Statistica Sinica Preprint No: SS-2020-0145	
Title	High-Dimensional Factor Regression for Heterogeneous Subpopulations
Manuscript ID	SS-2020-0145
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202020.0145
Complete List of Authors	Peiyao Wang, Quefeng Li, Dinggang Shen and Yufeng Liu
Corresponding Author	Yufeng Liu
E-mail	yfliu@email.unc.edu
Notice: Accepted version subject to English editing.	

HIGH-DIMENSIONAL FACTOR REGRESSION FOR HETEROGENEOUS SUBPOPULATIONS

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

[1]*University of North Carolina at Chapel Hill and* [2]*Korea University*

Abstract: In modern scientific research, data heterogeneity is commonly observed due 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 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 the group-specific and the 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 dataset from Alzheimer's Disease Neuroimaging Initiative and an aggregated microarray dataset 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, as 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). Besides 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 paper, 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. Due 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

the whole data may be restrictive because it ignores the group label information, while fitting distinct regression models in each group may not be optimal as well because it 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 in great needs to handle such heterogeneous data. In the literature, varying coefficient models (Hastie and Tibshirani, 1993) and mixed effects models (Pinheiro and Bates, 2000) can be useful to address data heterogeneity. However, those 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 variations. Meinshausen et al. (2015) proposed a maxmin effects approach under the mixture model. Zhao et al. (2016) proposed a partially linear regression framework to model massive heterogeneous data. Tang and Song (2016); Ma and Huang (2017) proposed fused penalties to estimate regression coefficients in order to identify subpopulations. Wang et al. (2018) proposed a locally-weighted penalized model by incorporating a progression score in the local kernels. However, those models are not designed to characterize the globally-shared and group-specific structures. 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 literature in studying data heterogeneity for unsupervised learning.

Principal component analysis (PCA) (Wold et al., 1987) techniques are popular, due 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 (2019). These methods can be easily extended 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 to handle data heterogeneity is factor models. Factor models are useful unsupervised learning tools to model dependence among 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 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.

In this paper, 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 variation has been removed, the residual signals are assumed to be homogeneous across subgroups, i.e. 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 homogeneous signals are shared across groups. We use PCA to estimate factors and homogeneous signals. Since the estimated factors and homogeneous signals are orthogonal, their coefficients can be estimated separately. The low-dimensional heterogeneous regression coefficients can be directly estimated by the ordinary least squares (OLS). 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 hard to be estimated. In light of penalization methods (Hoerl and Kennard, 1970; 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 homogeneous signals. As for prediction, we propose a data-driven trace maximization step to estimate factors and homogeneous signals in the test set before applying our model for prediction.

We establish estimation consistency for our proposed estimators using either an ℓ_2 or ℓ_1 penalty. In terms of 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 the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset as well as an aggregated microarray dataset 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 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 the 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 it with the global and the group-specific models. In Section 6, we apply our model to the ADNI data for predicting the clinical score. We conclude the paper with some discussion in Section 7.

2. Motivation and Model Framework

Factor models are useful to model dependence among 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 notations. 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 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 $\boldsymbol{Y} = (\boldsymbol{Y}_1', \ldots, \boldsymbol{Y}_G')'$ and $\mathbf{X} = (\mathbf{X}_1', \ldots, \mathbf{X}_G')'$. In this model, all the subgroups share the same intercept and regression coefficients. The global model ignores the het-

erogeneity 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 enough by ignoring 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 be both 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 $\mathbf{U}_g \in \mathbb{R}^{n_g\times p}$ is the homogeneous signals, also known as idiosyncratic errors in the factor model literature (Bai et al., 2008). The number of random factors K_g can vary among groups.

Denote the ith row of $\mathbf{X}_g$, $\mathbf{F}_g$, $\mathbf{U}_g$ 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 $\mathbb{E}(\mathbf{f}_{g,i}) = \mathbf{0}$, $\mathrm{cov}(\mathbf{f}_{g,i}) = \mathbf{I}_{K_g\times K_g}$, $\mathbb{E}(\mathbf{u}_{g,i}) = \mathbf{0}$ and $\mathrm{cov}(\mathbf{u}_{g,i}) = \boldsymbol{\Sigma}_u$. Hence, for each sample in group g,

we have $\text{cov}(\mathbf{x}_{g,i}) = \mathbf{\Lambda}_g'\mathbf{\Lambda}_g + \mathbf{\Sigma}_u$, which is the sum of the group-specific low-rank matrix $\mathbf{\Lambda}_g'\mathbf{\Lambda}_g$ capturing group-specific heterogeneity, and the matrix $\mathbf{\Sigma}_u$ that is homogeneous across different groups.

