Statistica Sinica **33** (2023), 27-53
doi:https://doi.org/10.5705/ss.202020.0145

HIGH-DIMENSIONAL FACTOR REGRESSION FOR HETEROGENEOUS SUBPOPULATIONS

Peiyao Wang[1], Quefeng Li[1], Dinggang Shen[2,3,4] and Yufeng Liu[1]

[1] *University of North Carolina at Chapel Hill,* [2] *ShanghaiTech University,*
[3] *Shanghai United Imaging Intelligence Co. and* [4] *Korea University*

Abstract: In modern scientific research, data heterogeneity is commonly observed owing to the abundance of complex data. We propose a factor regression model for data with heterogeneous subpopulations. The proposed model can be represented as a decomposition of heterogeneous and homogeneous terms. The heterogeneous term is driven by latent factors in different subpopulations. The homogeneous term captures common variation in the covariates and shares common regression coefficients across subpopulations. Our proposed model attains a good balance between a global model and a group-specific model. The global model ignores the data heterogeneity, while the group-specific model fits each subgroup separately. We prove the estimation and prediction consistency for our proposed estimators, and show that it has better convergence rates than those of the group-specific and global models. We show that the extra cost of estimating latent factors is asymptotically negligible and the minimax rate is still attainable. We further demonstrate the robustness of our proposed method by studying its prediction error under a misspecified group-specific model. Finally, we conduct simulation studies and analyze a data set from the Alzheimer's Disease Neuroimaging Initiative and an aggregated microarray data set to further demonstrate the competitiveness and interpretability of our proposed factor regression model.

Key words and phrases: Factor models, heterogeneity, penalized regression, prediction.

1. Introduction

Data heterogeneity is an important issue in modern complex data analysis. In practice, data heterogeneity may come from variables or samples. More specifically, multi-modality/source data have heterogeneity among the variables, because they may correspond to different types of measurements. For example, in biomedical imaging, people may acquire both MRI and PET images (Zhang et al. (2011)). In genomics studies, measurements are collected from different sources, such as mRNA and miRNA (Muniategui et al. (2013)). In addition to

Corresponding author: Yufeng Liu, Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA. E-mail: yfliu@email.unc.edu.

variable heterogeneity, data heterogeneity can also arise from samples. For example, there can be subpopulations, batch and clustering effects, or outliers in the data (Bühlmann (2016)), potentially violating the standard independent and identically distributed (i.i.d.) assumption. Ignoring such heterogeneity can lead to poor estimation and prediction. Hence, it is important to take data heterogeneity into account during the modeling process.

In this study, we are interested in data heterogeneity that comes from subgroup populations. For example, in the Alzheimer's Disease (AD) study, subjects can have five subtypes: Normal Control (NC), Significant Memory Concern (SMC), Early Mild Cognitive Impairment (eMCI), Late Mild Cognitive Impairment (lMCI), and AD, where these subtypes are ordered by disease severity. Owing to data heterogeneity, it can be difficult to build accurate and interpretable predictive models on such data using traditional statistical techniques. A global model that fits a single regression model to all the data may be restrictive because it ignores the group label information, whereas fitting distinct regression models in each group may not be optimal because this does not capture shared information across groups. Hence, a statistical regression model that can recover interpretable globally shared and group-specific signals in the data is required to handle such heterogeneous data. In the literature, varying coefficient models (Hastie and Tibshirani (1993)) and mixed-effects models (Pinheiro and Bates (2000)) are useful in addressing data heterogeneity. However, these models can be computationally expensive to use in practice, especially when the dimension is too high. More recently, Vicari and Vichi (2013) proposed a general regression model to account for both between-cluster and within-cluster variation. Meinshausen and Bühlmann (2015) proposed a maxmin-effects approach under the mixture model. Zhao, Cheng and Liu (2016) proposed a partially linear regression framework to model massive heterogeneous data. Tang and Song (2016) and Ma and Huang (2017) proposed fused penalties to estimate regression coefficients in order to identify subpopulations. Wang, Liu and Shen (2018) proposed a locally weighted penalized model by incorporating a progression score in the local kernels. However, these models are not designed to characterize the globally shared and group-specific structures. Thus, it is desirable to build a model that can identify such structures, quantify prediction errors, and draw interpretable and generalizable scientific conclusions.

There is a large body of literature on data heterogeneity for unsupervised learning. Principal component analysis (PCA) (Wold, Esbensen and Geladi (1987)) techniques are popular, owing to their computational simplicity and theoretical soundness. The joint and individual variations explained (JIVE) method

(Lock et al. (2013)) decomposes joint and individual low-rank signals across multiple sources of data. More recent extensions of JIVE can be found in Feng et al. (2018); Gaynanova and Li (2019); Park and Lock (2020). These methods can be extended easily to decompose data from multiple subgroups. Zhou et al. (2015) proposed a matrix factorization framework for common and individual feature extraction for multi-block data.

Closely related to PCA, another popular technique for handling data heterogeneity is factor models. Factor models are useful unsupervised learning tools that model the dependence between multiple variables. The relationship between PCA and factor models is well studied in the literature (Joliffe and Morgan (1992); Stock and Watson (2002); Bai and Ng (2002)). Factor models assume that the variations among the variables are driven by latent factors residing in a low-dimensional space. More recently, Fan et al. (2018) proposed a factor model framework to model the heterogeneity from different subgroups. They used the factor model in the context of Gaussian graphical models to estimate common and individual graphs from different groups. Their structural assumption on the data matrices can be generalizable to predictive modeling.

Here, we focus on supervised learning, and propose a novel factor regression model for heterogeneous data with jointly shared and group-specific structures. We assume that the leading factors in each group drive the majority of variation, which contributes to the heterogeneity effects. After the majority of the variation has been removed, the residual signals are assumed to be homogeneous across subgroups; that is, they have the same covariance matrix. Under this framework, the predictors in the proposed model can be decomposed into heterogeneous factors and homogeneous signals. Correspondingly, in our proposed model, the regression coefficients associated with the factors are group specific, whereas the regression coefficients associated with the homogeneous signals are shared across groups. We use PCA to estimate the factors and homogeneous signals. Because the estimated factors and homogeneous signals are orthogonal, their coefficients can be estimated separately. The low-dimensional heterogeneous regression coefficients can be estimated directly using the ordinary least squares (OLS) method. After projecting the responses on the estimated factors in each group, their residuals can be aggregated together to perform a global regression. When the dimension is high, the homogeneous signals' coefficients are difficult to estimate. Following given penalization methods (Hoerl and Kennard (2000); Tibshirani (1996); Zou and Hastie (2005)), we propose a flexible penalized least squares method to solve for the high-dimensional coefficients. In the least squares problem, we use the adaptive thresholding estimator (Cai and Liu (2011)) to estimate the covariance of the homogeneous signals. For prediction, we propose a data-driven trace maximization step to estimate the factors and homogeneous signals in the test set before applying our model for prediction.

We establish the estimation consistency for our proposed estimators using either an ℓ_2 or ℓ_1 penalty. In terms of the prediction accuracy, we study the prediction error of our method in both theoretical and simulation studies, and demonstrate that the proposed model attains a good balance between a global model and a group-specific model. Furthermore, we show that our method is robust when the underlying model is group specific, and has comparable prediction performance with respect to the group-specific model. We apply our method to an Alzheimer's Disease Neuroimaging Initiative (ADNI) data set and an aggregated microarray data set to show the competitiveness of our model in terms of model prediction and interpretability.

The rest of paper is organized as follows. In Section 2, we introduce the factor decomposition of heterogeneous and homogeneous signals and a corresponding regression model. In Section 3, we introduce the model estimation and a data-driven approach to estimate the factors in the testing data for prediction. In Section 4, we study the estimation and prediction consistency of our proposed method, and compare it with those of group-specific and global models under different scenarios. In Section 5, we conduct simulated experiments to evaluate the performance of our model under different settings, and compare them with that of the global and group-specific models. In Section 6, we apply our model to the ADNI data to predict the clinical score. We conclude the paper with a discussion in Section 7.

2. Motivation and Model Framework

Factor models are useful for modeling the dependence between multiple variables, if these variables are driven by some latent factors. For heterogeneous data, the subgroup heterogeneity can be captured by the group-specific latent factors. After removing such latent factors, different subgroups can be viewed as homogeneous samples for a joint analysis. In this section, we first motivate our proposed model by introducing two simple models in Section 2.1. Then, we briefly review the factor decomposition for heterogeneous data and propose our new factor regression model in Section 2.2.

2.1. Motivation

We first introduce some notation. Assume that the data come from G groups. There are n_g samples in the gth group, each having the same set of p explanatory variables. Let $\{\mathbf{X}_g, \boldsymbol{Y}_g\}_{g=1}^G$ be the observations from G groups, where $\mathbf{X}_g \in \mathbb{R}^{n_g \times p}$ is the data matrix and $\boldsymbol{Y}_g \in \mathbb{R}^{n_g}$ is the response vector.

There are two commonly used approaches in the regression setup for heterogeneous subpopulations. On the one hand, ignoring the group information, one can use a global model:

$$\boldsymbol{Y} = \boldsymbol{\mu}^* + \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}, \tag{2.1}$$

where $\boldsymbol{Y} = (\boldsymbol{Y}_1', \dots, \boldsymbol{Y}_G')'$ and $\mathbf{X} = (\mathbf{X}_1', \dots, \mathbf{X}_G')'$. In this model, all the subgroups share the same intercept and regression coefficients. The global model ignores the heterogeneity from subgroups and may be too restrictive. On the other hand, by modeling each group separately, one may consider a group-specific model:

$$\boldsymbol{Y}_g = \boldsymbol{\mu}_g^* + \mathbf{X}_g\boldsymbol{\beta}_g^* + \boldsymbol{\epsilon}_g. \tag{2.2}$$

However, this model may not be efficient because it ignores the shared information across subgroups. These global and group-specific models motivate us to consider a model in between, under which the group-specific heterogeneity and homogeneity across subgroups can both be accounted for. This can be achieved by using a factor model that decomposes covariates into the heterogeneous and homogeneous components.

2.2. Factor model framework

To model the heterogeneous effect introduced by groups, assume that the data matrix $\mathbf{X}_g$ can be decomposed as

$$\mathbf{X}_g = \mathbf{F}_g\boldsymbol{\Lambda}_g + \mathbf{U}_g, \tag{2.3}$$

where $\mathbf{F}_g \in \mathbb{R}^{n_g \times K_g}$ is the factor matrix, $\boldsymbol{\Lambda}_g \in \mathbb{R}^{K_g \times p}$ is the loading matrix, and $\mathbf{U}_g \in \mathbb{R}^{n_g \times p}$ denotes the homogeneous signals, also known as idiosyncratic errors in the factor model literature (Bai and Ng (2008)). The number of random factors K_g can vary among groups.

