Comput Stat (2015) 30:279–292
DOI 10.1007/s00180-014-0536-8

ORIGINAL PAPER

Generalized data-fitting factor analysis with multiple quantification of categorical variables

Naomichi Makino

Received: 29 June 2013 / Accepted: 23 September 2014 / Published online: 9 October 2014

Abstract In this study, a recently proposed data-fitting factor analysis (DFFA) procedure is generalized for categorical variable analysis. For generalized DFFA (GDFFA), we develop an alternating least squares algorithm consisting of a multiple quantification step and a model parameters estimation step. The differences between GDFFA and similar statistical methods such as multiple correspondence analysis and FACTALS are also discussed. The developed algorithm and its solution are illustrated with a real data example.

Keywords Data-fitting factor analysis · Categorical variables · Multiple quantification · Multiple correspondence analysis · FACTALS

1 Introduction

Exploratory factor analysis (EFA) is a time-honored statistical method that aims to explain the interrelationships among observed variables by latent variables called common factors. EFA assumes that factor scores can be treated as latent random variables (e.g., Anderson and Rubin 1956; Yanai and Ichikawa 2007). The model parameters of EFA are traditionally estimated by minimizing the differences between a model correlation structure and a sample covariance matrix using least squares or maximum likelihood discrepancy measures (e.g., Harman 1976; Mulaik 2010).

Recently, an alternative method for estimating model parameters has been developed, in which all parameters are treated as fixed and unknown (de Leeuw 2004, 2008; Unkel and Trendafilov 2010, 2011; Adachi 2012). In this study, we denote this estimation procedure as data-fitting factor analysis (DFFA) because the model parameters are directly fitted to the data matrix. If $\mathbf{X}$ is an n-observations $\times$ p-variables matrix of

N. Makino (✉)
Graduate School of Human Sciences, Osaka University, 1-2 Yamadaoka, Suita, Osaka 565-0871, Japan
e-mail: r_rm225@yahoo.co.jp

standardized observed data (with $n > p$), then the DFFA loss function is defined as minimization of the difference between the observed data matrix and the EFA model parameters:

$$\|\mathbf{X} - \mathbf{F}\mathbf{A}' - \mathbf{U}\boldsymbol{\Psi}\|^2. \tag{1}$$

Here, $\mathbf{F}$ is an n-observations $\times$ c-factors matrix of common factors, $\mathbf{A}$ is a $p \times c$ matrix of factor loadings, $\mathbf{U}$ is an $n \times p$ matrix of unique factors, and $\boldsymbol{\Psi}$ is a $p \times p$ diagonal matrix of uniqueness. We assume that the columns of $\mathbf{F}$ and $\mathbf{U}$ have a mean of zero and are scaled to have unit norm, and all common factors and unique factors are mutually uncorrelated. We also assume that the unique factors are orthogonal to the common factors. The idea behind these assumptions is that the common factors account for the correlation structure among the set of observed variables, while each unique factor corresponds to that portion of a particular observed variable that cannot be accounted for by the common factors. Thus, the model parameters are obtained by minimizing (1) over $\mathbf{F}$, $\mathbf{A}$, $\mathbf{U}$, and $\boldsymbol{\Psi}$, subject to the following constraints:

$$n^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_c,\ n^{-1}\mathbf{U}'\mathbf{U} = \mathbf{I}_p,\ \mathbf{F}'\mathbf{U} = \mathbf{O}_{c\times p},\ \boldsymbol{\Psi} \text{ diagonal}, \tag{2}$$

where $\mathbf{I}_p$ is a $p \times p$ identity matrix and $\mathbf{O}_{c\times p}$ is a $c \times p$ null matrix.

The DFFA procedures are limited to cases where all observed variables are numerical. However, social or behavioral research must frequently analyze multivariate categorical variables. For example, in questionnaires, participants are asked to choose one of the several response alternatives (categories) for each set of questions (variables). Quantification is a widely used statistical technique for analyzing categorical variables, in which the observed categorical variables are transformed into quantitative scores such that the data analysis model most closely matches the observations (Gifi 1990; Young 1981). Using the quantification technique, we propose a generalized DFFA (GDFFA) for categorical variables, which is defined as a problem that minimizes the difference between the quantified categorical data and the EFA model parameters.

Alternative methods such as multiple correspondence analysis (MCA) (Benzecri 1974, 1992; Greenacre 1984; Gifi 1990; Murakami et al. 1999) and FACTALS (Takane et al. 1979) also adopt the quantification technique. MCA is a famous technique to detect and represent underlying structures in categorical data, and FACTALS is a common factor analysis procedure for nonmetric datasets. The statistical models in MCA and FACTALS are quite similar to that in GDFFA. Later, in this study, we discuss the similarities and differences between GDFFA and these related methods.

The remaining parts of this paper are organized as follows. In the next section, we introduce the DFFA algorithms and propose a GDFFA procedure. In Sect. 3, GDFFA is compared with MCA and FACTALS. In Sect. 4, GDFFA is applied to categorical data. Concluding remarks are provided in Sect. 5.

2 Proposed method

We describe a DFFA algorithm in Sect. 2.1 and generalize the DFFA procedure in Sect. 2.2. An alternating least squares estimation algorithm is proposed in Sect. 2.3.

