Journal of the
American
Statistical
Association

Taylor & Francis
Taylor & Francis Group

Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: www.tandfonline.com/journals/uasa20

Generalized Factor Model for Ultra-High Dimensional Correlated Variables with Mixed Types

Wei Liu, Huazhen Lin, Shurong Zheng & Jin Liu

To cite this article: Wei Liu, Huazhen Lin, Shurong Zheng & Jin Liu (2023) Generalized Factor Model for Ultra-High Dimensional Correlated Variables with Mixed Types, Journal of the American Statistical Association, 118:542, 1385-1401, DOI: 10.1080/01621459.2021.1999818

To link to this article: https://doi.org/10.1080/01621459.2021.1999818

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2023, VOL. 118, NO. 542, 1385–1401: Theory and Methods
https://doi.org/10.1080/01621459.2021.1999818

Taylor & Francis
Taylor & Francis Group

Generalized Factor Model for Ultra-High Dimensional Correlated Variables with Mixed Types

Wei Liu[a], Huazhen Lin[a], Shurong Zheng[b], and Jin Liu[c]

[a]Center of Statistical Research and School of Statistics, Southwestern University of Finance and Economics, Chengdu, China; [b]School of Mathematics and Statistics, Northeast Normal University, Changchun, China; [c]Centre for Quantitative Medicine, Program in Health Services & Systems Research, Duke-NUS Medical School, Singapore, Singapore

ABSTRACT

As high-dimensional data measured with mixed-type variables gradually become prevalent, it is particularly appealing to represent those mixed-type high-dimensional data using a much smaller set of so-called factors. Due to the limitation of the existing methods for factor analysis that deal with only continuous variables, in this article, we develop a generalized factor model, a corresponding algorithm and theory for ultra-high dimensional mixed types of variables where both the sample size n and variable dimension p could diverge to infinity. Specifically, to solve the computational problem arising from the non-linearity and mixed types, we develop a two-step algorithm so that each update can be carried out in parallel across variables and samples by using an existing package. Theoretically, we establish the rate of convergence for the estimators of factors and loadings in the presence of nonlinear structure accompanied with mixed-type variables when both n and p diverge to infinity. Moreover, since the correct specification of the number of factors is crucial to both the theoretical and the empirical validity of factor models, we also develop a criterion based on a penalized loss to consistently estimate the number of factors under the framework of a generalized factor model. To demonstrate the advantages of the proposed method over the existing ones, we conducted extensive simulation studies and also applied it to the analysis of the NFBC1966 dataset and a cardiac arrhythmia dataset, resulting in more predictive and interpretable estimators for loadings and factors than the existing factor model.

ARTICLE HISTORY
Received October 2019
Accepted September 2021

KEYWORDS
Convergence rate; Generalized factor model; Mixed type of data; Nonlinear; Ultra-high dimension; Uniform consistency

1. Introduction

In the era of “Big Data,” factor models have been shown to be effective in simultaneously modeling the commonality and cross-sectional dependence of the observed data (Fan, Xue, and Yao 2017), reviving as a powerful framework in data compression and dimension reduction. Traditional factor models (Lawley 1940; Anderson and Rubin 1956; Amemiya, Fuller, and Pantula 1987) usually take the following assumptions: all manifest variables (observed variables) are normally distributed; the number of manifest variables is fixed; the number of latent variables is a known priori; the error term is homoscedastic. These assumptions are too stringent and can be easily violated. Since the pioneering work by Spearman (1904, 1927), many efforts have been made to relax these assumptions. For example, Browne (1984) and Amemiya and Anderson (1990) weakened the requirements by allowing all manifest variables to be continuous rather than normal. McDonald (1969), Olsson (1979), Bartholomew (1980), and Muthén (1984) further relaxed the restriction for categorical variables. Bock and Lieberman (1970), Christoffersson (1975), and Muthén (1978) considered the framework of threshold factor modeling for binary manifest variables. However, all the above works are limited to the fixed dimension and single-type manifest variables, and stuck in computation problem due to multiple integration. In practice, for example, in genetic studies with tens of millions of SNPs along the human genome, the dosage of SNPs is the number of reference alleles, taking values of $\{0, 1, 2\}$, while microbiome data from next-generation sequencing are the count data.

To deal with mixed-type manifest variables, most works convert the mixed types into a single type. For example, Bartholomew (1987) transformed the continuous variables into categorical variables, and Muthén (1984) introduced latent continuous surrogates of the discrete variables. Alternatively to those converting methods, the latent trait model has been developed by Moustaki (1996) for mixtures of continuous and binary variables, which is further extended to the generalized latent trait model (Moustaki and Knott 2000). All methods were summarized by Bartholomew, Knott, and Moustaki (2011) and related developments and future directions of the latent trait model were presented by Cudeck and MacCallum (2012). A limitation of these works is the requirement of fixed dimension of the manifest variables.

To break the restriction of fixed dimension, Bai and Ng (2002) introduced the high-dimensional factor model for continuous variables and made the factor models applicable

CONTACT Huazhen Lin linhz@swufe.edu.cn Center of Statistical Research and School of Statistics, Southwestern University of Finance and Economics, 555 Liutai road, Wenjiang district, Chengdu 611130, China.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

to the high-dimensional data. Along the line, Stock and Watson (2002) considered time series data. Bai (2003) proposed the inferential theory; Bai et al. (2012) further proposed a maximum likelihood estimation method, where they presented asymptotical properties for high-dimensional factor model. Bai and Ng (2013) studied conditions under which the latent factors can be estimated asymptotically by using the principal component analysis (PCA) to avoid rotation. Recently, Fan, Xue, and Yao (2017) and Jiang, Ma, and Wei (2019) considered supervised latent models, and Li et al. (2018) proved the properties of the blessing of dimensionality in factor models. However, all of these works are based on the linear relationship between the manifest variables and latent variables. The assumption for linearity is not technical but is crucial for the validity of their theories and computational algorithms. In particular, a closed form of the estimated factors, which can be derived only under the linear structure, is the key to develop related theories and algorithms. Moreover, the PCA-based estimation for the factors is only valid for the linear structure.

In applications, high-dimensional data with mixed-type beyond single-type continuous variables are collected rapidly. For instance, our motivating examples, the NFBC1966 dataset and a cardiac arrhythmia dataset, are the cases. NFBC1966 is a dataset for genome-wide association studies (GWAS), consisting of five metabolic traits and 364,590 SNPs from 5,402 individuals (Sabatti et al. 2009), where covariates contain continuous variables, for example, Body Mass Index (BMI), and the dosage of SNPs takes categorical values of $\{0, 1, 2\}$. The cardiac arrhythmia dataset contains 61 binary and 165 continuous features measured for 420 individuals. Not only using a linear relationship to extract factors from the categorical or discrete variables makes the model not interpretable, but also the corresponding models lose the power for further predictions. In particular, for the NFBC1966 dataset, the resulting cross-validated predicted normalized mean square errors (NMSEs; Torgo 2011) averaging over 100 random splits are 0.526 and 1.000 for the proposed generalized factor model (GFM) and the existing linear factor model (LFM) (Bai and Ng 2002), respectively, where $NMSE = \frac{\sum_{i=1}^{n_1}(\hat{Y}_i - Y_i)^2}{\sum_{i=1}^{n_1}(\bar{Y} - Y_i)^2}$, $\bar{Y} = n_1^{-1}\sum_{i=1}^{n_1} Y_i$ on the testing dataset. The average NMSE of LFM is about equal to 1, so the prediction error based on LFM is approximately equivalent to that based on the mean value of Y_i, indicating that the LFM model does not provide more predictive power for Y_i. Similarly, for the cardiac arrhythmia dataset, the performance of LFM is nearly equivalent to random guessing, while that of GFM is close to the method using all data information although only eight extracted factors by GFM, which leads to more concise, informative, and interpretable classification.

In this article, based on the framework for the exponential family, we propose the generalized factor model (GFM) for ultra-high dimensional data with mixed-type variables. Challenges for the GFM come from both theoretical and computational perspectives due to the following facts: (i) the closed form of the estimated factors are not available under the generalized model; (ii) the nonlinear structure varies with the type of the manifest variables; (iii) any linear approximation error to nonlinear structure in the high-dimensional environment is not ignorable. To address these issues, we develop a matrix blocking technique to establish the rate of convergence, and propose an iterative procedure in which each update can be conducted in parallel across variables and samples using the existing package. The matrix blocking technique is to separate the Hessian matrix of the log-likelihood function as a 2 × 2 symmetric matrix with respects to factors and loadings. Based on the matrix blocking technique, we hence can decompose the rules of factors and loadings, which is crucial to establish the rate of convergence of estimators. Moreover, since the correct specification of the number of factors is essential to both the theoretical and empirical validity of factor models (Bai and Ng 2002; Lam and Yao 2012), we also develop a formal statistical procedure that can consistently estimate the number of factors as both n and p diverge to infinity under the framework of the generalized factor model. To our knowledge, the study in the article would be the first attempt to discuss the theory and computation of a factor model for analyzing data with ultra-high dimensional variables of mixed types under a nonlinear framework.

The article is organized as follows. We begin in Section 2 with a description of the model and the proposed estimation procedure. The asymptotic properties of the proposed estimators are presented in Section 3. The performance of the proposed estimation procedure is assessed by simulation studies in Section 4. In Section 5, we apply the proposed method to analyze two real datasets, that is, the NFBC1966 dataset and a cardiac arrhythmia dataset. The results reveal that GFM can effectively extract more concise and informative factors than the existing methods so that follow-up analysis based on the extracted factors has higher prediction and classification accuracy and interpretability. A brief discussion is made to extend the proposed method in Section 6. Technical details for proofs are relegated to the supplementary materials. Moreover, we develop our proposed method into an efficient and user-friendly MATLAB toolbox called GFM, which is available at https://github.com/LinhzLab/GFM.

2. Model and Estimation

2.1. Model

Suppose that the observations $\mathbf{x}_i \in R^p, i = 1, \ldots, n$, are independent identically distributed (iid), where $n \to \infty$ and $p \to \infty$. We assume $\mathbf{x}_i$ are correlated because of sharing a latent factor $\mathbf{h}_i$, that is, we assume that $\mathbf{x}_i$'s are independent conditional on $\mathbf{h}_i$, and $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ may include continuous variables, binary variables, count variables, and other types of variables. Generalized linear model (McCullagh and Nelder 1989) provided a distributional framework accounting for the type of a variable. More specifically, we consider the conditional density function of x_{ij} given $\mathbf{h}_i$, $f_j(x_{ij}|\mathbf{h}_i)$, which belongs to a canonical exponential family model,

$$f_j(x|\mathbf{h}_i) = \exp[\{x(\mathbf{b}_j^T\mathbf{h}_i + \mu_j) - \nu_j(\mathbf{b}_j^T\mathbf{h}_i + \mu_j)\}/c_j + \phi_j(x, c_j)],$$
$$1 \le j \le p, 1 \le i \le n, \qquad (1)$$

where $\nu_j(\theta)$ is some known function, c_j is a known scale parameter, and both are determined by the type of variable j. For example, if x_{ij} is a count variable that follows a Poisson distribution, $\nu_j(\theta) = e^\theta$ and $c_j = 1$; if x_{ij} is a binary variable that follows a Bernoulli distribution, $\nu_j(\theta) = \ln(1+e^{\theta})$ and $c_j = 1$; if x_{ij} is a continuous variable that follows a normal distribution, $\nu_j(\theta) = \frac{\theta^2}{2}$ and c_j is the variance of x_{ij} that can be replaced by sample variance (McCullagh and Nelder 1989; Moustaki and Knott 2000). $\phi_j(x, c_j)$ is independent of the parameters interested and can be ignored. $\mathbf{h}_i = (h_{i1}, \dots, h_{iq})^T$ is a vector of q-dimensional latent factors with $q \ll p$ and $E(\mathbf{h}_i) = 0$, $\mathbf{b}_j$ and μ_j are the corresponding loading vector and intercept scalar of variable j, respectively. Denote $\mathbf{H} = (\mathbf{h}_1, \dots, \mathbf{h}_n)^T$, $\mathbf{B} = (\mathbf{b}_1, \dots, \mathbf{b}_p)^T$, $\boldsymbol{\mu} = (\mu_1, \dots, \mu_p)^T$, $\boldsymbol{\gamma}_j = (\mu_j, \mathbf{b}_j^T)^T$ and $\Upsilon = (\boldsymbol{\mu}, \mathbf{B}) \in R^{p\times(q+1)}$. Throughout the article, we focus on the estimation of factors $\mathbf{H}$, loadings $\mathbf{B}$, and intercepts $\boldsymbol{\mu}$. It is obvious that the model (1) is not identifiable. To make it identifiable, we impose three constraints:

(A1) $\frac{1}{n}\sum_{i=1}^{n} \mathbf{h}_i = 0$ and $\frac{1}{n}\mathbf{H}^T\mathbf{H} = \mathbf{I}_q$;
(A2) $\mathbf{B}^T\mathbf{B}$ is diagonal with decreasing diagonal elements;
(A3) the first nonzero element in each column of $\mathbf{B}$ is positive.

Condition (A1), a restriction on factors, is commonly used in the literature of factor analysis (McDonald 1985; Jolliffe 2002; Bai and Ng 2013; Fan, Liao, and Mincheva 2013; Li et al. 2018). Conditions (A2) and (A3) on loadings are a little different from Bai and Ng (2002), Bai (2003), and Bai et al. (2012) in which $\mathbf{B}$ and $\mathbf{H}$ can be only determined uniquely up to a sign matrix (diagonal matrix with 1 and -1 on the diagonals) transformation. When all of $\mathbf{x}_i$ are normal or continuous variables, the identifiability Condition (A3) is also adopted by Jiang, Ma, and Wei (2019).

We state the identifiability in the following proposition, and its proof is deferred to the supplementary materials.

Proposition 1. Let $(\mathbf{H}_1, \mathbf{B}_1, \boldsymbol{\mu}_1)$ and $(\mathbf{H}_2, \mathbf{B}_2, \boldsymbol{\mu}_2)$ be two set of parameters satisfying model (1) and Conditions (A1)–(A3), then we have $\mathbf{H}_1 = \mathbf{H}_2, \mathbf{B}_1 = \mathbf{B}_2$ and $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$.