Denote the ith row of $\mathbf{X}_g$, $\mathbf{F}_g$, and $\mathbf{U}_g$ by $\mathbf{x}_{g,i}$, $\mathbf{f}_{g,i}$, and $\mathbf{u}_{g,i}$ respectively. By (2.3), we have $\mathbf{x}_{g,i} = \boldsymbol{\Lambda}_g'\mathbf{f}_{g,i} + \mathbf{u}_{g,i}$. We assume $\mathbf{f}_{g,i}$ and $\mathbf{u}_{g,i}$ are uncorrelated and satisfy $\mathbb{E}(\mathbf{f}_{g,i}) = \mathbf{0}$, $\operatorname{cov}(\mathbf{f}_{g,i}) = \mathbf{I}_{K_g \times K_g}$, $\mathbb{E}(\mathbf{u}_{g,i}) = \mathbf{0}$, and $\operatorname{cov}(\mathbf{u}_{g,i}) = \boldsymbol{\Sigma}_u$. Hence, for each sample in group g, we have $\operatorname{cov}(\mathbf{x}_{g,i}) = \boldsymbol{\Lambda}_g'\boldsymbol{\Lambda}_g + \boldsymbol{\Sigma}_u$, which is the sum of the group-specific low-rank matrix $\boldsymbol{\Lambda}_g'\boldsymbol{\Lambda}_g$ capturing the group-specific heterogeneity, and the matrix $\boldsymbol{\Sigma}_u$ that is homogeneous across different groups.

We adopt the approximate factor model (Stock and Watson (2002)) by assuming that $\boldsymbol{\Sigma}_u$ is sparse. Its sparsity can be characterized by m_p, defined as

$$m_p = \max_{i \leq p} \sum_{j=1}^{p} I(\sigma_{u,ij} \neq 0),$$

which is the maximum number of nonzero entries in the row of $\boldsymbol{\Sigma}_u$.

Under the decomposition (2.3), we have the following regression model for the gth group:

$$\boldsymbol{Y}_g = \boldsymbol{\mu}_g^* + \mathbf{F}_g \boldsymbol{\gamma}_g^* + \mathbf{U}_g \boldsymbol{\beta}^* + \boldsymbol{\epsilon}_g. \tag{2.4}$$

Here, $\boldsymbol{\mu}_g^*$ is the true group mean vector, $\boldsymbol{\gamma}_g^* \in \mathbb{R}^{K_g}$ denotes the true group-specific coefficients for $\mathbf{F}_g$, $\boldsymbol{\beta}^* \in \mathbb{R}^p$ denotes the common coefficients shared across G groups for $\mathbf{U}_g$, and $\boldsymbol{\epsilon}_g$ is the noise term and has variance σ^2. In (2.4), $\boldsymbol{\gamma}_g^*$ vary across G groups, and they characterize the heterogeneity induced by the factors in the regression model. Moreover, the group mean term $\boldsymbol{\mu}_g^*$ contributes to the heterogeneity in the regression model (2.4). When the heterogeneous effect is removed from (2.4), we have the same coefficients $\boldsymbol{\beta}^*$ for $\mathbf{U}_g$ across G groups.

From (2.4), we can see that the heterogeneity is modeled by $\boldsymbol{\mu}_g^* + \mathbf{F}_g \boldsymbol{\gamma}_g^*$. After adjusting this heterogeneous term, the remainder term $\mathbf{U}_g \boldsymbol{\beta}^*$ is homogeneous. Model (2.4) implies that, for the response $y_{g,i}$ of the ith subject in group g, we have $\mathrm{var}(y_{g,i}) = \boldsymbol{\gamma}_g^{*\prime} \boldsymbol{\gamma}_g^* + \boldsymbol{\beta}^{*\prime} \boldsymbol{\Sigma}_u \boldsymbol{\beta}^* + \sigma^2$. This decomposition shows that the variance can be decomposed as the sum of a group-specific part $\boldsymbol{\gamma}_g^{*\prime} \boldsymbol{\gamma}_g^*$, a homogeneous part $\boldsymbol{\beta}^{*\prime} \boldsymbol{\Sigma}_u \boldsymbol{\beta}^*$, and the background noise σ^2. This decomposition allows us to account for the heterogeneity among subgroups, while also borrowing information across subgroups to model homogeneous effects.

One special case of our proposed model (2.4) is when there is no group-specific factor; that is, $\mathbf{F}_g = \mathbf{0}$. Then, it reduces to a mean-specific model:

$$\boldsymbol{Y}_g = \boldsymbol{\mu}_g^* + \mathbf{X}_g \boldsymbol{\beta}^* + \boldsymbol{\epsilon}_g. \tag{2.5}$$

This model lies between the global model (2.1) and the group-specific model (2.2). It is different from (2.1) because it adjusts the group mean. It is different from (2.2) because different groups share the common regression coefficients. We refer to (2.5) as the "Factor-0" model.

3. Model Estimation and Prediction

In this section, we introduce the model estimation procedure and a data-driven way to estimate the factors in the testing data for prediction. The overall training procedure consists of two steps. First, we estimate the factors and homogeneous signals from the training data. Second, we estimate the regression coefficients using the estimated factors and homogeneous signals. In Section 3.1, we introduce how the factors can be estimated using a PCA. In Section 3.2, we introduce our procedure for estimating the model parameters. After the model is trained, in Section 3.3, we propose a data-driven procedure to estimate the factors in the testing data in order to make predictions.

3.1. Factor model estimation

For group g, the estimation of $\mathbf{F}_g$ and $\mathbf{\Lambda}_g$ can be formulated into the following optimization problem:

$$\begin{aligned} &\min_{\mathbf{F}_g,\mathbf{\Lambda}_g} \|\mathbf{X}_g - \mathbf{F}_g\mathbf{\Lambda}_g\|_F, \\ &\text{s.t. } \mathbf{F}_g'\mathbf{F}_g = n_g\mathbf{I}, \quad \mathbf{\Lambda}_g\mathbf{\Lambda}_g' \text{ is diagonal,} \end{aligned} \tag{3.1}$$

where $\|\cdot\|_F$ denotes the matrix Frobenius norm. The solution to (3.1) can be obtained by performing the eigendecomposition of the matrix $\mathbf{X}_g\mathbf{X}_g'$. Following the standard PCA procedure, we estimate $\mathbf{F}_g$ by $\hat{\mathbf{F}}_g$, where the kth column of $\hat{\mathbf{F}}_g$ is $\sqrt{n_g}$ times the eigenvector corresponding to the kth largest eigenvalue of $\mathbf{X}_g\mathbf{X}_g'$. Then, the loading matrix $\mathbf{\Lambda}_g$ can be estimated by regressing $\mathbf{X}_g$ on $\hat{\mathbf{F}}_g$ to obtain $\hat{\mathbf{\Lambda}}_g = \hat{\mathbf{F}}_g^T\mathbf{X}_g/n_g$. The homogeneous signal matrix $\mathbf{U}_g$ can hence be estimated by the residual matrix $\hat{\mathbf{U}}_g = \mathbf{X}_g - \hat{\mathbf{F}}_g\hat{\mathbf{\Lambda}}_g$.

We now consider estimating the number of factors K_g. In the literature, several estimators have been proposed to solve this problem (Bai and Ng (2002); Lam and Yao (2012); Ahn and Horenstein (2013)). We consider the following estimator:

$$\hat{K}_g = \underset{k\le K_{\max}}{\operatorname{argmax}} \frac{\lambda_k(\mathbf{X}_g\mathbf{X}_g')}{\lambda_{k+1}(\mathbf{X}_g\mathbf{X}_g')}, \tag{3.2}$$

where $\lambda_k(\cdot)$ denotes the kth largest eigenvalue (Lam and Yao (2012)). Here, $K_{\max}$ is a pre-determined upper bound for the number of factors. This estimator has been shown to be a consistent estimator (Ahn and Horenstein (2013)) for the true K_g, and is simple to implement in practice.

3.2. Estimation of regression coefficients

Given $\hat{\mathbf{F}}_g$ and $\hat{\mathbf{U}}_g$, as discussed in Section 3.1, we can estimate the model parameters μ_g^*, γ_g^*, and $\boldsymbol{\beta}^*$. The factor decomposition (2.3) projects the original signals onto the low-dimensional space spanned by $\mathbf{F}_g$ and the space spanned by $\mathbf{U}_g$, which is orthogonal to $\mathbf{F}_g$. Owing to the properties of an eigendecomposition, we have $\hat{\mathbf{F}}_g$ and $\hat{\mathbf{U}}_g$ orthogonal to each other. Hence, we can estimate the regression coefficients $\boldsymbol{\gamma}_g^*$ and $\boldsymbol{\beta}^*$ in (2.4) separately. Given $\hat{\mathbf{F}}_g$, μ_g^* and $\boldsymbol{\gamma}_g^*$ can be estimated by the following OLS estimators:

$$\hat{\mu}_g = \bar{Y}_g, \quad \hat{\gamma}_g = \frac{\hat{\mathbf{F}}_g^T \boldsymbol{Y}_g}{n_g}, \tag{3.3}$$

where $\bar{Y}_g$ denotes the sample mean of the response in group g.

Note that the factor matrix $\mathbf{F}_g$ and the coefficients γ_g^* are not separately identifiable, because for any orthogonal matrix $\mathbf{H}_g$, we have $\mathbf{F}_g\gamma_g^* = \mathbf{F}_g\mathbf{H}_g'\mathbf{H}_g\gamma_g^*$. Hence, $(\mathbf{F}_g, \gamma_g^*)$ cannot be identified from $(\mathbf{F}_g\mathbf{H}_g', \mathbf{H}_g\gamma_g^*)$. In practice, it does not matter which one is used, because the linear space spanned by the columns of $\mathbf{F}_g\mathbf{H}_g'$ is the same as that spanned by those of $\mathbf{F}_g$.

For homogeneous regression coefficients $\boldsymbol{\beta}^*$, because they are shared across groups, we can aggregate the residuals from the response and the factor projection to perform a global regression to estimate $\boldsymbol{\beta}^*$. Denote the aggregated residual vectors from the response as $\tilde{\boldsymbol{Y}} = (\tilde{\mathbf{Y}}_1', \ldots, \tilde{\boldsymbol{Y}}_G')'$, where $\tilde{\boldsymbol{Y}}_g = \boldsymbol{Y}_g - \hat{\boldsymbol{\mu}}_g - \hat{\mathbf{F}}_g\hat{\gamma}_g$. Let $\mathbf{U} = (\mathbf{U}_1', \ldots, \mathbf{U}_G')'$ and $\hat{\mathbf{U}} = (\hat{\mathbf{U}}_1', \ldots, \hat{\mathbf{U}}_G')'$. We solve the following penalized quadratic minimization problem to estimate $\boldsymbol{\beta}^*$:

