{S}^{-2}\boldsymbol{L}\right)$ and $\boldsymbol{D}_2=\mathrm{diag}\left(\boldsymbol{L}'\left(\boldsymbol{S}^{-1}-\boldsymbol{\Psi}\boldsymbol{S}^{-2}\right)\boldsymbol{L}\right)$ with $\mathrm{diag}(\bullet)$ denoting the diagonal matrix whose diagonal elements are those of the parenthesized matrix. Further, we can rewrite (11) as $f\left(\boldsymbol{\Delta}\middle|\boldsymbol{L},\boldsymbol{\Psi}\right)=\boldsymbol{D}_1\left\|\boldsymbol{\Delta}-\boldsymbol{D}_1^{-1}\boldsymbol{D}_2\right\|^2+\mathrm{tr}\boldsymbol{D}_1^{-2}\boldsymbol{D}_2^2+c$ with $\|\bullet\|$ denoting the Frobenius norm. It shows that the function (11) is minimized for

$$\boldsymbol{\Delta}=\boldsymbol{D}_1^{-1}\boldsymbol{D}_2=\mathrm{diag}\left(\boldsymbol{L}'\boldsymbol{S}^{-2}\boldsymbol{L}\right)^{-1}\mathrm{diag}\left(\boldsymbol{L}'\left(\boldsymbol{S}^{-1}-\boldsymbol{\Psi}\boldsymbol{S}^{-2}\right)\boldsymbol{L}\right) \tag{12}$$

for fixed $\boldsymbol{L}$ and $\boldsymbol{\Psi}$.

Next, we consider minimizing (7) over $\boldsymbol{\Psi}$ with $\boldsymbol{L}$ and $\boldsymbol{\Delta}$ fixed. Summarizing the parts irrelevant to $\boldsymbol{\Psi}$ in (7) into $c^*=p-2\mathrm{tr}\boldsymbol{S}^{-1}\boldsymbol{L}\boldsymbol{\Delta}\boldsymbol{L}'+\mathrm{tr}\boldsymbol{S}^{-2}\boldsymbol{L}\boldsymbol{\Delta}^2\boldsymbol{L}'$ and using the fact of $\boldsymbol{\Psi}$ being a diagonal matrix, the loss function (7) can be rewritten as

$$\begin{aligned}f\left(\boldsymbol{\Psi}\middle|\boldsymbol{L},\boldsymbol{\Delta}\right)&=\mathrm{tr}\boldsymbol{S}^{-2}\boldsymbol{\Psi}^2-2\mathrm{tr}\left(\boldsymbol{S}^{-1}-\boldsymbol{S}^{-2}\boldsymbol{L}\boldsymbol{\Delta}\boldsymbol{L}'\right)\boldsymbol{\Psi}+c^*\\&=\mathrm{tr}\boldsymbol{D}_3\boldsymbol{\Psi}^2-2\mathrm{tr}\boldsymbol{D}_4\boldsymbol{\Psi}+c^*\\&=\boldsymbol{D}_3\left\|\boldsymbol{\Psi}-\boldsymbol{D}_3^{-1}\boldsymbol{D}_4\right\|^2+\mathrm{tr}\boldsymbol{D}_3^{-2}\boldsymbol{D}_4^2\end{aligned} \tag{13}$$

with $\boldsymbol{D}_3=\mathrm{diag}\left(\boldsymbol{S}^{-2}\right)$ and $\boldsymbol{D}_4=\mathrm{diag}\left(\boldsymbol{S}^{-1}-\boldsymbol{S}^{-2}\boldsymbol{L}\boldsymbol{\Delta}\boldsymbol{L}'\right)$. We can find that (13) is minimized for

$$\boldsymbol{\Psi}=\boldsymbol{D}_3^{-1}\boldsymbol{D}_4=\mathrm{diag}\left(\boldsymbol{S}^{-2}\right)^{-1}\mathrm{diag}\left(\boldsymbol{S}^{-1}-\boldsymbol{S}^{-2}\boldsymbol{L}\boldsymbol{\Delta}\boldsymbol{L}'\right) \tag{14}$$

for fixed $\boldsymbol{L}$ and $\boldsymbol{\Delta}$.