2.2. Estimation

Since the information about the dependence of $\mathbf{x}_i$ is fully captured by $\mathbf{h}_i$, $x_{i1}, \dots, x_{ip}$ are conditional independent given $\mathbf{h}_i$, for example, $f(\mathbf{x}_i|\mathbf{h}_i) = \Pi_{j=1}^{p} f_j(x_{ij}|\mathbf{h}_i)$. The conditional likelihood function given $\mathbf{H}$ is $L_{\text{con}}(\mathbf{X}; \mathbf{H}, \Upsilon) = \Pi_{i=1}^{n}\Pi_{j=1}^{p} f_j(x_{ij}|\mathbf{h}_i)$, where $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^T$. The observational likelihood is $L_{\text{obser}}(\mathbf{X}; \Upsilon) = \Pi_{i=1}^{n} \int \left\{\Pi_{j=1}^{p} f_j(x_{ij}|\mathbf{h}) f_{\mathbf{h}}(\mathbf{h})\right\} d\mathbf{h}$, where $f_{\mathbf{h}}(\mathbf{h})$ is the p.d.f. of $\mathbf{h}_i$. Since $\mathbf{H}$ is unobserved, Moustaki and Knott (2000) maximize $L_{\text{obser}}(\mathbf{X}; \Upsilon)$ with fixed p, which suffers from intensive computation due to multiple integral on $\mathbf{h}_i$. To make it computationally feasible, Stock and Watson (2002), Bai and Ng (2002), Fan, Liao, and Mincheva (2013), and Bai and Ng (2013) treated the factors $\mathbf{h}_i, i = 1, \dots, n$ as parameters and simultaneously estimated loadings and factors from the conditional likelihood, which is least-square error under a linear factor model. Furthermore, Stock and Watson (2002), Bai and Ng (2002), Fan, Liao, and Mincheva (2013), and Bai and Ng (2013) applied this idea to extract latent factors $\mathbf{h}_i$ by a principle component method based on the least-square criterion under a linear factor model. In this article, we borrow the idea that treating $\mathbf{h}_i$ as a parameter. Particularly, we estimate $(\mathbf{H}, \Upsilon)$ by maximizing the conditional log-likelihood function,

$$l(\mathbf{H}, \Upsilon) = \sum_{i=1}^{n} \sum_{j=1}^{p} L_j(\mathbf{h}_i, \boldsymbol{\gamma}_j; x_{ij}), \tag{2}$$

where $L_j(\mathbf{h}_i, \boldsymbol{\gamma}_j; x_{ij}) = \{x_{ij}(\mathbf{b}_j^T\mathbf{h}_i + \mu_j) - \nu_j(\mathbf{b}_j^T\mathbf{h}_i + \mu_j)\}/c_j + C$, and C being independent of the parameters of interest. If all of $\mathbf{x}_i$ are normal or continuous variables with homoscedasticity, problem (2) is the ordinary least-square estimate and reduced to Bai and Ng (2002) and Fan, Liao, and Mincheva (2013). If all of $\mathbf{x}_i$ are normal or continuous but with heteroscedasticity, then (2) is the weighted least-square estimate and reduced to (3.1) in Bai and Liao (2013).

Hence, when $\mathbf{x}_i$ is normally distributed or continuous, the estimators for Υ and $\mathbf{H}$ have closed-form solutions that are computationally feasible and theoretically establishable. Moreover, Bai and Ng (2013) proposed the principal component method to estimate the latent factors $\mathbf{H}$ so that the rotation can be avoided. A key for all of these methods is the linear structure $\nu_j'(x) = x$. When x_{ij} is a binary or count variable, that is, $\nu_j'(x) \neq x$, there exists no closed-form estimators for Υ and $\mathbf{H}$, and the principal component method does not work. Furthermore, the optimization for problem (2) is further complicated by the mixed types of $\mathbf{x}_i$. Particularly, for example, to estimate $\mathbf{h}_i$ given Υ, we maximize $\sum_{j=1}^{p} L_j(\mathbf{h}_i, \boldsymbol{\gamma}_j; x_{ij})$ respect to $\mathbf{h}_i$. It looks like the parameter estimation problem for response x_{ij} and covariates $\boldsymbol{\gamma}_j$ by taking index $j, j = 1, \dots, p$, as "samples". However, it is not straightforward for computation as x_{ij}'s have different distributions for different "samples" j's. In other words, the distributions of "samples" are individual-specific. So far, there exists no available packages for such problems.

In this article, we propose a two-step procedure to estimate Υ and $\mathbf{H}$. The advantage of our procedure is that the estimation can be conducted in parallel across all variables and individuals by using the existing package. The programming and computation are simple.

2.2.1. The First-Step Estimation for H and Υ

To make the presentation clear, we first alternately estimate Υ and $\mathbf{H}$ given the rest. Then, we present the iterative algorithm. When $\mathbf{H}$ is given, we estimate $\boldsymbol{\gamma}_j$ as

$$\tilde{\boldsymbol{\gamma}}_j = \text{argmax}_{\boldsymbol{\gamma}_j} n^{-1} \sum_{i=1}^{n} L_j(\mathbf{h}_i, \boldsymbol{\gamma}_j; x_{ij}), \tag{3}$$

where $L_j(\mathbf{h}_i, \boldsymbol{\gamma}_j; x_{ij}) = \log\{f_j(x_{ij}|\mathbf{h}_i)\}$. The estimator based on Equation (3) is equivalent to fit a generalized linear model based on the observations $(\mathbf{h}_i, x_{ij}), \forall i = 1, \dots, n$, where the response is x_{ij} and the covariate is $\mathbf{h}_i$ under the conditional distribution $f_j(x|\mathbf{h}_i)$. Thus, the computation can be implemented in parallel over $j = 1, \dots, p$ by using the *glmfit* function in MATLAB software.

Given Υ, denote $L_j(\boldsymbol{\gamma}_j, \mathbf{h}_i; x_{ij}) = \{x_{ij}(\mathbf{b}_j^T\mathbf{h}_i + \mu_j - \nu_j(\mathbf{b}_j^T\mathbf{h}_i + \mu_j)\}/c_j + C$, and A_s to be the indicator set of the variables belonging to type s. We assume that the components of A_s are large enough, and then we estimate $\tilde{\mathbf{h}}_i$ for $\mathbf{h}_i$ as follows:

$$\tilde{\mathbf{h}}_i = \arg\max_{\mathbf{h}_i} \sum_{j\in A_s} L_{P_s}(\boldsymbol{\gamma}_j, \mathbf{h}_i; x_{ij}). \tag{4}$$

Problem (4) is the estimation of coefficient in a generalized linear model with the common distribution f_{P_s} and the samples $(\boldsymbol{\gamma}_j, x_{ij}), j \in A_s$, where x_{ij} is the response, $\mathbf{b}_j$ is the covariate and μ_j is an offset term. Specifically, we use the *glmfit* function in MATLAB to fit the generalized linear model.

To start the proposed iterative algorithm, we first consider the initial value $(\mathbf{H}^{[0]}, \Upsilon^{[0]})$ for $(\mathbf{H}, \Upsilon)$. Recall $\Upsilon = (\boldsymbol{\mu}, \mathbf{B})$, we obtain $\boldsymbol{\mu}^{[0]} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$ and $(\mathbf{B}^{[0]}, \mathbf{H}^{(0)})$ from the linear factor model (Bai and Ng 2002) on the centered sample by performing PCA. Let $(\mathbf{H}^{[r-1]}, \Upsilon^{[r-1]})$ be the estimators of $(\mathbf{H}, \Upsilon)$ obtained after the $(r-1)$th iteration. In the rth iteration, we update the estimators as follows:

- Obtain $\Upsilon_n = (\boldsymbol{\mu}_n, \mathbf{B}_n)$ in Equation (3) with $\mathbf{H}$ replaced by $\mathbf{H}^{[r-1]}$. To adhere to the identification condition on $\mathbf{B}$, we further perform a singular value decomposition to get $\mathbf{B}_n = S_1\Lambda^{1/2}D_1$ and update $\Upsilon^{[r-1]}$ by $\Upsilon^{[r]} = (\boldsymbol{\mu}_n, \mathbf{B}^{[r]})$, where $\mathbf{B}^{[r]} = S_1\Lambda^{1/2}$ and the first nonzero element in each column of $\mathbf{B}^{[r]}$ is positive.
- Obtain $\mathbf{H}_n$ in Equation (4) with Υ replaced by $\Upsilon^{[r]}$. Likewise, by taking into account the identification condition on $\mathbf{H}$, we normalize $\mathbf{H}_n$ to get $\mathbf{H}^{[r]}$ so that the identifiability condition (A1) can be satisfied.

Repeat the iteration until convergence and denote the estimator as $(\widetilde{\mathbf{H}}, \widetilde{\Upsilon})$. In practice, the convergence is defined as $\sup \|\Upsilon^{[r]} - \Upsilon^{[r-1]}\| + \sup \|\mathbf{H}^{[r]} - \mathbf{H}^{[r-1]}\| < a_0$ or $|l(\mathbf{H}^{[r]}, \Upsilon^{[r]}) - l(\mathbf{H}^{[r-1]}, \Upsilon^{[r-1]})| < a_0$, where a_0 is a prespecified small positive number. The computation of Equations (3) and (4) can be performed in parallel over $i = 1, \ldots, n, j = 1, \ldots, p$ and hence the computation time is limited. In the supplementary materials, we show that the proposed iterative algorithm converges, which is stated in the following proposition. Let Ω be the parameter space of the parameter $(\mathbf{H}, \Upsilon)$ satisfying Conditions (A1)–(A3).

Proposition 2. If Assumptions (S1)–(S4) in the supplementary materials hold, given the proposed iterative algorithm based on Equations (3) and (4), we have that all the limit points of $(\mathbf{H}^{[r]}, \Upsilon^{[r]})$ are local maxima of $l(\mathbf{H}, \Upsilon)$ in the parameter space Ω, and $l(\mathbf{H}^{[r]}, \Upsilon^{[r]})$ converges monotonically to $L^* = l(\mathbf{H}^*, \Upsilon^*)$ for some $(\mathbf{H}^*, \Upsilon^*) \in \mathcal{G}$, where $\mathcal{G} = \{$set of local maxima of $l(\cdot)$ in the interior of $\Omega\}$.

2.2.2. The Second-Step Estimation for H and Υ

Although the computation for $(\widetilde{\mathbf{H}}, \widetilde{\Upsilon})$ is simple, the efficiency may be lost since $\widetilde{\mathbf{H}}$ is obtained not based on the log-likelihood function $l(\mathbf{H}, \Upsilon)$. To ensure the efficiency, we then conduct a one-step update based on score function and Hessian matrix of $l(\mathbf{H}, \Upsilon)$. Let the score functions $\nabla_{\mathbf{h}_i} l(\mathbf{H}, \Upsilon) = \partial l(\mathbf{H}, \Upsilon)/\partial \mathbf{h}_i$ and $\nabla_{\boldsymbol{\gamma}_j} l(\mathbf{H}, \Upsilon) = \partial l(\mathbf{H}, \Upsilon)/\partial \boldsymbol{\gamma}_j$, and the Hessian matrices $\nabla^2_{\mathbf{h}_i} l(\mathbf{H}, \Upsilon) = \partial^2 l(\mathbf{H}, \Upsilon)/\partial \mathbf{h}_i \partial \mathbf{h}_i^T$ and $\nabla^2_{\boldsymbol{\gamma}_j} l(\mathbf{H}, \Upsilon) = \partial^2 l(\mathbf{H}, \Upsilon)/\partial \boldsymbol{\gamma}_j \partial \boldsymbol{\gamma}_j^T$ regarding $\mathbf{h}_i$ and $\boldsymbol{\gamma}_j$, respectively. We update $\widetilde{\mathbf{h}}_i$ and $\widetilde{\boldsymbol{\gamma}}_j$ by

$$\hat{\mathbf{h}}_i = \widetilde{\mathbf{h}}_i - \left\{\nabla^2_{\mathbf{h}_i} l(\widetilde{\mathbf{H}}, \widetilde{\Upsilon})\right\}^{-1} \nabla_{\mathbf{h}_i} l(\widetilde{\mathbf{H}}, \widetilde{\Upsilon}), i = 1, \ldots, n. \quad (5)$$

$$\hat{\boldsymbol{\gamma}}_j \hat{=} (\hat{\mu}_j, \hat{\mathbf{b}}_j^T)^T = \widetilde{\boldsymbol{\gamma}}_j - \left\{\nabla^2_{\boldsymbol{\gamma}_j} l(\widetilde{\mathbf{H}}, \widetilde{\Upsilon})\right\}^{-1} \nabla_{\boldsymbol{\gamma}_j} l(\widetilde{\mathbf{H}}, \widetilde{\Upsilon}),$$
$$j = 1, \ldots, p. \quad (6)$$

Denote the final estimator still as $(\widehat{\mathbf{H}}, \widehat{\Upsilon})$. Again, the computation of Equations (5) and (6) for $\hat{\mathbf{h}}_i$ and $\hat{\boldsymbol{\gamma}}_j$ can be performed in parallel over $i = 1, \ldots, n, j = 1, \ldots, p$ and hence the computation time is largely reduced even for large n and p.

Remark 1. Since the estimators $(\tilde{\mathbf{H}}, \tilde{\mathbf{B}})$ from the first step are identifiable, the one-step correction in Step 2 based on Equations (5) and (6) has closed form and hence is identifiable. In the first step, we can also directly perform SVD decomposition on $\mathbf{H}^{[r]}(\mathbf{B}^{[r]})^T$ in each iterative step for identification, but this will increase computational cost. The reason is that the computation complexity of SVD on $\mathbf{HB}^T \in R^{n\times p}$, $O(\min(np^2, n^2p)$, is much higher than that for both $\mathbf{H} \in R^{n\times q}$ and $\mathbf{B} \in R^{p\times q}$, $O(nq^2+pq^2)$, where $q \ll \min(n, p)$ with both n and p tending to infinity. In the simulation studies, we also compared our method with the method that simultaneously exerts the identifiability condition on $\mathbf{HB}^T$. The resulting estimators suggest that the two methods are nearly equivalent and the proposed method performs a little bit better in terms of numerical performance; See Section 4.6.

Remark 2. It can be shown that the asymptotic properties of $(\mathbf{H}, \widehat{\Upsilon})$ is independent of the choice of $(\widetilde{\mathbf{H}}, \widetilde{\Upsilon})$ given the consistency of $(\widetilde{\mathbf{H}}, \widetilde{\Upsilon})$, which has been proved if the number of variables for type s is large enough. If we denote $\tilde{\mathbf{h}}_i^{(s)}$ to be the estimator from Equation (4) based on the variables of type s, these conclusions imply that we can set $\widetilde{\mathbf{h}}_i$ to be one of $\tilde{\mathbf{h}}_i^{(s)}, s = 1, \ldots, d$, providing the number of variables for type s is large enough. In practice, we choose the type that has strong signal, such as the type having the large number of variables or being continuous or counting type, to estimate $\widetilde{\mathbf{h}}_i$. In addition, to improve the robustness, we also can use several types, for example, $s = 1, \ldots, d_1$, to update $\mathbf{H}$ by $\widetilde{\mathbf{h}}_i = \frac{1}{d_1}\sum_{s=1}^{d_1} \tilde{\mathbf{h}}_i^{(s)}, i = 1, \ldots, n$.