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

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

which is the maximum number of non-zero entries in the row of $\mathbf{\Sigma}_u$.

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, $\boldsymbol{\gamma}_g^* \in \mathbb{R}^{K_g}$ is the true group-specific coefficients for $\mathbf{F}_g$, $\boldsymbol{\beta}^* \in \mathbb{R}^p$ is 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), $\boldsymbol{\gamma}_g^*$'s vary across G groups, and they characterize the heterogeneity induced by factors in the regression model. Moreover, the group mean term $\boldsymbol{\mu}_g^*$ contributes to the heterogeneity in the regression model (2.4) as well. 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 $\text{var}(y_{g,i}) = \boldsymbol{\gamma}_g^{*\prime}\boldsymbol{\gamma}_g^* + \boldsymbol{\beta}^{*\prime}\boldsymbol{\Sigma}_u\boldsymbol{\beta}^* + \sigma^2$. 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 heterogeneity among subgroups, and at the same time borrow information across subgroups to model homogeneous effects.

One special case of our proposed model (2.4) is when there is no group-specific factor, i.e. $\mathbf{F}_g = \mathbf{0}$. 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) since it adjusts the group mean. It is different from (2.2) since 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 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 from PCA. In Section 3.2, we introduce our procedure for estimating model parameters. After the model is trained, in Section 3.3, we propose a data-driven procedure to estimate factors in the testing data in order to make predictions.

3.1 Factor Model Estimation

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

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

where $\|\cdot\|_F$ denotes the matrix Frobenius norm. The solution to (3.1) can be obtained by performing the eigendecomposition of 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 et al., 2012; Ahn and Horenstein, 2013). In this paper, we consider the following estimator:

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

where $\lambda_k(\cdot)$ denote the kth largest eigenvalue (Lam et al., 2012). Here, $K_{\max}$ is a pre-determined upper bound for the number of factors. This estimator was 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 then estimate the model parameters μ_g^*, $\boldsymbol{\gamma}_g^*$ 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 $\mathbf{F}_g$. Due to the properties of 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:

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

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

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

For homogeneous regression coefficients $\boldsymbol{\beta}^*$, since 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 $\tilde{\boldsymbol{Y}} = (\tilde{\boldsymbol{Y}}_1', \ldots, \tilde{\boldsymbol{Y}}_G')'$, where $\tilde{\boldsymbol{Y}}_g = \boldsymbol{Y}_g - \hat{\boldsymbol{\mu}}_g - \hat{\mathbf{F}}_g\hat{\boldsymbol{\gamma}}_g$. 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 a penalized quadratic minimization problem to estimate $\boldsymbol{\beta}^*$:

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

where $P(\boldsymbol{\beta})$ is a penalty function and λ is a tuning parameter, whose optimal value is chosen by 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 materials, 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, instead of 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 materials.

We summarize the overall training procedure as follows:

1. For $g = 1, \ldots, G$,
   (a) Estimate K_g from (3.2).
   (b) Perform PCA on $\mathbf{X}_g\mathbf{X}_g'$ to obtain $\hat{\mathbf{F}}_g$. Estimate μ_g^* and $\boldsymbol{\gamma}_g^*$ from (3.3).
   (c) Compute 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 $\hat{\mathbf{U}} = \mathbf{HX}, \tilde{\boldsymbol{Y}} = \mathbf{H}(\boldsymbol{Y} - \hat{\boldsymbol{\mu}})$, 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 materials.

3.3 Prediction

After training the model, in order to make predictions on the testing data, we need to estimate 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 for $\mathbf{F}_{g,*}$ is the same as that of $\mathbf{F}_g$.

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 to estimate $\mathbf{F}_{g,*}$ 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, whose solution 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 have worse predictions than 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 notations. 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 $\|\mathbf{A}\|_F = \sqrt{\mathrm{tr}(\mathbf{A}'\mathbf{A})}$, $\|\mathbf{A}\| = \lambda_{\max}(\mathbf{A}'\mathbf{A})$, $\|\mathbf{A}\|_1 = \max_{j\le p}\sum_{i=1}^{p}|a_{ij}|$ 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 $\{j : b_j \neq 0\}$. In addition, 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 $\boldsymbol{\Sigma}$ is said to satisfy the restricted eigenvalue (RE) condition if and only if there exists a positive constant κ, such that $\boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta} \ge \kappa\|\boldsymbol{\beta}\|^2$