2.3. Whole Algorithm

The results in the last two subsections show that the proposed MD algorithm can be listed as follows:

Step 1. Initialize $\boldsymbol{L}$, $\boldsymbol{\Delta}$, and $\boldsymbol{\Psi}$.

Step 2. Update $\boldsymbol{L}$ with (9) l times.

Step 3. Update $\boldsymbol{\Delta}$ with (12).

Step 4. Update $\boldsymbol{\Psi}$ with (14).

Step 5. Finish with $\boldsymbol{\Lambda}$ set to (5) if convergence is reached; otherwise, return to Step 2.

It should be noted in Step 2 that the update of $\boldsymbol{L}$ by (9) does not minimize (7) but only decreases its value, which implies that that update can be replicated (l times) for further decreasing the value of (7). In this paper, we set $l=5$.

In Step 1, the initialization is performed using the principal component analysis of sample covariance matrix $\boldsymbol{S}$. That is, the initial $\boldsymbol{L}$ and $\boldsymbol{\Delta}$ are given by $\boldsymbol{V}$ and $\boldsymbol{\Omega}^{1/2}$, respectively, with $\boldsymbol{\Omega}$ the $m\times m$ diagonal matrix whose diagonal elements are the m largest eigenvalues of $\boldsymbol{S}$, and the columns of $\boldsymbol{V}$ $(p\times m)$ being the eigenvectors corresponding to $\boldsymbol{\Omega}$. The initial $\boldsymbol{\Psi}$ is set to $\mathrm{diag}\left(\boldsymbol{S}-\boldsymbol{V}\boldsymbol{\Omega}\boldsymbol{V}'\right)$.

In Step 5, we define the convergence as the decrease in the value of (7) or (4) from the previous round being less than $p^2\times 0.1^{10}$.

3. Illustration

In this section, we illustrate the performance of the MD algorithm with a 190-person × 25-item data matrix, which was collected by the author and publicly available at http://bm.osaka-u.ac.jp/data/big5/. This data set contains the self-ratings of the persons (university students) for to what extents they are characterized by the personalities described by the 25 items. According to a theory in personality psychology [20], the items can be classified into the five groups shown in the first column of **Table 1**. The 25 × 25 matrix of the correlation coefficients among those items was obtained from the data set.

We carried out the MD algorithm for the correlation matrix with the number of factors *m* set to five. **Figure 1** shows the change in the value of loss function (4) until the steps in Section 2.3 were iterated ten times and the change after the tenth iteration. There, we can find that the function value decreased monotonically with iteration, which was finally reached to convergence at the 542 nd iteration.

Table 1. Loadings and unique variances $\boldsymbol{\Psi 1}_p$ for personality rating data.

Group	Item	Loading matrix					$\boldsymbol{\Psi 1}_p$
Neurotic	Worry	**0.72**	−0.14	−0.12	0.22	0.13	0.32
	Sensitive	**0.74**	−0.06	−0.04	0.01	−0.08	0.38
	Pessimistic	**0.67**	−0.29	−0.13	−0.03	0.27	0.35
	Unrest	**0.60**	0.06	−0.15	−0.08	−0.34	0.33
	Careful	**0.62**	−0.09	−0.16	0.10	0.32	0.32
Extrovert	Sociable	−0.16	**0.81**	0.09	0.09	0.12	0.23
	Talkative	0.04	**0.82**	−0.05	−0.02	−0.09	0.25
	Voluntary	−0.10	**0.73**	0.17	0.10	0.13	0.29
	Cheerful	−0.21	**0.80**	0.08	0.16	0.00	0.23
	Showy	0.08	**0.66**	0.23	−0.04	−0.05	0.38
Open	Creative	−0.09	0.03	**0.85**	−0.04	−0.07	0.23
	Adventurous	−0.22	0.22	**0.67**	−0.02	−0.20	0.36
	Progressive	−0.19	0.24	**0.64**	0.08	0.10	0.40
	Flexible	−0.28	0.31	**0.37**	0.20	0.03	0.57
	Imaginative	0.19	0.09	**0.47**	0.11	−0.34	0.44
Agreeable	Mild	−0.15	−0.12	0.15	**0.63**	0.02	0.38
	Tenderhearted	0.09	0.20	0.13	**0.62**	−0.01	0.39
	Altruistic	0.13	0.01	0.00	**0.71**	0.15	0.39
	Cooperative	0.00	0.27	−0.14	**0.66**	0.08	0.25
	Sympathetic	0.11	0.12	−0.04	**0.73**	0.21	0.32
Conscious	Deliberate	0.14	−0.04	0.09	0.22	**0.60**	0.37
	Reliable	−0.13	0.32	0.08	0.26	**0.57**	0.36
	Diligent	−0.04	0.05	−0.07	0.15	**0.77**	0.29
	Systematic	0.09	−0.02	−0.09	0.01	**0.71**	0.40
	Methodical	0.21	0.02	−0.19	0.04	**0.74**	0.31