Remark 3. Our GFM can be extended to the regularized GFM. Regularized factor analysis (Lan et al. 2014; Carvalho et al. 2008) and sparse principal component analysis (Zou, Hastie, and Tibshirani 2006) have been proposed to improve the interpretability of factors and principal components in a linear factor model. We can also extend Equation (2) to the regularized version,

$$(\widehat{\Upsilon}^{Re}, \widehat{\mathbf{H}}^{Re}) = \arg\max_{\Upsilon, \mathbf{H}} \left\{ n^{-1} \sum_{i=1}^{n}\sum_{j=1}^{p} L_j(\mathbf{h}_i, \mathbf{b}_j, \mu_j; x_{ij}) - \sum_{j=1}^{p} P_\lambda(\mathbf{b}_j) \right\}, \quad s.t. \frac{1}{n}\mathbf{H}^T\mathbf{H} = \mathbf{I}_q, \mathbf{B}^T\mathbf{B} \text{ is diagonal.} \quad (7)$$

where $P_\lambda(\cdot)$ is a penalty function and λ is a penalty parameter. The optimization of the objective function (7) is the same as Equation (2) except for a slight modification of Equation (3) to,

$$\tilde{\boldsymbol{\gamma}}_j^{Re} = \text{argmax}_{\mathbf{b}_j} \left\{ n^{-1} \sum_{i=1}^{n} L_j(\mathbf{h}_i, \mathbf{b}_j, \mu_j; x_{ij}) - P_\lambda(\mathbf{b}_j) \right\}, \quad (8)$$

which is equivalent to fit a penalized maximum likelihood estimator based on data $\{x_{ij}, \mathbf{h}_i\}, i = 1, \ldots, n$ with the distribution $f_j(x|\mathbf{h}_i)$. The computation can also be performed in parallel over $j = 1, \ldots, p$ by the *glmnet* function to maximize the objective function in MATLAB software. As shown in the analysis of the NFBC1966 dataset, regularized GFM provides stronger interpretability for latent factor analysis.

2.3. Determining the Number of Factors

The correct specification of the number of factors is a key to both the theoretical and the empirical validity of factor models (Bai and Ng 2002). Under linear factor model, the criteria based on the penalized loss (PC) (Bai and Ng 2002) and eigenvalue ratio test (Lam and Yao 2012; Ahn and Horenstein 2013) are proposed to consistently estimate the number of factors. The PC criteria use penalized least-square error, while the eigenvalue ratio test is based on the ordered eigenvalues of covariance matrix of $\mathbf{x}_i$. Since the correlation coefficient or covariance is not suitable to describe the correlation of non-continuous random variables, in this article, we use the idea of PC criteria to select the number of factors.

Given the number of factors k, let $(\hat{\mathbf{H}}(k), \hat{\Upsilon}(k))$ be the estimator of $(\mathbf{H}, \Upsilon)$. Then we estimate the number of factors q by

$$\hat{q} = \arg\min_k \{-\frac{1}{np} l(\hat{\mathbf{H}}(k), \hat{\Upsilon}(k)) + k \times g(n,p)\}, \tag{9}$$

where $g(n,p)$ is a prespecified penalty function. Theorem 4 in Section 3 shows that $\hat{q}$ is a consistent estimator of q by choosing a "good" $g(n,p)$. In simulation studies and real data analysis, we choose the penalty $g(n,p) = \frac{n+p}{np}\ln(\frac{np}{n+p})$ which satisfies the conditions in Theorem 4 and is confirmed to work well.

3. Asymptotical Theory

We now establish the theoretical properties of the proposed estimators $\widehat{\mathbf{h}}_i$, $\widehat{\boldsymbol{\gamma}}_j$ and $\hat{q}$. Their proofs are deferred to the supplementary materials. Before giving the asymptotical properties, we introduce the following notations and assumptions.

Throughout, we use the subscript "0" for the true value. For example, $\mathbf{h}_{i0}$ is the true value of $\mathbf{h}_i$ and $\mathbf{B}_0$ is the true value of $\mathbf{B}$. Denote $m_j(x) = \nu_j'(x)$ to be the mean function for the jth variable, $\boldsymbol{\kappa}_i = (1, \mathbf{h}_i^T)^T$, $\mathbf{K} = (\boldsymbol{\kappa}_1, \ldots, \boldsymbol{\kappa}_n)^T$, $a_{ji} = a(\mathbf{h}_i, \boldsymbol{\gamma}_j) = \frac{x_{ij} - m_j(\boldsymbol{\gamma}_j^T \boldsymbol{\kappa}_i)}{c_j}$, $\mathbf{v}_{ji}(\mathbf{h}_i, \boldsymbol{\gamma}_j) = a_{ji}\mathbf{b}_j$, $\mathbf{w}_{ji}(\mathbf{h}_i, \boldsymbol{\gamma}_j) = a_{ji}\boldsymbol{\kappa}_i$, $m_{ji} = m_j(\boldsymbol{\gamma}_j^T \boldsymbol{\kappa}_i)$, $m_{ji}' = m_j'(\boldsymbol{\gamma}_j^T \boldsymbol{\kappa}_i)$, $\Lambda_i = \mathrm{diag}(\frac{m_{1i}'}{c_1}, \ldots, \frac{m_{pi}'}{c_p})$, $\tilde{\Lambda}_j = \mathrm{diag}(\frac{m_{j1}'}{c_j}, \ldots, \frac{m_{jn}'}{c_j})$. In addition, for any symmetric matrix $\mathbf{A}$, $\lambda_{\min}(\mathbf{A})$ denotes the minimum eigenvalue of $\mathbf{A}$. Denote $\boldsymbol{\theta} = (\mathbf{h}_1^T, \ldots, \mathbf{h}_n^T, \boldsymbol{\gamma}_1^T, \ldots, \boldsymbol{\gamma}_p^T)^T$ and $p_s = |A_s|$. Since the number of types for variables, d, is finite, there must be a s such that $p_s = O(p)$. Without loss of generality, we suppose $s = 1$. Then let $\mathbf{f}(\boldsymbol{\theta}) = \Big[\sum_{j=1}^{p_1} \mathbf{v}_{j1}(\mathbf{h}_1, \boldsymbol{\gamma}_j)^T, \ldots, \sum_{j=1}^{p_1} \mathbf{v}_{jn}(\mathbf{h}_n, \boldsymbol{\gamma}_j)^T, \sum_{i=1}^{n} \mathbf{w}_{1i}(\mathbf{h}_i, \boldsymbol{\gamma}_1)^T, \cdots, \sum_{i=1}^{n} \mathbf{w}_{pi}(\mathbf{h}_i, \boldsymbol{\gamma}_p)^T\Big]^T$, $\nabla \mathbf{f}(\boldsymbol{\theta}) = \frac{\partial \mathbf{f}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}$ be the Hessian matrix, $\nabla_s \mathbf{f}(\boldsymbol{\theta}) = \big(\mathrm{diag}(p^{-1}\mathbf{I}_{nq}, n^{-1}\mathbf{I}_{p(q+1)})\frac{\partial \mathbf{f}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\big)$ be the scaled Hessian matrix. Throughout the article, we denote $\|\cdot\|$ the l_2 norm of any vector and spectral norm of any matrix, and M, M_1, M_2 be some generic positive constants which can be varied in different situations. The following conditions are required for asymptotical theory.

(C1)—*Identifiable Condition*: $(\mathbf{H}_0, \mathbf{B}_0)$ satisfy (A1)–(A3);

(C2)—*Factors and Loadings*: (1) $\sup_j \|\boldsymbol{\gamma}_{j0}\| \le M$; (2) There exist $r_1 > 0$ and $s_1 > 0$, such that for any $t > 0, l \le q$, $P(|h_{il0}| > t) \le \exp(-(t/s_1)^{r_1})$, where h_{il0} is the lth element of $\mathbf{h}_{i0}$.

(C3)—Moments and tail probability: (1) $\mathrm{E}\|p_1^{-1/2}\sum_{j=1}^{p_1} \mathbf{v}_{ji}(\mathbf{h}_{i0}, \boldsymbol{\gamma}_{j0})\|^\tau \le M$ and $\mathrm{E}\|p^{-1/2}\sum_{j=1}^{p} \mathbf{v}_{ji}(\mathbf{h}_{i0}, \boldsymbol{\gamma}_{j0})\|^\tau \le M$ for some $\tau \ge 4$; (2) There exist $r_2 > 0$ and $s_2 > 0$, such that for any $t > 0, j \le p, P(|a_{ji0}| > t) \le \exp(-(t/s_2)^{r_2})$.

(C4)—Lower bounded eigenvalue: For any $\epsilon > 0$ and some $\lambda_2 > 0$, there exists M_1 and M_2 such that $P\Big(\lambda_{\min}(- \nabla_s \mathbf{f}(\boldsymbol{\theta})) > \lambda_2 : \boldsymbol{\theta} \in \Big\{\sup_j \|\boldsymbol{\gamma}_j - \boldsymbol{\gamma}_{j0}\| \le M_1, \sup_i \|\mathbf{h}_i - \mathbf{h}_{i0}\| \le M_2\Big\}\Big) > 1 - \epsilon$.

(C5)—Smoothing: For all j, $m_j''(x)$ is a continuous function on its corresponding support.

(C6)—Lower bound: $\inf_j c_j \ge c > 0$.

Condition (C1) ensures that the true values $(\mathbf{H}_0, \mathbf{B}_0)$ satisfy the identifiability condition. Condition (C2) is similar as Assumption 3.2 in Bai and Liao (2013) for the linear factor model. Particularly, Conditions (C2.1) requires the uniform bound of loading-intercept vectors, and Condition (C2.2), which is the generalized version of (iii) of Assumption 3.2 in Bai and Liao (2013), requires the exponential tail probability of factors. (C2.2) is required to deduce the uniformly convergent rate of $\widehat{\boldsymbol{\gamma}}_j$ and $\widehat{\mathbf{h}}_i$.

Condition (C3.1) and (C3.2) are the generalized versions of (iv) of Assumption 3.4 and (iii) of Assumption 3.2 in Bai and Liao (2013), respectively, to assist deducing the unform convergent rate for estimators.

Condition (C4) requires the minimum eigenvalue of the negative scaled hessian matrix to be lower bounded with high probability. If there is only one type of variables, as shown in proof, $-\nabla_s \mathbf{f}(\boldsymbol{\theta})$ asymptotically is a $(nq+pq+p) \times (nq+pq+p)$ block-diagonal matrix with two block entries $\frac{1}{p}\mathbf{B}^T\Lambda_i\mathbf{B}$ and $\frac{1}{n}\mathbf{K}^T\tilde{\Lambda}_j\mathbf{K}$. In the linear factor model, we have $\frac{1}{p}\mathbf{B}^T\Lambda_i\mathbf{B} = \frac{1}{p}\mathbf{B}^T\Lambda^{-1}\mathbf{B}$ for all i and $\frac{1}{n}\mathbf{K}^T\tilde{\Lambda}_j\mathbf{K} = \frac{1}{n}\mathbf{K}^T\mathbf{K}/c_j, \Lambda = \mathrm{diag}(c_1, \ldots, c_p)$, and then Condition (C4) is reduced to that the eigenvalues of $\frac{1}{p}\mathbf{B}^T\Lambda^{-1}\mathbf{B}$ and $\frac{1}{n}\mathbf{K}^T\mathbf{K}$ are bounded from zero, which is in accordance with Assumptions A and B in Bai (2003). Further, if all of $x_{ij}, j = 1, \ldots, p$ are continuous with homoscedasticity, for example, $c_j = c$, then $\frac{1}{p}\mathbf{B}^T\Lambda_i\mathbf{B} = \frac{1}{p}\mathbf{B}^T\mathbf{B}/c$, Condition (C4) becomes that the eigenvalues of $\frac{1}{p}\mathbf{B}^T\mathbf{B}$ and $\frac{1}{n}\mathbf{K}^T\mathbf{K}$ are bounded from zero, which is the counterpart of Condition (A3) in Jiang, Ma, and Wei (2019).

Condition (C5) is a regular mathematical condition and holds for the commonly used types of variables with exponential family distribution. We also give $m(x)$ for the three cases that satisfy this condition. (1) If x_{ij} is normal, $m(x) = x$; (2) If x_{ij} is Bernoulli, then $m(x) = \frac{1}{1+e^{-x}}$; (3) If x_{ij} is Poisson, then $m(x) = e^x$. Condition (C6) ensures $c_j \neq 0$ by imposing that the scale parameters are uniformly lower bounded and it easily holds for exponential family.

Denote $C_{np} = \min\{\sqrt{n}, \sqrt{p}\}$, then we have the following theorems.

Theorem 1. (Uniform convergence and convergent rate on factors). Under Conditions (C1)–(C6), we have that $\|\widehat{\mathbf{h}}_i - \mathbf{h}_{i0}\| = O_p(C_{np}^{-1})$ for each i. Furthermore, it holds that $\sup_i \|\widehat{\mathbf{h}}_i - \mathbf{h}_{i0}\|^2 = O_p(n^{-1} + p^{-1}n^{1/\tau})$, where $\tau \geq 4$ is defined in Condition (C3.2).

Similar results have been established for continuous variable. For example, (Bai and Ng 2002, theor. 1) gave the result $C_{np}^2\{\frac{1}{n}\sum_{i=1}^n \|\widehat{\mathbf{h}}_i - \mathbf{h}_{i0}\|^2\} = O_p(1)$, which is consistent with our result. They also pointed out that the result $C_{np}^2\|\widehat{\mathbf{h}}_i - \mathbf{h}_{i0}\|^2 = O_p(1)$ can be obtained with some extra assumptions. Bai and Ng (2013) showed that, for each i, $\|\widehat{\mathbf{h}}_i - \mathbf{h}_{i0}\| = O_p(\frac{1}{\sqrt{p}})$ if $\frac{\sqrt{p}}{n} = o(1)$ in their Theorem 1, which is in agreement with our result. The same uniform convergence rate of factors is also given in (Bai and Liao 2013, theor. 3.2).

Theorem 2. (Uniform convergence and convergent rate on loadings). Under Conditions (C1)–(C6), we have that $\|\widehat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_{j0}\| = O_p(C_{np}^{-1})$ for each j. Further, it holds that $\sup_j \|\widehat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_{j0}\|^2 = O_p(p^{-1} + n^{-1}\ln p)$.

Similar results also have been established for continuous variable. Particularly, Bai and Ng (2013) shown that, for each j, $\|\widehat{\mathbf{b}}_j - \mathbf{b}_{j0}\| = O_p(\frac{1}{\sqrt{n}})$ if $\frac{\sqrt{n}}{p} = o(1)$ in their Theorem 1, which is consistent with our result. The same uniform convergence rate of loadings is also given in Theorem 3.2 by Bai and Liao (2013). According to the results of Theorems 1 and 2, we can further establish the asymptotical normality for loadings and the consistency of the estimated factor number.

Theorem 3. (Asymptotical normality). Under Conditions (C1)–(C6) and $\frac{\sqrt{n}}{p} = o(1)$, it holds that

$$\sqrt{n}(\widehat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_{j0}) \xrightarrow{d} N(\mathbf{0}, \Gamma_j^{-1}),$$

where $\Gamma_j = var(a_{ji0}\boldsymbol{\kappa}_{i0})$.

Theorem 4. Under the Conditions (C1)–(C6), if $0 < g(n,p) \to 0$, $g(n,p)C_{np}^2 \to +\infty$, and $E(p^{-1}\sum_{j=1}^p a_{ji0})^4 < \infty$, we have $P(\hat{q} = q) \to 1$, as $n, p \to \infty$.