2.1 Estimation procedures for DFFA

There are two major algorithms in DFFA: de Leeuw's algorithm (de Leeuw 2004, 2008), and Unkel and Trendafilov's algorithm (Unkel and Trendafilov 2011). In both the algorithms, the factor score estimation and the factor loading estimation steps are iterated alternately, but different procedures are used in the former. In this paper, we briefly introduce Unkel and Trendafilov's algorithm, in which $\mathbf{F}$ and $\mathbf{U}$ are successively updated by solving the following orthogonal Procrustes problems.

First, find $\mathbf{F}$ that minimizes

$$\|(\mathbf{X} - \mathbf{U}\boldsymbol{\Psi}) - \mathbf{F}\mathbf{A}'\|^2, \tag{3}$$

subject to $n^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_c$ for given $\mathbf{A}$, $\mathbf{U}$, and $\boldsymbol{\Psi}$. The optimal solution of $\mathbf{F}$ is explicitly given by the singular value decomposition (SVD) of $(\mathbf{X} - \mathbf{U}\boldsymbol{\Psi})\mathbf{A}$.

Next, find $\mathbf{U}$ that minimizes

$$\|(\mathbf{I} - \mathbf{P}_F)(\mathbf{X} - \mathbf{U}\boldsymbol{\Psi})\|^2, \tag{4}$$

subject to $n^{-1}\mathbf{U}'\mathbf{U} = \mathbf{I}_p$ and $\mathbf{F}'\mathbf{U} = \mathbf{O}_{c\times p}$ for given $\mathbf{F}$, $\mathbf{A}$, and $\boldsymbol{\Psi}$. Here, $\mathbf{P}_F$ is a projection matrix for the column space of $\mathbf{F}$. The optimal $\mathbf{U}$ is found by the SVD of $\boldsymbol{\Psi}\mathbf{X}'(\mathbf{I}_n - \mathbf{P}_F)$.

After updating factor score matrices $\mathbf{F}$ and $\mathbf{U}$, $\mathbf{A}$, and $\boldsymbol{\Psi}$ are obtained by solving the regression problems for both algorithms:

$$\mathbf{A} = n^{-1}\mathbf{X}'\mathbf{F}, \tag{5}$$

$$\boldsymbol{\Psi} = n^{-1}\mathrm{diag}(\mathbf{X}'\mathbf{U}). \tag{6}$$

In EFA, common and unique factors cannot be uniquely identified. This form of indeterminacy is known as factor indeterminacy (e.g., Mulaik 2010). In other estimation methods, the factor score matrices $\mathbf{F}$ and $\mathbf{U}$ can be obtained after estimating $\mathbf{A}$ and $\boldsymbol{\Psi}$. However, DFFA allows simultaneous estimation of all model parameters, and non-uniqueness of the common and unique factor scores is not a problem from an algorithmic perspective as seen above.

2.2 Generalized DFFA (GDFFA)

We rewrite the loss function (1) as a problem that minimizes the difference between the column vector $\mathbf{x}_i = [x_{1i}, \ldots, x_{ni}]'$ $(n \times 1)$ and the corresponding EFA model parameters:

$$\|\mathbf{X} - \mathbf{F}\mathbf{A}' - \mathbf{U}\boldsymbol{\Psi}\|^2 = \sum_{i=1}^{p} \|\mathbf{x}_i - \mathbf{F}\mathbf{a}_i - \psi_i\mathbf{u}_i\|^2, \tag{7}$$

where $\mathbf{a}_i = [a_{i1}, \ldots, a_{ic}]'$ $(c \times 1)$, $\mathbf{u}_i = [u_{1i}, \ldots, u_{ni}]'$ $(n \times 1)$ and ψ_i is the ith diagonal element of $\boldsymbol{\Psi}$.

When the ith variable is categorical, $\mathbf{x}_i$ is transformed into quantitative scores by a quantification procedure. If K_i denotes the number of categories in the ith variable and $\mathbf{G}_i$ denotes the $n \times K_i$ indicator matrix, then $\mathbf{G}_i\mathbf{Q}_i$ represents multiple quantification of $\mathbf{x}_i$ where $\mathbf{Q}_i$ (K_i-categories $\times$ R_i-dimensions) is the quantification parameter matrix of the ith variable. The number of dimensions R_i is required to be pre-specified and must satisfy $1 \leq R_i \leq \min(K_i - 1, c)$ for each variable. The lower bound in the inequality is trivial; the upper bound will be discussed in Sect. 2.3. To avoid trivial solutions, the quantified data are assumed to be centered and orthonormal in a column-wise manner:

$$\mathbf{1}_n'\mathbf{G}_i\mathbf{Q}_i = \mathbf{0}_{K_i}', \; n^{-1}\mathbf{Q}_i'\mathbf{G}_i'\mathbf{G}_i\mathbf{Q}_i = \mathbf{I}_{R_i}, \tag{8}$$

where $\mathbf{1}_n$ is an n-dimensional vector whose elements are all ones and $\mathbf{0}_{K_i}$ is a K_i-dimensional vector whose elements are all zeros.

