Statistics

A Journal of Theoretical and Applied Statistics

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/gsta20

Integrated partially linear model for multi-centre studies with heterogeneity and batch effect in covariates

Lei Yang & Yongzhao Shao

To cite this article: Lei Yang & Yongzhao Shao (2023) Integrated partially linear model for multi-centre studies with heterogeneity and batch effect in covariates, Statistics, 57:5, 987-1009, DOI: 10.1080/02331888.2023.2258429

To link to this article: https://doi.org/10.1080/02331888.2023.2258429

Published online: 21 Sep 2023.

STATISTICS
2023, VOL. 57, NO. 5, 987–1009
https://doi.org/10.1080/02331888.2023.2258429

Taylor & Francis
Taylor & Francis Group

Integrated partially linear model for multi-centre studies with heterogeneity and batch effect in covariates

Lei Yang and Yongzhao Shao

Department of Population Health, New York University, New York, NY, USA

ABSTRACT

Multi-centre study is increasingly used for borrowing strength from multiple research groups to obtain reproducible study findings. Regression analysis is widely used for analysing multi-group studies, however, some of the regression predictors are nonlinear and/or often measured with batch effects. Also, the group compositions are potentially heterogeneous across different centres. The conventional pooled data analysis can cause biased regression estimates. This paper proposes an integrated partially linear regression model (IPLM) to account for predictor's nonlinearity, general batch effect, group composition heterogeneity, and potential measurement-error in covariates simultaneously. A local linear regression-based approach is employed to estimate the nonlinear component and a regularization procedure is introduced to identify the predictors' effects. The IPLM-based method has estimation consistency and variable-selection consistency. Moreover, it has a fast computing algorithm and its effectiveness is supported by simulation studies. A multi-centre Alzheimer's disease research project is provided to illustrate the proposed IPLM-based analysis.

ARTICLE HISTORY
Received 12 July 2020
Accepted 4 September 2023

KEYWORDS
Multi-centre study; data harmonization; partially linear regression model; general batch effects; group composition heterogeneity

1. Introduction

The design of multi-centre study becomes increasingly used because it enables researchers to obtain more generalizable and reproducible study findings in many fields, including cancer study [1–7] and Alzheimer's disease (AD) research [8–14]. Regression analysis is widely used to analyse multi-centre studies, however, in the current literature, there is a lack of a flexible and rigorous regression approach with an efficient computing algorithm to account for the complexities and many statistical challenges simultaneously in multi-centre or multi-group collaborative studies.

The first major challenge, as in studies of cancers and other complex human disorders, is that potential predictors are often numerous (i.e. high-dimensional) and many of them are nonlinear predictors. For example, in ovarian cancer, both the mean and variance effect of some DNA methylations are significant [15] in risk prediction. Thus it is desired to

CONTACT Yongzhao Shao ✉ shaoy01@nyu.edu Department of Population Health, New York University, New York, NY 10012-1126, USA

develop flexible regression models that can incorporate both linear and nonlinear predictors, e.g., using the partially linear model (PLM) [16], $Y = \boldsymbol{\beta}^{*\mathrm{T}}\boldsymbol{X} + f^*(W) + \epsilon$, where Y is the response variable, $\boldsymbol{X} \in R^{p_n}$ is the vector of p_n-dimensional linear predictors (p_n may grow with sample size n) with effect parameter vector $\boldsymbol{\beta}^*$, W is a nonlinear predictor, $f^*(\cdot)$ is a nonlinear function, and ϵ is a random error. Additionally, it is typically unknown *a priori* whether the effects of the large number of linear predictors are homogeneous across study centres. Thus, for multi-centre studies, a natural modelling choice is the following systems of partially linear models:

$$Y_k = (\boldsymbol{\beta}^* + \boldsymbol{\alpha}_k^*)^{\mathrm{T}}\boldsymbol{X}_k + f_k^*(W_k) + \epsilon_k, \quad \text{for} \quad 1 \leq k \leq K,$$

where K is the number of centres, $\boldsymbol{\beta}^*$ is the common effect of $\boldsymbol{X}$ between centres, and $\boldsymbol{\alpha}_k^*$ denotes the heterogeneous effects specific to the kth centre satisfying the constraint $\sum_{k=1}^{K} \boldsymbol{\alpha}_k^* = 0$.

Another common challenge in multi-centre studies is that the predictors are potentially measured with some general batch effects. For example, in genetics and genomics research, the microarray gene expression data is typically measured with batch effects [17,18]. In Alzheimer's disease (AD) research, the level of the cerebrospinal fluid (CSF) Aβ_{42} protein is a well-known risk factor for developing AD and have been used for decades without the knowledge that levels of CSF Aβ_{42} might be measured with major batch effects. Surprisingly, as shown by [11], the levels of CSF Aβ_{42} protein have a nonlinear cyclic seasonal pattern over measurement dates. Similarly, we examined a recent multi-centre AD research data and found identical nonlinear cyclic pattern (Figure 2 in Section 5). One may group levels of CSF Aβ_{42} protein per measurement calendar month as a batch to correct for the seasonal batch effect. When such batch effect is ignored, the regression estimates tend to be severely biased regardless of sample size which may lead to misleading and seemingly contradicting study findings among independent studies. More numerical demonstrations and details can be found in Section 4.

A further statistical challenge is that the group compositions of some key predictors in different study centres are heterogeneous. For example, in AD research, some study cohorts are younger while others are much older. According to [19], the levels of CSF Aβ_{42} is a nonlinear function of age. In a younger cohort, we may have a significant positive correlation while in older people we may observe a significantly negative correlation between age and levels of CSF Aβ_{42}. If a conventional linear model is applied in a younger versus an older centres independently to study the relationship between age and Aβ_{42}, a positive versus negative slope may be reported and cause confusions. Thus, due to the interplay between nonlinear effect and heterogeneity in group composition of some predictors (e.g., in age), conventional single-centre analyses can potentially lead to contradictory study findings among different centres in the presence of heterogeneity in group compositions. Instead of single centre analysis, to overcome the adverse impact of heterogeneous composition in the presence of nonlinear relationship, one can conduct integrated analysis by combining data from multi-centres via suitable frequency matching or propensity score matching. In this context, the analysis using the simple pooling of multi-centre data via the commonly-used z-score method can cause severe biases.

Additional common issues in multi-centre studies of biomedical research and many other applications include that the linear predictors might be high-dimensional but only a small set of predictors are truly informative or relevant. Thus variable selection is needed for robust and efficient regression analysis. It is also quite common that the linear predictor $\boldsymbol{X}_k$ might have measurement errors. For example, in studies of acquired immune deficiency syndrome (AIDs), virologic and immunologic markers including plasma concentrations of human immunodeficiency virus (HIV)-1 RNA and CD4+ cell counts are often measured with errors [20]. One popular choice to account for the measurement error in variable selection procedure is subtracting a bias correction term from the loss function [21,22,32]. There have been extensive research on both variable selection and measurement error in regression models. Therefore, regression analysis for multi-centre studies should also be able to effectively deal with variable selection with measurement error in addition to account for inter-plays of other common complexities.

In practice, it is quite common that a combination of the above complexities can occur in a single multi-centre study as demonstrated using the multi-centre study of AD in Section 5. However, in the current literature, there is a lack of flexible and integrated regression analysis that can account for the interplays of multiple complexities (e.g., heterogeneity and nonlinearity and batch effects) in multi-centre studies simultaneously. In Section 2, we propose the integrated partially linear model (IPLM) that can account for predictors' nonlinearity, general batch effects, group composition heterogeneity, high-dimensionality, and measurement error simultaneously. In particular, a local linear regression-based approach [23] is applied to estimate the nonlinear component and a regularized procedure is introduced to select informative predictors and estimate the predictors' effect that can be either homogeneous or heterogeneous across study centres. If all the predictors' effects are homogeneous across centres, the proposed IPLM can automatically reduce to one parsimonious partially linear regression model that is applicable to all centres while simultaneously account for nonlinearity, batch effect, group composition heterogeneity, high-dimensionality, and measurement errors. The integrated analysis facilitates generalizable and reproducible outcomes in multi-centre studies. Asymptotically, the proposed regularized method yields variable selection consistency and estimation consistency for the linear and nonlinear components, which are specified in Section 3 including the case where the number of predictors with non-zero effects in the model can be high-dimensional and increasing with the sample size. Also, for practical applications, efficient numerical implementation of the proposed model and analysis method is of crucial importance. Numerical studies are provided in Section 4 to demonstrate the effectiveness of the proposed IPLM-based analysis and illustrate the disadvantages or biases of the conventional within-group regression analysis and the direct data-pooling analysis without suitable batch effect adjustment. Section 5 includes the analysis of a multi-centre AD research project to illustrate the proposed procedures. A short summary is provided in Section 6.

2. Models

For the kth centre in a multi-centre study, let Y_k be the response variable, $\boldsymbol{X}_k$ the linear predictors and W_k the nonlinear predictor. However, W_k is potentially observed with batch effect, e.g., amyloid Aβ_{42} protein measured in different seasons in AD research [11] and $\boldsymbol{X}_k \in \mathcal{R}^{p_n}$ is potentially measured with error. Instead of observing $\boldsymbol{X}_k$ and W_k directly, we actually can only observe $\boldsymbol{Z}_k$ and V_k, $k = 1, 2, \ldots, K$. That is, we propose the following integrated partially linear regression model (IPLM) with heterogeneity and batch effect in covariates for a K-centre multi-centre study:

$$\begin{aligned} Y_k &= \boldsymbol{X}_k^{\mathrm{T}}(\boldsymbol{\beta}^* + \boldsymbol{\alpha}_k^*) + f_k^*(W_k) + \epsilon_k, \\ \boldsymbol{Z}_k &= \boldsymbol{X}_k + \boldsymbol{U}_k, \\ V_k &= W_k + g_k(m_k; \boldsymbol{\psi}_k^*) \quad \text{for} \quad 1 \le k \le K, \end{aligned}$$

with the constraint $\sum_{k=1}^{K} \boldsymbol{\alpha}_k^* = \mathbf{0}$, where $\boldsymbol{\beta}^*$ and $\boldsymbol{\alpha}_k^*$ are the mean and heterogeneous effect respectively, $f_k^*(\cdot)$ is the nonlinear function to be estimated, ϵ_k is the error term with mean 0 and variance σ_k^2, $\boldsymbol{U}_k$ is the measurement error independent with $(\boldsymbol{X}_k, W_k, m_k, \epsilon_k)$ with mean $\mathbf{0}$ and covariance matrix Σ_k, and $\Sigma_k = 0$ when there is no measurement error. $g_k(\cdot; \boldsymbol{\psi}_k^*)$ is the general batch effect. Note that the function form of $g_k(\cdot; \boldsymbol{\psi}_k)$ is assumed known while the parameters $\boldsymbol{\psi}_k$ need to be estimated. The batch effect $g_k(\cdot; \boldsymbol{\psi}_k)$ does not include intercept to ensure model identifiability and $g_k(\cdot; \cdot) = 0$ when there is no batch effect. The m_k is an observed covariate which can be part of $\boldsymbol{Z}_k$. As part of the data harmonization step in multi-centre studies, we can apply least square method to get the estimated batch effects $g_k(; \widehat{\boldsymbol{\psi}}_k)$ based on the observed $\boldsymbol{Z}_k$ and V_k, for $k = 1, 2, \ldots, K$. The above integrated partially linear model includes the batch effect and measurement error as part of model to increase mathematical rigour and reproducibility of the study findings.