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\left(\boldsymbol{\beta}'\hat{\boldsymbol{\Sigma}}_u\boldsymbol{\beta} - \frac{2}{n}\tilde{\boldsymbol{Y}}'\hat{\mathbf{U}}\boldsymbol{\beta}\right) + \lambda P(\boldsymbol{\beta}), \tag{3.4}$$

where $P(\boldsymbol{\beta})$ is a penalty function and λ is a tuning parameter, the optimal value of which is chosen using cross-validation. In particular, we consider an ℓ_1 penalty that $P(\boldsymbol{\beta}) = \sum_{j=1}^p |\beta_j|$ and an ℓ_2 penalty that $P(\boldsymbol{\beta}) = \sum_{j=1}^p \beta_j^2$, and denote the corresponding solutions of (3.4) as $\hat{\boldsymbol{\beta}}_\lambda^{lasso}$ and $\hat{\boldsymbol{\beta}}_\lambda^{ridge}$, respectively. In (3.4), $\hat{\boldsymbol{\Sigma}}_u$ is an estimator of $\boldsymbol{\Sigma}_u$. To obtain such an estimator, we use the adaptive thresholding method (Cai and Liu (2011)). More specifically, let $\hat{\sigma}_{ij} = (1/n)\sum_{g=1}^G \sum_{t=1}^{n_g} \hat{u}_{g,ti}\hat{u}_{g,tj}$ and $\hat{\theta}_{ij} = (1/n)\sum_{g=1}^G \sum_{t=1}^{n_g}(\hat{u}_{g,ti}\hat{u}_{g,tj} - \hat{\sigma}_{ij})^2$, where $\hat{u}_{g,ti}$ is the (t, i)th element of $\hat{\mathbf{U}}_g$. We have

$$\hat{\boldsymbol{\Sigma}}_u = (\hat{\sigma}_{ij}^{\mathcal{T}})_{p\times p}, \quad \hat{\sigma}_{ij}^{\mathcal{T}} = \begin{cases} \hat{\sigma}_{ii}, & i = j, \\ s_{ij}(\hat{\sigma}_{ij}), & i \neq j, \end{cases} \tag{3.5}$$

where $s_{ij}(\cdot)$ is any thresholding function that satisfies that for all $z \in \mathbb{R}$,

$$s_{ij}(z) = 0 \text{ when } |z| < \tau_{ij}, \text{ and } |s_{ij}(z) - z| \leq \tau_{ij} \text{ when } |z| \geq \tau_{ij}. \tag{3.6}$$

Here, $\tau_{ij} = D\omega_n\sqrt{\hat{\theta}_{ij}}$ is an adaptive threshold, where $\omega_n = 1/\sqrt{p} + \sqrt{\log p/n}$. The purpose of using such a thresholding estimator is to ensure $\mathbf{\Sigma}_u$ can be consistently estimated when $p > n$. In Section S3.1 of the Supplementary Material, we perform a sensitivity study on the choice of D, and find that the prediction performance of our method is not sensitive to D. Thus, we recommend choosing D to be a fixed number, rather than tuning it. When $p < n$, $\mathbf{\Sigma}_u$ does not have to be sparse. In this case, we find it is safe to choose $D = 0$; see Section S3.1 of the Supplementary Material.

We summarize the overall training procedure as follows:

1. For $g = 1, \ldots, G$:
   (a) Estimate K_g from (3.2).
   (b) Perform a PCA on $\mathbf{X}_g\mathbf{X}_g'$ to obtain $\hat{\mathbf{F}}_g$. Estimate μ_g^* and γ_g^* from (3.3).
   (c) Compute the projection matrix $\mathbf{P}_g = \hat{\mathbf{F}}_g\hat{\mathbf{F}}_g'/n_g$.
2. Let $\mathbf{H} = \text{diag}\{\mathbf{I} - \mathbf{P}_1, \ldots, \mathbf{I} - \mathbf{P}_G\}$ be the block diagonal matrix. Compute the aggregated signals $\hat{\mathbf{U}} = \mathbf{H}\mathbf{X}, \tilde{\boldsymbol{Y}} = \mathbf{H}(\boldsymbol{Y} - \hat{\boldsymbol{\mu}})$, where $\hat{\boldsymbol{\mu}} = (\hat{\boldsymbol{\mu}}_1', \ldots, \hat{\boldsymbol{\mu}}_G')'$. Estimate $\hat{\mathbf{\Sigma}}_u$ from $\hat{\mathbf{U}}$ using (3.5). Solve the optimization problem (3.4) to estimate $\boldsymbol{\beta}^*$.

In practice, it can be desirable to have an automatic way to choose between the proposed factor regression model (2.4) and the group-specific model (2.2). We provide an effective rule of thumb in Section S2 in the Supplementary Material.

3.3. Prediction

After training the model, in order to make predictions on the testing data, we need to estimate the factors and homogeneous signals in the testing data. In practice, they are not observable. We provide a data-driven procedure to estimate them based on the estimated loading matrix. Let $\mathbf{X}_{g,*} \in \mathbb{R}^{n_{g,*} \times p}$ denote the testing data matrix from group g. We aim to estimate the factor matrix $\mathbf{F}_{g,*} \in \mathbb{R}^{n_{g,*} \times K_g}$ and the homogeneous signal matrix $\mathbf{U}_{g,*} \in \mathbb{R}^{n_{g,*} \times p}$. Note that the number of columns in $\mathbf{F}_{g,*}$ is the same as that in $\mathbf{F}_g$.

Motivated by (3.1), we assume that the training and testing data from the same group have the same factor decomposition with the same loading matrix $\mathbf{\Lambda}_g$.

Hence, given $\hat{\mathbf{\Lambda}}_g$ from the training data, we propose estimating $\mathbf{F}_{g,*}$ by solving

$$\begin{aligned} &\min_{\mathbf{F}_{g,*}} \|\mathbf{X}_{g,*} - \mathbf{F}_{g,*}\hat{\mathbf{\Lambda}}_g\|_F, \\ &\text{s.t. } \mathbf{F}'_{g,*}\mathbf{F}_{g,*} = n_{g,*}\mathbf{I}. \end{aligned} \tag{3.7}$$

Note that (3.7) can be formulated as a trace maximization problem, the solution of which is given by $\hat{\mathbf{F}}_{g,*} = \sqrt{n_{g,*}}\tilde{\mathbf{V}}_g\tilde{\mathbf{U}}'_g$, where $\tilde{\mathbf{V}}_g$ and $\tilde{\mathbf{U}}_g$ come from a singular value decomposition with $\hat{\mathbf{\Lambda}}_g\mathbf{X}'_{g,*} = \tilde{\mathbf{U}}_g\mathbf{S}_g\tilde{\mathbf{V}}'_g$.

4. Theoretical Properties

We study the statistical properties of the proposed estimator. Without loss of generality, we assume that $\mu^*_g = 0$ for any $g \in \{1, \ldots, G\}$, so that (2.4) reduces to

$$\boldsymbol{Y}_g = \mathbf{F}_g\boldsymbol{\gamma}^*_g + \mathbf{U}_g\boldsymbol{\beta}^* + \boldsymbol{\epsilon}_g. \tag{4.1}$$

We establish the following theoretical results. First, we prove in Theorem 1 that the proposed estimators are consistent up to a rotation of the true parameters. As a corollary, we give an upper bound of the prediction error for the proposed method. Second, we show in Theorems 2 and 3 that if (4.1) is true, the group-specific model and the global model yield worse predictions than those of our proposed method. On the other hand, we show in Theorem 4 that even if one assumes each group has a distinct model, our method can have the same prediction error as the group-specific model when p is sufficiently large. Thus, our method is robust to model mis-specification.

First, we introduce some notation. For a matrix $\mathbf{A} \in \mathbb{R}^{p\times p}$, let $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ denote its minimum and maximum eigenvalues, respectively. Let $\|\mathbf{A}\|_F = \sqrt{\text{tr}(\mathbf{A}'\mathbf{A})}$, $\|\mathbf{A}\| = \lambda_{\max}(\mathbf{A}'\mathbf{A})$, $\|\mathbf{A}\|_1 = \max_{j\le p}\sum_{i=1}^p |a_{ij}|$, and $\|\mathbf{A}\|_{\max}$ $= \max_{i,j\le p}|a_{ij}|$ denote its Frobenius, ℓ_2, ℓ_1, and elementwise maximum norms, respectively. For a vector $\boldsymbol{b} \in \mathbb{R}^p$, let $\|\boldsymbol{b}\| = \sqrt{\sum_{j=1}^p b_j^2}$, $\|\boldsymbol{b}\|_1 = \sum_{j=1}^p |b_j|$, and $\|\boldsymbol{b}\|_\infty = \max_{j\le p}|b_j|$ denote its ℓ_2, ℓ_1, and maximum norms, respectively, and define its support as $\{j : b_j \neq 0\}$. Furthermore, we let $n_{\max} = \max_{g\le G} n_g$, $n = \sum_{g=1}^G n_g$, and $[m] = \{1, \ldots, m\}$ for a general positive integer m. In addition, we introduce the following definitions.

Definition 1. A vector $\boldsymbol{\beta} \in \mathbb{R}^p$ is called s-sparse if and only if its support's cardinality is at most s.

Definition 2 (RE Condition)**.** A matrix $\mathbf{\Sigma}$ is said to satisfy the restricted eigenvalue (RE) condition if and only if there exists a positive constant κ, such that

$\boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta} \geq \kappa\|\boldsymbol{\beta}\|^2$ for any $\boldsymbol{\beta} \in \mathbb{C}(S) = \{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta}_{S^c}\|_1 \leq 3\|\boldsymbol{\beta}_S\|_1\}$, where $S \subset [p]$ and S^c denotes its complement.

4.1. Consistency of the factor regression method

To establish the consistency of our proposed method, we need to impose the following conditions.

Assumption 1 (Pervasiveness)**.** *There exist positive constants $C_{\min}$ and $C_{\max} > 0$ such that, for any $g \in [G]$,*

$$C_{\min} < \lambda_{\min}(p^{-1}\boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g') < \lambda_{\max}(p^{-1}\boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g') < C_{\max}.$$

Assumption 2. *For any $g \in [G]$, assume that $\{\mathbf{f}_{g,i}\}_{i\leq n_g}$ and $\{\mathbf{u}_{g,i}\}_{i\leq n_g}$ are i.i.d. sub-Gaussian random variables with zero means and covariances $\mathbf{I}_{K_g\times K_g}$ and $\boldsymbol{\Sigma}_u$, respectively. More explicitly, assume for any $\boldsymbol{\alpha} \in \mathbb{R}^{K_g}$, $\boldsymbol{\gamma} \in \mathbb{R}^p$, and $s > 0$, there exists $C > 0$ such that $\mathbb{P}(|\boldsymbol{\alpha}'\mathbf{f}_{g,i}| > s) \leq \exp(-Cs^2/\|\boldsymbol{\alpha}\|^2)$ and $\mathbb{P}(|\boldsymbol{\gamma}'\mathbf{u}_{g,i}| > s) \leq \exp(-Cs^2/\|\boldsymbol{\gamma}\|^2)$. Morever, assume $\{\mathbf{f}_{g,i}\}_{i\leq n_g}$ are uncorrelated with $\{\mathbf{u}_{g,i}\}_{i\leq n_g}$.*

Assumption 3. *There exist positive constants c_1 and c_2 such that $\lambda_{\min}(\boldsymbol{\Sigma}_u) > c_1$ and $\|\boldsymbol{\Sigma}_u\|_1 < c_2$.*

Assumption 4. *For any $g \in [G]$, $j \in [p]$, and $i_1, i_2, i \in [n_g]$, there exists a positive constant M such that*

(a) $\|\boldsymbol{\lambda}_{g,j}\|_\infty < M$, where $\boldsymbol{\lambda}_{g,j}$ denotes the jth column of $\boldsymbol{\Lambda}_g$;

(b) $\mathbb{E}[p^{-1/2}\{\mathbf{u}_{g,i_1}'\mathbf{u}_{g,i_2} - \mathbb{E}(\mathbf{u}_{g,i_1}'\mathbf{u}_{g,i_2})\}]^4 < M$;

(c) $\mathbb{E}\|p^{-1/2}\sum_{j=1}^p \boldsymbol{\lambda}_{g,j}u_{g,ij}\|^4 < M$.