Combining the quantification technique with the loss function (7), we obtain the loss function of GDFFA for categorical variables. GDFFA is defined as the minimization of

$$\sum_{i=1}^{p} \|\mathbf{G}_i\mathbf{Q}_i - \mathbf{F}\mathbf{A}_i - \mathbf{U}_i\mathbf{\Psi}_i\|^2, \tag{9}$$

over $\mathbf{F}$, $\mathbf{A}_i$, $\mathbf{U}_i$, $\mathbf{\Psi}_i$, and $\mathbf{Q}_i$ ($i = 1, \ldots, p$), subject to constraints (2) and (8). Here, $\mathbf{A}_i$ ($c \times R_i$) is a loading matrix, $\mathbf{U}_i$ ($n \times R_i$) is a unique factor matrix, and $\mathbf{\Psi}_i$ ($R_i \times R_i$) is a diagonal matrix of uniqueness, which are the corresponding model parameters of the multidimensionally quantified ith variable. In a multiple quantification situation, the corresponding EFA model parameters are not vector-formed. The quantified ith variable $\mathbf{G}_i\mathbf{Q}_i$ is fitted to the corresponding model parameters in the ith variable $\mathbf{F}\mathbf{A}_i + \mathbf{U}_i\Psi_i$. The matrix-formed parameters $\mathbf{A}_i$, $\mathbf{U}_i$, $\mathbf{\Psi}_i$, and $\mathbf{Q}_i$ reduce to the vectors or the scalar $\mathbf{a}_i$, $\mathbf{u}_i$, $\boldsymbol{\psi}_i$, and $\mathbf{q}_i$ when R_i is set to one. Here, $\mathbf{q}_i$ is the quantification parameter vector ($K_i \times 1$). We refer to this situation as single quantification, which is a special case of multiple quantification.

When quantifying categorical variables, the number of dimensions of each variable must be determined. Single quantification is preferred, particularly when categories in variables have a clear ordinal interpretation (van Burg et al. 1988). However, cases are likely to exist for which categories in a nominal variable cannot be scaled unidimensionally (Adachi and Murakami 2011; Nishisato 2006). In other words, nominal variables are not necessary to have an ordinal interpretation. Therefore, single quantification is insufficient for analyzing nominal variables, and we perform multiple quantification for such variables.

2.3 Estimation of quantification and model parameters

In quantification processes, a loss function is optimized by alternating two major steps. The parameters in GDFFA can be estimated by alternate iterations of the operations that quantify the categorical variables and that obtain the EFA model parameters.

2.3.1 Update of quantification parameters

We reparametrize $\mathbf{G}_i\mathbf{Q}_i$ as

$$\begin{aligned}\mathbf{G}_i\mathbf{Q}_i &= \mathbf{G}_i(\mathbf{G}_i^{'}\mathbf{G}_i)^{-1/2}(\mathbf{G}_i^{'}\mathbf{G}_i)^{1/2}\mathbf{Q}_i \\ &= \mathbf{H}_i\mathbf{B}_i,\end{aligned}$$

where $\mathbf{H}_i = n^{1/2}\mathbf{G}_i(\mathbf{G}_i'\mathbf{G}_i)^{-1/2}$ and $\mathbf{B}_i = n^{-1/2}(\mathbf{G}_i'\mathbf{G}_i)^{1/2}\mathbf{Q}_i$. Substituting $\mathbf{H}_i$ and $\mathbf{B}_i$ in (9), the problem reduces to minimizing

$$\|\mathbf{H}_i\mathbf{B}_i - \mathbf{F}\mathbf{A}_i - \mathbf{U}_i\mathbf{\Psi}_i\|^2, \tag{10}$$

over $\mathbf{B}_i$, subject to (8) for fixed model parameters. This minimization can be viewed as an orthogonal Procrustes problem because $\mathbf{B}_i$ is column-orthonormal matrix where: $\mathbf{B}_i^{'}\mathbf{B}_i = n^{-1}\mathbf{Q}_i^{'}\mathbf{G}_i^{'}\mathbf{G}_i\mathbf{Q}_i = \mathbf{I}_{R_i}$. Let us define the SVD as $\mathbf{H}_i(\mathbf{F}\mathbf{A}_i - \mathbf{U}_i\mathbf{\Psi}_i) = \mathbf{K}\mathbf{\Delta}\mathbf{L}'$ with $\mathbf{K}'\mathbf{K} = \mathbf{L}'\mathbf{L} = \mathbf{L}\mathbf{L}' = \mathbf{I}_{R_i}$, and let $\mathbf{\Delta}$ be the $q \times q$ diagonal matrix whose diagonal elements are arranged in descending order. The optimal $\mathbf{B}_i$ is given by

$$\mathbf{B}_i = \mathbf{K}\mathbf{L}' \tag{11}$$

(ten Berge 1993), and then we have

$$\mathbf{Q}_i = n^{1/2}(\mathbf{G}_i^{'}\mathbf{G}_i)^{-1/2}\mathbf{K}\mathbf{L}'. \tag{12}$$

The rank of $\mathbf{H}_i(\mathbf{F}\mathbf{A}_i - \mathbf{U}_i\mathbf{\Psi}_i)$ at most is equal to the smaller of $K_i - 1$ or c. Consequently, the dimension of quantification R_i is upper bounded by $\min(K_i - 1, c)$.