Suppose $(y_{ki}, \boldsymbol{z}_{ki}, v_{ki})$ are the observations from the kth group with $1 \le i \le n_k$. For easy exposition, we first assume the covariance matrix Σ_k associated with measurement error is known. The situation where Σ_k is unknown can be similarly treated [21]. We adjust the batch effects and get bias-free observations $(y_{ki}, \boldsymbol{z}_{ki}, v'_{ki})$, where $v'_{ki} = v_{ki} - g_k(m_{ki}; \widehat{\boldsymbol{\psi}}_k)$ is the nonlinear predictor after batch effect adjustment. Denote $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^{\mathrm{T}}, \ldots, \boldsymbol{\alpha}_K^{\mathrm{T}})^{\mathrm{T}}$. In cancer studies, AD research and many other applications, the genetic and proteomic predictors are often high-dimensional. Thus dimension-reduction and variable selection methods are needed to identify informative variables for effective regression analysis. To estimate both the linear and nonlinear components as well as identify the truly informative mean effect and the heterogeneous effect, we proposed to use a regularized loss function. In particular, we denote the naive square-loss function as

$$l(\boldsymbol{\beta}, \boldsymbol{\alpha}, f_k(\cdot)) = \sum_{k=1}^{K} \sum_{i=1}^{n_k} \left(y_{ki} - \boldsymbol{z}_{ki}^{\mathrm{T}}\boldsymbol{\beta} - \boldsymbol{z}_{ki}^{\mathrm{T}}\boldsymbol{\alpha}_k - f_k(v'_{ki}) \right)^2,$$

with the constraint $\sum_{k=1}^{K} \boldsymbol{\alpha}_k = \mathbf{0}$. Then the regularized square loss function is defined as follows:

$$l_p(\boldsymbol{\beta}, \boldsymbol{\alpha}, f_k(\cdot)) = l(\boldsymbol{\beta}, \boldsymbol{\alpha}, f_k(\cdot)) - \sum_{k=1}^{K} n_k(\boldsymbol{\beta} + \boldsymbol{\alpha}_k)^{\mathrm{T}} \Sigma_k (\boldsymbol{\beta} + \boldsymbol{\alpha}_k) + p_{\lambda_\beta}(\boldsymbol{\beta}) + p_{\lambda_\alpha}(\boldsymbol{\alpha}_k), \quad (1)$$

subject to the constraint $\sum_{k=1}^{K} \boldsymbol{\alpha}_k = \mathbf{0}$, where $p_{\lambda_\beta}(\boldsymbol{\beta}) = \lambda_\beta \sum_{j=1}^{p_n} \pi_j |\beta_j|$ is the penalty term to identify the informative mean effect, $p_{\lambda_\alpha}(\boldsymbol{\alpha}_k) = \lambda_\alpha \sum_{k=1}^{K} \sum_{j=1}^{p_n} \pi_{kj} |\alpha_{kj}|$ is the penalty term to identify the informative heterogeneous effect, $\sum_{k=1}^{K} n_k(\boldsymbol{\beta} + \boldsymbol{\alpha}_k)^{\mathrm{T}} \Sigma_k (\boldsymbol{\beta} + \boldsymbol{\alpha}_k)$ is the penalty term to correct the measurement error, and π_j and π_{kj} are the adaptive Lasso weight [24]. Note that the adaptive Lasso weight can be achieved by setting $\pi_j = 1/|\widetilde{\beta}_j|$ and $\pi_{kj} = 1/|\widetilde{\alpha}_{kj}|$, where

$$(\widetilde{\boldsymbol{\beta}},\widetilde{\boldsymbol{\alpha}},\widetilde{f}_k(\cdot)) = \text{argmin}_{\boldsymbol{\beta},\boldsymbol{\alpha},f_k(\cdot)} l(\boldsymbol{\beta},\boldsymbol{\alpha},f(\cdot)) - \sum_{k=1}^{K} n_k(\boldsymbol{\beta}+\boldsymbol{\alpha}_k)^{\mathrm{T}}\Sigma_k(\boldsymbol{\beta}+\boldsymbol{\alpha}_k),$$

subject to the constraint $\sum_{k=1}^{K}\boldsymbol{\alpha}_k = \mathbf{0}$. Importantly, without the batch effect adjustment for V_k, the regression estimates tend to be biased, also can lead to misleading or contradictory study findings.

If the measurement error covariance matrix Σ_k is unknown, to estimate Σ_k, it is common to assume that there are replicated measurements [21], i.e., we observe $\boldsymbol{Z}_{kij} = \boldsymbol{X}_{ki} + \boldsymbol{U}_{kij}$ for $j = 1,\dots,J_{ki}$. Let $\bar{\boldsymbol{Z}}_{ki} = J_{ki}^{-1}\sum_{i=1}^{J_{ki}}\boldsymbol{Z}_{kij}$ to be the sample mean of J_{ki} replicates for ith subject in kth group. Then a consistent moments estimate of Σ_k is

$$\widehat{\Sigma}_k = \sum_{i=1}^{n_k}\sum_{j=1}^{J_{ki}}(\boldsymbol{Z}_{kij}-\bar{\boldsymbol{Z}}_{ki})(\boldsymbol{Z}_{kij}-\bar{\boldsymbol{Z}}_{ki})^{\mathrm{T}}/\sum_{i=1}^{n_k}J_{ki}.$$

Thus the penalized least square is defined as to minimize the following objective function:

$$l_p(\boldsymbol{\beta},\boldsymbol{\alpha},f_k(\cdot)) = l(\boldsymbol{\beta},\boldsymbol{\alpha},f(\cdot)) - \sum_{k=1}^{K} n_k(\boldsymbol{\beta}+\boldsymbol{\alpha}_k)^{\mathrm{T}}\widehat{\Sigma}_k(\boldsymbol{\beta}+\boldsymbol{\alpha}_k) + p_{\lambda_\beta}(\boldsymbol{\beta}) + p_{\lambda_\alpha}(\boldsymbol{\alpha}_k), \quad (2)$$

subject to the constraint $\sum_{k=1}^{K}\boldsymbol{\alpha}_k = \mathbf{0}$. We can estimate $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ by minimizing (2).

2.1. Constrained optimization

The optimization procedure for (1) and (2) are the same, thus in this section, we mainly focus on solving the linear and nonlinear components in (1). To minimize (1), we have $f_k(V'_k) = E(Y_k|V'_k) - E(\boldsymbol{Z}_k|V'_k)^{\mathrm{T}}(\boldsymbol{\beta}+\boldsymbol{\alpha}_k)$. Denote $m_{ky}(V'_k) = E(Y_k|V'_k)$ and $\boldsymbol{m}_{kz}(V'_k) = E(\boldsymbol{Z}_k|V'_k)$. Then the regularized loss function (1) can be rewritten as

$$l_p(\boldsymbol{\beta},\boldsymbol{\alpha}) = l(\boldsymbol{\beta},\boldsymbol{\alpha}) - \sum_{k=1}^{K} n_k(\boldsymbol{\beta}+\boldsymbol{\alpha}_k)^{\mathrm{T}}\Sigma_k(\boldsymbol{\beta}+\boldsymbol{\alpha}_k) + p_{\lambda_\beta}(\boldsymbol{\beta}) + p_{\lambda_\alpha}(\boldsymbol{\alpha}_k),$$

where $\sum_{k=1}^{K}\boldsymbol{\alpha}_k = \mathbf{0}$ and

$$l(\boldsymbol{\beta},\boldsymbol{\alpha}) = \sum_{k=1}^{K}\sum_{i=1}^{n_k}\left(y_{ki} - m_{ky}(v'_{ki}) - (\boldsymbol{z}_{ki} - \boldsymbol{m}_{kz}(v'_{ki}))^{\mathrm{T}}(\boldsymbol{\beta}+\boldsymbol{\alpha}_k)\right)^2.$$

In this article, we use local linear regression [23] to estimate both $m_{ky}(\cdot)$ and $\boldsymbol{m}_{kz}(\cdot)$ and the R package 'locpol' can be directly applied. Let $\widehat{m}_{ky}(\cdot)$ and $\widehat{\boldsymbol{m}}_{kz}(\cdot)$ be the estimates using local linear regression, $\widehat{y}_{ki} = y_{ki} - \widehat{m}_{ky}(v'_{ki})$ and $\widehat{\boldsymbol{z}}_{ki} = \boldsymbol{z}_{ki} - \widehat{\boldsymbol{m}}_{ky}(v'_{ki})$. Thus $l_p(\boldsymbol{\beta},\boldsymbol{\alpha})$ can be written as

$$l_p(\boldsymbol{\beta},\boldsymbol{\alpha}) = l(\boldsymbol{\beta},\boldsymbol{\alpha}) - \sum_{k=1}^{K} n_k(\boldsymbol{\beta}+\boldsymbol{\alpha}_k)^{\mathrm{T}}\Sigma_k(\boldsymbol{\beta}+\boldsymbol{\alpha}_k) + p_{\lambda_\beta}(\boldsymbol{\beta}) + p_{\lambda_\alpha}(\boldsymbol{\alpha}_k), \tag{3}$$

where $l(\boldsymbol{\beta},\boldsymbol{\alpha}) = \sum_{k=1}^{K}\sum_{i=1}^{n_k}(\widehat{y}_{ki} - \widehat{\boldsymbol{z}}_{ki}^{\mathrm{T}}(\boldsymbol{\beta}+\boldsymbol{\alpha}_k))^2$. Next we will solve $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ alternately. Given $\boldsymbol{\alpha}$, the unknown parameter $\boldsymbol{\beta}$ can be solved by R package 'glmnet'. Given $\boldsymbol{\beta}$, we can apply ADMM [25] to solve $\boldsymbol{\alpha}$ under linear constraint $\sum_{k=1}^{K}\boldsymbol{\alpha}_k = \mathbf{0}$. The details are showed in the next section. Given estimated $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\alpha}}_k$, the estimated nonlinear function is

$$\widehat{f}_k(\cdot) = \widehat{m}_{ky}(\cdot) - \widehat{\boldsymbol{m}}_{kz}(\cdot)^{\mathrm{T}}(\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\alpha}}_k).$$