Assumption 1 is a typical pervasiveness assumption to ensure that the latent factors can be well estimated by the PCA method (Bai and Ng (2013); Fan, Liao and Mincheva (2013)). Such an assumption assumes that the latent factors affect a large proportion of variables, and is commonly used in the factor analysis literature. Assumption 2 is a typical sub-Gaussian assumption on the latent factors and the idiosyncratic components. Assumption 3 is a regularity condition on $\boldsymbol{\Sigma}_u$. Assumption 4 is a collection of technical conditions needed to establish the factor estimation consistency. Such conditions are commonly used in the factor analysis literature (Bai (2003); Bai and Ng (2008); Fan, Liao and Mincheva (2013)). Given these conditions, we show that under model (4.1), the proposed estimators are consistent.

Theorem 1. *Suppose Assumptions* 1–3 *hold,* $\log p = o(n^{2/39})$*,* $n = o(p^2)$*, and* $m_p\omega_n = o(1)$*. Then, it follows that*

(a) $\|\hat{\gamma}_g - \mathbf{H}_g\gamma_g^*\| = O_P(1/\sqrt{n_g} + 1/\sqrt{p})$*, where* $\hat{\gamma}_g$ *is as defined in* (3.3)*,* $\mathbf{H}_g = \hat{\mathbf{D}}_g^{-1}\hat{\mathbf{F}}_g'\mathbf{F}_g\mathbf{\Lambda}_g\mathbf{\Lambda}_g'$*, and* $\hat{\mathbf{D}}_g$ *is a* $\hat{K}_g \times \hat{K}_g$ *diagonal matrix consisting of the* $\hat{K}_g$ *largest eigenvalues of* $\mathbf{X}_g\mathbf{X}_g'$*.*

(b) In (3.4)*, if we choose an* ℓ_2 *penalty and* $\lambda = C\max\{n_{\max}^{3/4}/n, \sqrt{n_{\max}p}/n\}$*, for some large enough constant* C*, we have*

$$\|\hat{\boldsymbol{\beta}}_\lambda^{ridge} - \boldsymbol{\beta}^*\| = O_P\biggl(\frac{n_{\max}^{3/4}}{n} + \frac{\sqrt{n_{\max}p}}{n} + m_p\omega_n\frac{n_{\max}}{n}\biggr). \tag{4.2}$$

(c) Assuming that $\boldsymbol{\beta}^*$ *is* s*-sparse,* $\mathbf{\Sigma}_u$ *satisfies the RE condition, and* $s\omega_n = o(1)$*, if we choose an* ℓ_1 *penalty in* (3.4) *and* $\lambda = C\omega_n\bigl(m_p + \sqrt{n_{\max}/n}\bigr)$*, for some large enough constant* C*, then we have*

$$\|\hat{\boldsymbol{\beta}}_\lambda^{lasso} - \boldsymbol{\beta}^*\| = O_P\Biggl(\sqrt{s}\biggl(m_p\omega_n + \sqrt{\frac{n_{\max}}{n}}\omega_n\biggr)\Biggr). \tag{4.3}$$

Statement (a) shows that $\hat{\gamma}_g$ is consistent to γ_g^* up to a rotation given by $\mathbf{H}_g$. When the latent factors are known, the oracle convergence rate of $\hat{\gamma}_g$ is $O_P(1/\sqrt{n_g})$. Compared with this oracle rate, the extra term of $O_P(1/\sqrt{p})$ is essentially due to the estimation error of the latent factors; see Lemma 1 (a). When $p \gg n$, such a term is ignorable and the oracle rate can be attained. This is because, in that situation, many variables can be used to estimate the latent factors. The error in estimating the latent factors is so small that it does not affect the convergence rate of $\hat{\gamma}_g$. This is essentially due to the blessing of dimensionality property of a factor analysis, which has been studied in Li et al. (2018). Statements (b) and (c) show that the proposed penalized estimator in (3.4) is consistent to $\boldsymbol{\beta}^*$, regardless of whether an ℓ_1 or ℓ_2 penalty is imposed. To simplify the discussion, assume that $n_1 = \cdots = n_G$, m_p, and G are bounded. Then, the convergence rates in (4.2) and (4.3) reduce to

$$\|\hat{\boldsymbol{\beta}}_\lambda^{ridge} - \boldsymbol{\beta}^*\| = O_P\biggl(\frac{1}{n^{1/4}} + \sqrt{\frac{p}{n}}\biggr), \quad \|\hat{\boldsymbol{\beta}}_\lambda^{lasso} - \boldsymbol{\beta}^*\| = O_P\biggl(\sqrt{\frac{s}{p}} + \sqrt{\frac{s\log p}{n}}\biggr). \tag{4.4}$$

Hsu, Kakade and Zhang (2014) show that the minimax rate of a Ridge estimator in a linear regression model is $O_P(\sqrt{p/n})$ if no sparsity assumption is assumed. Compared with this minimax rate, our method has an extra term of $O_P(1/n^{1/4})$, which is again due to the error when estimating the latent factors; see Lemma

4. However, when $p \gg n$, such a term is ignorable and the minimax rate can be obtained. A similar conclusion can be drawn for the Lasso estimator. In (4.4), the term of $O_P(\sqrt{s \log p/n})$ agrees with the minimax rate of the standard Lasso problem (Raskutti, Wainwright and Yu (2011)). The extra term of $O_P(\sqrt{s/p})$ comes from the estimation error $\hat{\boldsymbol{\Sigma}}_u$; see Fan, Liao and Mincheva (2013). This term is ignorable when $p \gg n$, in which case the minimax rate is attained.

Let $\hat{\mathbf{Y}}_{g,\lambda}^{ridge} = \hat{\mathbf{F}}_g \hat{\gamma}_g + \hat{\mathbf{U}}_g \hat{\boldsymbol{\beta}}_{\lambda}^{ridge}$ and $\hat{\mathbf{Y}}_{g,\lambda}^{lasso} = \hat{\mathbf{F}}_g \hat{\gamma}_g + \hat{\mathbf{U}}_g \hat{\boldsymbol{\beta}}_{\lambda}^{lasso}$ denote the predicted values of $\boldsymbol{Y}_g$, where $\hat{\gamma}_g$ is given in (3.3), $\hat{\boldsymbol{\beta}}_{\lambda}^{ridge}$ and $\hat{\boldsymbol{\beta}}_{\lambda}^{lasso}$ are the Ridge and Lasso estimators, respectively, solved from (3.4), and $\hat{\mathbf{F}}_g$ and $\hat{\mathbf{U}}_g$ are as described in Section 3.1. The following corollary gives the upper bounds of the corresponding in-sample prediction errors.

Corollary 1. *Under the assumptions of Theorem* 1, *we have*

$$\left\| \frac{1}{n_g} \{ \hat{\mathbf{Y}}_{g,\lambda}^{ridge} - \mathbb{E}(\boldsymbol{Y}_g | \mathbf{F}_g, \mathbf{U}_g) \} \right\| =$$
$$O_P\left(\frac{n_{\max}^{3/4}}{n\sqrt{n_g}} + \frac{1}{n}\sqrt{\frac{n_{\max} p}{n_g}} + m_p \omega_n \frac{n_{\max}}{n\sqrt{n_g}} \right) + O_P\left(\frac{\sqrt{\log n_g \log p}}{n_g} + \frac{1}{n_g^{1/4}\sqrt{p}} \right), \tag{4.5}$$

$$\left\| \frac{1}{n_g} \{ \hat{\mathbf{Y}}_{g,\lambda}^{lasso} - \mathbb{E}(\boldsymbol{Y}_g | \mathbf{F}_g, \mathbf{U}_g) \} \right\| =$$
$$O_P\left(\sqrt{\frac{s}{n_g}} (m_p \omega_n + \sqrt{\frac{n_{\max}}{n}} \omega_n) \right) + O_P\left(\frac{\sqrt{\log n_g \log p}}{n_g} + \frac{1}{n_g^{1/4}\sqrt{p}} \right). \tag{4.6}$$

Again, if we assume $n_1 = \cdots = n_G$, m_p, and G are bounded, these results reduce to

$$\left\| \frac{1}{n_g} \{ \hat{\mathbf{Y}}_{g,\lambda}^{ridge} - \mathbb{E}(\boldsymbol{Y}_g | \mathbf{F}_g, \mathbf{U}_g) \} \right\| = O_P\left(\frac{1}{n^{1/4}\sqrt{p}} + \frac{\sqrt{p}}{n} \right), \tag{4.7}$$

$$\left\| \frac{1}{n_g} \{ \hat{\mathbf{Y}}_{g,\lambda}^{lasso} - \mathbb{E}(\boldsymbol{Y}_g | \mathbf{F}_g, \mathbf{U}_g) \} \right\| \tag{4.8}$$
$$= O_P\left(\frac{1}{n^{1/4}\sqrt{p}} + \sqrt{\frac{s}{np}} + \frac{\sqrt{\log n \log p}}{n} + \frac{\sqrt{s \log p}}{n} \right).$$

In (4.7), the term of $O_P(\sqrt{p}/n)$ agrees with the minimax rate of the prediction error given by the Ridge estimator in a standard linear regression problem (Dicker (2016); Dobriban and Wager (2018)). In (4.8), the term of $O_P(\sqrt{s \log p}/n)$ agrees with the prediction error given by the Lasso estimator in the standard setting (Bickel, Ritov and Tsybakov (2009)). All other terms are ignorable when $p \gg n$.

In conclusion, these results show that our proposed estimators can have the same convergence rates as the Ridge and Lasso estimators have under the standard homogeneous linear regression model, which is simpler than the heterogeneous model we have considered.

4.2. Consistency of group-specific and global models

We study the statistical properties of the group-specific and global models when the underlying model follows (4.1). We show that, in this case, our proposed method has an advantage over these two models in terms of the prediction error. We rewrite (4.1) as

$$\boldsymbol{Y}_g = \tilde{\mathbf{X}}_g\boldsymbol{\beta}^* + \mathbf{F}_g\boldsymbol{\delta}_g + d_p\mathbf{U}_g\boldsymbol{\beta}^* + \boldsymbol{\epsilon}_g, \tag{4.9}$$

where $\tilde{\mathbf{X}}_g = p^{-1/2}\mathbf{X}_g$, $\boldsymbol{\delta}_g = \boldsymbol{\gamma}_g^* - p^{-1/2}\boldsymbol{\Lambda}_g\boldsymbol{\beta}^*$, and $d_p = 1 - p^{-1/2}$. Here, we standardize $\mathbf{X}_g$ by dividing it by $p^{1/2}$. This is because the pervasiveness assumption means that $\|\mathbf{X}_g\|$ is unbounded, which is different from the typical linear regression model. Therefore, we rescale it to be $\tilde{\mathbf{X}}_g$. Then, the group-specific model seeks to solve

$$\hat{\boldsymbol{\beta}}_{g,\lambda} = \operatorname*{argmin}_{\boldsymbol{\beta}} \frac{1}{2n_g}\|\boldsymbol{Y}_g - \tilde{\mathbf{X}}_g\boldsymbol{\beta}\|^2 + \lambda P(\boldsymbol{\beta}), \tag{4.10}$$

whereas the global model seeks to solve

$$\hat{\boldsymbol{\beta}}_{\lambda,global} = \operatorname*{argmin}_{\boldsymbol{\beta}} \frac{1}{2n}\|\boldsymbol{Y} - \tilde{\mathbf{X}}\boldsymbol{\beta}\|^2 + \lambda P(\boldsymbol{\beta}), \tag{4.11}$$

where $\tilde{\mathbf{X}} = (\tilde{\mathbf{X}}_1', \ldots, \tilde{\mathbf{X}}_G')'$, λ is a tuning parameter and $P(\boldsymbol{\beta})$ is a general penalty function. Similar to (3.4), we choose either an ℓ_1 or an ℓ_2 penalty, and denote the corresponding solutions as $\hat{\boldsymbol{\beta}}_{g,\lambda}^{lasso}$, $\hat{\boldsymbol{\beta}}_{\lambda,global}^{lasso}$ and $\hat{\boldsymbol{\beta}}_{g,\lambda}^{ridge}$, $\hat{\boldsymbol{\beta}}_{\lambda,global}^{ridge}$ respectively. Next, we give the convergence rates of the estimators in the group-specific and global models in Theorems 2 and 3, respectively.