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 both $\{\mathbf{f}_{g,i}\}_{i\leq n_g}$ and $\{\mathbf{u}_{g,i}\}_{i\leq n_g}$ are independent and identically distributed sub-Gaussian random variables with zero means and covariance as $\mathbf{I}_{K_g\times K_g}$ 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 et al., 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 et al., 2008; Fan et al., 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) $\|\hat{\boldsymbol{\gamma}}_g - \mathbf{H}_g\boldsymbol{\gamma}_g^*\| = O_P(1/\sqrt{n_g} + 1/\sqrt{p})$, *where* $\hat{\boldsymbol{\gamma}}_g$ *is as defined in* (3.3), $\mathbf{H}_g = \hat{\mathbf{D}}_g^{-1}\hat{\mathbf{F}}'_g\mathbf{F}_g\boldsymbol{\Lambda}_g\boldsymbol{\Lambda}'_g$, *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*

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

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

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

Statement (a) shows that $\hat{\boldsymbol{\gamma}}_g$ is consistent to $\boldsymbol{\gamma}_g^*$ up to a rotation given by $\mathbf{H}_g$. When the latent factors are known, the oracle convergence rate of $\hat{\boldsymbol{\gamma}}_g$ 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 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 latent factors is so small that it will not affect the convergence rate of $\hat{\boldsymbol{\gamma}}_g$. This is essentially due to a blessing of dimensionality property of 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 $\boldsymbol{\beta}^*$, no matter whether an ℓ_1 or ℓ_2 penalty is imposed. To simply the discussion, suppose we 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

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

It was studied in Hsu et al. (2012) 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 for estimating 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 et al., 2011). The extra term of $O_P(\sqrt{s/p})$ comes from the estimation error $\hat{\mathbf{\Sigma}}_u$; see Fan et al. (2013). This term is ignorable when $p \gg n$, in which case the minimax rate is attained.

Let $\hat{\boldsymbol{Y}}_{g,\lambda}^{ridge} = \hat{\mathbf{F}}_g \hat{\boldsymbol{\gamma}}_g + \hat{\mathbf{U}}_g \hat{\boldsymbol{\beta}}_\lambda^{ridge}$ and $\hat{\boldsymbol{Y}}_{g,\lambda}^{lasso} = \hat{\mathbf{F}}_g \hat{\boldsymbol{\gamma}}_g + \hat{\mathbf{U}}_g \hat{\boldsymbol{\beta}}_\lambda^{lasso}$ denote the predicted values of $\boldsymbol{Y}_g$, where $\hat{\boldsymbol{\gamma}}_g$ is given in (3.3), $\hat{\boldsymbol{\beta}}_\lambda^{ridge}$ and $\hat{\boldsymbol{\beta}}_\lambda^{lasso}$ are the Ridge and Lasso estimators 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*

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

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

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

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

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

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 et al., 2016; Dobriban et al., 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 et al., 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 considered.

4.2 Consistency of Group-Specific and Global Models

We study statistical properties of the group-specific and the 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 $p^{1/2}$. The reason is that due to the pervasiveness assumption, $\|\mathbf{X}_g\|$ 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

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

whereas the global model seeks to solve

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

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. Similarly as in (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 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*

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

(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

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

Let $\hat{\boldsymbol{Y}}^{ridge}_{g,\lambda} = \tilde{\mathbf{X}}_g\hat{\boldsymbol{\beta}}^{ridge}_{g,\lambda}$ 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 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*

$$\|\frac{1}{n_g}\{\hat{\boldsymbol{Y}}^{ridge}_{g,\lambda} - \mathbb{E}(\boldsymbol{Y}_g|\mathbf{F}_g,\mathbf{U}_g)\}\| = O_P\Big(\sqrt{\frac{p}{n_g}}\|\boldsymbol{\delta}_g\| + d_p(\frac{1}{\sqrt{n_g}}+\frac{\sqrt{p}}{n_g})+\frac{\sqrt{p}}{n_g}\Big), \tag{4.14}$$

$$\|\tfrac{1}{n_g}\{\hat{\boldsymbol{Y}}^{lasso}_{g,\lambda} - \mathbb{E}(\boldsymbol{Y}_g|\mathbf{F}_g,\mathbf{U}_g)\}\| = O_P\Big(\sqrt{\tfrac{s}{n_g}}\{(1+\sqrt{\tfrac{\log p}{n_g}})(d_p+\|\boldsymbol{\delta}_g\|)+\sqrt{\tfrac{\log p}{n_g}}\}\Big). \tag{4.15}$$