Similar results also have been established for continuous variables in Bai and Ng (2002). Any penalty satisfying the conditions that $0 < g(n,p) \to 0$ and $g(n,p)C_{np}^2 \to +\infty$ can be used. Particularly, the condition $g(n,p)C_{np}^2 \to +\infty$ ensures that the probability of choosing the factor number greater than the true factor number q tends to zero, and the condition $0 < g(n,p) \to 0$ ensures that the probability of choosing the factor number less than q tends to zero. Thus, under the two conditions, we can consistently choose the true factor number. The penalty function $g(n,p) = \frac{n+p}{np}\ln(\frac{np}{n+p})$ used by Bai and Ng (2002) satisfies the conditions in Theorem 4 and hence is used in our simulation studies and real data analysis.

Remark 4. To ensure the sample-wise consistency of factors or variable-wise consistency of loadings, we require that p diverges at any rate, including exponential rate of n. Our simulation studies show that the performance of proposed method gets better as p increases. That is, the high dimension p beings blessing of dimensionality, instead of curse. This is a direct result of the factor model, where p actually plays the role of the number of observations for estimating $\mathbf{H}$ through which influences the estimator of $\mathbf{B}$. This is also reflected by the convergent rate on factors and loadings in Theorems 1 and 2. While, to guarantee the uniform consistency of factors and loadings, we require p diverges with the constraint that $n^{1/\tau} \prec p \preceq \exp(n^\alpha), 0 < \alpha < 1$, where $a \prec b$ represent that a is less than b in order, $a \preceq b$ represent that a is not greater than b in order. Since the factor number is finite, more variables means that more information can be used to estimate loadings and factors. From the theoretical aspect, when p is sufficiently large so that $n/p \to 0$, Theorem 2 shows that $\|\widehat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_{j0}\| = O_p(n^{-1/2})$ and $\sup_j \|\widehat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_{j0}\|^2 = O_p(n^{-1}\ln p)$, both achieve the corresponding rate when all latent factors are observable. The issue that the high dimension p is the blessing of dimensionality instead of the curse also has been carefully discussed for linear factor model in Li et al. (2018). In addition, if the number of variables is fixed or limited, it may be not necessary to compress the original variables to extract factors and we can analyze the raw data directly.

4. Numerical Studies

In Sections 4.1–4.4, we conducted simulation studies to assess the finite-sampling performance of the proposed method (GFM) by comparing it with the linear factor model (LFM) and the low-rank models. The performance of the estimator was assessed via the computational time and the accuracy of the estimator. The accuracy of an estimator is evaluated by canonical correlation (Bai et al. 2012), which has been widely used to measure the performance of factor model (Goyal, Pérignon, and Villa 2008; Doz, Giannone, D., and Reichlin 2012; Bai et al. 2012), and defined by the smallest nonzero canonical correlation between $\widehat{\mathbf{H}}$ and $\mathbf{H}_0$, denoted by $\mathbf{ccor}(\widehat{\mathbf{H}}, \mathbf{H}_0)$. In Sections 4.5 and 4.6, we examined the performance of formula (9) for selecting the number of factors and investigated the efficiency gain of the one-step correction in Step 2, respectively.

4.1. Simulation Setting

We first generated $\check{\mathbf{h}}_i$ from a multivariate normal distribution with mean zero and covariance matrix $(\sigma_{ij})_{q\times q}$ with $\sigma_{ij} = 0.5^{|i-j|}$ and denoted it by $\check{\mathbf{H}} = (\check{\mathbf{h}}_1, \dots, \check{\mathbf{h}}_n)^T \in R^{n\times q}$. Based on $\check{\mathbf{H}}$, we calculated the sample mean and sample covariance of $\check{\mathbf{h}}_i$, denoted by $\bar{\mathbf{h}} = n^{-1}\sum_{i=1}^n \check{\mathbf{h}}_i$ and $\widehat{\mathbf{S}} = n^{-1}\sum_{i=1}^n(\check{\mathbf{h}}_i - \bar{\mathbf{h}})(\check{\mathbf{h}}_i - \bar{\mathbf{h}})^T$. Then we generated $\mathbf{H} = (\mathbf{h}_1, \dots, \mathbf{h}_n)^T$, where $\mathbf{h}_i = \widehat{\mathbf{S}}^{-1/2}(\check{\mathbf{h}}_i - \bar{\mathbf{h}}_i)$. In this way, it is easy to check that $\mathbf{H}$ satisfies the identification condition (A1) as described in Section 2. To construct the matrix $\mathbf{B}$, we first generated n samples $\mathbf{z}_i, i = 1, \dots, n$ from a multivariate normal vector with mean zero and covariance matrix $(\sigma_{ij})_{p\times p}$ with $\sigma_{ij} = \sigma^2 \times 0.5^{|i-j|}$. We performed eigen decomposition on the matrix $\mathbf{Z}\mathbf{Z}^T$, where $\mathbf{Z} = (\mathbf{z}_1, \dots, \mathbf{z}_n)^T$. We further denoted an $n \times q$ orthogonal matrix $\mathbf{E}$ that corresponds to the first q largest eigenvalues and $\mathbf{B} = \sqrt{1/6}\mathbf{Z}^T\mathbf{E}$. Obviously, $\mathbf{B}^T\mathbf{B}$ is a diagonal matrix with decreasing diagonal entries. $\mu_j = 0.4w_j, w_j \overset{\text{iid}}{\sim} N(0,1), j = 1, \dots, p$. $\mathbf{B}$ and $\boldsymbol{\mu}$ were fixed after generation. We considered six settings with different q's. For each of the simulation settings, a total of 1000 replications were conducted.

- Example 1: The setting was the same as Bai et al. (2012) where $\sigma^2 = 1, q = 1$ and the combinations of $n = 30, 50, 100$ and $p = 30, 50, 100, 150$ were considered. For each j, $x_{ij} \sim N(\mathbf{h}_i^T\mathbf{b}_j + \mu_j, 1)$, indicating that $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ were p-dimensional Gaussian variables with homoscedasticity.
- Example 2 was similar with Example 1 except that $q = 2$, and for each j, $x_{ij} \sim N(\mathbf{h}_i^T\mathbf{b}_j + \mu_j, \tau_j^2)$ with $\tau_j^2 = 0.1 + 2U_j$, $U_j, j = 1, \ldots, p$ were iid from $U[0, 1]$. The 0.1 was added to avoid the variance close to zero. In this example, $\mathbf{x}_i$ were Gaussian variables with heteroscedasticity.
- Example 3: We set $\sigma^2 = 4, q = 3$ and considered the combination of $n = 50, 100, 150$ and $p = 100, 150, 200, 250$. For each j, $x_{ij} \sim \text{Poisson}(\exp(\mathbf{h}_i^T\mathbf{b}_j + \mu_j))$. This example was designed for nonlinear dimension reduction, for example, pixels take integers between 0 and 255 in the image analysis.
- Example 4: We set $\sigma^2 = 6, q = 4$ and considered the combination of $n = 100, 200, 300$ and $p = 100, 200, 300, 400$. For $j = 1, \ldots, [p/2]$, $x_{ij} \sim \text{Poisson}(\exp(\mathbf{h}_i^T\mathbf{b}_j + \mu_j))$; for $j = [p/2]+1, \ldots, p$, $x_{ij} \sim \text{Bernoulli}(1/(1+\exp(-\mathbf{h}_i^T\mathbf{b}_j - \mu_j))$. In this example, we considered the mixture of binary and count variables.
- Example 5: We set $\sigma^2 = 4, q = 5$ and considered the combination of $n = 100, 200, 300$ and $p = 100, 200, 300, 400$. For $j = 1, \ldots, [p/2]$, $x_{ij} \sim N(\mathbf{h}_i^T\mathbf{b}_j + \mu_j, 1)$; for $j = [p/2] + 1, \ldots, p$, $x_{ij} \sim \text{Poisson}(\exp(\mathbf{h}_i^T\mathbf{b}_j + \mu_j))$. Example 5 considers the mixture of continuous and count variables.
- Example 6: We set $\sigma^2 = 4, q = 6$ and considered the combination of $n = 100, 200, 300$ and $p = 100, 200, 300, 400$. For $j = 1, \ldots, [p/3]$, $x_{ij} \sim N(\mathbf{h}_i^T\mathbf{b}_j + \mu_j, 1)$; for $j = [p/3] + 1, \ldots, [(2p)/3]$, $x_{ij} \sim \text{Poisson}(\exp(\mathbf{h}_i^T\mathbf{b}_j + \mu_j))$; for $j = [(2p)/3]+1, \ldots, p$, $x_{ij} \sim \text{Bernoulli}(1/(1+\exp(-\mathbf{h}_i^T\mathbf{b}_j - \mu_j))$. In this example, we considered the mixture of continuous, count and binary variables.

Note that (σ^2, n, p) jointly determine the strength of the signal, where larger values indicate more strong signal.

4.2. Comparison With the Linear Factor Models

Tables 1–3 show the average of $\mathbf{ccor}(\widehat{\mathbf{H}}, \mathbf{H}_0)$ and $\mathbf{ccor}(\widehat{\Upsilon}, \Upsilon_0)$ using the LFM and GFM based on 1000 repetitions. Examples 1 and 2 satisfy the conditions required by both GFM and LFM. Theoretically, the LFM and GFM are equivalent under homoscedasticity, which is consistent with the approximately comparable results for Example 1 in Table 1. Since our method uses the information for heteroscedasticity, the GFM outperforms LFM as shown in the bottom panel of Table 1. It shows that the principal component method using the linear factor model cannot account for heteroscedastic manifest variables. Moreover, Table 1 shows that the precision of $\widehat{\mathbf{H}}$ and $\widehat{\Upsilon}$ increases from top-left to bottom-right for Examples 1 and 2, respectively. These results indicate that the precision of $\widehat{\mathbf{H}}$ and $\widehat{\Upsilon}$ increases as n or p increases, which is consistent with the theoretical results in Theorems 1 and 2. The precision of the GFM for $\widehat{\mathbf{H}}$ is more sensitive to p than to n, and the precision of the GFM for $\widehat{\Upsilon}$ is

Table 1. The results of Examples 1 and 2 for normal variables

		LFM				GFM			
	$n \setminus p$	30	50	100	150	30	50	100	150
		Example 1 with homoscedasticity							
$\mathbf{ccor}(\widehat{\mathbf{H}}, \mathbf{H}_0)$	30	0.8635	0.8882	0.9063	0.9209	0.8671	0.8914	0.9032	0.9252
	50	0.8739	0.8922	0.9190	0.9328	0.8775	0.8983	0.9157	0.9391
	100	0.8823	0.9043	0.9324	0.9478	0.8810	0.9026	0.9304	0.9441
$\mathbf{ccor}(\widehat{\Upsilon}, \Upsilon_0)$	30	0.8268	0.8273	0.8349	0.8364	0.8262	0.8277	0.8328	0.8365
	50	0.8477	0.8684	0.8819	0.9062	0.8464	0.8662	0.8875	0.9043
	100	0.8901	0.9066	0.9348	0.9478	0.8890	0.9052	0.9346	0.9474
		Example 2 with heteroscedasticity							
$\mathbf{ccor}(\widehat{\mathbf{H}}, \mathbf{H}_0)$	30	0.6935	0.7868	0.8517	0.8694	0.7305	0.8533	0.9140	0.9237
	50	0.7791	0.8197	0.8668	0.8844	0.8513	0.8678	0.9213	0.9345
	100	0.8032	0.8221	0.8781	0.9102	0.8556	0.8919	0.9231	0.9474
$\mathbf{ccor}(\widehat{\Upsilon}, \Upsilon_0)$	30	0.6038	0.7215	0.7623	0.7666	0.6336	0.7719	0.7852	0.7865
	50	0.7251	0.7878	0.8008	0.8128	0.7849	0.8131	0.8288	0.8388
	100	0.8328	0.8507	0.8739	0.8769	0.8601	0.8812	0.9044	0.9131

Table 2. The results of Example 3 for Poisson variables

		LFM				GFM			
	$n \setminus p$	100	150	200	250	100	150	200	250
$\mathbf{ccor}(\widehat{\mathbf{H}}, \mathbf{H}_0)$	50	0.6057	0.7258	0.7036	0.7279	0.9604	0.9714	0.9782	0.9790
	100	0.7279	0.8166	0.8807	0.8616	0.9682	0.9772	0.9786	0.9821
	150	0.7825	0.8816	0.8653	0.8690	0.9726	0.9784	0.9817	0.9835
$\mathbf{ccor}(\widehat{\Upsilon}, \Upsilon_0)$	50	0.5657	0.6596	0.6051	0.6740	0.8999	0.9031	0.9054	0.9183
	100	0.7001	0.7515	0.8165	0.7675	0.9509	0.9525	0.9555	0.9569
	150	0.8219	0.8339	0.8244	0.8119	0.9624	0.9624	0.9639	0.9692

Table 3. The results of Examples 4-6 for the variables with mixed types

		LFM				GFM			
		Example 4 for the mixture of Poisson and binary variables							
	$n \setminus p$	100	200	300	400	100	200	300	400
ccor$(\widehat{\mathbf{H}}, \mathbf{H}_0)$	100	0.2782	0.6077	0.5485	0.6470	0.8642	0.9625	0.9705	0.9754
	200	0.2527	0.6050	0.6341	0.6997	0.9097	0.9582	0.9708	0.9768
	300	0.4570	0.7420	0.7789	0.8149	0.9330	0.9582	0.9709	0.9769
ccor$(\widehat{\Upsilon}, \Upsilon_0)$	100	0.1224	0.4296	0.2980	0.3823	0.8332	0.8756	0.8785	0.8827
	200	0.1438	0.3601	0.4684	0.4223	0.8817	0.8994	0.9249	0.9262
	300	0.2455	0.5065	0.4695	0.4936	0.9206	0.9406	0.9466	0.9472
		Example 5 for the mixture of normal and Poisson variables							
	$n \setminus p$	100	200	300	400	100	200	300	400
ccor$(\widehat{\mathbf{H}}, \mathbf{H}_0)$	100	0.5719	0.8352	0.8417	0.8245	0.9536	0.9704	0.9753	0.9777
	200	0.7113	0.7166	0.8633	0.9052	0.9575	0.9725	0.9769	0.9816
	300	0.7325	0.7646	0.9232	0.9248	0.9600	0.9750	0.9799	0.9839
ccor$(\widehat{\Upsilon}, \Upsilon_0)$	100	0.3917	0.4481	0.5708	0.5213	0.9240	0.9384	0.9395	0.9422
	200	0.5145	0.5757	0.5388	0.5653	0.9540	0.9573	0.9596	0.9631
	300	0.4818	0.4975	0.5606	0.5635	0.9605	0.9670	0.9680	0.9698
		Example 6 for the mixture of normal, Poisson and binary variables							
	$n \setminus p$	200	300	400	500	200	300	400	500
ccor$(\widehat{\mathbf{H}}, \mathbf{H}_0)$	200	0.5284	0.6734	0.6687	0.6823	0.9176	0.9482	0.9664	0.9733
	300	0.5035	0.6433	0.6549	0.6613	0.9204	0.9499	0.9677	0.9744
	400	0.4697	0.6234	0.6525	0.6591	0.9215	0.9508	0.9682	0.9750
ccor$(\widehat{\Upsilon}, \Upsilon_0)$	200	0.3674	0.3462	0.4350	0.4282	0.9240	0.9327	0.9354	0.9368
	300	0.3589	0.3397	0.4372	0.4264	0.9472	0.9544	0.9563	0.9571
	400	0.3472	0.3315	0.4403	0.4292	0.9593	0.9652	0.9669	0.9676

more sensitive to n than to p. The reason is that the estimator for $\mathbf{h}_i$ is based on p variables of individual i, but the estimator for $\boldsymbol{\gamma}_j$ is based on variable j of n individuals.