Theorem 2. *Suppose Assumptions* 1–3 *hold and* $\log p = o(n)$. *Then, it follows that*

(a) If we use an ℓ_2 *penalty in* (4.10) *and choose* $\lambda = C/\sqrt{p}$, *for some large enough constant* C, *we have*

$$\|\hat{\boldsymbol{\beta}}_{g,\lambda}^{ridge} - \boldsymbol{\beta}^*\| = O_P\left(\sqrt{p}\|\boldsymbol{\delta}_g\| + d_p\left(1 + \sqrt{\frac{p}{n_g}}\right) + \sqrt{\frac{p}{n_g}}\right). \tag{4.12}$$

(b) Assuming that $\boldsymbol{\beta}^*$ *is s-sparse,* $\boldsymbol{\Lambda}_g'\boldsymbol{\Lambda}_g/\sqrt{p}$ *satisfies the RE condition, and* $s\sqrt{\log p/(n_g p)} = o(1)$, *if we use an* ℓ_1 *penalty in* (4.10) *and choose* $\lambda =$

$C\{(1+\sqrt{\log p/n_g})(d_p+\|\boldsymbol{\delta}_g\|)+\sqrt{\log p/n_g}\}/\sqrt{p}$, for some large enough constant C, we have

$$\|\hat{\boldsymbol{\beta}}^{lasso}_{g,\lambda}-\boldsymbol{\beta}^*\| = O_P\biggl(\sqrt{s}\biggl\{\biggl(1+\sqrt{\frac{\log p}{n_g}}\biggr)(d_p+\|\boldsymbol{\delta}_g\|)+\sqrt{\frac{\log p}{n_g}}\biggr\}\biggr). \quad (4.13)$$

Let $\hat{\mathbf{Y}}^{ridge}_{g,\lambda}=\tilde{\mathbf{X}}_g\hat{\boldsymbol{\beta}}^{ridge}_{g,\lambda}$ and $\hat{\boldsymbol{Y}}^{lasso}_{g,\lambda}=\tilde{\mathbf{X}}_g\hat{\boldsymbol{\beta}}^{lasso}_{g,\lambda}$ be the predicted values of $\boldsymbol{Y}_g$, where $\hat{\boldsymbol{\beta}}^{ridge}_{g,\lambda}$ and $\hat{\boldsymbol{\beta}}^{lasso}_{g,\lambda}$ are the Ridge and Lasso solutions, respectively, to (4.10). We have the following upper bounds of their in-sample prediction errors.

Corollary 2. *Under the assumptions of Theorem 2, we have*

$$\begin{aligned}&\|\frac{1}{n_g}\{\hat{\boldsymbol{Y}}^{ridge}_{g,\lambda}-\mathbb{E}(\mathbf{Y}_g|\mathbf{F}_g,\mathbf{U}_g)\}\|\\ &=O_P\biggl(\sqrt{\frac{p}{n_g}}\|\boldsymbol{\delta}_g\|+d_p\biggl(\frac{1}{\sqrt{n_g}}+\frac{\sqrt{p}}{n_g}\biggr)+\frac{\sqrt{p}}{n_g}\biggr), \quad (4.14)\\ &\|\frac{1}{n_g}\{\hat{\boldsymbol{Y}}^{lasso}_{g,\lambda}-\mathbb{E}(\mathbf{Y}_g|\mathbf{F}_g,\mathbf{U}_g)\}\|\\ &=O_P\biggl(\sqrt{\frac{s}{n_g}}\{(1+\sqrt{\frac{\log p}{n_g}})(d_p+\|\boldsymbol{\delta}_g\|)+\sqrt{\frac{\log p}{n_g}}\biggr\}\biggr). \quad (4.15)\end{aligned}$$

Theorem 3. *Suppose Assumptions* 1–3 *hold and* $\log p = o(n)$*. Then, it follows that*

(a) If we use an ℓ_2 penalty in (4.11) *and choose $\lambda = C/\sqrt{p}$, for some large enough constant C, we have*

$$\|\hat{\boldsymbol{\beta}}^{ridge}_{\lambda,global}-\boldsymbol{\beta}^*\| = O_P\biggl(\frac{n_{\max}\sqrt{p}}{n}\sum_{g=1}^{G}\|\boldsymbol{\delta}_g\|+d_p\biggl(\frac{n_{\max}}{n}+\frac{\sqrt{n_{\max}p}}{n}\biggr)+\frac{\sqrt{n_{\max}p}}{n}\biggr). \quad (4.16)$$

(b) Assuming that $\boldsymbol{\beta}^*$ is s-sparse, $\boldsymbol{\Lambda}_g'\boldsymbol{\Lambda}_g/\sqrt{p}$ satisfies the RE condition, and $s\sqrt{\log p/(n_g p)} = o(1)$ for any $g \in [G]$, if we use an ℓ_1 penalty in (4.11) *and choose $\lambda = C\bigl[\{n_{\max}/(n\sqrt{p})+(1/n)\sqrt{n_{\max}\log p/p}\}\bigl(d_p+\sum_{g=1}^{G}\|\boldsymbol{\delta}_g\|\bigr)+(1/n)\sqrt{n_{\max}\log p/p}\bigr]$, for some large enough constant C, we have*

$$\begin{aligned}&\|\hat{\boldsymbol{\beta}}^{lasso}_{\lambda,global}-\boldsymbol{\beta}^*\| \quad (4.17)\\ &=O_P\biggl(\sqrt{s}\biggl\{\biggl(\frac{n_{\max}}{n}+\frac{\sqrt{n_{\max}\log p}}{n}\biggr)\biggl(d_p+\sum_{g=1}^{G}\|\boldsymbol{\delta}_g\|\biggr)+\frac{\sqrt{n_{\max}\log p}}{n}\biggr\}\biggr).\end{aligned}$$

Let $\hat{\boldsymbol{Y}}^{ridge}_{g,\lambda} = \tilde{\mathbf{X}}_g \hat{\boldsymbol{\beta}}^{ridge}_{\lambda,global}$ and $\hat{\boldsymbol{Y}}^{lasso}_{g,\lambda} = \tilde{\mathbf{X}}_g \hat{\boldsymbol{\beta}}^{lasso}_{\lambda,global}$ be the predicted values of $\boldsymbol{Y}_g$, where $\hat{\boldsymbol{\beta}}^{ridge}_{\lambda,global}$ and $\hat{\boldsymbol{\beta}}^{lasso}_{\lambda,global}$ are the Ridge and Lasso solutions, respectively, (4.11). We have the following upper bounds for their in-sample prediction errors.

Corollary 3. *Under the assumptions of Theorem* 3, *we have*

$$\left\| \frac{1}{n_g} \{ \hat{\boldsymbol{Y}}^{ridge}_{g,\lambda} - \mathbb{E}(\boldsymbol{Y}_g | \mathbf{F}_g, \mathbf{U}_g) \} \right\| \tag{4.18}$$

$$= O_P\left(\frac{n_{\max}}{n} \sqrt{\frac{p}{n_g}} \sum_{g'=1}^{G} \|\boldsymbol{\delta}_{g'}\| + \frac{1}{n} \sqrt{\frac{n_{\max} p}{n_g}} + d_p \left(\frac{n_{\max}}{n\sqrt{n_g}} + \frac{1}{n} \sqrt{\frac{n_{\max} p}{n_g}} \right) \right),$$

$$\left\| \frac{1}{n_g} \left\{ \hat{\boldsymbol{Y}}^{lasso}_{g,\lambda} - \mathbb{E}(\boldsymbol{Y}_g | \mathbf{F}_g, \mathbf{U}_g) \right\} \right\| \tag{4.19}$$

$$= O_P\left(\sqrt{\frac{s}{n_g}} \left\{ \frac{\sqrt{n_{\max} \log p}}{n} + \left(\frac{n_{\max}}{n} + \frac{\sqrt{n_{\max} \log p}}{n} \right) \left(d_p + \sum_{g'=1}^{G} \|\boldsymbol{\delta}_{g'}\| \right) \right\} \right).$$

Under (4.1), $\|\boldsymbol{\delta}_g\| \le \|\boldsymbol{\gamma}^*_g\| + p^{-1/2} \|\boldsymbol{\Lambda}_g\| \|\boldsymbol{\beta}^*\| = O(1)$, for all $g \in [G]$ and $d_p = O(1)$. Thus, if we assume that $n_1 = \cdots = n_G$ and G is bounded, then (4.14) and (4.18) further reduce to $\|(1/n_g)\{\hat{\mathbf{Y}}^{ridge}_{g,\lambda} - \mathbb{E}(\boldsymbol{Y}_g | \mathbf{F}_g, \mathbf{U}_g)\}\| = O_P(\sqrt{p/n})$ for the Ridge estimator. Compared with the predictor error of our Ridge estimator, which is $O_P(\sqrt{p}/n)$, these two methods are worse by a factor of $\sqrt{n}$, owing to the mis-specified model (4.1). Similarly for the Lasso estimator, when $n_1 = \cdots = n_G$ and G is bounded, (4.15) and (4.19) reduce to $\|(1/n_g)\{\hat{\mathbf{Y}}^{lasso}_{g,\lambda} - \mathbb{E}(\boldsymbol{Y}_g | \mathbf{F}_g, \mathbf{U}_g)\}\| = O_P(\sqrt{s/n} + \sqrt{s \log p}/n)$. Compared with our Lasso estimator, they have an extra term of $\sqrt{s/n}$, which also comes from the model mis-specification and is non-ignorable.

4.3. Robustness

In this section, we assume each group follows a distinct model

$$\boldsymbol{Y}_g = \tilde{\mathbf{X}}_g \boldsymbol{\beta}^*_g + \boldsymbol{\epsilon}_g, \tag{4.20}$$

and examine how well our method performs under this model assumption. In other words, we study how robust our method is under model mis-specification. Here, we still use the rescaled $\tilde{\mathbf{X}}_g$ as the design matrix. We rewrite (4.20) as $\boldsymbol{Y}_g = p^{-1/2} \mathbf{F}_g \boldsymbol{\Lambda}_g \boldsymbol{\beta}^*_g + p^{-1/2} \mathbf{U}_g \boldsymbol{\beta}^*_g + \boldsymbol{\epsilon}_g$. Compared with (4.1), we see that $p^{-1/2} \boldsymbol{\Lambda}_g \boldsymbol{\beta}^*_g$ and $p^{-1/2} \boldsymbol{\beta}^*_g$ can be viewed as $\boldsymbol{\gamma}^*_g$ and $\boldsymbol{\beta}^*$, respectively, in our model. Under the model assumption in (4.20), we have the following results.