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

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

(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 $\lambda = C\big[\{n_{\max}/(n\sqrt{p})+(1/n)\sqrt{n_{\max}\log p/p}\}\big(d_p+\sum_{g=1}^{G}\|\boldsymbol{\delta}_g\|\big)+(1/n)\sqrt{n_{\max}\log p/p}\big]$ for some large enough con-

stant C, we have

$$\|\hat{\boldsymbol{\beta}}^{lasso}_{\lambda,global} - \boldsymbol{\beta}^*\| = O_P\Big(\sqrt{s}\big\{\big(\tfrac{n_{\max}}{n} + \tfrac{\sqrt{n_{\max}\log p}}{n}\big)\big(d_p + \textstyle\sum_{g=1}^{G}\|\boldsymbol{\delta}_g\|\big) + \tfrac{\sqrt{n_{\max}\log p}}{n}\big\}\Big). \quad (4.17)$$

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 to (4.11). We have the following upper bounds for their in-sample prediction errors.

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

$$\begin{aligned}\|\frac{1}{n_g}\big\{\hat{\boldsymbol{Y}}^{ridge}_{g,\lambda} - \mathbb{E}(\boldsymbol{Y}_g|\mathbf{F}_g,\mathbf{U}_g)\big\}\| = O_P\Big(&\frac{n_{\max}}{n}\sqrt{\frac{p}{n_g}}\sum_{g'=1}^{G}\|\boldsymbol{\delta}_{g'}\| + \frac{1}{n}\sqrt{\frac{n_{\max}p}{n_g}} + \\ &d_p\big(\frac{n_{\max}}{n\sqrt{n_g}} + \frac{1}{n}\sqrt{\frac{n_{\max}p}{n_g}}\big)\Big),\end{aligned} \quad (4.18)$$

$$\begin{aligned}\|\frac{1}{n_g}\big\{\hat{\boldsymbol{Y}}^{lasso}_{g,\lambda} - \mathbb{E}(\boldsymbol{Y}_g|\mathbf{F}_g,\mathbf{U}_g)\big\}\| = O_P\Big(&\sqrt{\frac{s}{n_g}}\big\{\frac{\sqrt{n_{\max}\log p}}{n} + \\ &\big(\frac{n_{\max}}{n} + \frac{\sqrt{n_{\max}\log p}}{n}\big)\big(d_p + \sum_{g'=1}^{G}\|\boldsymbol{\delta}_{g'}\|\big)\big\}\Big).\end{aligned} \quad (4.19)$$

Since 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 $d_p = O(1)$, if we assume that $n_1 = \cdots = n_G$ and G is bounded, (4.14) and (4.18) further reduce to $\|(1/n_g)\big\{\hat{\boldsymbol{Y}}^{ridge}_{g,\lambda} - \mathbb{E}(\boldsymbol{Y}_g|\mathbf{F}_g,\mathbf{U}_g)\big\}\| = O_P(\sqrt{p/n})$ for the Ridge estimator. Compared with the predictor error of our Ridge estimator, which is in the order of $O_P(\sqrt{p}/n)$, these two methods are worse by a factor of $\sqrt{n}$, which is due 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 $\|(1/n_g)\big\{\hat{\boldsymbol{Y}}^{lasso}_{g,\lambda} - \mathbb{E}(\boldsymbol{Y}_g|\mathbf{F}_g,\mathbf{U}_g)\big\}\| = O_P(\sqrt{s/n} + \sqrt{s\log p}/n)$. Com-

pared with our Lasso estimator, they have an extra term of $\sqrt{s/n}$, which also comes from model mis-specification and is non-ignorable.

4.3 Robustness

In this section, we study the problem that if each group follows a distinct model

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

how well does our method perform 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 $\boldsymbol{\beta}^*$ 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) $\|\hat{\boldsymbol{\gamma}}_g - p^{-1/2}\mathbf{H}_g\boldsymbol{\Lambda}_g\boldsymbol{\beta}_g^*\| = O_P(1/\sqrt{n_g} + 1/\sqrt{p})$, *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*

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

(c) Assuming that $\boldsymbol{\beta}_g^*$ *is* s*-sparse and* $\boldsymbol{\Sigma}_u$ *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