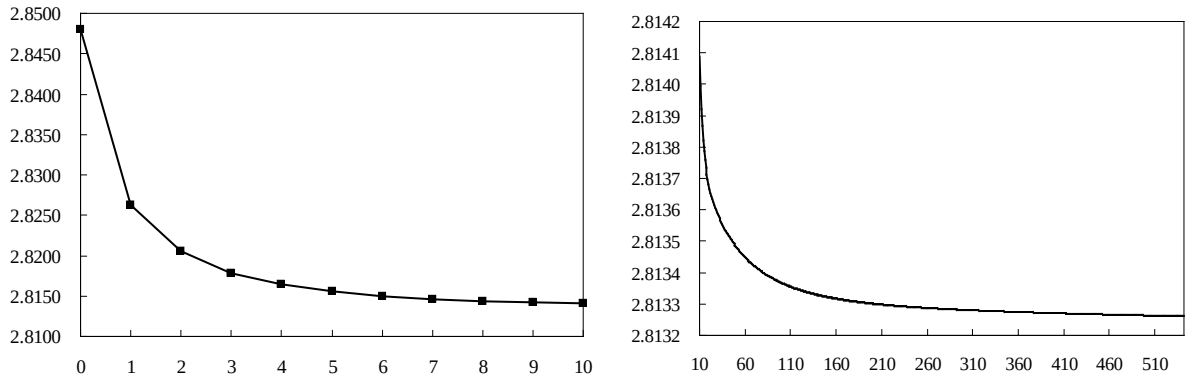

Figure 1. Change in the GLS loss function value as a function of the number of iteration.

As the resulting loading matrix has rotational freedom, that is, the $\boldsymbol{\Lambda}$ post-multiplied by arbitrary orthonormal matrix satisfies (1) and (2), the loading matrix was rotated by the varimax method [21]. The solution is pre-

sented in **Table 1**. There, bold font is used for the loadings whose absolute values are greater than 0.35. They show that the 25 items are clearly classified into the five groups as predicted by the theory in personality psychology [20], which demonstrates that the MD algorithm provided the reasonable solution.

4. Simulation Study

A simulation study was performed in order to assess how well parameter matrices are recovered by the proposed MD algorithm and compare it with the existing algorithms for the GLS estimation in the goodness of the recovery. We first describe the procedure for synthesizing the data to be analyzed, which is followed by results.

An *n*-observations × *p*-variables data matrix $\boldsymbol{Y}$ was synthesized according to the matrix versions of the FA model (1) and the assumptions in (2):

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\Lambda}' + \boldsymbol{E} = \boldsymbol{X}\boldsymbol{\Lambda}' + \boldsymbol{U}\boldsymbol{\Psi}^{1/2} = \left[\boldsymbol{X}, \boldsymbol{U}\right]\left[\boldsymbol{L}, \boldsymbol{\Psi}^{1/2}\right]' = \boldsymbol{Z}\boldsymbol{\Gamma}' \tag{15}$$

$$\mathbf{1}'_{p+m}\boldsymbol{Z} = \mathbf{0}'_{p+m} \quad \text{and} \quad \boldsymbol{Z}'\boldsymbol{Z} = \boldsymbol{I}_{p+m} \tag{16}$$