From Tables 2 and 3, we can see that the performance of GFM is much better than that of LFM for the Poisson variables or the mixture of Poisson, binary, and normal variables. Particularly, the canonical correlation of GFM for both $\widehat{\Upsilon}$ and $\widehat{\mathbf{H}}$ increases as n or p increases, but that of LFM does not. It is not surprising as the LFM requires the variables to be continuous, but the variables in Examples 3 – 6 do not follow the requirement. Again, Tables 2 and 3 show that the precision of the GFM for $\widehat{\mathbf{H}}$ is more sensitive to p than to n, but the precision of the GFM for $\widehat{\Upsilon}$ is more sensitive to n than to p.

4.3. *Comparison With the Low-Rank Models*

The low-rank method is a powerful tool in data compression. Denoting $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)^T$, low-rank methods aim to seek an unknown low-rank matrix $\boldsymbol{\Theta} = \mathbf{H}\mathbf{B}^T$ so that $\mathbf{X} = \boldsymbol{\Theta} + \mathbf{E}$, where $\mathbf{E}$ represents the random noise. Many efforts have been made to estimate $\boldsymbol{\Theta}$. For example, the unbiased risk estimates for singular value thresholding (SURE) by Candes, Sing-Long, and Trzasko (2013), the optimal shrinkage of singular values (OPT) by Gavish and Donoho (2017), and the iterated stable autoencoder (ISA) method by Josse and Wager (2016). The details of these methods are referred to Josse, Sardy, and Wager (2016) and their implementations can be performed by using R package `denoiseR`. In addition, Chen, Dong, and Chan (2013) considered the adaptive nuclear-norm penalization (CRRR) to estimate reduced rank regression with continuous response matrix for the model $\mathbf{X} = \mathbf{Z}\boldsymbol{\Theta} + \mathbf{E}$, and Luo et al. (2018) extended Chen, Dong, and Chan (2013) to consider the mixed responses by assuming exponential families, which was termed mixed-response reduced-rank regression (MRRR) here. Without covariates, that is, $\mathbf{Z} = \mathbf{I}_n$, Chen, Dong, and Chan (2013) and Luo et al. (2018) are low-rank methods for continuous response matrix and mixed-response matrix, respectively. The low-rank regressions in Chen, Dong, and Chan (2013) and Luo et al. (2018) can be implemented by R package `rrpack` (Luo et al. 2018).

Our methods can be regarded as a generalized low-rank method with nonlinear link between $\boldsymbol{\Theta} = \mathbf{H}\mathbf{B}^T$ and $\mathbf{X}$. In the Section, we compare the proposed method with these low-rank methods mentioned. Since the low-rank methods estimated $\boldsymbol{\Theta}$ as a whole, following Chen, Dong, and Chan (2013) and Luo et al. (2018), we assessed the performance of the estimators by the quantity $PE = \|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0\|_F^2/(np)$, where $\boldsymbol{\Theta}_0$ is the true value of $\boldsymbol{\Theta}$. Table 7 shows the average of PE using the proposed GFM, LFM, CRRR, MRRR, SURE, OPT, and ISA based on 1000 repetitions for Examples 2, 3, and 5, which correspond to the continuous variables, count variables and the mixture of continuous and count variables, respectively.

From Table 7, we can see that GFM outperformed other methods in terms of PE in all of the cases considered and the performance of GFM increased with increasing n or p, which is consistent with the results measured by canonical correlation in Section 4.2. Particularly, the GFM was slight better than the others even for the continuous variables (Example 2) where the linear structure holds. This is attributed to the likelihood framework that the proposed GFM used, which accounts for heteroscedastic variables. However, all other low-rank methods did not consider the heteroscedasticity in variables. For the count variables (Example 3) and the mixture of continuous and count variables (Example 5), the GFM was significantly better than the others. We also noted that the MRRR was much better than CRRR, LFM and the other regularized low-rank matrix methods since the MRRR considered the mixed structure of the data.

4.4. Running Time

We carried out the computation on a single 14 core machine with 378GB of RAM. Table 8 shows the precision of estimation on $(\mathbf{H}, \Upsilon)$ and average running time of repetition for Example 6 with $(n, p) = (500, 10{,}000)$ and $(5000, 100{,}000)$. The summaries displayed in Table 8 were based on 10 repetitions. As we can see from Table 8, it averagely took 18.8 and 897.8 seconds for the data of (n, p)=(500, 10,000) and (5000, 100,000), respectively, and the precision of estimation on $(\mathbf{H}, \Upsilon)$ was gradually getting better as n and p increased.

4.5. Model Selection

In this section, we examined the performance of formula (9) for selecting the number of factors. Similar to Bai and Ng (2002), we reported the average of $\hat{q}$ based on 100 repetitions, where the candidates of q were the integers from 1 to 10. In Table 4, we can see that the PC-type criterion (9) worked well in all six examples considered with different q. Especially for larger n and p, the performance was more significant. We also compared with the PC criteria using LFM in Bai and Ng (2002), and found that for the continuous cases of Examples 1 and 2, the results based on the PC criteria of by using GFM approximately had same performance as those using LFM, and they both worked well. However, for other cases not all being continuous variables, results of GFM were significantly better than those of LFM. Moreover, GFM could precisely select the true number of factors, but LFM tended to select more number of factors.

4.6. Efficiency Gain of the One-Step Correction and Two Identifiability Methods

To investigate the efficiency gain of the one-step correction in Step 2 with example 5, we used continuous variables, count variables, and both of them to obtain the estimator in Step 1, termed by S1N, S1P and S1NP, respectively. From Table 5, we can see that (i) the first step estimators became better as n or p increased. The S1NP is the best and S1P performed worse. It is consistent with our expectation because S1NP used the most information and continuous variables had stronger signals than count variables; (ii) the one-step updating could significantly improve the performance of the estimators from Step 1, and the improvement was almost independent of selection of the first step estimators.

Now we compared the results with two identifiability methods in the first step of the algorithm: simultaneously meeting the identifiability condition described in Remark 1 and the

Table 4. The average of the estimated number of factors for all examples by using our proposed PC criteria based on GFM and the PC criteria in Bai and Ng (2002) based on LFM from 100 repetitions.

	Example 1: $q = 1$					Example 2: $q = 2$					Example 3: $q = 3$				
GFM	$n \setminus p$	30	50	100	150	$n \setminus p$	100	150	200	250	$n \setminus p$	100	150	200	250
	30	2.94	1.10	1.00	1.00	100	2.00	2.00	1.91	1.94	100	3.02	2.99	2.94	2.96
	50	1.09	1.61	1.00	1.00	150	2.00	2.00	2.00	2.00	150	3.03	3.03	2.99	2.98
	100	1.00	1.00	1.00	1.00	200	2.00	2.00	2.00	2.00	200	3.01	3.00	3.00	3.00
LFM	$n \setminus p$	30	50	100	150	$n \setminus p$	100	150	200	250	$n \setminus p$	100	150	200	250
	30	3.00	1.20	1.00	1.00	100	2.17	2.00	1.82	1.99	100	9.54	7.94	7.44	7.32
	50	1.20	2.04	1.00	1.00	150	2.00	2.06	2.00	1.96	150	9.99	9.00	7.30	6.40
	100	1.00	1.00	1.00	1.00	200	1.99	2.00	2.00	2.00	200	8.29	9.71	7.33	6.43
	Example 4: $q = 4$					Example 5: $q = 5$					Example 6: $q = 6$				
GFM	$n \setminus p$	100	200	300	400	$n \setminus p$	100	200	300	400	$n \setminus p$	200	300	400	500
	100	3.95	4.00	3.97	3.21	100	5.02	5.03	5.02	4.98	200	6.04	6.02	6.01	6.02
	200	3.17	3.96	4.00	4.00	200	5.02	5.01	5.01	5.00	300	5.99	6.02	6.03	6.04
	300	3.94	3.99	4.00	4.00	300	5.03	5.01	5.00	5.00	400	6.00	6.02	5.97	6.01
LFM	$n \setminus p$	100	200	300	400	$n \setminus p$	100	200	300	400	$n \setminus p$	200	300	400	500
	100	10.0	10.0	9.96	7.60	100	10.00	8.46	6.70	6.60	200	10.0	10.0	10.0	10.0
	200	10.0	10.0	10.0	9.12	200	9.84	10.0	7.59	7.13	300	10.0	10.0	10.0	10.0
	300	10.0	10.0	10.0	10.0	300	8.02	7.93	8.49	6.08	400	10.0	10.0	10.0	10.0

Table 5. Results of Example 5. The estimators from the first step based on only Normal variables, only Poisson variables, and both of them were simplified as S1N, S1P and S1NP, respectively. And the estimators from one-step update in the second step based on S1N, S1P and S1NP were simplified as OSN, OSP, and OSNP. **ccorH** and **ccorB** were the average of $\mathbf{ccor}(\widehat{\mathbf{H}}, \mathbf{H}_0)$ and $\mathbf{ccor}(\widehat{\Upsilon}, \Upsilon_0)$, respectively, from 1000 repeats in three combination of (n, p).

	$(n, p) = (50, 50)$			$(n, p) = (50, 100)$			$(n, p) = (100, 100)$		
	S1N	S1P	S1NP	S1N	S1P	S1NP	S1N	S1P	S1NP
ccorH	0.7455	0.5039	0.8889	0.8733	0.7384	0.9348	0.9074	0.8573	0.9423
ccorB	0.8390	0.5833	0.8771	0.8806	0.8562	0.8926	0.9013	0.8853	0.9007
	OSN	OSP	OSNP	OSN	OSP	OSNP	OSN	OSP	OSNP
ccorH	0.9211	0.9093	0.9380	0.9420	0.9408	0.9526	0.9515	0.9509	0.9536
ccorB	0.8496	0.8165	0.8786	0.9018	0.8826	0.9110	0.9185	0.9126	0.9240

Table 6. The comparison of results from 1000 repeats based on two identifiability methods in the first step of the algorithm.

		Example 1			Example 5		
	(n, p)	(50,50)	(50,100)	(100,100)	(50,50)	(50,100)	(100,100)
$\mathbf{ccor}(\widehat{\mathbf{H}}, \mathbf{H}_0)$	Prop	0.8983	0.9026	0.9304	0.9472	0.9504	0.9536
	Simul	0.8983	0.9024	0.9304	0.9341	0.9442	0.9467
$\mathbf{ccor}(\widehat{\Upsilon}, \Upsilon_0)$	Prop	0.8662	0.9052	0.9346	0.8983	0.9051	0.9240
	Simul	0.8662	0.9051	0.9346	0.9072	0.9118	0.9233

Table 7. The average $\|\widehat{\Theta} - \Theta_0\|_F^2/(np)$ from 1000 repeats for seven different methods, which were divided into three groups, where the first group is based on latent factor models (GFM, LFM), the second is based on regularized reduced rank regression (CRRR, MRRR), and the third is based on regularized low-rank matrix estimation (SURE, OPT, ISA).

Cases	(n, p)	GFM	LFM	CRRR	MRRR	SURE	OPT	ISA
Example 2	$(100, 100)$	0.0460	0.0655	0.0657	0.1070	0.1251	0.0756	0.1519
	$(100, 200)$	0.0358	0.0475	0.0479	0.0716	0.0642	0.0405	0.0335
	$(200, 200)$	0.0224	0.0326	0.0326	0.0566	0.1033	0.0573	0.1376
Example 3	$(100, 100)$	0.1077	2.0061	3.3623	0.4942	3.7558	3.7319	3.9498
	$(100, 200)$	0.0546	0.6051	0.8003	0.3428	0.9187	0.9024	0.9359
	$(200, 200)$	0.0353	0.5036	0.6364	0.3069	0.7416	0.7381	0.8570
Example 5	$(100, 100)$	0.2439	3.3898	5.0797	0.8621	5.1484	5.1002	5.3341
	$(100, 200)$	0.0918	0.6423	1.8293	0.6294	1.8752	1.8445	1.8959
	$(200, 200)$	0.0678	1.3553	0.9736	0.5141	1.0387	1.0216	1.1432

Table 8. Average running time and the precision of estimation of GFM for Example 6.

n	p	Sta.	$\mathbf{ccor}(\widehat{\mathbf{H}}, \mathbf{H}_0)$	$\mathbf{ccor}(\widehat{\Upsilon}, \Upsilon_0)$	Time (sec.)
500	10000	Mean	0.9993	0.9788	18.7463
		SD	2.4302-05	2.8829e-04	4.8526
5000	100000	Mean	0.9999	0.9979	897.7951
		SD	8.5372e-07	7.5396e-06	14.5334

proposed method, termed by "Simul" and "Prop," respectively. Table 6 shows that the proposed and Simul methods were approximately equivalent and the proposed method performed a little bit better.

5. Real Data Analysis

In this section, we applied our method to the NFBC1966 dataset and a cardiac arrhythmia dataset.

5.1. NFBC1966 Dataset

The NFBC1966 dataset consists of n=5123 individuals. For each individual, we have the observations of 364, 490 SNPs and features include gender, body mass index (BMI), C-reactive protein (CRP), glucose, insulin, high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C), triglycerides (TG), total cholesterol (TC), systolic blood pressure (SysBP), and diastolic blood pressure (DiaBP). Each SNP takes one of 0, 1, and 2. We first performed strict quality control using PLINK (Purcell et al. 2007) and GCTA (Yang et al. 2011). We excluded individuals with missing rates larger than 5% across all SNPs and further removed SNPs with minor allele frequency larger than 5%, missing rates larger than 1%, or p-values less than 0.0001 for Hardy-Weinberg equilibrium test. Pairs of subjects with estimated relatedness greater than 0.025 were identified and one of the pair is removed. After quality control, we have $n = 5,123$ individuals with $319,147$ SNPs available for analysis. Here, we considered the risk prediction and association mapping for BMI (Y_i) based on GFM. Similar to the screening method, we first select the most related SNPs with BMI by the marginal regression analysis. In our analysis, we considered 43 controlling covariates, including eight individual features such as gender, LDL, etc., and 35 principal components from SNP data. Those covariates were included to control population stratification and remove confounding variables or batch effects. Then, we selected the SNPs, a total of 33,878, with p-value less than 0.1 in marginal regression analysis into our follow-up analysis, and Figure 2(a) showed the number of SNPs on each chromosome. We also added the 5 top PCs (continuous variables) extracted by principal component analysis on 319,147 SNPs to the GFM model, and then we have a total of $p = 33,883$ manifest variables ($\mathbf{x}_i$) in the model.