Theorem 4. *Suppose Assumptions* 1–3 *hold,* $\log p = o(n^{2/39})$, $n = o(p^2)$, *and*

$m_p\omega_n = o(1)$. Then, for any $g \in [G]$, it follows that

(a) $\|\hat{\gamma}_g - p^{-1/2}\mathbf{H}_g\mathbf{\Lambda}_g\boldsymbol{\beta}_g^*\| = O_P(1/\sqrt{n_g} + 1/\sqrt{p})$, where $\mathbf{H}_g$ is as defined in Theorem 1.

(b) If an ℓ_2 penalty in (3.4) is used and $\lambda = O(\max\{n_{\max}^{3/4}\sqrt{p}/n, \sqrt{n_{\max}}p/n\})$, then

$$\left\|\hat{\boldsymbol{\beta}}_\lambda^{ridge} - \frac{1}{\sqrt{p}}\boldsymbol{\beta}_g^*\right\| = O_P\left(\frac{\sqrt{n_{\max}p}}{n} + \frac{n_{\max}^{3/4}}{n}\right) + \sum_{g'=1}^{G} O_P\left(\frac{n_{g'}}{n\sqrt{p}}\|\boldsymbol{\beta}_{g'}^* - \boldsymbol{\beta}_g^*\|\right).$$

(c) Assuming that $\boldsymbol{\beta}_g^*$ is s-sparse and $\mathbf{\Sigma}_u$ satisfies the RE condition, if we use an ℓ_1 penalty in (3.4) and choose $\lambda = C\{\omega_n\sqrt{n_{\max}/n}+n_{\max}/(n\sqrt{p})\sum_{g'=1}^{G}\|\boldsymbol{\beta}_g^*-\boldsymbol{\beta}_{g'}^*\|\}$, for some large enough constant C, we have

$$\left\|\hat{\boldsymbol{\beta}}_\lambda^{lasso} - \frac{1}{\sqrt{p}}\boldsymbol{\beta}_g^*\right\| = O_P\left(\sqrt{s}\left(\sqrt{\frac{n_{\max}}{n}}\omega_n + \frac{n_{\max}}{n\sqrt{p}}\sum_{g'=1}^{G}\|\boldsymbol{\beta}_g^* - \boldsymbol{\beta}_{g'}^*\|\right)\right).$$

Let $\hat{\boldsymbol{Y}}_{g,\lambda}^{ridge}$ and $\hat{\boldsymbol{Y}}_{g,\lambda}^{lasso}$ be the same as in Corollary 1. Using Theorem 4, we give the upper bounds of the in-sample prediction errors given by our proposed method, when the underlying model follows (4.20).

Corollary 4. *Under the assumptions of Theorem 4, for each $g \in [G]$, we have*

$$\begin{aligned}&\left\|\frac{1}{n_g}\{\hat{\boldsymbol{Y}}_{g,\lambda}^{ridge} - \mathbb{E}(\boldsymbol{Y}_g|\tilde{\mathbf{X}}_g)\}\right\| \\ &= O_P\left(\frac{1}{n_g}\right) + O_P\left(\frac{1}{\sqrt{n_g p}}\right) + O_P\left(\frac{1}{\sqrt{n_g}}\|\hat{\boldsymbol{\beta}}_\lambda^{ridge} - \frac{1}{\sqrt{p}}\boldsymbol{\beta}_g^*\|\right),\end{aligned} \tag{4.21}$$

$$\begin{aligned}&\left\|\frac{1}{n_g}\{\hat{\boldsymbol{Y}}_{g,\lambda}^{lasso} - \mathbb{E}(\boldsymbol{Y}_g|\tilde{\mathbf{X}}_g)\}\right\| \\ &= O_P\left(\frac{1}{n_g}\right) + O_P\left(\frac{1}{\sqrt{n_g p}}\right) + O_P\left(\frac{1}{\sqrt{n_g}}\|\hat{\boldsymbol{\beta}}_\lambda^{lasso} - \frac{1}{\sqrt{p}}\boldsymbol{\beta}_g^*\|\right).\end{aligned} \tag{4.22}$$

When $n_1 = \cdots = n_G$ and G is bounded, (4.21) and (4.22) further reduces to

$$\begin{aligned}&\left\|\frac{1}{n_g}\{\hat{\boldsymbol{Y}}_{g,\lambda}^{ridge} - \mathbb{E}(\boldsymbol{Y}_g|\tilde{\mathbf{X}}_g)\}\right\| \\ &= O_P\left(\sum_{g'=1}^{G}\frac{1}{\sqrt{np}}\|\boldsymbol{\beta}_g^* - \boldsymbol{\beta}_{g'}^*\|\right) + O_P\left(\frac{\sqrt{p}}{n}\right) = O_P\left(\frac{\sqrt{p}}{n}\right),\end{aligned} \tag{4.23}$$

$$\left\|\frac{1}{n_g}\{\hat{\boldsymbol{Y}}_{g,\lambda}^{lasso} - \mathbb{E}(\boldsymbol{Y}_g|\tilde{\mathbf{X}}_g)\}\right\|$$

$$= O_P\left(\sum_{g'=1}^{G}\sqrt{\frac{s}{np}}\|\boldsymbol{\beta}_g^* - \boldsymbol{\beta}_{g'}^*\|\right) + O_P\left(\sqrt{\frac{s}{np}} + \frac{\sqrt{s\log p}}{n}\right). \tag{4.24}$$

We compare these convergence rates with those given by the group-specific model. Because the true model (4.20) is a special case of (4.9), by treating $d_p = 0$ and $\boldsymbol{\delta}_g = \mathbf{0}$, it follows from Theorem 2 that the prediction errors of the group-specific model are $O_P(\sqrt{p}/n_g)$ and $O_P(\sqrt{s\log p}/n_g)$, when using a Ridge or a Lasso estimator, respectively. Comparing then with (4.23) and (4.24), we find that the Ridge estimator of our model has the same rate as the group-specific Ridge estimators; see (4.23). For the Lasso estimator, when p is small, our model converges at a rate of $\sqrt{s/(np)}$, which is slower than that of the group-specific model by a factor of $\sqrt{n/(p\log p)}$. The reason is that our model estimates $G^{-1}\sum_{g'=1}^{G}\boldsymbol{\beta}_{g'}^*$, instead of $\boldsymbol{\beta}_g^*$, and needs to estimate $\boldsymbol{\Sigma}_u$, which introduces an extra error of $O_P(\sqrt{s/(np)})$. However, when $p \gg n$, all these terms are negligible, and our model has the same convergence as the group-specific model. In conclusion, we have shown that even if the true model is group-specific, our method still provides comparable prediction to that of the group-specific model, especially when the dimension p is high.

5. Simulation Studies

In this section, we perform two simulation studies to compare our proposed model with the global, group-specific, and Factor-0 models. In both studies, we choose $G = 3, p = 200, K_g = 3$, and $n_g = 100$, for any $g \in [G]$, generate 600 training samples to train all four models, and evaluate their mean squared error (MSE) in an independent test set of 600 samples. Additional simulation studies on other choices of K_g can be found in Section S3.4 in the Supplementary Material. We repeat the simulations 50 times. In setting 1, we generate data from our proposed model. In setting 2, we generate different models for different groups.

5.1. Setting 1: under proposed model

We first generate data from the proposed model in (2.4). For any $g \in [G]$, we generate $\{\mathbf{f}_{g,i}\}_{i\le n_g}$ as i.i.d. samples from $\mathcal{N}(\mathbf{0}, \mathbf{I}_{K_g\times K_g})$. We set

$$\boldsymbol{\Lambda}_g = \begin{bmatrix} \boldsymbol{\Lambda}_g^{1\prime}\boldsymbol{\Lambda}_g^1 \ \boldsymbol{\Lambda}_g^{1\prime}\boldsymbol{\Lambda}_g^2 \\ \boldsymbol{\Lambda}_g^{2\prime}\boldsymbol{\Lambda}_g^1 \ \boldsymbol{\Lambda}_g^{2\prime}\boldsymbol{\Lambda}_g^2 \end{bmatrix}.$$

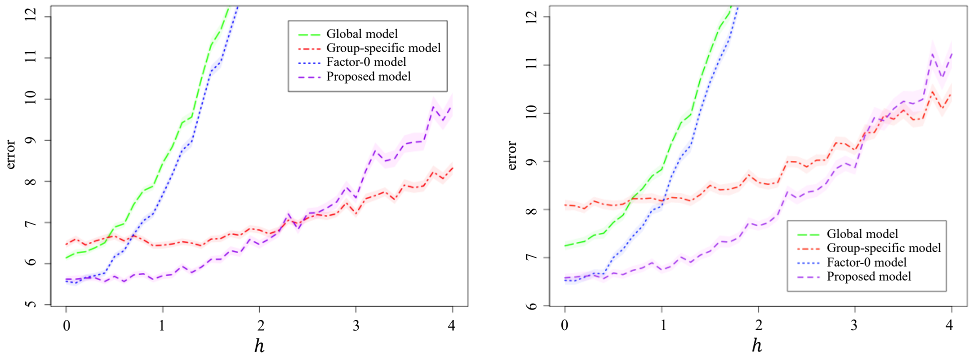

Figure 1. The MSE curves given by the four models. The left panel represents the results for a sparse $\boldsymbol{\beta}^*$, and the right panel represents the results for a dense $\boldsymbol{\beta}^*$.

To ensure $\boldsymbol{\Lambda}_g$ satisfies the pervasiveness assumption (Assumption 1), we first choose a positive-definite matrix $\mathbf{R} * \mathbf{s}_g\mathbf{s}_g'$, where $\mathbf{R} = (r_{ij})$ with $r_{ij} = 0.1^{|i-j|}$, $\mathbf{s}_g = (\sqrt{\lambda_{g,1}}, \ldots, \sqrt{\lambda_{g,K_g}})'$, $(\lambda_{1,1}, \lambda_{1,2}, \lambda_{1,3}) = (7.0, 3.5, 1.2)$, $(\lambda_{2,1}, \lambda_{2,2}, \lambda_{2,3}) = (10, 3.9, 1.2)$, $(\lambda_{3,1}, \lambda_{3,2}, \lambda_{3,3}) = (13, 3.9, 1.1)$, and $*$ denotes elementwise matrix multiplication. Additional simulation studies on other choices of $\lambda_{g,1}, \ldots, \lambda_{g,K_g}$ can be found in Section S3.2 in the Supplementary Material. Then, we perform an eigendecomposition on it to obtain $\mathbf{R} * \mathbf{s}_g\mathbf{s}_g' = \mathbf{V}_g\mathbf{D}_g\mathbf{V}_g'$, where $\mathbf{D}_g$ is the diagonal matrix consisting of its eigenvalues. Next, we set $\boldsymbol{\Lambda}_g^1 = \mathbf{Q}_g\mathbf{D}_g^{1/2}\mathbf{V}_g'$, where $\mathbf{Q}_g$ is a random orthonormal matrix, and $\boldsymbol{\Lambda}_g^2 = \mathbf{Q}_g\boldsymbol{T}_g$, where $\boldsymbol{T}_g$ is a $K_g \times (p - K_g)$ matrix with elements randomly generated from Unif$(-1/20, 1/20)$. This construction of $\boldsymbol{\Lambda}_g$ ensures that it has spiked eigenvalues, as required by the pervasiveness assumption, and its rank is K_g. We further generate $\{\mathbf{u}_{g,i}\}_{i \le n_g}$ as i.i.d. samples from $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_u)$, where $\boldsymbol{\Sigma}_u$ is a diagonal matrix with diagonal elements all equal to 0.03. For the coefficients in (2.4), we choose $\mu_g^* = g$ for $g = 1, 2, 3$. We set $\gamma_1^* = (h, h, 2h)'$, $\gamma_2^* = (h, 2h, h)'$, and $\gamma_3^* = (2h, h, h)'$, where we let h change so that, as it increases, the between-group heterogeneity increases accordingly. We consider two settings of $\boldsymbol{\beta}^*$. For a sparse $\boldsymbol{\beta}^*$, we set $\boldsymbol{\beta}^* = (\mathbf{2}_{10}, \mathbf{0}_{90}, -\mathbf{2}_{10}, \mathbf{0}_{90})'$, where $\boldsymbol{m}_L$ denotes an L-dimensional vector with elements all equal to m; for a dense $\boldsymbol{\beta}^*$, we set $\boldsymbol{\beta}^* = (\mathbf{1}_{80}, \mathbf{0}_{20}, -\mathbf{1}_{80}, \mathbf{0}_{20})'$. Finally, we generate the error term $\boldsymbol{\epsilon}$ as i.i.d samples from $\mathcal{N}(0, 4)$.