Here, $\mathbf{1}_{p+m}$ denotes the $(p+m)\times 1$ vector of ones, $\boldsymbol{Z} = \left[\boldsymbol{X}, \boldsymbol{U}\right]$ is an $n \times (m+p)$ block matrix whose right $n \times p$ block $\boldsymbol{U}$ is post-multiplied by $\boldsymbol{\Psi}^{1/2}$ to give the error matrix $\boldsymbol{E} = \boldsymbol{U}\boldsymbol{\Psi}^{1/2}$, and $\boldsymbol{\Gamma} = \left[\boldsymbol{\Lambda}, \boldsymbol{\Psi}^{1/2}\right]$ is an $(m+p)\times p$ block matrix including the loading matrix and the square roots of unique variances. It should be noticed that each row of $\boldsymbol{Y}$, $\boldsymbol{X}$, and $\boldsymbol{E}$ corresponds to $\boldsymbol{y}'$, $\boldsymbol{x}'$, and $\boldsymbol{e}'$, respectively, whose transposed vectors appear in (1), and the five equations in (2) can be summarized into the two matrix expressions in (16). The data synthesis procedure follows the next steps:

Step 1. Draw m from $DU(1,5)$, p from $DU(4m,7m)$, and n from $DU(9p,14p)$, with $DU(I,J)$ denoting the discrete uniform distribution defined for the integers within the range $[I,J]$.

Step 2. Draw each loading in $\boldsymbol{\Lambda}$ from $U(-1,1)$ and each unique variance in $\boldsymbol{\Psi}$ from $U(0.1,0.7)$ with $U(\alpha,\beta)$ denoting the uniform distribution over the range $[\alpha,\beta]$.

Step 3. Draw each elements of $\boldsymbol{Z}$ in (15) from $U(-1,1)$ which is followed by centering $\boldsymbol{Z}$ and post-multiplying it by the matrix that allows the resulting $\boldsymbol{Z}$ to satisfy (16).

Step 4. Form $\boldsymbol{Y}$ with (15) and obtain the covariance matrix $\boldsymbol{S} = n^{-1}\boldsymbol{Y}'\boldsymbol{Y}$.

In Step 3 we have used a uniform distribution for $\boldsymbol{Z}$, rather than the normal distribution typically used for such a matrix, as a feature of the GLS estimation is that it does not need the normality assumption required in the ML estimation. We replicated the above steps to have 2000 sets of $\boldsymbol{S}$. For them, the MD and the existing algorithms were carried out, where the latter are the two gradient algorithms [9] [10], as described in Section 1. We refer to the ones in [9] and [10] as the Newton-Raphson (NR) and Gauss-Newton (GN) algorithms, respectively. In the NR one, we obtained the gradient vector in [9], Equation (32), by pre-multiplying the vector of first derivatives by the Moore Penrose inverse of the corresponding Hessian matrix. Also in the NR and GN algorithms, we used the same initialization and definition of convergence as in Section 2.3.

Let us express the true $[\boldsymbol{\Lambda}, \boldsymbol{\Psi}]$ simply as $[\boldsymbol{\Lambda}, \boldsymbol{\Psi}]$ and use $[\boldsymbol{\Lambda}_{\#}, \boldsymbol{\Psi}_{\#}]$ for the solution given by the NR, GN, or MD algorithm. For assessing the recovery of the loading matrix, the averaged absolute difference (AAD) of the elements in $\boldsymbol{\Lambda}$ to the corresponding estimates, *i.e.*,

$$\mathrm{AAD}_{\Lambda}\left(\boldsymbol{\Lambda}, \boldsymbol{\Lambda}_{\#}\right) = (pm)^{-1}\left\|\boldsymbol{\Lambda} - \boldsymbol{\Lambda}_{\#}\right\|_{l1}, \tag{17}$$