To apply GFM and LFM to the SNPs data, we first chose factor number q using PC criteria, with the estimated $\hat{q} = 4$ and $\hat{q} = 1$ for GFM and LFM, respectively. Then, we extracted the latent factors $\widehat{\mathbf{h}}_i$ by GFM and LFM, followed by linear regression of Y_i on $\widehat{\mathbf{h}}_i$ for risk prediction, respectively, and these two methods were simplified as RGFM and RLFM(1). For fairness, we also compared with LFM with four factors, denoted as RLFM(4). Moreover, we also compared them with ridge regression and Lasso regression of Y_i on $\mathbf{x}_i$ that are widely used for risk prediction in GWAS (Campos, Gianola, and Allison 2010). We used two-fold cross-validations over 100 random splits to evaluate the performance of risk prediction of RGFM, RLFM(4), RLFM(1), ridge regression (Ridge), Lasso regression (Lasso) and MRRR (Luo et al. 2018), where Lasso is the degenerated version of generalized trace regression via regularization in Fan, Gong, and Zhu (2019). As shown in Figure 2(b), the resulting cross-validated prediction NMSE averaging over 100 random splits were 0.526, 0.542, 1.000, 0.576, 0.955 and 1.000 for the RGFM, RLFM(4), RLFM(1), Ridge, Lasso, and MRRR, respectively, where NMSE $= \frac{\sum_{i=1}^{n_1}(\hat{Y}_i - Y_i)^2}{\sum_{i=1}^{n_1}(\bar{Y} - Y_i)^2}$, $\bar{Y} = \frac{1}{n_1}\sum_{i=1}^{n_1} Y_i$ on the testing

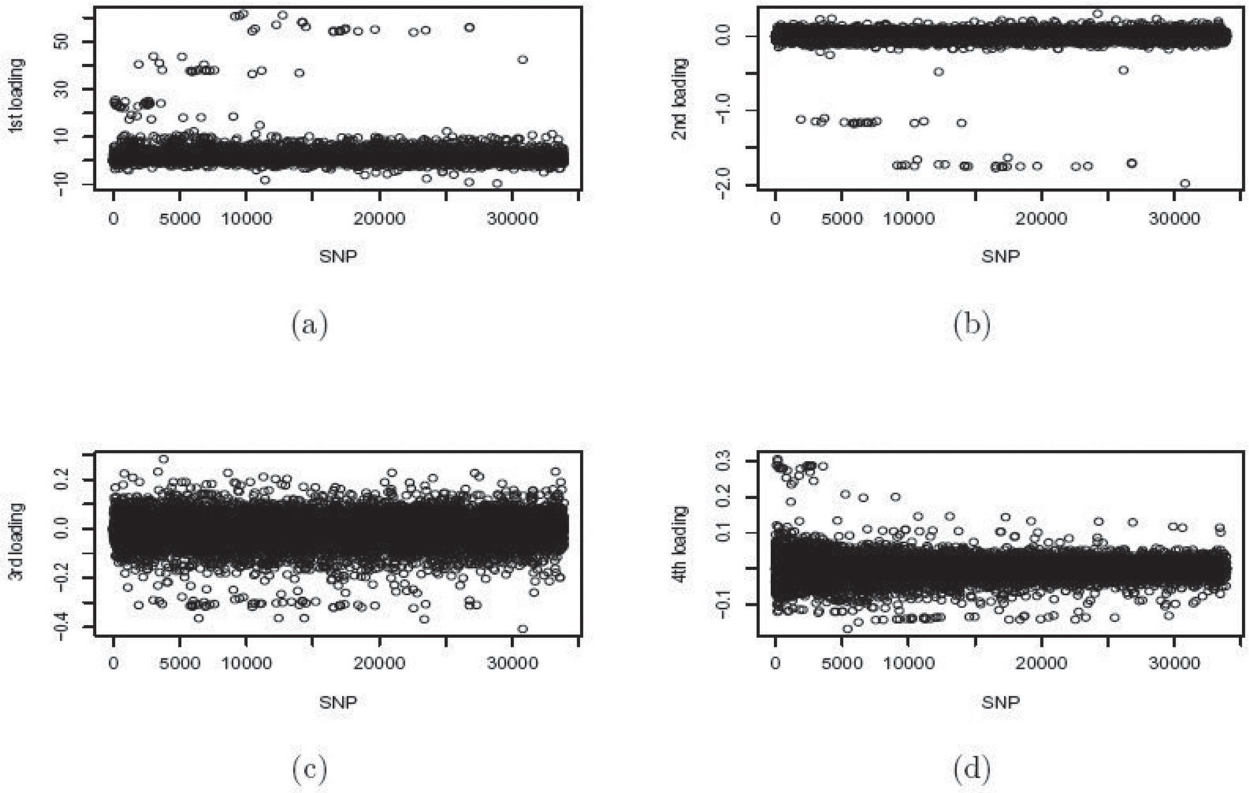

Figure 1. Sparse loading estimates from the regularized GFM with L_1 penalty as described in Section 6 to the GWAS dataset. (a): The estimated values of the first loading vectors across the 33878 SNPs; And (b), (c) and (d) are the estimated values of the second, third and forth loading vectors across the 33878 SNPs, repectively.

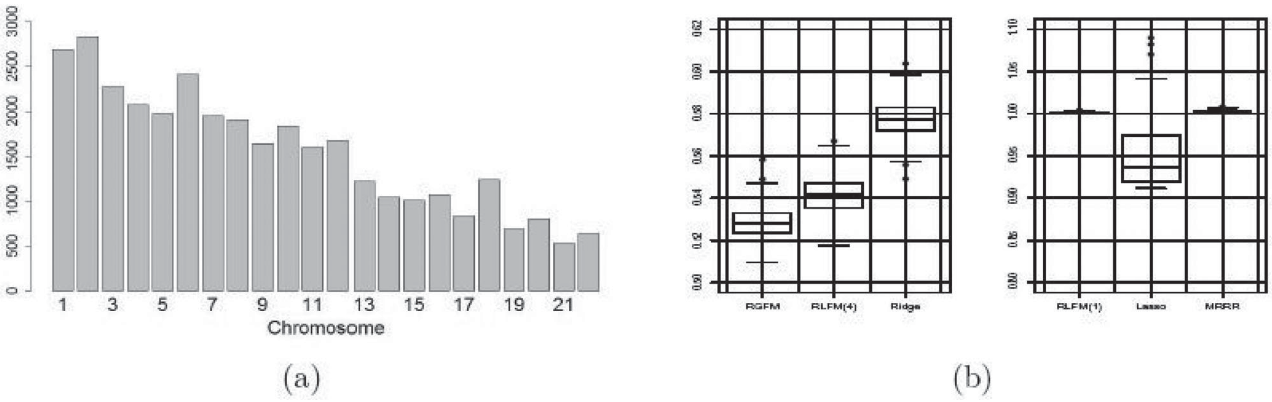

Figure 2. (a): The number of selected SNPs on each chromosome, a total of 33878. (b): Left panel: Risk prediction of RGFM, RLFM(4) and Ridge regression (Ridge); Right panel: Risk prediction of RLFM(1), Lasso regression (Lasso) and MRRR-based regression.

dataset. The average NMSE for our method was smaller than other alternative methods. We also noted that the average NMSE of RLFM(1) and MRRR were about equal to 1, approximately equivalent to the predictor of the mean value of Y_i, indicating that the RLFM(1) and MRRR model did not provide more predictive power for Y_i. We also calculated canonical correlation between four factors for LFM and four factors for GFM (**ccor** = 0.096), showing that the factors from LFM were significantly different from those from GFM. In addition, we calculated the scaled log-likelihood $l/(np)$ for GFM and LFM to measure the goodness of fit (−1.3261 and −1.6040 for GFM and LFM with four factors, respectively). Clearly, GFM had better goodness of fit for the data. To better understand the performances of Lasso and Ridge, we calculated the pairwise Pearson's correlation of the variables and we found many SNPs had strong correlation. Particularly, there were about 10,120 absolute pair correlations exceeding 0.9 and 29.34% of the maximum absolute pair correlations, defined as $\alpha_j = \max_{k<j} |\text{corr}(X_{ik}, X_{ij})|$, exceeding 0.9 (see Figure 6(a)). Since the Ridge regression was designed to overcome the strong correlation among the covariates in linear regression, the Ridge regression performed better for prediction than Lasso in this scenario.

To better understand the underlying mechanism, we applied the regularized GFM with L_1 penalty as described in Remark 3 to the GWAS dataset. In this way, we can produce sparse loading vectors as shown in Figure 1. Also, by matching SNPs with

Table 9. Gene set enrichment analysis for the first and second loadings with nonzero values

Pathway ID	# Genes (K)	Description	# Overlap (k)	k/K	*p*-value	FDR q-value
The first loading						
hsa04080	272	Neuroactive ligand-receptor interaction	34	0.1250	1.79E-13	3.32E-11
hsa04510	201	Focal adhesion	26	0.1294	5.65E-11	5.25E-9
hsa04512	84	ECM-receptor interaction	17	0.2024	9.95E-11	6.17E-9
hsa05412	76	Arrhythmogenic right ventricular cardiomyopathy	16	0.2105	1.89E-10	8.78E-9
hsa04360	129	Axon guidance	20	0.1550	3.56E-10	1.33E-8
hsa05414	92	Dilated cardiomyopathy	17	0.1848	4.48E-10	1.39E-8
hsa05410	85	Hypertrophic cardiomyopathy (HCM)	16	0.1882	1.09E-9	2.9E-8
hsa04514	134	Cell adhesion molecules (CAMs)	19	0.1418	4.54E-9	1.06E-7
hsa00230	159	Purine metabolism	18	0.1132	3.81E-7	7.87E-6
hsa04810	216	Regulation of actin cytoskeleton	21	0.0972	5.57E-7	1.04E-5
hsa04020	178	Calcium signaling pathway	18	0.1011	2E-6	3.38E-5
hsa04010	267	MAPK signaling pathway	22	0.0824	4.78E-6	7.41E-5
The second loading						
hsa04510	201	Focal adhesion	37	0.1841	1.34E-15	2.5E-13
hsa05414	92	Dilated cardiomyopathy	25	0.2717	4.36E-15	4.06E-13
hsa04080	272	Neuroactive ligand-receptor interaction	40	0.1471	2.31E-13	1.43E-11
hsa05412	76	Arrhythmogenic right ventricular cardiomyopathy	21	0.2763	4.66E-13	2.17E-11
hsa04020	178	Calcium signaling pathway	31	0.1742	1.23E-12	4.56E-11
hsa04810	216	Regulation of actin cytoskeleton	34	0.1574	2.02E-12	6.25E-11
hsa04512	84	ECM-receptor interaction	21	0.2500	3.93E-12	1.04E-10
hsa04730	70	Long-term depression	19	0.2714	8.49E-12	1.97E-10
hsa04514	134	Cell adhesion molecules (CAMs)	25	0.1866	3.84E-11	7.85E-10
hsa05410	85	Hypertrophic cardiomyopathy (HCM)	20	0.2353	4.22E-11	7.85E-10
hsa00230	159	Purine metabolism	26	0.1635	3.28E-10	5.26E-9
hsa04270	115	Vascular smooth muscle contraction	22	0.1913	3.39E-10	5.26E-9
hsa04540	90	Gap junction	19	0.2111	9.36E-10	1.34E-8
hsa04360	129	Axon guidance	22	0.1705	3.28E-9	4.36E-8
hsa04010	267	MAPK signaling pathway	32	0.1199	1.02E-8	1.26E-7
hsa04062	190	Chemokine signaling pathway	26	0.1368	1.59E-8	1.83E-7
hsa04520	75	Adherens junction	16	0.2133	1.67E-8	1.83E-7
hsa05200	328	Pathways in cancer	35	0.1067	4.22E-8	4.36E-7
hsa04070	76	Phosphatidylinositol signaling system	15	0.1974	1.39E-7	1.36E-6
hsa04670	118	Leukocyte transendothelial migration	18	0.1525	4.87E-7	4.53E-6
hsa04530	134	Tight junction	19	0.1418	7.53E-7	6.67E-6
hsa04912	101	GnRH signaling pathway	16	0.1584	1.24E-6	1.05E-5
hsa04144	183	Endocytosis	22	0.1202	1.85E-6	1.49E-5
hsa04960	42	Aldosterone-regulated sodium reabsorption	10	0.2381	2.81E-6	2.18E-5
hsa04260	80	Cardiac muscle contraction	13	0.1625	9.09E-6	6.76E-5
hsa04660	108	T cell receptor signaling pathway	15	0.1389	1.38E-5	9.89E-5

overlapped genes, we then conducted the gene set enrichment analysis for each loading with nonzero entries (Subramanian et al. 2005). Tables 9–11 show the significant KEGG pathways with FDR q-value $< 1e - 1 \times 10^{-4}$. In all significant KEGG pathways, our results replicated many of the top findings in (Zhao et al. 2015), for example, GnRH signaling pathway, Long-term potentiation, Vascular smooth muscle contraction, gap junction, Wnt signaling pathway. Many studies have reported the associations of BMI with psychiatric disorders, for example, schizophrenia (Coodin 2001), depression (De Wit et al. 2009). As we expected, we identified the long-term depression pathway in the third loading. However, it is still elusive to understand the underlying mechanism behind these associations. In our comprehensive analysis, we identified the neuroactive ligand-receptor interaction pathway as the most top in the first loading, which suggests how BMI affects psychiatric disorders through the neuroactivities. On the other hand, we also identified pathways from wnt signaling and type II diabetes mellitus. Previous studies found that wnt signaling links to the regulation of adipogenesis (Christodoulides et al. 2009), while obesity is an increased risk of developing type II diabetes (Hjartåker, Langseth, and Weiderpass 2008).

5.2. *Arrhythmia Dataset*

The cardiac arrhythmia dataset used in this study is publicly available in machine learning repository *https://archive.ics.uci.edu/ml/datasets*. Cardiac arrhythmia, or called disorders of cardiac rhythm, may indicate the susceptibility of severe heart disease, stroke, or sudden cardiac death. This dataset is ambiguous bio-signal data collected from a total of 452 patient cases. In this dataset, there are 245 normal individuals, 278 feature variables ($\mathbf{x}_i$), and a response variable of interest (Y_i) that is a nominal variable with 16 levels representing 16 types of individuals, including normal and another 15 ill individuals. After preprocessing the data, such as dropping out the observations with missing values and the variables without information, 226 features across 420 observations were remained, including 165 continuous features and 61 binary features.