2.3.2 Update of model parameters

After updating the quantification parameters, the model parameters are updated using the DFFA algorithms introduced in Sect. 2.1. That is, the model parameters can be obtained if the observed numerical data are replaced by quantified data in the DFFA algorithms.

2.3.3 Complete algorithm

The alternating procedure described in the previous section is continued until a prespecified convergence criterion ε, say 10^{-6}, is met. The GDFFA algorithm is summarized as follows:

Step 1. Initial values are chosen for $\mathbf{Q}_i$, $\mathbf{F}$, $\mathbf{U}_i$, $\mathbf{A}_i$ and $\mathbf{\Psi}_i$. Take arbitrary matrix $\mathbf{Q}_i$ which satisfies (8). Values for $\mathbf{F}$ and $\mathbf{U}_i$ can be chosen randomly, and they should satisfy the constraints in (2). Then, $\mathbf{A}_i$ and $\mathbf{\Psi}_i$ are given as $\mathbf{A} = n^{-1}\mathbf{Q}_i'\mathbf{G}_i'\mathbf{F}$, $\mathbf{\Psi} = n^{-1}\mathrm{diag}(\mathbf{Q}_i'\mathbf{G}_i'\mathbf{U}_i)$, respectively.

Step 2. Update the quantification parameters by solving the orthogonal Procrustes problems.

Step 3. Update the EFA model parameters using the DFFA algorithm.
Step 4. Finish if the decrease in (9) from the previous step is less than ε; otherwise, return to Step 2.

The proposed algorithm monotonically decreases the loss function. Since the loss function is bounded below, it converges to a solution that is at least a local optimum (Young 1981). To increase the chance of finding the global maximum, the algorithm should be run several times, with different initial values. In Step 4, either de Leeuw's algorithm (de Leeuw 2004, 2008) or Unkel and Trendafilov's algorithm (Unkel and Trendafilov 2011) can be used.

3 Comparison between GDFFA and related methods

GDFFA is compared with MCA and FACTALS in Sects. 3.1 and 3.2, respectively. The relationships among all three methods are summarized in Sect. 3.3.

3.1 Multiple correspondence analysis

Multiple correspondence analysis (MCA) is a useful technique to find underlying structures inherent in categorical data. There are various approaches to formulate an MCA, but they have proved to give essentially equivalent solutions originating in different theoretical foundations (Tenenhaus and Young 1985). Among these formulations, we introduce Murakami, Kiers and ten Berge's formulation (Murakami et al. 1999), in which MCA is defined as minimization of the difference between the quantified data and the corresponding MCA model parameters. That is, the loss function

$$\sum_{i=1}^{p} \|\mathbf{G}_i\mathbf{Q}_i - \mathbf{F}\mathbf{A}_i\|^2, \tag{13}$$

is minimized over $\mathbf{F}$, $\mathbf{A}_i$, and $\mathbf{Q}_i$ ($i = 1, \ldots, p$), subject to the constraint $n^{-1}\mathbf{F}'\mathbf{F} = \mathbf{I}_c$.

In ordinary MCA, the number of quantification dimensions is fixed to the smaller of the number of factors or categories minus one and can be solved explicitly. For this MCA formulation, the number of dimensions can be set in $1 \leq R_i \leq \min(K_i - 1, c)$. This type of MCA is referred to as rank-restricted MCA (Murakami et al. 1999), which includes ordinary MCA as a special case. When all categorical variables are unidimensionally quantified, rank-restricted MCA reduces to nonmetric principal component analysis (NPCA) (Adachi and Murakami 2011; Murakami et al. 1999). The loss function is defined as follows:

$$\sum_{i=1}^{p} \|\mathbf{G}_i\mathbf{q}_i - \mathbf{F}\mathbf{a}_i\|^2. \tag{14}$$

NPCA also aims to find underlying structures inherent in categorical variables, but categorical variables are assumed to have a unidimensional structure in NPCA.

It has been previously shown that MCA and NPCA are each a generalization of principal component analysis to nonmetric data (Adachi and Murakami 2011), and

that the unique term distinguishes the EFA model from the PCA model (de Leeuw 2004; Trendafilov et al. 2013). Another difference, however, occurs between GDFFA and MCA. Let $\mathbf{N}$ be a nonsingular matrix ($c \times c$) and $\mathbf{T}_i$ ($R_i \times R_i$) be an orthonormal matrix ($i = 1, \ldots, p$). Then, the parameters can be transformed without changing the fitness of the MCA loss function:

$$\sum_{i=1}^{p} \|\mathbf{G}_i\mathbf{Q}_i - \mathbf{F}\mathbf{A}_i'\|^2 = \sum_{i=1}^{p} \|\mathbf{G}_i\mathbf{Q}_i\mathbf{T}_i - \mathbf{F}\mathbf{N}'^{-1}\mathbf{N}'\mathbf{A}_i'\mathbf{T}_i\|^2. \tag{15}$$

This non-uniqueness is called rotational indeterminacy or rotational freedom. In MCA, two rotational indeterminacies exist in the quantification parameter and loading matrices. Generally, to simplify the interpretations, the loading matrix is rotated in such a way that when considering one variable, few squared loadings are large and as many as possible are close to zero. Murakami (1999) proposed a modified orthomax criterion in which the loading matrix is rotated towards the simple structure by pre- and post-multiplied rotation matrices $\mathbf{N}$ and $\mathbf{T}_i$ ($i = 1, \ldots, p$), respectively. Oblique rotation is probably more appropriate in most practical situations, because fewer constraints are imposed in oblique rotation and it is generally possible to obtain a solution more easily than in orthogonal rotation (Browne 2001). However, the orthogonal rotations alone are permitted by Murakami's criterion, and thus the rotated solutions may be difficult to interpret.