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

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*

$$\|\tfrac{1}{n_g}\{\hat{\boldsymbol{Y}}_{g,\lambda}^{ridge} - \mathbb{E}(\boldsymbol{Y}_g|\tilde{\mathbf{X}}_g)\}\| = O_P(\tfrac{1}{n_g}) + O_P(\tfrac{1}{\sqrt{n_g p}}) + O_P\Big(\tfrac{1}{\sqrt{n_g}}\|\hat{\boldsymbol{\beta}}_{\lambda}^{ridge} - \tfrac{1}{\sqrt{p}}\boldsymbol{\beta}_g^*\|\Big), \quad (4.21)$$

$$\|\tfrac{1}{n_g}\{\hat{\boldsymbol{Y}}_{g,\lambda}^{lasso} - \mathbb{E}(\boldsymbol{Y}_g|\tilde{\mathbf{X}}_g)\}\| = O_P(\tfrac{1}{n_g}) + O_P(\tfrac{1}{\sqrt{n_g p}}) + O_P\Big(\tfrac{1}{\sqrt{n_g}}\|\hat{\boldsymbol{\beta}}_{\lambda}^{lasso} - \tfrac{1}{\sqrt{p}}\boldsymbol{\beta}_g^*\|\Big). \quad (4.22)$$

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

$$\|\tfrac{1}{n_g}\{\hat{\boldsymbol{Y}}_{g,\lambda}^{ridge} - \mathbb{E}(\boldsymbol{Y}_g|\tilde{\mathbf{X}}_g)\}\| = O_P\Big(\textstyle\sum_{g'=1}^{G}\tfrac{1}{\sqrt{np}}\|\boldsymbol{\beta}_g^* - \boldsymbol{\beta}_{g'}^*\|\Big) + O_P\Big(\tfrac{\sqrt{p}}{n}\Big) = O_P\Big(\tfrac{\sqrt{p}}{n}\Big), \quad (4.23)$$

$$\|\tfrac{1}{n_g}\{\hat{\boldsymbol{Y}}_{g,\lambda}^{lasso} - \mathbb{E}(\boldsymbol{Y}_g|\tilde{\mathbf{X}}_g)\}\| = O_P\Big(\textstyle\sum_{g'=1}^{G}\sqrt{\tfrac{s}{np}}\|\boldsymbol{\beta}_g^* - \boldsymbol{\beta}_{g'}^*\|\Big) + O_P\Big(\sqrt{\tfrac{s}{np}} + \tfrac{\sqrt{s\log p}}{n}\Big). \quad (4.24)$$

We compare these convergence rates with the ones given by the group-specific model. As 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 either a Ridge or a Lasso estimator. 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). As for the Lasso

estimator, when p is small, our model converges in 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 our model 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 still 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 has comparable prediction as 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, the group-specific, and the Factor-0 models. In both studies, we choose $G = 3, p = 200, K_g = 3, n_g = 100$ for any $g \in [G]$, generate 600 training samples to train all four models and evaluate their Mean Squared Errors (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 materials. We repeat simulations for 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

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

To ensure $\mathbf{\Lambda}_g$ 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 materials. 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 $\mathbf{\Lambda}_g^1 = \mathbf{Q}_g\mathbf{D}_g^{1/2}\mathbf{V}_g'$, where $\mathbf{Q}_g$ is a random orthonormal matrix, and $\mathbf{\Lambda}_g^2 = \mathbf{Q}_g\boldsymbol{T}_g$, where $\boldsymbol{T}_g$ is a $K_g \times (p - K_g)$ matrix whose elements are randomly generated from Unif$(-1/20, 1/20)$. This construction of $\mathbf{\Lambda}_g$ 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 $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_u)$, where $\mathbf{\Sigma}_u$ is a diagonal matrix whose diagonal elements all equal to 0.03. As for the coefficients in (2.4), we choose $\mu_g^* = g$ for $g = 1, 2, 3$. We set $\boldsymbol{\gamma}_1^* = (h, h, 2h)'$, $\boldsymbol{\gamma}_2^* = (h, 2h, h)'$ and $\boldsymbol{\gamma}_3^* = (2h, h, h)'$, 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 a 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 $\mathcal{N}(0, 4)$.