can be used with $\|\bullet\|_{l1}$ denoting the $l1$ norm. Here, it should be noted that $\boldsymbol{\Lambda}_{\#}$ has rotational freedom and must be rotated so that the resulting $\boldsymbol{\Lambda}_{\#}$ is optimally matched to $\boldsymbol{\Lambda}$. Such a rotated $\boldsymbol{\Lambda}_{\#}$ can be obtained by the orthogonal Procrustes method [22] with $\boldsymbol{\Lambda}$ a target matrix. The loading matrix $\boldsymbol{\Lambda}_{\#}$ in (17) thus stands for the one rotated by the Procrustes method. The recovery of unique variances can also be assessed with the AAD index $\mathrm{AAD}_{\Psi}\left(\boldsymbol{\Psi}, \boldsymbol{\Psi}_{\#}\right) = p^{-1}\left\|\boldsymbol{\Psi} - \boldsymbol{\Psi}_{\#}\right\|_{l1}$, where the unique variances are uniquely determined, thus the additional procedure as for $\boldsymbol{\Lambda}_{\#}$ is unnecessary. Smaller values of those AAD indices stand for better recovery.

The statistics of AAD values over 2000 data sets are presented in **Table 2**. There, the averages show that the recovery by the MD algorithm is the best and that for the NR one is the worst. It should be noted that the 50 and 75 percentiles for the NR algorithm are zero, while the maximum and 99 percentile are very large. That is, the recovery by the NR algorithm was perfect for more than 75 percent of the 2000 data sets, but for a few percent of them, recovery was considerably bad, which increased the averages for the NR one. In contrast, the maximum AAD of loadings and unique variances for the MD algorithm are 0.0041 and 0.0013, respectively, which are

Table 2. Statistics for the differences between the true parameter values and their estimated counterparts.

Parameter	Loadings			Unique variances		
Algorithm	NR	GN	MD	NR	GN	MD
Average	0.0026	0.0020	0.0005	0.0030	0.0007	0.0000
50 percentile	0.0000	0.0013	0.0005	0.0000	0.0001	0.0000
75 percentile	0.0000	0.0031	0.0007	0.0000	0.0005	0.0000
95 percentile	0.0050	0.0061	0.0011	0.0103	0.0035	0.0000
99 percentile	0.0821	0.0095	0.0014	0.0857	0.0105	0.0001
Maximum	0.2199	0.0203	0.0041	0.1406	0.0323	0.0013

small enough to be ignored. That is, the proposed MD algorithm well recovered the true parameter values for all of the 2000 data sets. We can thus conclude that the MD algorithm is superior to the existing ones in the goodness of recovery.

5. Discussion

We proposed the majorizing-diagonal (MD) algorithm for the GLS estimation in FA. In the algorithm, the loading matrix is reparameterized as the product of a column-orthonormal matrix and a diagonal one, and the former one is updated with Kiers and ten Berge's [18] majorization technique, while the latter diagonal matrix and another diagonal one including unique variances are updated so that their quadratic functions are minimized. The iteration of those updates decreases monotonically the GLS loss function. The simulation study demonstrated the exact recovery of loadings and unique variances by the MD algorithm and its superiority to the existing gradient algorithms in the recovery.

One of the tasks remaining for the MD algorithm is to study its mathematical properties as have been done for the algorithms in the other estimation procedures. For example, it has been found that the EM algorithm for the ML estimation [12] can never give an improper solution under a certain condition [23], where the improper solution refers to the one including a negative unique variance. Whether such special features are possessed by the MD algorithm is considered to be found by studying the properties of the matrix update formulas in the algorithm.

References

[1] Harman, H.H. (1976) Modern Factor Analysis. 3rd Edition, The University of Chicago Press, Chicago.

[2] Mulaik, S.A. (2010) Foundations of Factor Analysis. 2nd Edition, CRC Press, Boca Raton.

[3] Yanai, H. and Ichikawa, M. (2007) Factor Analysis. In: Rao, C.R. and Sinharay, S., Eds., *Handbook of Statistics*, Vol. 26: *Psychometrics*, Elsevier, Amsterdam, 257-296.

[4] Anderson, T.W. and Rubin, H. (1956) Statistical Inference in Factor Analysis. In: Neyman, J., Ed., *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 5, University of California Press, Berkeley, 111-150.