In contrast, the unique term in GDFFA eliminates rotational indeterminacy of the quantification matrix, as seen in the following loss function:

$$\sum_{i=1}^{p} \|\mathbf{G}_i\mathbf{Q}_i - \mathbf{S}_i\mathbf{\Lambda}_i\|^2 = \sum_{i=1}^{p} \|\mathbf{G}_i\mathbf{Q}_i - [\mathbf{F}\mathbf{N}^{'-1}\ \mathbf{U}_i][\mathbf{A}_i^{'}\mathbf{N}\ \ \mathbf{\Psi}_i]^{'}\|^2. \tag{16}$$

Consequently, the difficulties inherent in pre- and post-multiplied rotation of the MCA solution are precluded in GDFFA. Although the two techniques are very similar, the unique term in GDFFA creates a very different structure from that in MCA.

3.2 FACTALS

FACTALS is a nonmetric factor analysis procedure for analyzing categorical variables (Takane et al. 1979). FACTALS and GDFFA are both nonmetric factor analysis procedures, but the assumptions made for quantification are different. The difference between FACTALS and GDFFA, is most clearly seen by expressing both models as follows:

$$\text{GDFFA: } \mathbf{G}_i\mathbf{Q}_i \cong \mathbf{F}\mathbf{A}_i + \mathbf{U}_i\Psi_i, \tag{17}$$

$$\text{FACTALS: } \mathbf{G}_i\mathbf{q}_i \cong \mathbf{F}\mathbf{a}_i + \psi_i\mathbf{u}_i, \tag{18}$$

where $\cong$ denotes a least squares approximation. Thus, FACTALS loss function is defined as follows:

$$\sum_{i=1}^{p} \|\mathbf{G}_i\mathbf{q}_i - \mathbf{F}\mathbf{a}_i - \psi_i\mathbf{u}_i\|^2. \tag{19}$$

Table 1 Relationships among GDFFA, MCA, FACTALS, NPCA, PCA, and DFFA

	Model	
	PCA model	EFA model
Quantification		
Numerical	PCA	DFFA
Single	NPCA	FACTALS
Multiple	MCA	GDFFA

Essentially, while only single quantification is applied to all categorical variables in FACTALS, multiple quantification can be applied in GDFFA. Then, GDFFA and FACTALS are equivalent when all categorical variables are unidimensionally quantified. Hence, GDFFA can be viewed as a generalization of FACTALS to multiple quantification.

3.3 Relationships among GDFFA, MCA, and FACTALS

As described above, GDFFA, MCA, NPCA and FACTALS are chiefly distinguished by their model parameters and the quantification dimensions. In single quantification, quantification parameters are assumed to be $\mathbf{1}_n'\mathbf{G}_i\mathbf{q}_i = \mathbf{0}_{K_i}'$, $n^{-1}\mathbf{q}_i'\mathbf{G}_i'\mathbf{G}_i\mathbf{q}_i = 1$. In the case of numerical variables, standardized numerical variables also satisfy the constraints above. In other words, standardization of numerical variables is a restricted version of single quantification, in which quantification parameters are already known. Thus, FACTALS and NPCA are equivalent to DFFA and PCA respectively when all variables are numerical. GDFFA and MCA are also generalizations of DFFA and PCA as well as FACTALS and NPCA. The hierarchical relationships among GDFFA, MCA, FACTALS, NPCA, PCA, and DFFA are summarized in Table 1.

The purpose of PCA, NPCA and MCA is to reduce large sets of variables into smaller sets of components that summarize the information contained in the data. However, DFFA, FACTALS and GDFFA are aimed at depicting the relationships between variables and latent factors, and the rank of model parts is greater than that of the observed data. The factor scores and principal components are close to each other in some conditions (Schneeweiss and Mathes 1995), but PCA and EFA models have fundamentally different bases and standpoints.

4 Real data example

The proposed algorithm is illustrated using Japanese baseball data (Adachi 2012). We also compared GDFFA with MCA. This dataset describes the scores of 62 batters in Japanese professional baseball in 2010 under the following six variables: batting average, runs, doubles, home runs, runs batted in, and strikeouts. All variables in this dataset are numerical, and we categorized as higher or lower. Next, we added the batting order to Adachi's categorized data: group 1 comprising first and second (Nos.

Table 2 Factor loadings, unique variances (UV) and factor correlation obtained by FACTALS with $R_7 = 1$

	Factor 1	Factor 2	UV
Batting average	**0.68**	0.11	0.54
Runs	**0.99**	−0.02	0.01
Doubles	**0.67**	−0.11	0.52
Home runs	0.02	**0.85**	0.29
Runs batted in	−0.15	**0.91**	0.11
Strikeouts	0.13	**0.59**	0.65
Batting order	0.36	**0.45**	0.71
	Factor correlation		
Factor 1	1	0.11	
Factor 2		1	

1 and 2), group 2 (Nos. 3–5), and group 3 (Nos. 6–9). The analyzed data are presented in "Appendix".