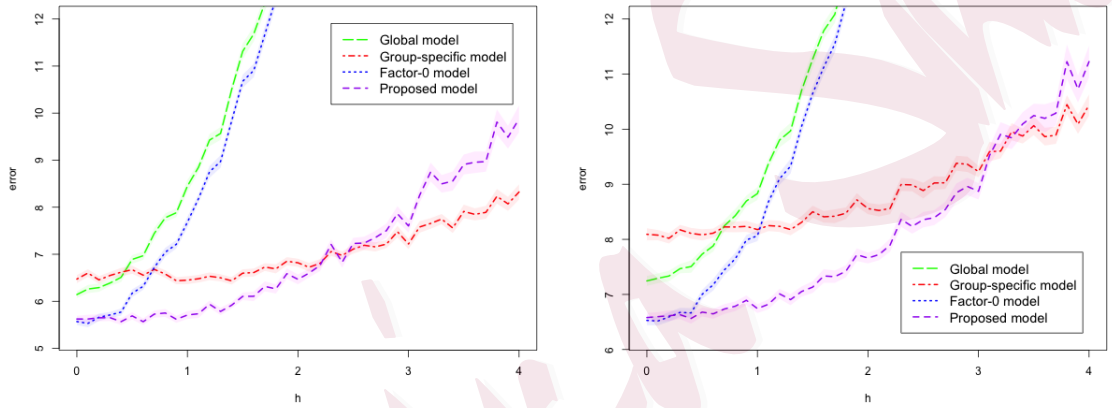

Figure 1: The MSE curves given by the four models. The left panel represents results for a sparse $\boldsymbol{\beta}^*$ and the right panel represents results for a dense $\boldsymbol{\beta}^*$.

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 MSEs in the 50 simulations. The optimal tuning parameters in these methods are chosen by 10-fold cross-validation. It is clearly seen that for most h, our model performs the best. Due to 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 utilized 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 $\boldsymbol{\gamma}_g^*$ and $\boldsymbol{\beta}^*$ are reported in Tables S2 and S3 in the supplementary materials.

5.2 Setting 2: Under Group-specific Model

We generate different models for different groups and inspect how robust our model can be under such a scenario. We generate $\mathbf{f}_{g,i}$ the same way 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 materials. For $\boldsymbol{\Lambda}_g$, we set $\boldsymbol{\Lambda}_g = \tilde{\mathbf{Q}}_g * \mathbf{s}_g$, where $\mathbf{s}_g$ is the same 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 among 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, even though our model is not too far behind it. Compared with these two models, the global and the Factor-0 models are much worse as they fail to recognize the group difference. For a dense $\boldsymbol{\beta}^*$, when h is small, 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 as discussed in the sparse case. But 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 various groups are different. The estimation errors on $\boldsymbol{\beta}_g^*$ are reported in Table S4 in the supplementary materials.

6. Application to ADNI Data Analysis

Alzheimer's Disease is an irreversible neurodegenerative disease that results in a loss of mental functions caused by the deterioration of brain. It is the most common cause of dementia among people over the age of 65, affecting an estimated 5.5 million Americans,

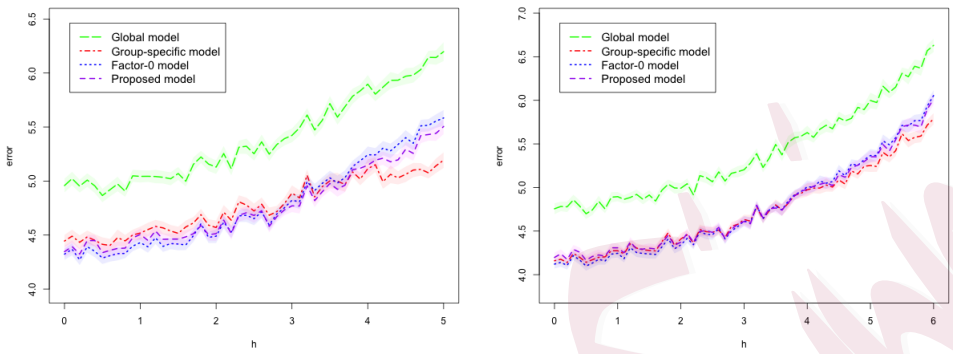

Figure 2: The MSE curves given by the four models. The left panel represents results for sparse $\boldsymbol{\beta}_g^*$ and the right panel represents results for dense $\boldsymbol{\beta}_g^*$.

yet no prevention methods or cures have been discovered. The ADNI was started in 2004 with a goal to track the progression of the disease using biomarkers, and use 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 by structural magnetic resonance imaging (MRI) scans. All subjects in our analysis are from the ADNI2 phase of the study. There are in total 697 subjects in our analysis and 5 groups: NC, SMC, eMCI, lMCI, and AD, ordered by the 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 the MRI data with 93 regions of interest (ROIs). For 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 dataset into two parts: 75% for training the model and the rest for testing the performance. We repeat the random split for 100 times. The testing mean squared errors (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.