[5] Lange, K. (2010) Numerical Analysis for Statisticians. 2nd Edition, Springer, New York.

[6] ten Berge, J.M.F. (1993) Least Squares Optimization in Multivariate Analysis. DSWO Press, Leiden.

[7] Jöreskog, K.G. (1967) Some Contributions to Maximum Likelihood Factor Analysis. *Psychometrika*, **32**, 443-482. http://dx.doi.org/10.1007/BF02289658

[8] Jennrich, R.I. and Robinson, S.M. (1969) A Newton-Raphson Algorithm for Maximum Likelihood Factor Analysis. *Psychometrika*, **34**, 111-123. http://dx.doi.org/10.1007/BF02290176

[9] Jöreskog, K.G. and Goldberger, A.S. (1972) Factor Analysis by Generalized Least Squares. *Psychometrika*, **37**, 243-250. http://dx.doi.org/10.1007/BF02306782.

[10] Lee, S.Y. (1978) The Gauss-Newton Algorithm for the Weighted Least Squares Factor Analysis. *Journal of the Royal Statistical Society*: *Series D* (*The Statistician*), **27**, 103-114. http://dx.doi.org/10.2307/2987906

[11] Harman, H.H. and Jones, W.H. (1966) Factor Analysis by Minimizing Residuals (Minres). *Psychomerika*, **31**, 351-369. http://dx.doi.org/10.1007/BF02289468

[12] Rubin, D.B. and Thayer, D.T. (1982) EM Algorithms for ML Factor Analysis. *Psychometrika*, **47**, 69-76. http://dx.doi.org/10.1007/BF02293851

[13] Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B*, **39**, 1-38.

[14] Groenen, P.J.F. (1993) The Majorization Approach to Multidimensional Scaling: Some Problems and Extensions. DSWO Press, Leiden.

[15] Unkel, S. and Trendafilov, N.T. (2010) A Majorization Algorithm for Simultaneous Parameter Estimation in Robust Exploratory Factor Analysis. *Computational Statistics and Data Analysis*, **54**, 3348-3358. http://dx.doi.org/10.1016/j.csda.2010.02.003

[16] Unkel, S. and Trendafilov, N.T. (2010) Simultaneous Parameter Estimation in Exploratory Factor Analysis: An Expository Review. *International Statistical Review*, **78**, 363-382. http://dx.doi.org/10.1111/j.1751-5823.2010.00120.x

[17] Adachi, K. (2012) Some Contributions to Data-Fitting Factor Analysis with Empirical Comparisons to Covariance-Fitting Factor Analysis. *Journal of the Japanese Society of Computational Statistics*, **25**, 25-38. http://dx.doi.org/10.5183/jjscs.1106001_197

[18] Kiers, H.A.L. and ten Berge, J.M.F. (1992) Minimization of a Class of Matrix Trace Functions by Means of Refined Majorization. *Psychometrika*, **57**, 371-382. http://dx.doi.org/10.1007/BF02295425

[19] Kiers, H.A.L. (1990) Majorization as a Tool for Optimizing a Class of Matrix Functions. *Psychometrika*, **55**, 417-428. http://dx.doi.org/10.1007/BF02294758

[20] Costa, P.T. and McCrae, R.R. (1992) NEO PI-R Professional Manual: Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI). Psychological Assessment Resources, Odessa, FL.

[21] Kaiser, H.F. (1958) The Varimax Criterion for Analytic Rotation in Factor Analysis. *Psychometrika*, **23**, 187-200. http://dx.doi.org/10.1007/BF02289233

[22] Gower, J.C. and Dijksterhuis, G.B. (2004) Procrustes Problems. Oxford University Press, Oxford. http://dx.doi.org/10.1093/acprof:oso/9780198510581.001.0001

[23] Adachi, K. (2013) Factor Analysis with EM Algorithm Never Gives Improper Solutions When Sample Covariance and Initial Parameter Matrices Are Proper. *Psychometrika*, **78**, 380-394. http://dx.doi.org/10.1007/s11336-012-9299-8