To apply GFM and LFM to the cardiac arrhythmia data, we first chose the number of factors using PC criteria, with the estimated $\hat{q} = 8$ and $\hat{q} = 1$ for GFM and LFM, respectively. The different numbers of factors chosen for GFM and LFM could be due to the loss of information on binary manifest variables by LFM. To further compare the prediction performance, we conducted classification on the response variable Y_i using five methods, GFM-RF, LFM-RF(1), LFM-RF(8), MRRR-RF, and

Table 10. Gene set enrichment analysis for the third and fourth loadings with nonzero values

Pathway ID	# Genes (K)	Description	# Overlap (k)	k/K	*p*-value	FDR q-value
		The third loading				
hsa04730	70	Long-term depression	22	0.3143	6.68E-15	1.24E-12
hsa04020	178	Calcium signaling pathway	33	0.1854	3.73E-14	2.82E-12
hsa05412	76	Arrhythmogenic right ventricular cardiomyopathy	22	0.2895	4.55E-14	2.82E-12
hsa04270	115	Vascular smooth muscle contraction	26	0.2261	1.47E-13	6.85E-12
hsa05414	92	Dilated cardiomyopathy	23	0.2500	3.8E-13	1.41E-11
hsa05410	85	Hypertrophic cardiomyopathy (HCM)	22	0.2588	5.71E-13	1.77E-11
hsa05200	328	Pathways in cancer	43	0.1311	1.65E-12	4.39E-11
hsa04540	90	Gap junction	22	0.2444	2.01E-12	4.67E-11
hsa04530	134	Tight junction	26	0.1940	6.26E-12	1.23E-10
hsa04510	201	Focal adhesion	32	0.1592	6.62E-12	1.23E-10
hsa04360	129	Axon guidance	25	0.1938	1.63E-11	2.76E-10
hsa04912	101	GnRH signaling pathway	21	0.2079	1.69E-10	2.62E-9
hsa04010	267	MAPK signaling pathway	35	0.1311	1.89E-10	2.7E-9
hsa04810	216	Regulation of actin cytoskeleton	31	0.1435	2.09E-10	2.78E-9
hsa04520	75	Adherens junction	18	0.2400	2.8E-10	3.47E-9
hsa04514	134	Cell adhesion molecules (CAMs)	23	0.1716	1.29E-9	1.5E-8
hsa04080	272	Neuroactive ligand-receptor interaction	33	0.1213	4.49E-9	4.92E-8
hsa00500	52	Starch and sucrose metabolism	14	0.2692	5.29E-9	5.47E-8
hsa00053	25	Ascorbate and aldarate metabolism	10	0.4000	1.13E-8	1.1E-7
hsa04070	76	Phosphatidylinositol signaling system	16	0.2105	2.06E-8	1.91E-7
hsa04666	97	Fc gamma R-mediated phagocytosis	18	0.1856	2.23E-8	1.97E-7
hsa00860	41	Porphyrin and chlorophyll metabolism	12	0.2927	2.33E-8	1.97E-7
hsa00983	51	Drug metabolism—other enzymes	13	0.2549	3.82E-8	3.09E-7
hsa04310	151	Wnt signaling pathway	22	0.1457	6.41E-8	4.97E-7
hsa04720	70	Long-term potentiation	14	0.2000	3.06E-7	2.28E-6
hsa04670	118	Leukocyte transendothelial migration	18	0.1525	4.91E-7	3.52E-6
hsa04512	84	ECM-receptor interaction	15	0.1786	5.5E-7	3.69E-6
hsa00040	28	Pentose and glucuronate interconversions	9	0.3214	5.55E-7	3.69E-6
hsa00230	159	Purine metabolism	21	0.1321	6.78E-7	4.35E-6
hsa04144	183	Endocytosis	22	0.1202	1.86E-6	1.16E-5
hsa04930	47	Type II diabetes mellitus	10	0.2128	8.36E-6	5.02E-5
hsa04260	80	Cardiac muscle contraction	13	0.1625	9.14E-6	5.31E-5
hsa00982	72	Drug metabolism—cytochrome P450	12	0.1667	1.54E-5	8.63E-5
hsa05222	84	Small cell lung cancer	13	0.1548	1.58E-5	8.63E-5
hsa05416	73	Viral myocarditis	12	0.1644	1.78E-5	9.44E-5
		The fourth loading				
hsa04080	272	Neuroactive ligand-receptor interaction	18	0.0662	2.7E-7	5.02E-5

Table 11. Gene set enrichment analysis for the fourth loading with nonzero values

Pathway ID	# Genes (K)	Description	# Overlap (k)	k/K	*p*-value	FDR q-value
hsa04080	272	Neuroactive ligand-receptor interaction	18	0.0662	2.7E-7	5.02E-5

RAW-RF, namely three factor models (GFM with $\hat{q} = 8$, LFM with $\hat{q} = 1$ and LFM with $\hat{q} = 8$), mixed reduced-rank regression model (Luo et al. 2018) and original data (RAW) followed by applying random forest for classification. In details, we first used GFM, LFM and MRRR (covariates matrix was regarded as identity matrix) to extract factors, respectively. Then we applied the random forest on the extracted factors to classify the outcome Y_i. RAW-RF means that the classification was directly conducted on raw data ($\mathbf{x}_i$) without applying factor analysis. Finally, we evaluated the classification using 10-fold cross-validation. Classification accuracy is the average accuracy of these 10 runs. The results are summarized in Figure 4. The classification accuracy of GFM-RF was closer to that of RAM-RF. Since the GFM only used eight factors for classification, hence it is more concise and informative than the method based on the original data. The difference of performance between GFM-RF and LFM-RF(8) was not significant, since this data included 165 continuous variables and 61 binary variables, the ratio of variance between continuous and binary variables was 12,187.74, that is, the continuous variables absolutely dominated the binary variables in signal strength. Thus, GFM was nearly degenerated to LFM.

Next, we investigated the meanings of factors. The majority of variables in this data were measured from different channels, representing different positions near the heart. To measure the heart's electrical activity accurately, proper channels (electrode placement) is crucial in ECG records. From Figure 3, we knew factor 1 absorbed the information of variables $x_9, x_{146}, x_{139}, x_{145}, x_{153}$ that represent the information measured from V3, V4 and V5 channel (Guvenir et al. 1997) in the ECG records; factor 2 absorbed the information of variables $x_{138}, x_{139}, x_9, x_{130}, x_{131}$ that mainly represent the information measured from V2 channel (Guvenir et al. 1997); factor 3 absorbed the information of variables $x_{122}, x_{11}, x_{27}, x_{139}, x_{31}$ that mainly represent the information measured from VAR channel (Guvenir et al. 1997); factor 4 absorbed the information of $x_5, x_8, x_7, x_{131}, x_{139}$ that mainly represent the information of average duration between two consecutive waves in msec; factor 5 absorbed the information from AVR, AVF and DII channels; factor 6 absorbed the information from V1 and V4 channels; factor 7 absorbed the

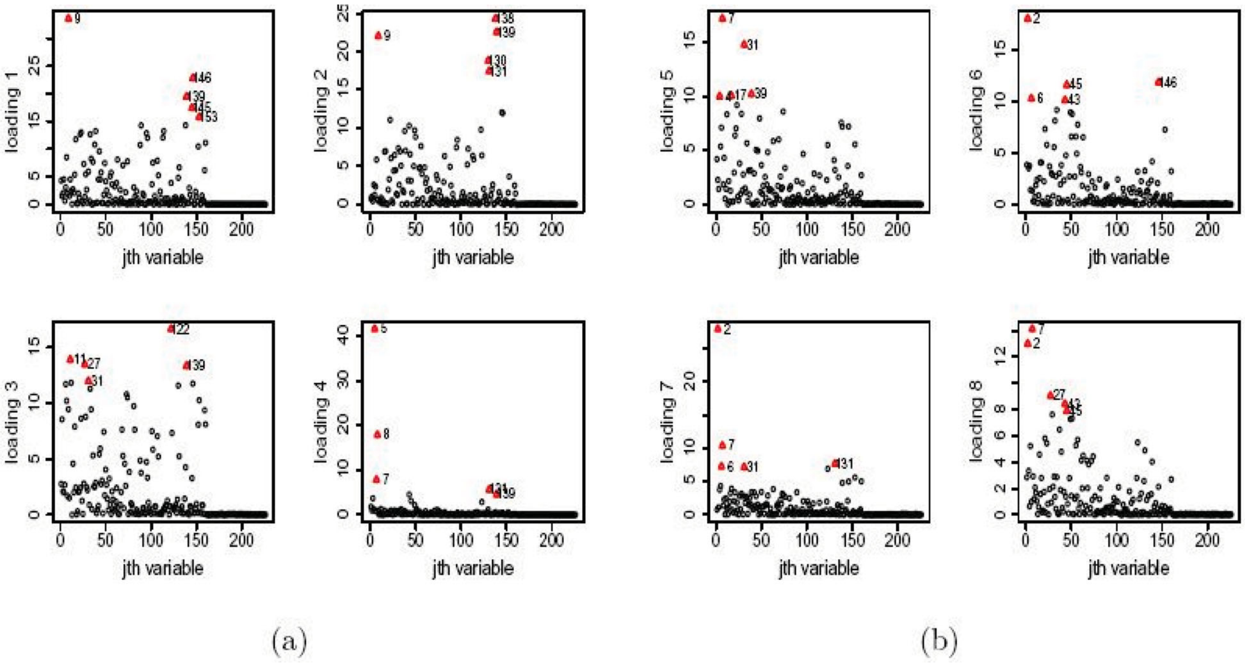

Figure 3. The absolute values among the eight loadings, where the red triangular dots are the first eight largest absolute values of loading.

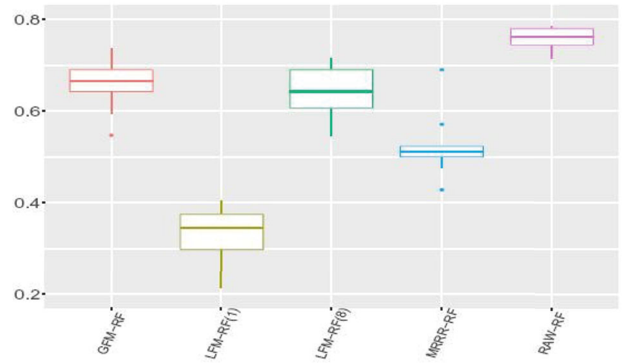

Figure 4. Comparison of classification accuracy for the 16 types of individuals, where GFM-RF, LFM-RF(1), LFM-RF(8), MRRR-RF, and RAW-RF represent three factor models (GFM with eight factors, LFM with one factor and LFM with eight factors same as that of GFM), mixed rank-reduced regression (MRRR) model and original data (RAW) followed by applying random forest for classification.

information from V2 and AVR channels; and factor 8 absorbed the information from V1 and AVR channels (Guvenir et al. 1997).

To further evaluate the classification performance using the receiver operating characteristic (ROC) curve, we collapsed the multiple levels of response into two levels, 0 for normal individuals, 1 for ill individuals, with 237 and 183 observations, respectively. Then, we randomly divided the observations into the training sample and testing sample in a ratio of roughly 2 : 1. Six methods, including the above five methods and logistic regression with l_1 penalty (Logistic), were applied to the training sample to fit the model, and the ROC curves were calculated for testing samples, as shown in Figure 5, where Logistic is also the degenerated version of generalized trace regression via regularization in Fan, Gong, and Zhu (2019). Clearly, the performance of LFM-RF(1) and MRRR-RF were nearly close to random guessing while the performance of GFM-RF was little better than LFM-RF(8)'s. Moreover, AUC of GFM-RF was close to the methods using all data information, that is, Logisitc and RAW-RF. To understand the result, we also examined the correlations of the raw variables, and found that all of absolute pairwise Pearson's correlations are less than 0.93 and only 2.21% of the maximum absolute pairwise correlation exceeds 0.9 (see Figure 6(b)). In this case, logistic can work better.

6. Discussion

In this article, we developed a generalized factor model (GFM) for the mixed-type variables in the case of large n and large p. Meanwhile, a computationally efficient algorithm has been developed to estimate the factors and loadings from GFM. Under some regularity conditions, we have established the consistency and convergent rates for the estimated factors and loadings. The comprehensive simulation studies using the GFM and the linear factor model (LFM) demonstrate the advantages of GFM in the presence of mixed-type variables. Particularly, the real data analysis of GFM on a GWAS dataset and an arrhythmia dataset show that GFM has better prediction and classification performance in the presence of non-continuous predictors.

By moving forward, the model and estimation strategy admit several potential extensions. Following the literature on factor analysis (Bai and Ng 2002; Fan, Liao, and Mincheva 2013; Bai and Ng 2013; Li et al. 2018), we assume the loading matrix $\mathbf{B}$ as well as the factor matrix $\mathbf{H}$ has finite rank, it is interesting to consider the case with growing rank as n or p increases.

In this article, we consider extracting the latent factors from the ultra-high dimensional variables of mixed types. In practice, we may be interested in the relationship between the response

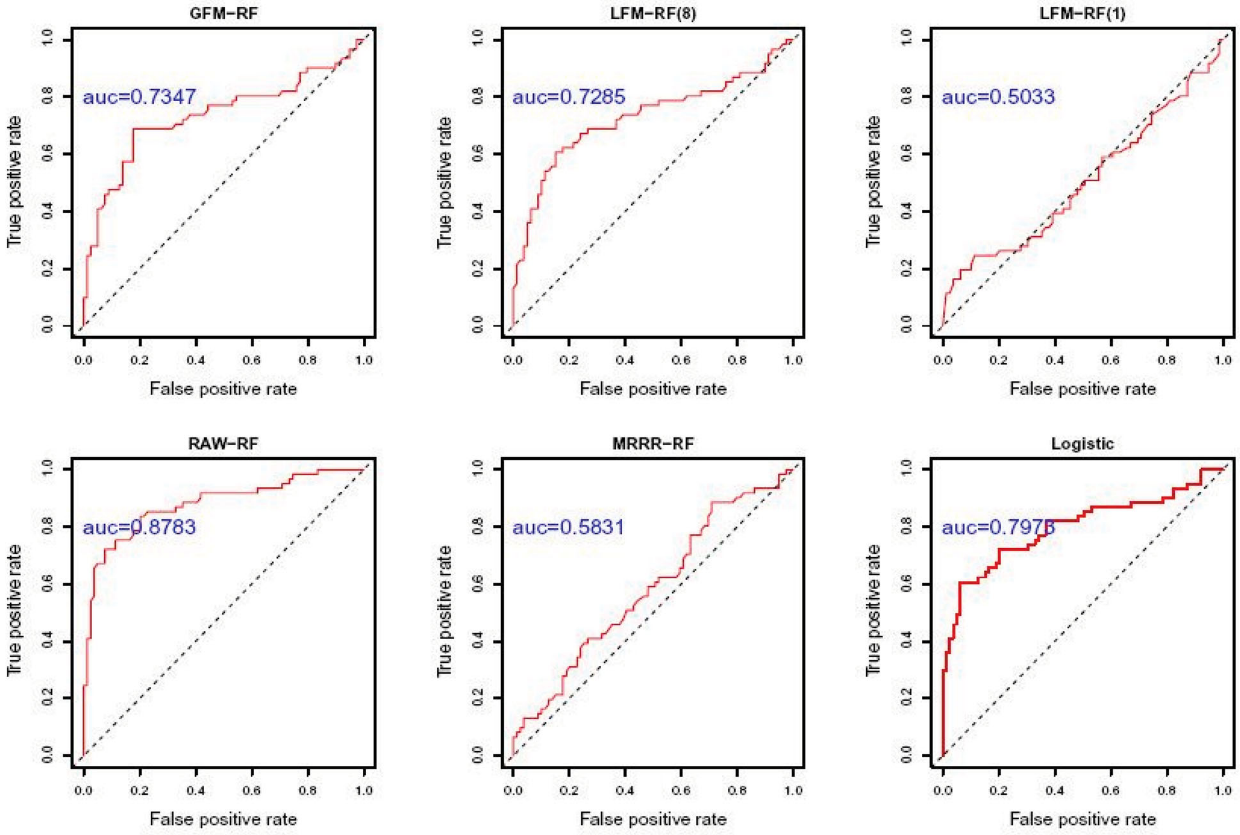

Figure 5. Comparison of ROC curves with AUC value in Arrhythmia data analysis, where GFM-RF, LFM-RF(1), LFM-RF(8), MRRR-RF, and RAW-RF represent three factor models (GFM with eight factors, LFM with one factor and LFM with eight factors same as that of GFM), mixed rank-reduced regression (MRRR) model and original data (RAW) followed by applying random forest for classification; Logistic represents the logistic regression with l_1 penalty.