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)

As shown in Tables 1 and 2, our proposed models achieve promising performance in most cases. Global model performs the worst, since it does not utilize the label information at all. 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 heterogeneous factors identified in the AD group are much stronger than those in other groups. This appears to be reasonable, since the brain structure of AD patients is significantly more impaired.

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)

Our model has good interpretations. In this real dataset, 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

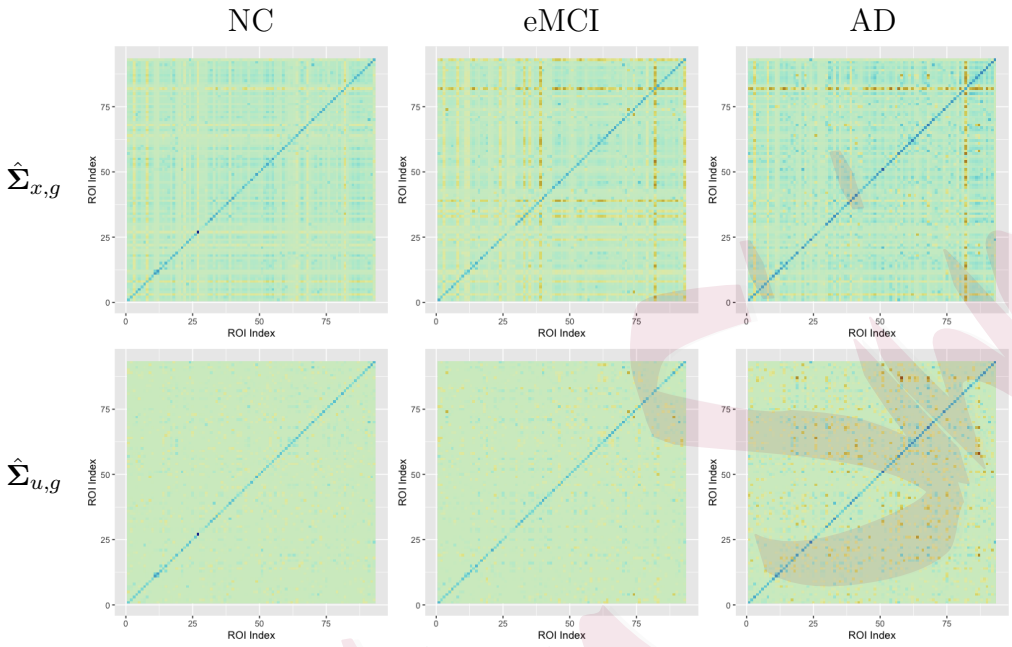

Figure 3: Heatmaps of $\hat{\mathbf{\Sigma}}_{x,g}$ and $\hat{\mathbf{\Sigma}}_{u,g}$ in NC, eMCI and AD groups.

the heatmaps of $\hat{\mathbf{\Sigma}}_{x,g} = (1/n_g)\mathbf{X}_g'\mathbf{X}_g$ (the top row), where $\mathbf{\Sigma}_{x,g} = \text{cov}(\mathbf{x}_{g,i})$, 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 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 et al., 2011). See details in Section S4 in the supplementary materials.

7. Conclusion

In this paper, we propose a factor regression model for heterogeneous data with sub-populations. Our proposed model decomposes the predictors into heterogeneous com-

ponents 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. As factors are unobserved, we first estimate them using the standard PCA procedure. We use OLS to directly estimate the group-specific coefficients. For the homogeneous regression coefficients, we propose a flexible penalized least square solution. For model prediction, we also propose a data-driven procedure to estimate factors for 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 great balance between the two. Finally, we apply the proposed method to the ADNI dataset for clinical score prediction and demonstrate our model has good prediction power and meaningful interpretation. One interesting future direction is to extend the method for other outcomes such as categorical or count data.

Acknowledgments

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

Supplementary Materials

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 dataset.

References

Ahn, S. C. and A. R. Horenstein (2013). Eigenvalue ratio test for the number of factors. *Econometrica 81*(3), 1203–1227.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica 71*(1), 135–171.

Bai, J. and S. Ng (2002). Determining the number of factors in approximate factor models. *Econometrica 70*(1), 191–221.

Bai, J. and S. Ng (2013). Principal components estimation and identification of static factors. *Journal of Econometrics 176*(1), 18–29.

Bai, J., S. Ng, et al. (2008). Large dimensional factor analysis. *Foundations and Trends® in Econometrics 3*(2), 89–163.