Under this model generation scheme, Figure 1 shows how the MSEs of these four methods change as h varies. When $\boldsymbol{\beta}^*$ is sparse, all methods use an ℓ_1 penalty; when $\boldsymbol{\beta}^*$ is dense, all methods use an ℓ_2 penalty. The shaded areas represent the standard errors of the MSEs in the 50 simulations. The optimal tuning parameters in these methods are chosen using 10-fold cross-validation. It is clearly seen that for most h, our model performs best. Owing to the model mis-specification, the group-specific model loses some efficiency in estimating the homogeneous part of (4.9) separately, and the global model entirely ignores the heterogeneity. The Factor-0 model adjusts for group means; therefore, it is better than the global model. However, it is still worse than the proposed full model, indicating that some additional heterogeneity has not been fully taken into account in the Factor-0 model. When h increases, the true model (2.4) becomes more group-specific, and less homogeneity can be used to estimate the common $\boldsymbol{\beta}^*$. In this case, the group-specific model gradually outperforms our method. They both become much better than the global and the Factor-0 models. The estimation errors on γ_g^* and $\boldsymbol{\beta}^*$ are reported in Tables S2 and S3 in the Supplementary Material.

5.2. Setting 2: under group-specific model

We generate different models for different groups and inspect how robust our model is under such a scenario. We generate $\mathbf{f}_{g,i}$ as we did in the first study and $\mathbf{u}_{g,i}$ as i.i.d samples from $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_u)$, where $\boldsymbol{\Sigma}_u = (\sigma_{u,ij})$, with $\sigma_{u,ij} = 0.1^{|i-j|}0.03$ if $|i-j| \leq 2$, and $\sigma_{u,ij} = 0$ otherwise. Additional simulation studies on $\{\mathbf{f}_{g,i}\}_{i \leq n_g}$ and $\{\mathbf{u}_{g,i}\}_{i \leq n_g}$ generated from more general sub-Gaussian distributions for both settings can be found in Section S3.3 in the Supplementary Material. For $\boldsymbol{\Lambda}_g$, we set $\boldsymbol{\Lambda}_g = \tilde{\mathbf{Q}}_g * \mathbf{s}_g$, where $\mathbf{s}_g$ is as in the first study, and $\tilde{\mathbf{Q}}_g$ is a random $K_g \times p$ orthonormal matrix. Then, we use these elements to generate $\boldsymbol{X}_g$ according to (2.3) and normalize it to obtain the design matrix $\tilde{\boldsymbol{X}}_g$. Given $\tilde{\boldsymbol{X}}_g$, for any $g \in [G]$, we generate $\boldsymbol{Y}_g$ from (4.20) by setting $\mu_g = g$ for $g \in [G]$, generating $\boldsymbol{\epsilon}$ as i.i.d. samples from $N(0, 4)$, and choosing two kinds of $\boldsymbol{\beta}_g^*$. For sparse $\boldsymbol{\beta}_g^*$, we set $\boldsymbol{\beta}_1^* = (10h, 10h, -10h, \mathbf{10}_5, \mathbf{0}_{187}, \mathbf{10}_5)$, $\boldsymbol{\beta}_2^* = (10h, -10h, 10h, \mathbf{10}_5, \mathbf{0}_{187}, \mathbf{10}_5)$, and $\boldsymbol{\beta}_3^* = (-10h, 10h, 10h, \mathbf{10}_5, \mathbf{0}_{187}, \mathbf{10}_5)$. For dense $\boldsymbol{\beta}_g^*$, we set $\boldsymbol{\beta}_1^* = (10h, 10h, -10h, \mathbf{1}_{80}, \mathbf{0}_{37}, \mathbf{1}_{80})$, $\boldsymbol{\beta}_2^* = (10h, -10h, 10h, \mathbf{1}_{80}, \mathbf{0}_{37}, \mathbf{1}_{80})$, and $\boldsymbol{\beta}_3^* = (-10h, 10h, 10h, \mathbf{1}_{80}, \mathbf{0}_{37}, \mathbf{1}_{80})$.

Under this model generation scheme, Figure 2 shows the MSE curves of the four methods, which are computed the same way as in the first study. For sparse $\boldsymbol{\beta}_g^*$, when h is small, the differences between the group-specific, the Factor-0, and our method are marginal, which agrees with what we proved in Corollary 4. When h gets larger, the group difference dominates. In this case, the group-specific model gives the best prediction, although our model is not far off. Compared with these two models, the global and the Factor-0 models are much worse because they fail to recognize the group difference. For a dense $\boldsymbol{\beta}^*$, when h is small,

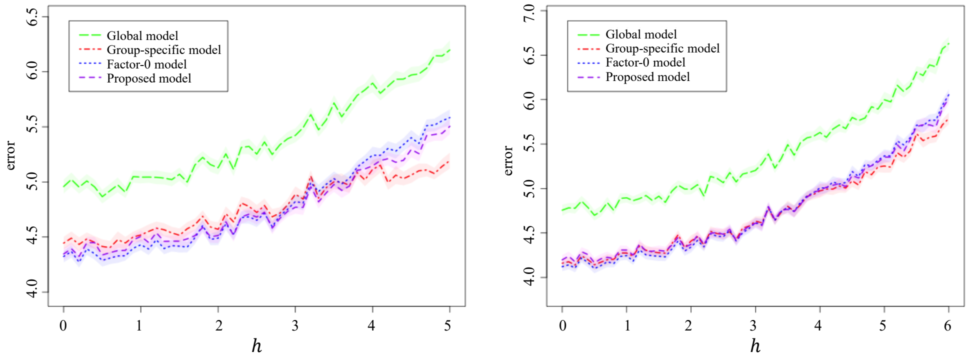

Figure 2. The MSE curves given by the four models. The left panel represents the results for sparse $\boldsymbol{\beta}_g^*$, and the right panel represents the results for dense $\boldsymbol{\beta}_g^*$.

all other models have similar performance, except for the global model. As h gets larger, our model becomes slightly worse than the group-specific model for the same reason discussed in the sparse case. However, the performance of the Factor-0 model deteriorates much faster. In conclusion, this study shows that our method's performance is still acceptable, even when the underlying models in the various groups are different. The estimation errors on $\boldsymbol{\beta}_g^*$ are reported in Table S4 in the Supplementary Material.

6. Application to ADNI Data Analysis

AD is an irreversible neurodegenerative disease that results in a loss of mental functions caused by a deterioration of the brain. It is the most common cause of dementia among people over the age of 65, affecting an estimated 5.5 million Americans, yet no prevention methods or cures have been discovered. The ADNI was started in 2004 with the goal of tracking the progression of the disease using biomarkers, and using clinical measures to assess the brain's function over the course of the disease states. In this section, we apply our method to the ADNI data. We are interested in predicting the ADAS-Cog scores using structural magnetic resonance imaging (MRI) scans. All subjects in our analysis are from the ADNI2 phase of the study. In total, there are 697 subjects in our analysis and five groups: NC, SMC, eMCI, lMCI, and AD, ordered by disease severity. The MRI images were preprocessed using anterior commissure-posterior commissure correction, intensity inhomogeneity correction, skull stripping, cerebellum removal based on registration with atlas, spatial segmentation, and registration. After registration, we obtain MRI data with 93 regions of interest (ROIs). For

Table 1. Overall MSEs for the four models.

Penalty	Global	Group-specific	Factor-0	Proposed
Ridge	27.52 (0.33)	15.70 (0.19)	15.17 (0.18)	**15.04** (0.18)
EN	28.23 (0.33)	16.26 (0.21)	15.47 (0.18)	**15.40** (0.18)
Lasso	28.27 (0.34)	16.39 (0.23)	15.49 (0.19)	**15.45** (0.18)

each of the 93 ROIs, we compute the volume of gray matter as a feature. As a result, for each subject, we finally obtain 93 MRI features. Our goal is to predict the ADAS-Cog scores using the 93 MRI features, together with the group information.

We randomly partition the whole data set into two parts: 75% for training the model, and the rest for testing the performance. We repeat the random split 100 times. The testing MSEs and the corresponding standard errors are reported in Table 1 (overall performance) and Table 2 (groupwise performance). We compare four models: the global model (2.1), the group-specific model (2.2), the Factor-0 model (2.5), and our proposed model, as shown in (2.4). For each model, we use three penalty functions, the ℓ_2 penalty (Ridge), the ℓ_1 penalty (Lasso), and the Elastic Net (EN) penalty with the bridging parameter 0.5.

As shown in Tables 1 and 2, our proposed models achieve promising performance in most cases. The global model performs worst, because it does not use the label information at all. The group-specific model does not perform as well as our proposed models, because it does not borrow information across different groups. Note that the Factor-0 model achieves great improvement over the global model, which demonstrates that the difference on group means is the main source of the heterogeneous effect on the clinical scores across the five groups. It is seen in Table 2 that our model achieves the greatest improvement on the AD patients over the other models, which indicates that the effects of the heterogeneous factors identified in the AD group are much stronger than those in other groups. This appears to be reasonable, because the brain structure of AD patients is significantly more impaired.

Our model has good interpretations. In this real data set, we can interpret variations due to identified factors as disease-specific variations, and the variation due to the homogeneous signals as the disease-shared variation among all groups. Figure 3 gives heatmaps of $\hat{\mathbf{\Sigma}}_{x,g} = (1/n_g)\mathbf{X}_g'\mathbf{X}_g$ (the top row), where $\mathbf{\Sigma}_{x,g} = \mathrm{cov}(\mathbf{x}_{g,i})$, and $\hat{\mathbf{\Sigma}}_{u,g}$ (the bottom row), which is obtained by applying an adaptive soft threshold to $\hat{\mathbf{\Sigma}}_{x,g} - \hat{\mathbf{\Lambda}}_g'\hat{\mathbf{\Lambda}}_g$. The left, middle, and right columns of Figure 3 are for the NC, eMCI, and AD groups, respectively. From Figure 3, we can

Table 2. Groupwise MSEs for the four models.