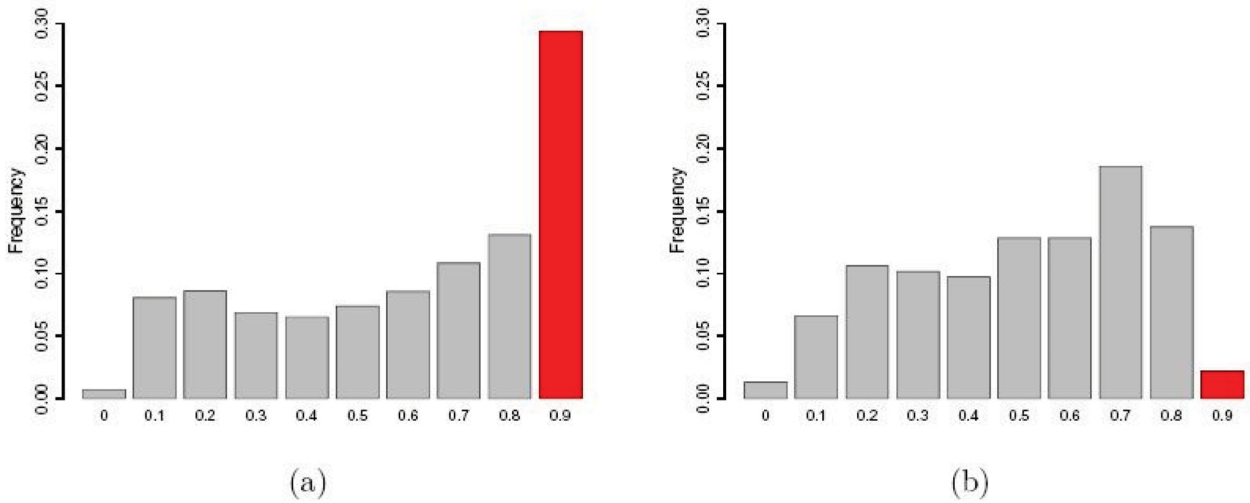

Figure 6. (a): The frequency of the maximum absolute pairwise correlation, defined by $\alpha_j = \max_{k<j} |\text{corr}(X_{ik}, X_{ij})|$, for the NFBC1966 dataset. (b): The frequency of the maximum absolute pairs correlation for the arrthythmia dataset. And the red bar represents the frequency of the maximum absolute pairwise correlation exceeding 0.9.

and the covariates. Our estimation procedures are unsupervised in the sense that the latent factors are extracted without respect to the response. Hence, the principal features captured by the latent variables are the most important directions concerning the information of covariates alone rather than under the supervision of the response. It is still an open but essential question on how to extract factors from mixed-type covariates in the direction supervised by a response variable.

Supplementary Materials

In the supplementary materials, we give the detailed proofs of Propositions 1–2 and Theorems 1–4 and all codes used in simulation studies and real data examples.

Funding

Lin's work was supported by National Natural Science Foundation of China (Nos. 11931014 and 11829101) and Fundamental Research Funds for the Central Universities (No. JBK1806002) of China. Zheng's work was supported by NSFC grants 12071066 and 11690012. Jin Liu's work was supported by grant R-913-200-098-263 from the Duke-NUS Medical School, and AcRF Tier 2 (MOE2018-T2-1-046 and MOE2018-T2-2-006) from the Ministry of Education, Singapore.

References

Ahn, S. C., and Horenstein, A. R. (2013), "Eigenvalue Ratio Test for the Number of Factors," *Econometrica*, 81, 1203–1227. [1389]

Amemiya, Y., and Anderson, T. W. (1990), "Asymptotic Chi-Square Tests for a Large Class of Factor Analysis Models," *The Annals of Statistics*, 1453–1463. [1385]

Amemiya, Y., Fuller, W. A., and Pantula, S. G. (1987), "The Asymptotic Distributions of Some Estimators for a Factor Analysis Model," *Journal of Multivariate Analysis*, 22, 51–64. [1385]

Anderson, T. W., and Rubin, H. (1956), "Statistical Inference in Factor Analysis," in *Berkeley Symposium on Mathematicalstatistics & Probability*. [1385]

Bai, J. (2003), "Inferential Theory for Factor Models of Large Dimensions," *Econometrica*, 71, 135–171. [1386,1387,1389]

Bai, J., and Li, K. (2012), "Statistical Analysis of Factor Models of High Dimension," *The Annals of Statistics*, 40, 436–465. [1386,1387,1390,1391]

Bai, J., and Liao, Y. (2013), "Statistical Inferences Using Large Estimated Covariances for Panel Data and Factor Models," arXiv:1307.2662. [1387,1389,1390]

Bai, J., and Ng, S. (2002), "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 70,191–221. [1385,1386,1387,1388,1389,1390,1393,1398]

——— (2013), "Principal Components Estimation and Identification of Static Factors," *Journal of Econometrics*, 176, 18–29. [1386,1387,1390,1398]

Bartholomew, D. J. (1980), "Factor Analysis for Categorical Data," *Journal of the Royal Statistical Society*, Series B, 42, 293–321. [1385]

——— (1987), *Latent Variable Models and Factors Analysis*, Oxford: Oxford University Press, Inc. [1385]

Bartholomew, D. J., Knott, M., and Moustaki, I. (2011), *Latent Variable Models and Factor Analysis: A Unified Approach*, Vol. 904. Wiley. [1385]

Bock, R. D., and Lieberman, M. (1970), "Fitting a Response Model Forn Dichotomously Scored Items," *Psychometrika*, 35, 179–197. [1385]

Browne, M. W. (1984), "Asymptotically Distribution-Free Methods for the Analysis of Covariance Structures," *British Journal of Mathematical and Statistical Psychology*, 37, 62–83. [1385]

Campos, G. D. L., Gianola, D., and Allison, D. B. (2010), "Predicting Genetic Predisposition in Humans: The Promise of Whole-Genome Markers," *Nature Reviews Genetics*, 11, 880–886. [1394]

Candes, E. J., Sing-Long, C. A., and Trzasko, J. D. (2013), "Unbiased Risk Estimates for Singular Value Thresholding and Spectral Estimators," *IEEE Transactions on Signal Processing*, 61, 4643–4657. [1392]

Carvalho, C. M., Chang, J. T., Lucas, J. E., Nevins, J. R., Wang, Q., and West, M. A. (2008), "High-Dimensional Sparse Factor Modeling: Applications in Gene Expression Genomics," *Journal of the American Statistical Association*, 103, 1438–1456. [1388]

Chen, K., Dong, H., and Chan, K.-S. (2013), "Reduced Rank Regression Via Adaptive Nuclear Norm Penalization," *Biometrika*, 100, 901–920. [1392]

Christodoulides, C., Lagathu, C., Sethi, J. K., and Vidal-Puig, A. (2009), "Adipogenesis and WNT Signalling," *Trends in Endocrinology & Metabolism*, 20, 6–24. [1396]

Christoffersson, A. (1975), "Factor Analysis of Dichotomized Variables," *Psychometrika*, 40, 5–32. [1385]

Coodin, S. (2001), "Body Mass Index in Persons With Schizophrenia," *The Canadian Journal of Psychiatry*, 46, 549–555. [1396]

Cudeck, R., and MacCallum, R. C. (2012), *Factor Analysis at 100: Historical Developments and Future Directions*, Routledge. [1385]

De Wit, L. M., Van Straten, A., Van Herten, M., Penninx, B. W., and Cuijpers, P. (2009), "Depression and Body Mass Index, a U-Shaped Association," *BMC Public Health*, 9, 14. [1396]

Doz, C., Giannone, D., and Reichlin, L. (2012), "A Quasi–Maximum Likelihood Approach for Large, Approximate Dynamic Factor Models," *Review of Economics and Statistics*, 94, 1014–1024. [1390]

Fan, J., Gong, W., and Zhu, Z. (2019), "Generalized High-Dimensional Trace Regression Via Nuclear Norm Regularization," *Journal of Econometrics*, 212, 177–202. [1394,1398]

Fan, J., Liao, Y., and Mincheva, M. (2013), "Large Covariance Estimation by Thresholding Principal Orthogonal Complements," *Journal of the Royal Statistical Society*, Series B, 75, 603–680. [1387,1398]

Fan, J., Xue, L., and Yao, J. (2017), "Sufficient Forecasting Using Factor Models," *Journal of Econometrics*, 201, 292–306. [1385,1386]

Gavish, M., and Donoho, D. L. (2017), "Optimal Shrinkage of Singular Values," *IEEE Transactions on Information Theory*, 63, 2137–2152. [1392]

Goyal, A., Pérignon, C., and Villa, C. (2008), "How Common Are Common Return Factors Across the NYSE and NASDAQ?" *Journal of Financial Economics*, 90, 252–271. [1390]

Guvenir, H. A., Acar, B., Demiroz, G., and Cekin, A. (1997), "A Supervised Machine Learning Algorithm for Arrhythmia Analysis," in *Computers in Cardiology*, 1997. pp. 433–436. Lund, Sweden: IEEE. [1397,1398]

Hjartåker, A., Langseth, H., and Weiderpass, E. (2008), "Obesity and Diabetes Epidemics," in *Innovative Endocrinology of Cancer*, eds. Lev M. Berstein, Richard J. Santen, New York, NY: Springer, pp. 72–93. [1396]

Jiang, F., Ma, Y., and Wei, Y. (2019), "Sufficient Direction Factor Model and Its Application to Gene Expression Quantitative Trait Loci Discovery," *Biometrika*, 106, 417–432. [1386,1387,1389]

Jolliffe, I. T. (2002), *Principal Component Analysis*, vol. 29 in *Springer Series in Statistics*, Berlin: Springer. [1387]

Josse, J., Sardy, S., and Wager, S. (2016), "denoiser: A Package for Low Rank Matrix Estimation," arXiv:1602.01206. [1392]

Josse, J., and Wager, S. (2016), "Bootstrap-Based Regularization for Low-Rank Matrix Estimation," *The Journal of Machine Learning Research*, 17, 4227–4255. [1392]

Lam, C., and Yao, Q. (2012), "Factor Modeling for High-Dimensional Time Series: Inference for the Number of Factors," *The Annals of Statistics*, 40, 694–726. [1386,1389]

Lan, A. S., Waters, A. E., Studer, C., and Baraniuk, R. G. (2014), "Sparse Factor Analysis for Learning and Content Analytics," *Journal of Machine Learning Research*, 15, 1959–2008. [1388]

Lawley, D. (1940), "The Estimation of Factor Loadings by the Method of Maximum Likelihood," *Proceedings of the Royal Society of Edinborough*, 60, 64–82. [1385]

Li, Q., Cheng, G., Fan, J., and Wang, Y. (2018), "Embracing the Blessing of Dimensionality in Factor Models," *Journal of the American Statistical Association*, 113, 380–389. [1386,1387,1390,1398]

Luo, C., Liang, J., Li, G., Wang, F., Zhang, C., Dey, D. K., and Chen, K. (2018), "Leveraging Mixed and Incomplete Outcomes Via Reduced-Rank Modeling," *Journal of Multivariate Analysis*, 167, 378–394. [1392,1394,1397]

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models*, Vol. 37. Boca Raton, FL: CRC Press. [1386,1387]

McDonald, R. P. (1969), "The Common Factor Analysis of Multicategory Data," *British Journal of Mathematical and Statistical Psychology*, 22, 165–175. [1385]

——— (1985), *Factor Analysis and Related Methods*, Hillsdaie. [1387]

Moustaki, I. (1996), "A Latent Trait and a Latent Class Model for Mixed Observed Variables," *British Journal of Mathematical and Statistical Psychology*, 49, 313–334. [1385]

Moustaki, I., and Knott, M. (2000), "Generalized Latent Trait Models," *Psychometrika*, 65, 391–411. [1385,1387]

Muthén, B. (1978), "Contributions to Factor Analysis of Dichotomous Variables," *Psychometrika*, 43, 551–560. [1385]

——— (1984), "A General Structural Equation Model With Dichotomous, Ordered Categorical, and Continuous Latent Variable Indicators," *Psychometrika*, 49, 115–132. [1385]

Olsson, U. (1979), "On the Robustness of Factor Analysis Against Crude Classification of the Observations," *Multivariate Behavioral Research*, 14, 485–500. [1385]

Purcell, S., Neale, B. M., Toddbrown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., De Bakker, P. I. W., Daly, M. J. (2007), "Plink: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses," *American Journal of Human Genetics*, 81, 559–575. [1394]

Sabatti, C., Hartikainen, A., Pouta, A., Ripatti, S., Brodsky, J., Jones, C. G., Zaitlen, N., Varilo, T., Kaakinen, M., Sovio, U., Ruokonen, A., Laitinen, J., Jakkula, E., Coin, L., Hoggart, C., Collins, A., Turunen, H., Gabriel, S., Elliot, P., McCarthy, M. I., Daly, M. J., Järvelin, M. R., Freimer, N. B., Peltonen, L. (2009), "Genome-Wide Association Analysis of Metabolic Traits in a Birth Cohort From a Founder Population," *Nature Genetics*, 41, 35–46. [1386]

Spearman, C. (1904), "General Intelligence," Objectively Determined and Measured," *The American Journal of Psychology*, 15, 201–292. [1385]

Spearman, C. (1927), *The Abilities of Man: Their Nature and Measurement*, Vol. 8. New York: Macmillan. [1385]

Stock, J. H., and Watson, M. W. (2002), "Forecasting Using Principal Components From a Large Number of Predictors," *Journal of the American Statistical Association*, 97, 1167–1179. [1386,1387]

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., Mesirov, J. P. (2005), "Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles," *Proceedings of the National Academy of Sciences*, 102, 15545–15550. [1396]

Torgo, L. (2011), *Data Mining With R: Learning With Case Studies*, Chapman and Hall/CRC. [1386]

Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011), "Gcta: A Tool for Genome-Wide Complex Trait Analysis," *American Journal of Human Genetics*, 88, 76–82. [1394]

Zhao, X., Gu, J., Li, M., Xi, J., Sun, W., Song, G., and Liu, G. (2015), "Pathway Analysis of Body Mass Index Genome-Wide Association Study Highlights Risk Pathways in Cardiovascular Disease," *Scientific Reports*, 5, 13025. [1396]

Zou, H., Hastie, T., and Tibshirani, R. (2006), "Sparse Principal Component Analysis," *Journal of Computational and Graphical Statistics*, 15, 265–286. [1388]