2.2. Details of ADMM

Given $\boldsymbol{\beta}$, the optimization (3) is reduced to minimize

$$l(\boldsymbol{\beta},\boldsymbol{\alpha}) - \sum_{k=1}^{K} n_k(\boldsymbol{\beta}+\boldsymbol{\alpha}_k)^{\mathrm{T}}\Sigma_k(\boldsymbol{\beta}+\boldsymbol{\alpha}_k) + \lambda_\alpha\sum_{k=1}^{K}\sum_{j=1}^{p_n}\pi_{kj}|\alpha_{kj}|, \tag{4}$$

with the constraint $\sum_{k=1}^{K}\boldsymbol{\alpha}_k = \mathbf{0}$. Then the sub-optimization (4) is equivalent to minimize

$$l(\boldsymbol{\beta},\boldsymbol{\alpha}) - \sum_{k=1}^{K} n_k(\boldsymbol{\beta}+\boldsymbol{\alpha}_k)^{\mathrm{T}}\Sigma_k(\boldsymbol{\beta}+\boldsymbol{\alpha}_k) + \lambda_\alpha\sum_{k=1}^{K}\sum_{j=1}^{p_n}\pi_{kj}|\eta_{kj}|$$

with constraint $\boldsymbol{\alpha}_k = \boldsymbol{\eta}_k$ and $\sum_{k=1}^{K}\boldsymbol{\alpha}_k = \mathbf{0}$. Denote $\boldsymbol{\eta} = (\boldsymbol{\eta}_1^{\mathrm{T}},\ldots,\boldsymbol{\eta}_K^{\mathrm{T}})^{\mathrm{T}}$. The linear constraint $\sum_{k=1}^{K}\boldsymbol{\alpha}_k = \mathbf{0}$ and $\boldsymbol{\alpha}_k = \boldsymbol{\eta}_k$ can be rewritten as $A\boldsymbol{\alpha} + B\boldsymbol{\eta} = \mathbf{0}$, where $A = [\mathbf{1}_K \otimes I_{p_n}, I_{p_nK}]^{\mathrm{T}}$ and $B = [0_{p_nK\times p_n}, -I_{p_nK}]^{\mathrm{T}}$. Note that $\mathbf{1}_K$ is the K-dimensional vector of ones, $0_{p_nK\times p_n}$ is the $p_nK \times p_n$-dimensional matrix of zeros, I_{p_n} and I_{p_nK} are the $p_n \times p_n$ and $p_nK \times p_nK$ identify matrices and $\otimes$ represents the Kronecker product. Then the augmented Lagrange multiplier is

$$\begin{aligned}&\operatorname{argmin}_{\boldsymbol{\alpha},\boldsymbol{\eta},\boldsymbol{\Lambda}} l(\boldsymbol{\beta},\boldsymbol{\alpha}) - \sum_{k=1}^{K} n_k(\boldsymbol{\beta}+\boldsymbol{\alpha}_k)^{\mathrm{T}}\Sigma_k(\boldsymbol{\beta}+\boldsymbol{\alpha}_k) + \lambda_\alpha\sum_{k=1}^{K}\sum_{j=1}^{p_n}\pi_{kj}|\eta_{kj}|\\ &\quad - \boldsymbol{\Lambda}^{\mathrm{T}}(A\boldsymbol{\alpha} + B\boldsymbol{\eta}) + \frac{\delta}{2}\|A\boldsymbol{\alpha} + B\boldsymbol{\eta}\|_2^2,\end{aligned}$$

and the unknown parameters $\boldsymbol{\alpha}$, $\boldsymbol{\eta}$ and Lagrange multiplier parameter $\boldsymbol{\Lambda}$ can be updated alternately. Let $\boldsymbol{\alpha}^{(t)}$, $\boldsymbol{\eta}^{(t)}$ and $\boldsymbol{\Lambda}^{(t)}$ denote the current estimated at iteration t. Given $\boldsymbol{\eta}^{(t)}$ and $\boldsymbol{\Lambda}^{(t)}$, $\boldsymbol{\alpha}^{(t+1)}$ can be solved using the Newton–Raphson method. Given $\boldsymbol{\alpha}^{(t+1)}$ and $\boldsymbol{\Lambda}^{(t)}$, $\boldsymbol{\eta}^{(t+1)}$ can be solved using the R package 'glmnet'. Given $\boldsymbol{\alpha}^{(t+1)}$ and $\boldsymbol{\eta}^{(t+1)}$, $\boldsymbol{\Lambda}^{(t+1)}$ can be updated by $\boldsymbol{\Lambda}^{(t+1)} = \boldsymbol{\Lambda}^{(t)} - \delta^{-1}(A\boldsymbol{\alpha}^{(t+1)} + B\boldsymbol{\eta}^{(t+1)})$. We can stop iteration at the convergence.

2.3. Parsimonious model

In biological mechanistic studies and many other real applications, it is frequently assumed that all the predictors' effects are homogeneous across study centres, i.e., $\sum_{k=1}^{K}\sum_{j=1}^{p_n}\alpha_{kj}^2 = 0$ and $f_k(\cdot) = f_1(\cdot)$ for any $k = 1,\dots,K$, and thus the IPLM automatically reduces to single partially linear regression model with batch effects in covariates. Then the penalized least square will be reduced to

$$l_p(\boldsymbol{\beta}, f(\cdot)) = l(\boldsymbol{\beta}, f(\cdot)) - \sum_{k=1}^{K} n_k \boldsymbol{\beta}^{\mathrm{T}} \widehat{\Sigma}_k \boldsymbol{\beta} + p_{\lambda_\beta}(\boldsymbol{\beta}), \tag{5}$$

where

$$l(\boldsymbol{\beta}, f(\cdot)) = \sum_{k=1}^{K}\sum_{i=1}^{n_k}\left(y_{ki} - \boldsymbol{z}_{ki}^{\mathrm{T}}\boldsymbol{\beta} - f(v'_{ki})\right)^2.$$

It is clear that we can build the unified model by minimizing (5) and solve $\boldsymbol{\beta}$ and $f(\cdot)$ as in Sections 2.1 and 2.2.

3. Asymptotic properties

In this section, we will establish both the estimation and variable selection consistency of the proposed IPLM-based inferential method. We also establish asymptotic normality of the estimates of the parameters in the linear component of the IPLMs. We assume the covariance matrix Σ_k associated with measurement error is known. The situation of Σ_k being unknown can be similarly proved as in Liang and Li [21]. Without loss of generality, we assume that $\beta_j^* = 0$ for $j > p_{n,0}$ and $\alpha_{kj}^* = 0$ for $j > p_{n,0}$, where $p_{n,0}$ is some integers smaller than p_n that may diverge to infinity as $n \to \infty$. The following eight technical assumptions are made first.

Assumption 3.1: The support for W_k and m_k are bounded for $k = 1,\dots,K$.

Assumption 3.2: The bandwidth in estimating $m_{ky}(\cdot)$ and $\boldsymbol{m}_{kz}(\cdot)$ are of order $n^{-\frac{1}{5}}$.

Assumption 3.3: The covariance matrix of $\boldsymbol{X}_k$ given W_k and Σ_k are positive definite and have constant eigenvalues for $k = 1,\dots,K$.

Assumption 3.4: The density function of W_k and the density function of (Y_k, W_k) are bounded away from 0 and have bounded continuous second derivatives.

Assumption 3.5: $m_{ky}(\cdot)$ and $\boldsymbol{m}_{kz}(\cdot)$ have bounded and continuous second derivative.

Assumption 3.6: The batch effect $g_k(\cdot;\boldsymbol{\psi}_k)$ is a continuous function and continuously differentiable over $\boldsymbol{\psi}_k$ for $k = 1,\dots,K$.

Assumption 3.7: $n_k = O(n)$ for $k = 1,\dots,K$.

Assumption 3.8: $p_n = p[n^a]$ where $0 \leq a < 1/3, p$ is a positive integer $[n^a]$ is the integer part of n^a.

Assumptions 3.1–3.5 are included and discussed in Liang and Li [21]. Assumption 3.6 ensures that the batch effect adjustment using least square method is adequate. Assumption 3.7 ensures that the sample size across different centres is comparable. No centre has a dominating or negligible sample size. Assumption 3.8 indicates that dimension p_n may grow to infinity but with a smaller order than the sample size n. Note that the newly proposed IPLM is more general and more complex than the linear models commonly studied in the literature [26]. To ensure the consistency of the estimates of nonparametric component we need $p_n = p[n^a]$ with $0 \leq a < 1/3$. The proof also works for $a = 0$ where p_n is fixed. Under this assumption, we also establish asymptotic normality of the estimates of the parameters in the linear component of the IPLMs in Theorem 3.4.

Theorem 3.1: *Under Assumptions* 3.1–3.8*, we have* $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\| = O_p((n/p_n)^{-\frac{1}{2}})$, $\|\widehat{\boldsymbol{\alpha}}_k - \boldsymbol{\alpha}_k^*\| = O_p((n/p_n)^{-\frac{1}{2}})$ *and* $E|\widehat{f}_k(\cdot) - f_k^*(\cdot)| = O_p(\max\{n^{-\frac{1}{4}}, (n/p_n^3)^{-\frac{1}{2}}\})$ *for any* $k = 1, \ldots, K$ *if* $\lambda_\beta (n/p_n)^{-\frac{1}{2}} \to 0$ *and* $\lambda_\alpha (n/p_n)^{-\frac{1}{2}} \to 0$ *as* $n \to \infty$.