Group	Global	Group-specific	Factor-0	Proposed
		Penalty = Ridge		
NC	16.66 (0.38)	6.24 (0.09)	6.50 (0.10)	**6.19** (0.10)
SMC	14.52 (0.31)	6.68 (0.15)	**6.43** (0.15)	6.54 (0.15)
eMCI	18.37 (0.41)	10.26 (0.19)	9.84 (0.19)	**9.82** (0.19)
lMCI	19.17 (0.38)	16.75 (0.32)	**15.61** (0.30)	15.92 (0.32)
AD	73.55 (0.38)	41.25 (0.32)	40.00 (0.30)	**39.28** (0.32)
		Penalty = Elastic Net		
NC	16.79 (0.38)	6.45 (0.09)	6.40 (0.11)	**6.37** (0.09)
SMC	15.46 (0.38)	7.12 (0.09)	**6.78** (0.11)	6.96 (0.09)
eMCI	18.65 (0.38)	10.59 (0.09)	**10.13** (0.11)	10.22 (0.09)
lMCI	20.26 (0.38)	18.32 (0.09)	**16.14** (0.11)	16.43 (0.09)
AD	75.00 (0.38)	41.49 (0.09)	40.54 (0.11)	**39.64** (0.09)
		Penalty = Lasso		
NC	16.69 (0.38)	6.49 (0.09)	6.41 (0.11)	**6.37** (0.09)
SMC	15.57 (0.38)	7.16 (0.09)	**6.84** (0.11)	7.05 (0.09)
eMCI	18.44 (0.38)	10.73 (0.09)	**10.17** (0.11)	10.26 (0.09)
lMCI	20.36 (0.38)	18.53 (0.09)	**16.21** (0.11)	16.50 (0.09)
AD	75.40 (0.38)	41.73 (0.09)	40.47 (0.11)	**39.68** (0.09)

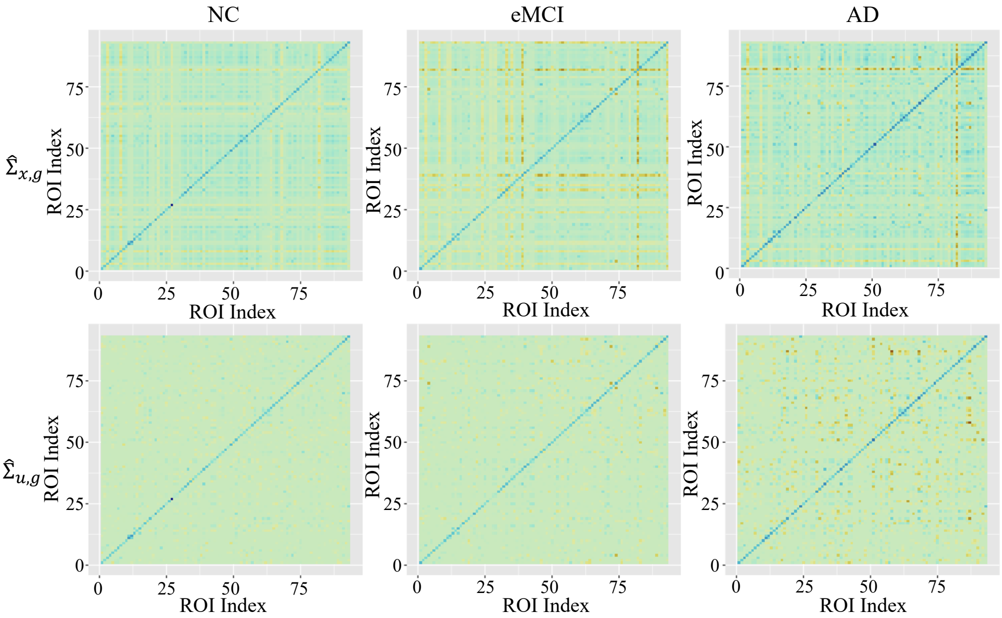

Figure 3. Heatmaps of $\hat{\boldsymbol{\Sigma}}_{x,g}$ and $\hat{\boldsymbol{\Sigma}}_{u,g}$ in NC, eMCI and AD groups.

see that the bottom row looks more homogeneous than the top row. We further represent brain connections using precision matrices estimated from Gaussian graphical models (Cai, Liu and Luo (2011)). See Section S4 in the Supplementary Material.

7. Conclusion

We have proposed a factor regression model for heterogeneous data with subpopulations. Our proposed model decomposes the predictors into heterogeneous components driven by latent factors and homogeneous components. We assume the group-specific latent factors explain the main heterogeneous variations and, consequently, their associated coefficients can differ by groups. The homogeneous components share the same covariance matrix and, as a result, they share the same regression coefficients. Because the factors are unobserved, we first estimate them using a standard PCA procedure. We use an OLS to directly estimate the group-specific coefficients. For the homogeneous regression coefficients, we propose a flexible penalized least squares solution. For model prediction, we also propose a data-driven procedure to estimate the factors for the testing data. Theoretical studies on the estimation and prediction consistency under ℓ_2 and ℓ_1 penalties are established. We show that our proposed model is robust under the group-specific model. Extensive simulation studies further demonstrate the competitive performance of our proposed model over the global model and the group-specific model, and our proposed model achieves a good balance between the two. Finally, we apply the proposed method to an ADNI data set for clinical score prediction, and demonstrate that our model has good prediction power and meaningful interpretation. One interesting future direction is to extend the method to include other outcomes, such as categorical or count data.

Supplementary Material

Section S1 gives proofs of Theorems 1–4, Corollaries 1.1–4.1, and the supporting lemmas. Section S2 provides a rule of thumb to choose between our proposed model and the group-specific model in practice. Section S3 presents additional simulation results. Section S4 contains additional results from the ADNI data analysis. Section S5 shows the analysis results when we apply our method to a combined microarray data set.

Acknowledgments

The authors would like to thank the editor, associate editor, and reviewers for their helpful comments and suggestions. This research was supported in part by NSF grant DMS-1821231, NIH grants R01GM126550 and R01AG073259.

References

Ahn, S. C. and Horenstein, A. R. (2013). Eigenvalue ratio test for the number of factors. *Econometrica* **81**, 1203–1227.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica* **71**, 135–171.

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* **70**, 191–221.

Bai, J. and Ng, S. (2013). Principal components estimation and identification of static factors. *Journal of Econometrics* **176**, 18–29.

Bai, J. and Ng, S. (2008). Large dimensional factor analysis. *Foundations and Trends® in Econometrics* **3**, 89–163.

Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics* **37**, 1705–1732.

Bühlmann, P. (2016). Partial least squares for heterogeneous data. In *The Multiple Facets of Partial Least Squares and Related Methods* (Edited by DAbdi, H., Esposito Vinzi, V., Russolillo, G., Saporta, G. and Trinchera, L.), 3–15. Springer International Publishing, Cham.

Cai, T. and Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association* **106**, 672–684.

Cai, T., Liu, W. and Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association* **106**, 594–607.

Dicker, L. H. (2016). Ridge regression and asymptotic minimax estimation over spheres of growing dimension. *Bernoulli* **22**, 1–37.

Dobriban, E. and Wager, S. (2018). High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics* **46**, 247–279.

Fan, J., Liao, Y. and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75**, 603–680.

Fan, J., Liu, H., Wang, W. and Zhu, Z. (2018). Heterogeneity adjustment with applications to graphical model inference. *Electronic Journal of Statistics* **12**, 3908–3952.

Feng, Q., Jiang, M., Hannig, J. and Marron, J. (2018). Angle-based joint and individual variation explained. *Journal of Multivariate Analysis* **166**, 241–265.

Gaynanova, I. and Li, G. (2019). Structural learning and integrative decomposition of multi-view data. *Biometrics* **75**, 1121–1132.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological)* **55**, 757–796.

Hoerl, A. E. and Kennard, R. W. (2000). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **42**, 80–86.

Hsu, D., Kakade, S. M. and Zhang, T. (2014). Random design analysis of ridge regression. *Foundations of Computational Mathematics* **14**, 569–600.

Joliffe, I. and Morgan, B. (1992). Principal component analysis and exploratory factor analysis. *Statistical Methods in Medical Research* **1**, 69–95.

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.

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.

Lock, E. F., Hoadley, K. A., Marron, J. S. and Nobel, A. B. (2013). Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *The Annals of Applied Statistics* **7**, 523–542.

Ma, S. and Huang, J. (2017). A concave pairwise fusion approach to subgroup analysis. *Journal of the American Statistical Association* **112**, 410–423.

Meinshausen, N. and Bühlmann, P. (2015). Maximin effects in inhomogeneous large-scale data. *The Annals of Statistics* **43**, 1801–1830.

Muniategui, A., Pey, J., Planes, F. J. and Rubio, A. (2013). Joint analysis of miRNA and mRNA expression data. *Briefings in Bioinformatics* **14**, 263–278.

Park, J. Y. and Lock, E. F. (2020). Integrative factorization of bidimensionally linked matrices. *Biometrics* **76**, 61–74.

Pinheiro, J. C. and Bates, D. M. (2000). Linear mixed-effects models: Basic concepts and examples. *Mixed-Effects Models in S and S-Plus*, 3–56.

Raskutti, G., Wainwright, M. J. and Yu, B. (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory* **57**, 6976–6994.

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.

Tang, L. and Song, P. X. (2016). Fused lasso approach in regression coefficients clustering: Learning parameter heterogeneity in data integration. *The Journal of Machine Learning Research* **17**, 3915–3937.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Vicari, D. and Vichi, M. (2013). Multivariate linear regression for heterogeneous data. *Journal of Applied Statistics* **40**, 1209–1230.

Wang, P., Liu, Y. and Shen, D. (2018). Flexible locally weighted penalized regression with applications on prediction of alzheimer's disease neuroimaging initiative's clinical scores. *IEEE Transactions on Medical Imaging* **38**, 1398–1408.

Wold, S., Esbensen, K. and Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems* **2**, 37–52.

Zhang, D., Wang, Y., Zhou, L., Yuan, H. and Shen, D. (2011). Multimodal classification of Alzheimer's disease and mild cognitive impairment. *Neuroimage* **55**, 856–867.

Zhao, T., Cheng, G. and Liu, H. (2016). A partially linear framework for massive heterogeneous data. *The Annals of Statistics* **44**, 1400–1437.

Zhou, G., Cichocki, A., Zhang, Y. and Mandic, D. P. (2015). Group component analysis for multiblock data: Common and individual feature extraction. *IEEE Transactions on Neural Networks and Learning Systems* **27**, 2426–2439.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**, 301–320.

Peiyao Wang

Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA.

E-mail: peiyaow76@gmail.com

Quefeng Li

Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA.

E-mail: quefeng@email.unc.edu

Dinggang Shen

School of Biomedical Engineering, ShanghaiTech University, Shanghai, China.
Shanghai United Imaging Intelligence Co., Ltd., Shanghai, China.
Department of Artificial Intelligence, Korea University, Seoul 02841, Republic of Korea.

E-mail: dinggang_shen@gmail.com

Yufeng Liu

Department of Statistics and Operations Research, Department of Genetics, Department of Biostatistics, Carolina Center for Genome Sciences, Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA.

E-mail: yfliu@email.unc.edu

(Received April 2020; accepted April 2021)