In Adachi (2012), these data are best-fitted by an EFA model with two common factors ($c = 2$), which can be interpreted as expressing whether batters hit for average (table setters) or power (sluggers), respectively. In general, the batting order is mainly determined by two factors. Members of the group 1 are expected to hit consistently but not necessarily with great strength because their goal is to ensure the team has base runners for more powerful hitters who come to bat later. In contrast, members of group 2 are expected to consistently hit with power because their goal is to "drive in" base-runners. Although members of group 3 are postulated to be better at defense than hitting, some members of this group will hopefully fit more home runs than those of the group 1. In other words, the batting order categories are assumed to be closely related to the two extracted factors and not expected to be scaled unidimensionally. In GDFFA, the number of dimensions must be pre-specified. In this dataset, since $\min(K_7 - 1, c) = 2$, possible dimensions in batting order are one or two.

In Tables 2 and 3, we briefly report the matrices $\mathbf{A}$ and $\boldsymbol{\Psi}^2$ obtained from the one- and two-dimensional quantification analyses, as well as the factor correlation matrices. The MCA solution is reported in Table 4. The estimated quantification parameters are presented in Tables 5, 6 and 7. The resulting loading matrices are rotated by the geomin method (Browne 2001, p. 119) to yield a simple structure. That is, because we expect fewer and smaller cross-loadings in this data example and geomin might be used in such a situation (Schimitt and Sass 2011). Although two rotational indeterminacies exist in the MCA solution as discussed in Sect. 3.1, we only rotate the loading matrix in MCA for the comparison between GDFFA and MCA. Absolute values with coefficients exceeding 0.40 are listed in bold font.

Factor 1 is positively correlated with batting average, runs, and doubles, which are associated with reliable batters, i.e. players whose main concern is to reach bases and hit constantly. Factor 2 is positively correlated with home runs, runs batted in, and strikeouts. This second factor characterizes sluggers who frequently hit home runs and drive in a lot of runs, but whose long swings lead to many strikeouts. Thus, the same factors previously identified in Adachi (2012) are extracted in this analysis.

Table 3 Factor loadings, unique variances (UV), and factor correlation obtained by GDFFA with $R_7 = 2$

	Factor 1	Factor 2	UV
Batting average	**0.69**	0.12	0.53
Runs	**0.97**	0.01	0.05
Doubles	**0.69**	−0.11	0.48
Home runs	0.02	**0.82**	0.33
Runs batted in	−0.17	**0.93**	0.08
Strikeouts	0.12	**0.61**	0.63
Batting order (DIM1)	−0.15	**−0.59**	0.66
Batting order (DIM2)	**−0.42**	0.06	0.81
	Factor correlation		
Factor 1	1	−0.11	
Factor 2		1	

Table 4 Component loadings obtained by MCA with $R_7 = 2$

	Component 1	Component 2
Batting average	**0.78**	0.15
Runs	**0.91**	−0.01
Doubles	**0.81**	−0.11
Home runs	−0.02	**0.85**
Runs batted in	−0.24	**0.87**
Strikeouts	0.11	**0.75**
Batting order (DIM1)	−0.08	**−0.73**
Batting order (DIM2)	**−0.59**	−0.01
	Component correlation	
Component 1	1	0.06
Component 2		1

Table 5 Estimated quantification parameters in FACTALS ($R_7 = 1$)

	DIM 1
Group 1	1.50
Group 2	−0.43
Group 3	−1.17

Table 6 Estimated quantification parameters in GDFFA ($R_7 = 2$)

	DIM1	DIM2
Group 1	1.49	−0.48
Group 2	−0.82	−0.46
Group 3	0.02	2.15

Table 7 Estimated quantification parameters in MCA ($R_7 = 2$)

	DIM1	DIM2
Group 1	1.44	−0.62
Group 2	−0.86	−0.38
Group 3	0.23	2.14

The GDFFA solutions differ widely between single and multiple quantification. In the two-dimensional solution ($R_7 = 2$), the obtained batting order dimensions are strongly associated with both Factors 1 and 2. However, in the one-dimensional solution ($R_7 = 1$), which reduces to FACTALS, a large loading is assigned only to Factor 2. According to Table 6, the estimated quantification parameter in the two-dimensional solution complies with the assumptions, but the quantification parameter in the one-dimensional solution is irrational to estimated factors. Considering the roles of each group, the two-dimensional solution seems reasonable because batting order is presumably related to both factors.

In MCA, component 1 produces large loadings to batting average, runs, and doubles, and component 2 shows large loadings to home runs, runs batted in, and strikeouts. This can be interpreted as expressing table setters and sluggers, respectively. Comparing the solutions in GDFFA with $R_7 = 2$ and MCA, we get similar results. Although both solutions are quite similar, they cannot be directly compared because the purposes of GDFFA and MCA are fundamentally different, as discussed in Sect. 3.3. Trendafilov et al. (2013) pointed out that differences between PCA and EFA solutions have often been obscured because both techniques produce very similar solutions in a number of practical cases. It is possible that the same thing mentioned in Trendafilov et al. (2013) occurred in this real data example.