Theorem 3.1 ensures that we can estimate both the linear and nonlinear components consistently even though there are batch effect in nonlinear predictor and can also have measurement error in the linear predictor vector. More importantly, from the theoretical proof in the appendix, it is clear that the estimates for both the linear and nonlinear components are inconsistent without batch effect adjustment. Thus in real applications, we must correct the batch effect before model fitting, e.g., adjusting the cyclic seasonal pattern of Aβ protein in AD research. Otherwise, the study findings are potentially biased.

Theorem 3.2: *Under Assumptions* 3.1–3.8*, we have* $\lim_{n\to\infty} P(\widehat{\beta}_j = 0) = 1$ *and* $\lim_{n\to\infty} P(\widehat{\alpha}_{kj} = 0) = 1$ *for* $k = 1, \ldots, K$ *and* $j > p_{n,0}$ *if* $\lambda_\beta (n/p_n)^{-\frac{1}{2}} \to 0$, $\lambda_\alpha (n/p_n)^{-\frac{1}{2}} \to 0$, $\lambda_\beta p_n^{-1} \to \infty$ *and* $\lambda_\alpha p_n^{-1} \to \infty$ *as* $n \to \infty$.

Theorem 3.2 ensures that, in probability, all the informative mean and heterogeneous effects can be identified while all non-informative predictors can be excluded in the presence of general batch effect and measurement error in covariates despite the presence of combinations of nonlinearity, general batch effects, measurement errors, heterogeneity of group compositions between centres, and high-dimensional predictors.

Similar to Theorem 3.1 and Theorem 3.2, both the estimation and variable selection consistency hold when dimension p_n is fixed, i.e., $a = 0, p_n = p$. Without loss of generality, we assume that $\beta_j^* = 0$ for $j > p_0$ and $\alpha_{kj}^* = 0$ for $j > p_0$, where p_0 is some integers smaller than p.

Theorem 3.3: *Under Assumptions* 3.1–3.7 *with* $p_n = p$*, we have* $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\| = O_p(n^{-\frac{1}{2}})$, $\|\widehat{\boldsymbol{\alpha}}_k - \boldsymbol{\alpha}_k^*\| = O_p(n^{-\frac{1}{2}})$ *and* $E|\widehat{f}_k(\cdot) - f_k^*(\cdot)| = O_p(n^{-\frac{1}{4}})$, $\lim_{n\to\infty} P(\widehat{\beta}_j = 0) = 1$ *and* $\lim_{n\to\infty} P(\widehat{\alpha}_{kj} = 0) = 1$ *for any* $k = 1, \ldots, K$ *and* $p > p_0$ *if* $\lambda_\beta n^{-\frac{1}{2}} \to 0$, $\lambda_\alpha n^{-\frac{1}{2}} \to 0$, $\lambda_\beta \to \infty$ *and* $\lambda_\alpha \to \infty$ *as* $n \to \infty$.

Denote the linear predictors' effect for each centre as $\boldsymbol{\beta}_k = \boldsymbol{\beta} + \boldsymbol{\alpha}_k$. Let $\boldsymbol{\beta}_{kI}$, $\boldsymbol{X}_{kI,}$ and $\boldsymbol{U}_{kI}$ to be the first p_0 elements of $\boldsymbol{\beta}_k$, $\boldsymbol{X}_k$ and $\boldsymbol{U}_k$, Σ_{kI} to be the (p_0, p_0) left upper matrix of Σ_k, $\Sigma^k_{X|W} = cov(\boldsymbol{X}_{kI} - E(\boldsymbol{X}_{kI}|W_k))$. As $n \to \infty$, $\widehat{\boldsymbol{\beta}}_{kI}$ are asymptotic normally distributed.

Theorem 3.4: *Under Assumptions* 3.1–3.7 *with* $p_n = p$, *if* $\lambda_\beta n^{-\frac{1}{2}} \to 0, \lambda_\alpha n^{-\frac{1}{2}} \to 0, \lambda_\beta \to \infty$ *and* $\lambda_\alpha \to \infty$ *as* $n \to \infty$, *for any* $k = 1, \dots, K$, *we have* $\sqrt{n_k}\Sigma^k_{X|W}(\widehat{\boldsymbol{\beta}}_{kI} - \boldsymbol{\beta}^*_{kI}) \to N(\boldsymbol{0}, \Gamma_k)$, *where* $\Gamma_k = E\{(\boldsymbol{X}_{kI} - E(\boldsymbol{X}_{kI}|W_k))(\epsilon_k - \boldsymbol{U}^{\mathrm{T}}_{kI}\boldsymbol{\beta}^*_{kI}) + \epsilon_k\boldsymbol{U}_{kI} + (\Sigma_{kI} - \boldsymbol{U}_{kI}\boldsymbol{U}^{\mathrm{T}}_{kI})\boldsymbol{\beta}^*_{kI}\}^{\otimes 2}$

The proofs of the above theorems can be found in the Appendix while the proof of Theorem 3.3 was essentially the same as proof of Theorem 3.1 and 3.2, which was omitted.

4. Numerical studies

In this section, we report numerical studies conducted to demonstrate the effectiveness of the proposed IPLM and its associated analysis algorithms. We studied numerical performance in cases where covariates can have either homogeneous or heterogeneous effects across centres with covariate dimensions comparable to the sample size. Also, the conventional pooled data analysis (e.g., using z-scores) or individual-centre based analysis ignores the interplay between nonlinearity and group composition heterogeneity, batch effect, measurement error, and other data incoherence in multi-centre setting thus can cause biased regression estimates and misleading outcomes as illustrated in numerical examples and graphically displayed in Figure 1.

4.1. Covariates with homogeneous effects across groups

In this section, we present numerical studies considering a two group IPLM, i.e., $K = 2$, and

$$\begin{aligned} Y_k &= \boldsymbol{X}^{\mathrm{T}}_k(\boldsymbol{\beta}^* + \boldsymbol{\alpha}^*_k) + f^*_k(W_k) + \epsilon_k, \\ \boldsymbol{Z}_k &= \boldsymbol{X}_k + \boldsymbol{U}_k, \\ V_k &= W_k + g_k(m_k; \boldsymbol{\psi}^*_k) \quad \text{for} \quad 1 \le k \le 2. \end{aligned}$$

Here $\boldsymbol{X}$ is from a p_n-dimensional normal distribution, i.e., $\boldsymbol{X}_{1i} \sim N(\boldsymbol{0}, \Sigma)$, $\boldsymbol{X}_{2i} \sim N(\boldsymbol{0}, \Sigma)$, and Σ is the AR(1) matrix with parameter 0.5, i.e., the (j, l)th element of Σ is $0.5^{|j-l|}$, which implies that the linear predictors are dependent. The measurement error $\boldsymbol{U}_k \sim N(\boldsymbol{0}, \Sigma_k)$, where $\Sigma_k = \Delta_k/3$ and Δ_k are AR(1) matrix with parameter ρ for $k = 1$, 2, i.e., the (j, l)th element of Δ_k is $\rho^{|j-l|}$. To estimate Σ_k, two replicates of $\boldsymbol{Z}_k$, i.e., $\boldsymbol{Z}_k$ and $\boldsymbol{Z}^R_k$, are generated. Linear coefficients $\boldsymbol{\beta}^*$, $\boldsymbol{\alpha}^*_1$ and $\boldsymbol{\alpha}^*_2$ are p_n dimensional vectors, i.e., $\boldsymbol{\beta}^* = (0.5, 0.5, 0.5, 0, \dots, 0)$, $\boldsymbol{\alpha}^*_1 = \boldsymbol{\alpha}^*_2 = (0, 0, \dots, 0)$, which implies that all predictors' effects are homogeneous. ϵ_k is the error term with normal distribution, i.e., $\epsilon_{1i} \sim N(0, \sigma^2)$, $\epsilon_{2i} \sim N(0, \sigma^2)$ and the nonlinear functions for the two centres are:

$$f^*_1(W) = f^*_2(W) = (W - 1)^2.$$

Moreover, W_k has uniform distributions:

$$W_{1i} \sim U(0, 1), \quad W_{2i} \sim U(1, 2),$$

Table 1. Variable selection performance for simulation scenario 4.1.

(n, p_n, σ, ρ)	Method	NM	ZM	NH	ZH
$(500, 100, 0.25, 0.25)$	Ada Lasso	100%	0.00%	0.00	0.00
	Lasso	100%	0.00%	0.00	0.00
$(500, 100, 0.5, 0.5)$	Ada Lasso	100%	0.00%	0.00	0.00
	Lasso	100%	0.00%	0.00	0.00
$(500, 250, 0.25, 0.25)$	Ada Lasso	100%	0.01%	0.00	0.00
	Lasso	100%	0.00%	0.00	0.00
$(500, 250, 0.5, 0.5)$	Ada Lasso	100%	0.00%	0.00	0.00
	Lasso	100%	0.00%	0.00	0.00

which indicates that the group compositions are heterogeneous. For batch effect, we have

$$V_{1i} = W_{1i} + 1.6\sin(m_{1i}), \quad V_{2i} = W_{2i} + 1.6\sin(m_{2i}),$$

where $m_{1i} \sim U(0, \pi)$ and $m_{2i} \sim U(\pi, 2\pi)$.

We generate n observations from each group. Instead of directly observing $(Y_k, \boldsymbol{X}_k, W_k)$, we only observe $(Y_k, \boldsymbol{Z}_k, \boldsymbol{Z}_k^R, m_k, V_k)$ in each group. There are multiple statistical challenges in this simulated example. First, the effect of predictor W_k is nonlinear, e.g., effect of some DNA methylations in ovarian cancer risk prediction [15]. Second, the nonlinear predictor V_k contains batch effects, e.g., the cyclic seasonal pattern of Aβ protein [11]. Third, the group compositions are heterogeneous over W_k. In addition, the linear predictors are measured with measurement error.