Bickel, P. J., Y. Ritov, A. B. Tsybakov, et al. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics 37*(4), 1705–1732.

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

Cai, T. and W. Liu (2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association 106*(494), 672–684.

Cai, T., W. Liu, and X. Luo (2011). A constrained $\ell 1$ minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association 106*(494), 594–607.

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

Dobriban, E., S. Wager, et al. (2018). High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics 46*(1), 247–279.

Fan, J., Y. Liao, and M. Mincheva (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75*(4), 603–680.

Fan, J., H. Liu, W. Wang, Z. Zhu, et al. (2018). Heterogeneity adjustment with applications to graphical model inference. *Electronic Journal of Statistics 12*(2), 3908–3952.

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

Gaynanova, I. and G. Li (2019). Structural learning and integrative decomposition of multi-view data. *Biometrics 75*(4), 1121–1132.

Hastie, T. and R. Tibshirani (1993). Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological) 55*(4), 757–779.

Hoerl, A. E. and R. W. Kennard (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics 12*(1), 55–67.

Hsu, D., S. M. Kakade, and T. Zhang (2012). Random design analysis of ridge regression. In *Conference on learning theory*, pp. 9–1.

Joliffe, I. and B. Morgan (1992). Principal component analysis and exploratory factor analysis. *Statistical methods in medical research 1*(1), 69–95.

Lam, C., Q. Yao, et al. (2012). Factor modeling for high-dimensional time series: inference for the number of factors. *The Annals of Statistics 40*(2), 694–726.

Li, Q., G. Cheng, J. Fan, and Y. Wang (2018). Embracing the blessing of dimensionality in factor models. *Journal of the American Statistical Association 113*(521), 380–389.

Lock, E. F., K. A. Hoadley, J. S. Marron, and A. B. Nobel (2013). Joint and individual variation explained (jive) for integrated analysis of multiple data types. *The annals of applied statistics 7*(1), 523.

Ma, S. and J. Huang (2017). A concave pairwise fusion approach to subgroup analysis. *Journal of the American Statistical Association 112*(517), 410–423.

Meinshausen, N., P. Bühlmann, et al. (2015). Maximin effects in inhomogeneous large-scale data. *The Annals of Statistics 43*(4), 1801–1830.

Muniategui, A., J. Pey, F. J. Planes, and A. Rubio (2013). Joint analysis of mirna and mrna expression data. *Briefings in bioinformatics 14*(3), 263–278.

Park, J. Y. and E. F. Lock (2019). Integrative factorization of bidimensionally linked matrices. *Biometrics*.

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

Raskutti, G., M. J. Wainwright, and B. Yu (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE transactions on information theory 57*(10), 6976–6994.

Stock, J. H. and M. W. Watson (2002). Forecasting using principal components from a large number of predictors. *Journal of the American statistical association 97*(460), 1167–1179.

Tang, L. and P. X. Song (2016). Fused lasso approach in regression coefficients clustering: learning parameter heterogeneity in data integration. *The Journal of Machine Learning Research 17*(1), 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 M. Vichi (2013). Multivariate linear regression for heterogeneous data. *Journal of Applied Statistics 40*(6), 1209–1230.

Wang, P., Y. Liu, and D. Shen (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*(6), 1398–1408.

Wold, S., K. Esbensen, and P. Geladi (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems 2*(1-3), 37–52.

Zhang, D., Y. Wang, L. Zhou, H. Yuan, D. Shen, A. D. N. Initiative, et al. (2011). Multimodal classification of alzheimer's disease and mild cognitive impairment. *Neuroimage 55*(3), 856–867.

Zhao, T., G. Cheng, and H. Liu (2016). A partially linear framework for massive heterogeneous data. *Annals of statistics 44*(4), 1400.

Zhou, G., A. Cichocki, Y. Zhang, and D. P. Mandic (2015). Group component analysis for multiblock data: Common and individual feature extraction. *IEEE transactions on neural networks and learning systems 27*(11), 2426–2439.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67*(2), 301–320.

Department of Statistics and Operations Research, University of North Carolina at Chapel Hill

E-mail: peiyaow76@gmail.com

Department of Biostatistics, University of North Carolina at Chapel Hill

E-mail: quefeng@email.unc.edu

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

E-mail: yfliu@email.unc.edu

Department of Radiology and Biomedical Research Imaging Center, University of North Carolina at Chapel Hill, and Department of Brain and Cognitive Engineering, Korea University

E-mail: dgshen@med.unc.edu