5 Conclusion

The GDFFA procedure has been proposed for multivariate categorical data analysis in this study. The proposed procedure is a nonmetric factor analysis model that allows multiple quantification of categorical variables. GDFFA is shown to be a generalization of the FACTALS procedure, which is designed solely for single quantification. It should be noted that DFFA, which is applicable to completely numerical variables, is a special case of nonmetric factor analysis. That is, GDFFA and FACTALS reduce to DFFA when all observed variables are numerical. Consequently, hierarchical relationships can be established among the GDFFA, FACTALS, and DFFA methods.

Although MCA and GDFFA are algebraically similar, substantial differences exist between them. In the numerical case (PCA and DFFA), it has already been mentioned that the unique term $\mathbf{U\Psi}$ distinguishes the EFA model from the PCA model (de Leeuw 2004; Trendafilov et al. 2013). In the nonmetric case (MCA and GDFFA), we have proved that the unique term in GDFFA eliminates the rotational indeterminacy of the quantification parameter, although it remains in MCA. That is, the unique term not only engenders differences between the PCA and the EFA models

but also determines whether the rotational indeterminacy exists in the quantification parameters.

As illustrated in the real data example, single quantification may be inadequate for some categories of a nominal variable. Nishisato (2006) noted that single quantification is sometimes insufficient, and recommended the use of multiple quantification for categorical data analysis . For instance, Adachi and Murakami (2011) presented examples that nominal variables cannot be unidimensionally quantified, and Murakami (2001) applied multiple quantification to categorical variables and investigated the justifiability of the Likert scale by means of (rank-restricted) MCA and NPCA. The findings of these earlier studies indicate that GDFFA can be a better alternative to MCA and NPCA.

Although we advocate multiple quantification, two problems are still present. First, it is difficult to determine the optimal number of dimensions in quantification. This difficulty is also encountered in rank-restricted MCA and other statistical methods with multiple quantification. In practice, an appropriate number of dimensions should be determined through fitting and parsimony. Considering applications, the dimension of quantification for nominal variables is set to be the same as that of conventional MCA or lower to facilitate interpretation. Second, when the number of variables is more, many iterations and significant computation time may be required for convergence of the ALS algorithm. In this paper, we applied GDFFA to the smaller dataset. For the application to GDFFA to very large datasets, acceleration techniques like NPCA (e.g., Kuroda et al. 2012) or algorithms for large datasets like rank-restricted MCA (Murakami et al. 1999; Murakami 1999) needed to be considered. Dimensions setting in multiple quantification and the application to large datasets, however, remain subjects of further discussion.

Acknowledgments The author would like to thank Prof. Kohei Adachi for his very helpful comments on previous versions of this paper.

Appendix

Japanese baseball data in 2010

The first six variables are the same as those reported in Adachi (2012), while the additional nominal variable was obtained from http://www.baseball-data.com/10/lineup/. The first six variables are as follows: [1] batting average (BA), defined as the proportion of hits to at-bats multiplied by one thousand; [2] runs (R), referring to the number of times a batter scored; [3] doubles (D), indicating the number of two-base hits; [4] home runs (HR), i.e., the number of homers hit; [5] runs batted in (RBI), referring to the number of times the batter was responsible for runs scored; and [6] strikeouts (SO) denoting the number of times the batter struck out. The six variables are dichotomized into lower (1) and higher (2). For batting order (BO), batters are categorized into the following three groups: those hitting in the first two slots (1), those hitting in the middle three slots (2), and those hitting in the bottom four slots (3) (Table 8).

Table 8 Six scores achieved by 62 batters and their batting order in Japanese professional baseball, 2010

Batter	BA	R	D	HR	RBI	S	BO	Batter	BA	R	D	HR	RBI	S	BO
1	2	2	2	1	1	1	1	32	2	2	2	1	1	2	1
2	2	2	1	1	1	1	1	33	2	2	1	1	1	1	1
3	2	2	2	2	2	1	1	34	2	2	2	1	2	1	2
4	2	2	2	2	2	1	2	35	2	1	1	2	2	2	2
5	2	2	2	2	2	1	2	36	2	2	2	2	2	2	2
6	2	2	2	1	1	1	2	37	2	2	2	1	1	1	2
7	2	2	2	2	2	2	2	38	2	2	2	1	1	1	1
8	2	1	2	1	1	1	2	39	2	1	1	1	1	2	3
9	2	2	1	2	2	2	2	40	2	2	2	2	2	2	2
10	2	2	2	1	1	1	1	41	2	2	2	1	2	2	2
11	2	2	2	2	2	2	2	42	2	2	2	1	2	1	1
12	2	2	2	2	2	1	3	43	2	2	2	1	1	2	2
13	2	2	2	2	2	2	2	44	2	2	2	1	1	1	1
14	2	1	1	1	1	1	1	45	2	2	1	1	1	1	1
15	2	2	1	2	2	2	2	46	1	2	2	1	1	1	1
16	1	1	1	1	1	1	2	47	1	2	2	1	2	2	2
17	1	1	1	1	1	2	1	48	1	2	2	2	2	2	2
18	1	1	2	1	1	1	1	49	1	1	1	1	1	2	1
19	1	1	1	1	1	1	3	50	1	1	2	1	2	1	2
20	1	1	1	2	1	1	3	51	1	1	2	2	2	2	2
21	1	1	1	1	1	1	3	52	1	1	1	1	1	1	2
22	1	2	2	2	2	2	2	53	1	1	1	1	1	1	1
23	1	2	2	2	2	1	1	54	1	1	1	2	2	2	2
24	1	1	1	1	1	1	3	55	1	1	1	2	2	2	2
25	1	1	1	1	1	1	3	56	1	1	1	1	1	1	2
26	1	1	1	2	1	1	3	57	1	1	1	2	2	2	3
27	1	1	1	1	1	1	2	58	1	1	1	1	1	2	2
28	1	1	1	1	1	1	1	59	1	1	1	1	1	2	3
29	1	1	1	2	2	2	2	60	1	1	1	2	2	2	3
30	1	2	2	2	2	2	2	61	1	1	1	2	2	2	2
31	1	1	2	2	2	2	2	62	1	1	1	2	2	2	2