We first apply the linear regression model

$$v_{ki} = a_{ki} + b_k\sin(m_{ki})$$

to get estimates $\widehat{b}_k$ and correct the batch effects by $v'_{ki} = v_{ki} - \widehat{b}_k\sin(m_{ki})$ to get bias-free nonlinear predictor V'_k. Next we use the two replicates of $\boldsymbol{Z}_k$, i.e., $(\boldsymbol{Z}_k, \boldsymbol{Z}_k^R)$, to estimate Σ_k for $k = 1, 2$. Then we fit IPLM (1) using $(Y_k, \boldsymbol{Z}_k, V'_k, \widehat{\Sigma}_k)$ to select the informative variables. We apply the Bayesian information criteria [27] (BIC) to select the best tuning parameter and set $n^{-\frac{1}{5}}$ as the bandwidth in local linear regression when solving the nonlinear components. In fact, both the variable selection, linear and nonlinear components estimation are stable around the selected bandwidth $n^{-\frac{1}{5}}$ in our numerical study.

Each scenario is duplicated for $B = 50$ times and the variable selection by Lasso and Adaptive Lasso approach for both mean and heterogeneous effects are summarized in Table 1. Specifically, NM indicates the average percentage of the selected nonzero entries in the mean vector $\boldsymbol{\beta}^*$, ZM indicates the average percentage of the selected zero entries in the mean vector $\boldsymbol{\beta}^*$, NH indicates the average number of the selected nonzero entries in the heterogeneous vector $\boldsymbol{\alpha}_k^*$, ZH indicates the average number of the selected zero entries in the heterogeneous vector $\boldsymbol{\alpha}_k^*$. It is obvious that the variable selection performance by Lasso and Adaptive Lasso is both excellent because it is very close to the oracle where NM $= 100\%$, ZM $= 0.00\%$, NH $= 0.00$ and ZH $= 0.00$.

Using the above simulation set up, the average mean squared error (MSE) of estimated effect for group 1 and group 2 (i.e., $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\beta}}_2$) was summarized in Table 2. Specifically, the

Table 2. Estimation performance for simulation scenario 4.1.

(n, p_n, σ, ρ)	Method	MSE_{β_1}	MSE_{β_2}
$(500, 100, 0.25, 0.25)$	Ada Lasso	0.007	0.007
	Lasso	0.007	0.007
$(500, 100, 0.5, 0.5)$	Ada Lasso	0.008	0.007
	Lasso	0.007	0.007
$(500, 250, 0.25, 0.25)$	Ada Lasso	0.006	0.006
	Lasso	0.006	0.006
$(500, 250, 0.5, 0.5)$	Ada Lasso	0.008	0.008
	Lasso	0.008	0.008

average MSE of estimated group 1 effect $\widehat{\boldsymbol{\beta}}_1$ and group 2 effect $\widehat{\boldsymbol{\beta}}_2$ was defined as

$$MSE_{\beta_1} = \frac{1}{B}\sum_{b=1}^{B}\left\|\widehat{\boldsymbol{\beta}}_1^b - \boldsymbol{\beta}_1^*\right\|_2^2 \quad and \quad MSE_{\beta_2} = \frac{1}{B}\sum_{b=1}^{B}\left\|\widehat{\boldsymbol{\beta}}_2^b - \boldsymbol{\beta}_2^*\right\|_2^2,$$

where $\widehat{\boldsymbol{\beta}_1}^b$ and $\widehat{\boldsymbol{\beta}}_2^b$ were the estimated effect from bth simulated data set. Moreover, the mean, empirical standard deviation (denoted as SD1) and average standard deviation estimated using Theorem 3.4 (denoted as SD2) of $\widehat{\beta}_{11}$ (estimated value of first element of $\boldsymbol{\beta}_1$) and $\widehat{\beta}_{21}$ (estimated value of first element of $\boldsymbol{\beta}_2$), and the coverage of the 95% confidence interval estimated using Theorem 3.4 for β_{11} and β_{21} were summarized in Table 3. Note that coverage of the 95% CI is estimated by replicating each scenario for 1000 times given the status of truly informative/non-informative variables to achieve computational efficiency. It is reasonable to assume knowing status of the truly informative/non-informative variables given variable selection performance is almost perfect in Table 1. As is well known, both Lasso and adaptive Lasso are widely used in practice. We also used both adaptive Lasso and Lasso in our simulation studies. We mostly focus on adaptive Lasso in our theorems because the adaptive Lasso has desirable selection and estimation properties asymptotically as established by Zou (2006) and others for generalized linear models. In terms of estimation, when some true regression coefficient is much larger compared to other coefficients of the informative predictors, the extremely large coefficient can be heavily penalized in Lasso leading to potentially excessive bias for the Lasso estimate of the particular parameter. In comparison, the adaptive Lasso estimation can avoid the severe bias due to excessive Lasso penalty for such large parameter values. Of course, one would need to identify some preliminary consistent estimate to properly use the adaptive Lasso. An initial preliminary consistent estimate may not be easy to find in some applications. Also, when all informative variables have the same effect sizes (same coefficients), the adaptive Lasso estimates would have no essential advantages over the ordinary Lasso as demonstrated by our numerical simulation and summarized in Table 2. It is clear that the estimation performance of the adaptive Lasso approach is essentially equivalent to the Lasso in terms of similar MSE in this setting. Moreover, from Table 3, the empirical standard deviation SD1 and estimated standard deviation SD2 estimated using Theorem 3.4 are quite close, and the coverage probability of the 95% CI predicted by the asymptotic normality theory for both β_{11} and β_{21} are close to 95%.

Table 3. Variance and coverage for simulation scenario 4.1.

(n, p_n, σ, ρ)	Parameter	Method	Mean	SD1	SD2	95% CI Coverage*
(500, 100, 0.25, 0.25)	$\widehat{\beta}_{11}$	Ada Lasso	0.505	0.048	0.040	94.7%
		Lasso	0.503	0.047	0.040	94.7%
	$\widehat{\beta}_{21}$	Ada Lasso	0.501	0.049	0.042	95.1%
		Lasso	0.503	0.047	0.040	95.1%
(500, 100, 0.5, 0.5)	$\widehat{\beta}_{11}$	Ada Lasso	0.496	0.044	0.051	97.3%
		Lasso	0.496	0.044	0.051	97.3%
	$\widehat{\beta}_{21}$	Ada Lasso	0.496	0.044	0.048	97.3%
		Lasso	0.496	0.044	0.050	97.3%
(500, 250, 0.25, 0.25)	$\widehat{\beta}_{11}$	Ada Lasso	0.503	0.046	0.040	95.2%
		Lasso	0.503	0.047	0.040	95.2%
	$\widehat{\beta}_{21}$	Ada Lasso	0.503	0.046	0.040	94.9%
		Lasso	0.503	0.047	0.040	94.9%
(500, 250, 0.5, 0.5)	$\widehat{\beta}_{11}$	Ada Lasso	0.498	0.055	0.050	96.8%
		Lasso	0.498	0.055	0.050	96.8%
	$\widehat{\beta}_{21}$	Ada Lasso	0.498	0.055	0.052	96.8%
		Lasso	0.498	0.055	0.051	96.8%

*Coverage estimated from 1000 replicates given status of the truly informative/non-informative variables.

Individual-group analysis can lead to contradictory findings: If we analyse data for each group (or centre) separately, we find that the effect of V is negative in centre 1 while positive in centre 2 in Figure 1. Thus we get contradictory study findings from different individual group analyses. This is due to the heterogeneous group compositions of W_k. More importantly, the contradictory and misleading study findings cannot be avoided as the sample size increases. This demonstrates the disadvantages of single group study, i.e. , the single centre/group model study findings cannot always be generalized to other study groups. In Figure 1, we only provide the results for one scenario, other seven scenarios show similar pattern of results. In real applications, W_k is typically related to another measurable variable η_k such as age in AD research. Thus we might be able to use frequency matching or using propensity score matching on W_k over η_k to remove the impact of heterogeneous group composition.

Direct pooled-data analysis can lead to misleading pattern: If we simply pool the data from two groups together without suitable batch effect adjustment, the estimated nonlinear curve (marked by the dashed line) is provided in Figure 1. The estimated nonlinear curve, which first shows a positive upward trend and then shows a negative downward trend, is opposite to the true pattern which is a strictly convex curve. This happens due to a simple pooling of data without adjusting for batch effects. Thus in real applications, we must account for the batch effects of the predictors, e.g., the seasonal cyclic pattern of CSF Aβ in AD research. Otherwise, we may get biased estimates and misleading study findings.

IPLM leading to superior predictive performance: Because all predictors' effects are homogeneous, our method produces the unified model for two groups together. We first apply linear regression model to adjust the batch effects of V_k and get batch effect adjustment nonlinear predictor V'_k within each group. Then we combine two groups together and get a unified model. The estimated nonlinear curve is shown in Figure 1, marked by the solid line. The estimation performance of our proposed method is much better than other competitors because our estimated nonlinear curve is very close to the true curve, which is marked by dot-dash line in the figure. This indicates that the estimation performance of our proposed IPLM is superior to other competitors.

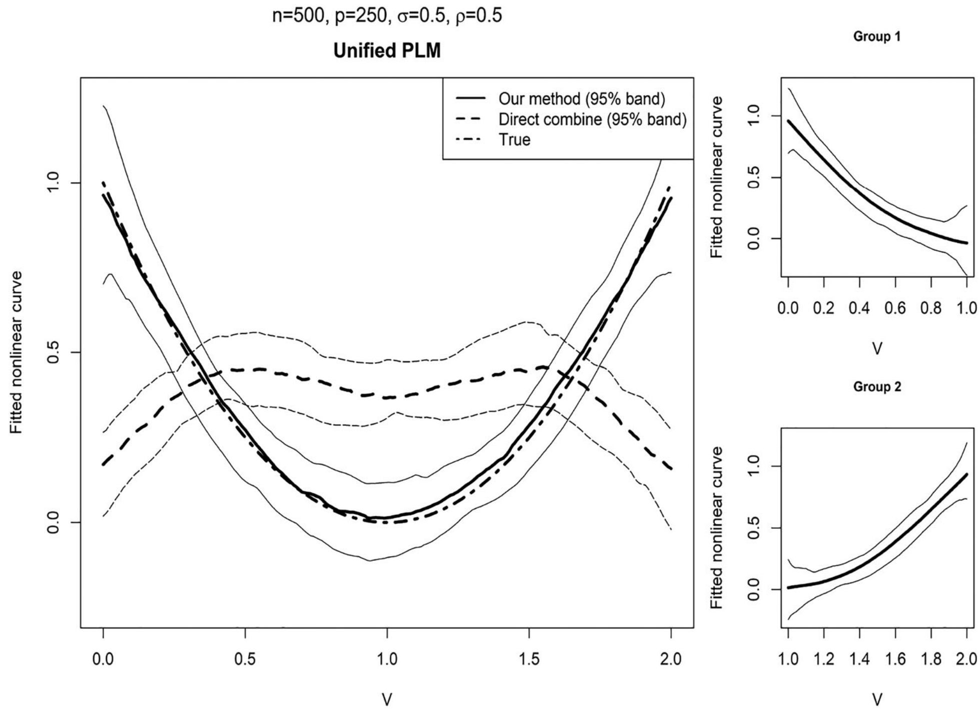

Figure 1. Nonlinear curve estimation in simulation study.

Table 4. Variable selection performance for simulation scenario 4.2.

(n, p_n, σ, ρ)	Method	NM	ZM	NH	ZH
(500, 100, 0.25, 0.25)	IPLM Ada Lasso	100%	0.00%	2.00	0.00
	IPLM Lasso	100%	0.00%	2.00	0.00
(500, 100, 0.5, 0.5)	IPLM Ada Lasso	100%	0.00%	2.00	0.00
	IPLM Lasso	100%	0.02%	2.00	0.00
(500, 250, 0.25, 0.25)	IPLM Ada Lasso	100%	0.00%	2.00	0.00
	IPLM Lasso	100%	0.00%	2.00	0.00
(500, 250, 0.5, 0.5)	IPLM Ada Lasso	100%	0.00%	2.00	0.00
	IPLM Lasso	100%	0.00%	2.00	0.00

4.2. Covariates with heterogeneous effects across groups

The data generating process is the same as scenario 4.1, except that $\boldsymbol{\beta}^* = (3, 3, 3, 0, \ldots, 0)$, $\boldsymbol{\alpha}_1^* = (2, 2, 0, \ldots, 0)$, $\boldsymbol{\alpha}_2^* = (-2, -2, 0, \ldots, 0)$, implying that some predictors' effects are heterogeneous. The variable selection, parameter estimation and 95% CI coverage performance are showed in Table 4, 5, and 6. It is obvious that the variable selection performance by Lasso and Adaptive Lasso are both excellent in Table 4 because it is very close to the oracle where NM $=$ 100%, ZM $=$ 0.00%, NH $=$ 2.00 and ZH $=$ 0.00. For parameter estimation and coverage of 95% CI predicted by the asymptotic normality theory in scenario 4.2, the performance of both Lasso and Adaptive Lasso is quite good as reported in Table 6, which is consistent with the findings in scenario 4.1 when all effects are homogeneous.

Table 5. Estimation performance for simulation scenario 4.2.

(n, p_n, σ, ρ)	Method	MSE_{β_1}	MSE_{β_2}
(500, 100, 0.25, 0.25)	Ada Lasso	0.259	0.090
	Lasso	0.259	0.090
(500, 100, 0.5, 0.5)	Ada Lasso	0.366	0.101
	Lasso	0.376	0.111
(500, 250, 0.25, 0.25)	Ada Lasso	0.279	0.060
	Lasso	0.279	0.060
(500, 250, 0.5, 0.5)	Ada Lasso	0.352	0.113
	Lasso	0.352	0.113

Table 6. Variance and coverage for simulation scenario 4.2.

(n, p_n, σ, ρ)	Parameter	Method	Mean	SD1	SD2	95% CI Coverage*
(500, 100, 0.25, 0.25)	$\widehat{\beta}_{11}$	Ada Lasso	5.057	0.302	0.327	96.3%
		Lasso	5.057	0.302	0.327	96.3%
	$\widehat{\beta}_{21}$	Ada Lasso	0.988	0.167	0.130	93.4%
		Lasso	0.988	0.167	0.130	93.4%
(500, 100, 0.5, 0.5)	$\widehat{\beta}_{11}$	Ada Lasso	5.060	0.375	0.370	93.8%
		Lasso	5.060	0.375	0.370	93.8%
	$\widehat{\beta}_{21}$	Ada Lasso	0.982	0.164	0.143	92.6%
		Lasso	0.982	0.164	0.143	92.6%
(500, 250, 0.25, 0.25)	$\widehat{\beta}_{11}$	Ada Lasso	4.984	0.349	0.324	94.7%
		Lasso	4.984	0.349	0.324	94.7%
	$\widehat{\beta}_{21}$	Ada Lasso	1.012	0.125	0.130	92.9%
		Lasso	1.012	0.125	0.130	92.9%
(500, 250, 0.5, 0.5)	$\widehat{\beta}_{11}$	Ada Lasso	5.014	0.416	0.359	94.5%
		Lasso	5.014	0.416	0.359	94.5%
	$\widehat{\beta}_{21}$	Ada Lasso	0.996	0.166	0.143	94.8%
		Lasso	0.996	0.166	0.143	94.8%

*Coverage is estimated from 1000 replicates given status of the truly informative/non-informative variables.

5. Real-data illustration

In this section, we illustrate how to apply the proposed IPLM-based analysis to rigorously analyse biomarker data in a multi-centre Alzheimer's disease (AD) research project. The hallmarks of AD are the inter-neuron plaques and within-neuron neurofibrillary tangles (NFT) in patients' brain as discovered originally by Dr. Alzheimer in 1906. As is well known, the amyloid beta 42 (Aβ_{42}) and tau proteins in brain underlie the plaques and NFT, respectively. Moreover, the existence of within-neuron NFT indicates the dysfunction and/or death of neuron cells, thus high CSF tau is one of the most important biomarkers of neurodegeneration and risk biomarkers for AD [28,29]. Additionally, among elderly persons who are at risk for AD, reductions in the CSF Aβ_{42} are associated with brain Aβ_{42} deposition and often precede elevations in CSF tau levels. Therefore, there has been persistent interest to investigate the relationship between CSF Aβ_{42} and CSF tau protein and CSF tau can be used as a surrogate outcome variable in these AD research. Moreover, the biological function of Aβ_{42} in AD is complicated and thus we include it as a nonlinear predictor of the CSF tau. Some other widely used variables, e.g., age, gender, and APoE4ε status, are included as linear predictors of the partially linear model.

In our analysis, two of the study centres used in de Leon et al. [30], i.e., New York University (NYU) database and Alzheimer's Disease Neuroimaging Initiative (ADNI) database, are included. The NYU database contains 331 observations and ADNI database contains

335 observations. Because the nonlinear predictor Aβ_{42} is measured with batch effects, i.e., cyclic seasonal pattern over measurement time as displayed in Figure 2, we must adjust the batch effects of Aβ_{42} first for data harmonization. In the NYU and ADNI databases, we fit the observed Aβ_{42} values over its measurement time t (in month) with a sine wave A$\beta_{42} = \gamma + \gamma_1 \sin(t + \theta)$ to identify the cyclic seasonal pattern. We use least square method to estimate γ_1 and θ within each study centre and both the sine waves in NYU and ADNI are statistically significant. We then correct the cyclic seasonal pattern to remove batch effects and obtain the corrected Aβ_{42} values via:

$$\mathrm{A}\beta'_{42} = \mathrm{A}\beta_{42} - \widehat{\gamma_1}\sin(t + \widehat{\theta}).$$

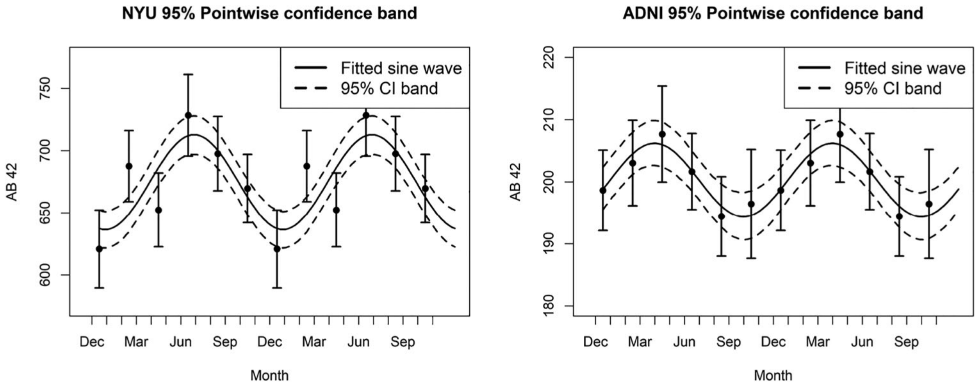

Figure 2. Batch effect of Aβ_{42}.

If we investigated the relationship between CSF Aβ_{42} and tau protein within each centre alone separately, we will get seemingly contradictory findings as seen in Figure 3. From the upper left sub-figure of Figure 3, the NYU centre data mainly shows an increasing trend while from the lower left sub-figure we can see that the ADNI centre data mainly shows a decreasing trend. This seemingly increasing versus decreasing contradictory pattern is mainly due to the heterogeneous age composition in the NYU and ADNI study cohorts. In fact, the NYU cohort has a large number of young adults and small portion of old adults while the ADNI cohort has only older adults. Based on data from one study centre only, e.g., the ADNI cohort, one might easily reach the conclusion of an monotone relationship between CSF tau and CSF Aβ_{42} which is not generally true and clearly does not apply to the NYU cohort. This real-data example demonstrates the limitation and disadvantage of commonly used single-centre analysis in producing not generalizable and even misleading findings.