References

Adachi K (2012) Some contributions to data-fitting factor analysis with empirical comparisons to covariance-fitting factor analysis. J Jpn Soc Comput Stat 25:25–38

Adachi K, Murakami T (2011) Nonmetric multivariate analysis: MCA, NPCA, and PCA. Asakura Shoten, Tokyo (in Japanese)

Anderson TW, 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, pp 111–150

Benzecri JP (1974) L'analyses des donnees: Tome (VoL) 1. La taxinomie: Tome. 2 La'analyses des correspondances. Dunod, Paris
Benzecri JP (1992) Correspondence analysis handbook. Marcel Dekker, New York
Browne MW (2001) An overview of analytic rotation in exploratory factor analysis. Multivar Behav Res 36:111–150
de Leeuw J (2004) Least squares optimal scaling of partially observed linear systems. In: van Montfort K, Oud J, Satorra A (eds) Recent developments on structural equation models: theory and applications. Kluwer, Dordrecht, pp 121–134
de Leeuw J (2008) Factor analysis as matrix decomposition. Preprint series: Department of Statistics, University of California, Los Angeles
Greenacre MJ (1984) Theory and application of correspondence analysis. Academic Press, London
Gifi A (1990) Nonlinear multivariate analysis. Wiley, Chichester
Harman HH (1976) Mordan factor analysis, 3rd edn. University of Chicago Press, Chicago
Kuroda M, Mori Y, Iizuka M, Sakakihara M (2012) Acceleration of convergence of the alternating least squares algorithm for nonlinear principal components analysis. In Sanguansat P (ed) Principal component analysis. InTech, Winchester
Mulaik SA (2010) Foundation of factor analysis, 2nd edn. Chapman and Hall/CRC, Boca Raton
Murakami T (1999) A psychometrics study on principal component analysis of categorical data. Technical report (in Japanese)
Murakami T (2001) Investigation of justifiability of Likert scaling using nonmetric principal components analysis and multiple correspondence analysis. Technical report (in Japanese)
Murakami T, Kiers HAL, ten Berge JMF (1999) Non-metric principal component analysis for categorical variables with multiple quantifications (unpublished manuscript)
Nishisato S (2006) Multidimensional nonlinear descriptive analysis. Chapman and Hall/CRC, Boca Raton
Schimitt TA, Sass DA (2011) Rotation criteria and hypothesis testing for exploratory factor analysis: implications for factor pattern loadings and interfactor correlations. Educ Psychol Measur 71:95–113
Schneeweiss H, Mathes H (1995) Factor analysis and principal components. J Multivar Anal 55:105–124
Takane Y, de Young FW, Leeuw J (1979) Nonmetric common factor anaysis: an alternating least square method with optimal scaling features. Behaviormetrika 6:45–56
ten Berge JMF (1993) Least squares optimazation in multivariate analysis. DSNO Press, Leiden
Tenenhaus M, Young YW (1985) An analysis and synthesis of multiple correspondence analysis, optimal scaling, dualscaling, homogeneity analysis and other methods for quantifying categorical multivariate data. Psychometrika 50:91–119
Trendafilov NT, Unkel S, Krzanowski W (2013) Exploratory factor and principal component analyses: some new aspects. Stat Comput 23:209–220
Unkel S, Trendafilov NT (2010) Simultaneous parameter estimation in exploratory factor analysis: an expository reviews. Int Stat Rev 78:363–382
Unkel S, Trendafilov NT (2011) Zig-zag exploratory factor analysis with more variables than observations. Comput Stat 28:107–125
van der Burg E, de Leeuw Y, Verdegaal R (1988) Homogeneity analysis with k sets of variables: An alternating least squares method with optimal scaling features. Psychometrika 53:177–197
Yanai H, Ichikawa M (2007) Factor analysis. In: Rao CR, Sinharay S (eds) HandBook of statistics vol 26: Psychometrics. Elsevier, Amsterdam, pp 257–296
Young FW (1981) Quantitative analysis of qualitative date. Psychometrika 46:357–388