We apply the proposed IPLM (1) and find that all predictors' effects, including age, gender and APoE4ε status, can be regarded as homogeneous between the two centres: NYU and ADNI. Also, the biological relationship between CSF tau and Aβ proteins and the mechanism should be largely identical across different centres [30]. Thus it is natural to combine two study centres together and build a unified model. However, the two centres have different age distribution. More specifically, the NYU group is younger than the ADNI group. It is known from the literature that the relationship of CSF Aβ_{42} and tau protein between younger adults and older adults differs [19]. Due to the heterogeneous

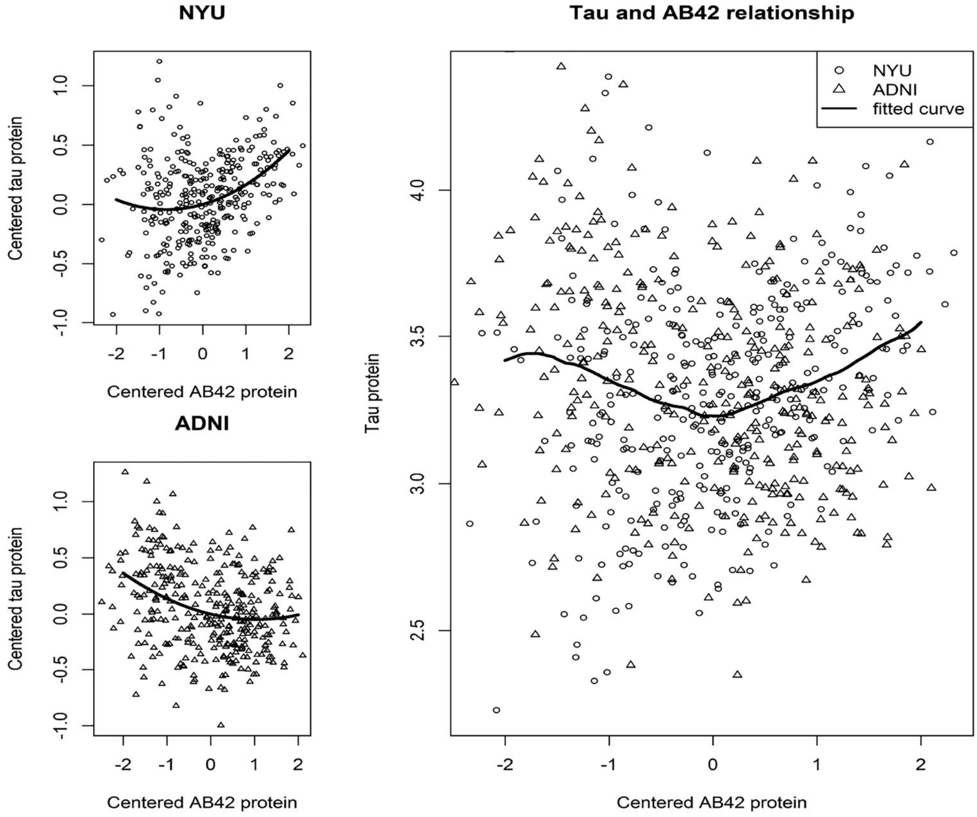

Figure 3. Individual group and combined group analysis.

age distribution between the study centres, if we use simple pooled analysis, i.e. pooling *z*-scores from the two centres as commonly used in the existing literature, the results would be biased. To overcome the impact of the differential age distribution between NYU and ADNI, we use an age-based frequency matching method to combine the biomarker data together. After combining data, we fit a unified model:

$$\text{tau}_i = \beta_1 \text{Age}_i + \beta_2 \text{APoE4}\varepsilon_i + \beta_3 \text{Gender}_i + f^*(\text{A}\beta'_{42i}) + \epsilon_i.$$

The fitted nonlinear curve $f^*(\cdot)$, i.e. the effect of Aβ_{42} on tau protein, is shown in Figure 3. From Figure 3, we find that the effect of Aβ_{42} on tau protein is clearly nonlinear. More specifically, as the value of Aβ_{42} increases, the value of tau protein decreases first and then increases. Moreover, the study findings, i.e. the nonlinear relationship between Aβ_{42} and tau protein, can be expected to be more robust and achieve higher generalizability and reproducibility because we used data from two independent study centres and the nonlinear curve fit both cohorts quite well as displayed in Figure 3. It should also be pointed out that nonlinearity of predictors and batch effects in measuring biomarkers widely exists between study centres. As extra examples, in AD research, Chen et al. [31] showed that the effect of APoE4ϵ on the rate of decline from subjects with mild cognitive impairment (MCI) to AD is not linear and a segmented linear model is used to model the longitudinal trend instead. Also the procedure to obtain the CSF Aβ_{42} values is more invasive than measuring the blood-based Aβ_{42} level. Therefore, AD researchers across different centres have started using plasma Aβ_{42} as non-invasive biomarkers. As reviewed in Figure 1 of Pannee et al. [13], the plasma Aβ_{42} values have very low between-centre correlations indicating major batch effects across centres that need to be harmonized carefully to produce reproducible findings [13]. In short, this AD-research real-data example demonstrates that the commonly used single-centre analysis can produce non-generalizable findings while our newly proposed IPLM-based approach can be used in multi-centre studies to automatically account for nonlinear predictors and adjust for batch effect and heterogeneity in centre compositions yielding generally valid findings.

6. Summary

Achieving generalizable and reproducible findings in multi-centre studies is of great importance in research. However, a general, rigorous and flexible statistical analysis method to account for combinations of common complexities associated with multi-centre studies has been lacking. In this manuscript, we introduced the integrated partially linear model (IPLM) and associated analysis methods. The proposed IPLM-based analysis can account for interplays of multiple complexities commonly exist in modern multi-centre studies, e.g., predictors having potentially nonlinear effects and heterogeneous group compositions, being measured with batch effects and/or potential measurement errors. We proposed the removal of batch effect in the data-harmonization step of the multi-centre study. We suggested a local linear regression-based constrained regularization estimation method with a computationally fast implementation of the newly proposed IPLM. The proposed regularized optimization method can automatically identify the predictors' effects that can be either homogeneous and/or heterogeneous, and can naturally yield a unified parsimonious model when all predictors' effects are homogeneous across study centres. We provided simulation examples to demonstrate the effectiveness of proposed IPLM and analysis method for variable selection and parameter estimation when covariates can have either homogeneous or heterogeneous effects across study centres. We illustrated the major biases and misleading findings from the conventional individual-group based analysis and the commonly used z-score-based data-pooling method without effective batch-effect adjustments and accounting for composition heterogeneity. Importantly, we have established estimation consistency and variable-selection consistency for the proposed method in our theorems where the covariate dimension can diverge as the sample size increases. We have also established asymptotic normality for the regression parameters under some suitable regularity conditions. The real-data application in a multi-centre Alzheimer's disease research project is used to illustrate the utility and effectiveness of proposed IPLM-based analysis in practice. Specifically, the AD-research real-data example was used to demonstrate that the commonly used individual-centre based analysis can produce misleading findings while our newly proposed IPLM-based approach can be used in multi-centre studies to automatically account for nonlinear predictors, heterogeneity in centre compositions, and batch effects in covariates and yield generally valid findings. Also, the IPLM can increase reproducibility by integrating potential batch-effect and/or measurement-error removal as part of the careful regression modelling procedure while, in the existing literature, neglected or casual batch-effect removal in data pooling before any careful statistical modelling often contributes to non-reproducible findings.

Acknowledgements

The authors would like to thank the reviewers and the associate editor for careful reading and for many constructive suggestions. The authors would like to thank Drs. Mony de Leon, Ricardo Osorio, and Elizabeth Pirraglia for sharing with us the NYU Alzheimer's disease data sets used in Section 5 for the illustration of our proposed model and analysis. The NYU study data are available from figshare (https://figshare.com/s/16d233d4822b810bcd9b, DOI: 10.6084/m9.figshare.5758554). One part of the data used in the preparation of the example in Section 5 of this article was obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu/data-samples/access-data/). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in the design, analysis or writing of this report. A complete list of ADNI investigators is at: http://adni.loni.usc.edu/wpcontent/uploads/how to apply/ADNI Acknowledgement List.pdf.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research was partially supported by the United States National Institute of Health grants (NIA grants P30AG066512, P01AG060882, NCI grants P50CA225450, P30CA016087) and Center for Disease Control and Prevention (CDC) grant U01OH012486.

References

[1] Arslan A, Tuminello S, Yang L, et al. Genome-wide dna methylation profiles in community members exposed to the world trade center disaster. Int J Environ Res Public Health. 2020;17(1):5493. doi:10.3390/ijerph17155493

[2] Cruz C, Llop-Guevara A, Garber JE, et al. Multicenter phase II study of lurbinectedin in BRCA-mutated and unselected metastatic advanced breast cancer and biomarker assessment substudy. J Clin Oncol. 2018;36(31):3134. doi:10.1200/JCO.2018.78.6558

[3] He X, Sun X, Shao Y. Network-based survival model to discover target genes for developing cancer immunotherapies and predicting patient survival. J Appl Stat. 2021;48:1352–1373. doi:10.1080/02664763.2020.1812543

[4] Rahbar K, Ahmadzadehfar H, Kratochwil C, et al. German multicenter study investigating 177lu-psma-617 radioligand therapy in advanced prostate cancer patients. J Nucl Med. 2017;58(1):85–90. doi:10.2967/jnumed.116.183194

[5] Roach PJ, Francis R, Emmett L, et al. The impact of 68ga-psma pet/ct on management intent in prostate cancer: results of an Australian prospective multicenter study. J Nucl Med. 2018;59(1):82–88. doi:10.2967/jnumed.117.197160

[6] Sturdza A, Pötter R, Fokdal LU, et al. Image guided brachytherapy in locally advanced cervical cancer: improved pelvic control and survival in retroembrace, a multicenter cohort study. Radiother Oncol. 2016;120(3):428–433. doi:10.1016/j.radonc.2016.03.011

[7] Sun X, Liu X, Xia M, et al. Multicellular gene network analysis identifies a macrophage-related gene signature predictive of therapeutic response and prognosis of gliomas. J Transl Med. 2019;17(1):159. doi:10.1186/s12967-019-1908-1

[8] Boada M, Anaya F, Ortiz P, et al. Efficacy and safety of plasma exchange with 5% albumin to modify cerebrospinal fluid and plasma amyloid-β concentrations and cognition outcomes in Alzheimer's disease patients: a multicenter, randomized, controlled clinical trial. J Alzheimers Dis. 2017;56(1):129–143. doi:10.3233/JAD-160565

[9] Ewers M, Mattsson N, Minthon L, et al. Csf biomarkers for the differential diagnosis of Alzheimer's disease: a large-scale international multicenter study. Alzheimers Dement. 2015;11(11):1306–1315. doi:10.1016/j.jalz.2014.12.006

[10] Khan W, Giampietro V, Banaschewski T, et al. A multi-cohort study of apoe4 and amyloid-β effects on the hippocampus in Alzheimer's disease. J Alzheimers Dis. 2017;56(3):1159–1174. doi:10.3233/JAD-161097

[11] Lim AS, Gaiteri C, Yu L, et al. Seasonal plasticity of cognition and related biological measures in adults with and without Alzheimer disease: analysis of multiple cohorts. PLoS Med. 2018;15(9):Article ID e1002647. doi:10.1371/journal.pmed.1002647

[12] Niemantsverdriet E, Ribbens A, Bastin C, et al. A retrospective Belgian multi-center MRI biomarker study in Alzheimer's disease (remember). J Alzheimers Dis. 2018;63(4):1509–1522. doi:10.3233/JAD-171140

[13] Pannee J, Shaw L, Korecka ea, et al. The global Alzheimer's association round robin study on plasma amyloid beta methods. Alzheimer's Dement. 2021;13:Article ID e12242. doi:10.1002/dad2.v13.1

[14] Van Steenoven I, Aarsland D, Weintraub D, et al. Cerebrospinal fluid Alzheimer's disease biomarkers across the spectrum of Lewy body diseases: results from a large multicenter cohort. J Alzheimers Dis. 2016;54(1):287–295. doi:10.3233/JAD-160322

[15] Ahn S, Wang T. A powerful statistical method for identifying differentially methylated markers in complex diseases. In: 'Biocomputing 2013'. World Scientific; 2013. p. 69–79 .

[16] Härdle W, Liang H, Gao J. Partially linear models. New York, NY: Springer Science & Business Media; 2012.

[17] Chen C, Grennan K, Badner J, et al. Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. PLoS ONE. 2011;6(2):Article ID e17238. doi:10.1371/journal.pone.0017238

[18] Scherer A. Batch effects and noise in microarray experiments: sources and solutions. Vol. 868. New York, NY: John Wiley & Sons; 2009.

[19] Shoji M, Kanai M, Matsubara E, et al. The levels of cerebrospinal fluid aβ40 and aβ42 (43) are regulated age-dependently. Neurobiol Aging. 2001;22(2):209–215. doi:10.1016/S0197-4580(00)00229-3

[20] Hessell A, Li L, Malherbe D, et al. Virus control in vaccinated rhesus macaques is associated with neutralizing and capturing antibodies against the shiv challenge virus but not with v1v2 vaccine-induced anti-v2 antibodies alone. J Immunol. 2021;206:1266–1283. doi:10.4049/jimmunol.2001010

[21] Liang H, Li R. Variable selection for partially linear models with measurement errors. J Am Stat Assoc. 2009;104(485):234–248. doi:10.1198/jasa.2009.0127

[22] Zhao P, Xue L. Variable selection for semiparametric varying coefficient partially linear errors-in-variables models. J Multivar Anal. 2010;101(8):1872–1883. doi:10.1016/j.jmva.2010.03.005

[23] Fan J, Gijbels I. 'Local polynomial modelling and its applications'. 1996.

[24] Zou H. The adaptive lasso and its oracle properties. J Am Stat Assoc. 2006;101(476):1418–1429. doi:10.1198/016214506000000735

[25] Boyd S, Parikh N, Chu E, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trendső in Mach Learn. 2011;3(1):1–122.

[26] Zou H, Zhang HH. On the adaptive elastic-net with a diverging number of parameters. Ann Stat. 2009;37(4):1733–1751. doi:10.1214/08-AOS625

[27] Schwarz G. Estimating the dimension of a model. Ann Stat. 1978;6(2):461–464. doi:10.1214/aos/1176344136

[28] Hansson O, Zetterberg H, Buchhave P, et al. Association between CSF biomarkers and incipient Alzheimer's disease in patients with mild cognitive impairment: a follow-up study. Lancet Neurol. 2006;5(3):228–234. doi:10.1016/S1474-4422(06)70355-6

[29] Herukka S-K, Helisalmi S, Hallikainen M, et al. Csf ab42, tau and phosphorylated tau, apoe4 allele and mci type in progressive mci. Neurobiol Aging. 2007;28(4):507–514. doi:10.1016/j.neurobiolaging.2006.02.001

[30] de Leon MJ, Pirraglia E, Osorio RS, et al. The nonlinear relationship between cerebrospinal fluid ab42 and tau in preclinical Alzheimer's disease. PLoS ONE. 2018;13(2):Article ID e0191240. doi:10.1371/journal.pone.0191240

[31] Chen X, Shao Y, Sadowski M. Segmented linear mixed model analysis reveals association of the apoe e4 allele with faster rate of Alzheimer dementia progression. J Alzheimers Dis. 2021;82(3):921–937. doi:10.3233/JAD-210434

[32] Zhou Z, Jiang R, Qian W. Variable selection for additive partially linear models with measurement error. Metrika. 2011;74(2):185–202. doi:10.1007/s00184-009-0296-6

Appendix: Technical proofs

Proof of Theorem 3.1: First, we adjust the batch effect using linear regression model. By Assumption 3.6, we have

$$|w_{k(i)} - v'_{k(i)}| = |g(m_{ki};\widehat{\boldsymbol{\psi}}_k) - g(m_{ki};\boldsymbol{\psi}^*_k)| \le |g'(m_{ki};\widetilde{\boldsymbol{\psi}}_k)||\widehat{\boldsymbol{\psi}}_k - \boldsymbol{\psi}^*_k| = O_p(n^{-\frac{1}{2}}) \qquad \text{(A1)}$$

for any $i = 1,\dots,n_k$ and $k = 1,\dots,K$, where $\widetilde{\boldsymbol{\psi}}_k \in (\boldsymbol{\psi}^*_k, \widehat{\boldsymbol{\psi}}_k)$. Then by Assumption 3.1– 3.5, the local polynomial estimates satisfy

$$\sup_{V'_k} |\widehat{m}_{ky}(V'_k) - m_{ky}(V'_k)| = o_p(n^{-\frac{1}{4}}) \quad \text{and} \quad \sup_{V'_k} |\widehat{m}_{kzj}(V'_k) - m_{kzj}(V'_k)| = o_p(n^{-\frac{1}{4}}) \qquad \text{(A2)}$$

for $j = 1,\dots,p_n$, where $\widehat{m}_{kzj}(\cdot)$ and $m_{kzj}(\cdot)$ are the jth element of $\widehat{\boldsymbol{m}}_{kz}(\cdot)$ and $\boldsymbol{m}_{kz}(\cdot)$.

By re-parametrization, let $\boldsymbol{\beta}_k = \boldsymbol{\beta} + \boldsymbol{\alpha}_k$. Thus we have $\boldsymbol{\beta} = \sum_{k=1}^K \boldsymbol{\beta}_k/K$ and $\boldsymbol{\alpha}_k = \boldsymbol{\beta}_k - \boldsymbol{\beta}$. Denote $\boldsymbol{\Theta} = (\boldsymbol{\beta}_1^{\mathrm{T}},\dots,\boldsymbol{\beta}_K^{\mathrm{T}})^{\mathrm{T}}$. Then we can rewrite $l_p(\boldsymbol{\beta},\boldsymbol{\alpha})$ (3) as

$$l_p(\boldsymbol{\Theta}) = l(\boldsymbol{\Theta}) - \sum_{k=1}^K n_k \boldsymbol{\beta}_k \Sigma_k \boldsymbol{\beta}_k + p_{\lambda_\beta}(\boldsymbol{\Theta}) + p_{\lambda_\alpha}(\boldsymbol{\Theta}),$$

where $l(\boldsymbol{\Theta}) = \sum_{k=1}^K \sum_{i=1}^{n_k} (\widehat{y}_{ki} - \widehat{\boldsymbol{z}}_{ki}^{\mathrm{T}} \boldsymbol{\beta}_k)^2$, $p_{\lambda_\beta}(\boldsymbol{\Theta}) = \lambda_\beta \sum_{j=1}^{p_n} |K^{-1} \sum_{k=1}^K \beta_{kj}|$ and $p_{\lambda_\alpha}(\boldsymbol{\Theta}) = \lambda_\alpha \sum_{k=1}^K \sum_{j=1}^{p_n} \pi_{kj} |\beta_{kj} - K^{-1} \sum_{k=1}^K \beta_{kj}|$. Denote $\boldsymbol{\theta} = (\boldsymbol{\xi}_1^{\mathrm{T}},\dots,\boldsymbol{\xi}_K^{\mathrm{T}})^{\mathrm{T}}$. Let $r_n = (n/p_n)^{-\frac{1}{2}}$. We show that for any given ζ, there exists a large enough constant C such that

$$P\Big\{ \inf_{\|\boldsymbol{\theta}\|=C} l_p(\boldsymbol{\Theta}^* + r_n\boldsymbol{\theta}) > l_p(\boldsymbol{\Theta}^*) \Big\} \ge 1 - \zeta.$$

Because π_j and π_{kj} are adaptive Lasso weight, we get $|\pi_j| > 0$ and $|\pi_{kj}| > 0$ for any $j \le p_{n,0}$ and $k = 1,\dots,K$, $|\pi_j| = O_p(r_n^{-1})$ and $|\pi_{kj}| = O_p(r_n^{-1})$ for any $j > p_{n,0}$ and $k = 1,\dots,K$. For the penalty terms $p_{\lambda_\beta}(\boldsymbol{\Theta})$ and $p_{\lambda_\alpha}(\boldsymbol{\Theta})$, it is easy to verify that

$$\begin{aligned} p_{\lambda_\beta}(\boldsymbol{\Theta}^* + r_n\boldsymbol{\theta}) - p_{\lambda_\beta}(\boldsymbol{\Theta}^*) &\ge \lambda_\beta \sum_{j=1}^{p_{n,0}} \pi_j \left| \sum_{k=1}^K \beta^*_{kj}/K + r_n \sum_{k=1}^K \xi_{kj}/K \right| - \lambda_\beta \sum_{j=1}^{p_{n,0}} \pi_j \left| \sum_{k=1}^K \beta^*_{kj}/K \right| \\ &\ge -\lambda_\beta r_n \sum_{j=1}^{p_{n,0}} \pi_j \left| \sum_{k=1}^K \xi_{kj}/K \right|. \end{aligned} \qquad \text{(A3)}$$

Similarly, we have

$$p_{\lambda_\alpha}(\boldsymbol{\Theta}^* + r_n\boldsymbol{\theta}) - p_{\lambda_\alpha}(\boldsymbol{\Theta}^*) \ge -\lambda_\alpha r_n \sum_{k=1}^K \sum_{j=1}^{p_{n,0}} \pi_{kj} \left| \xi_{kj} - \sum_{k=1}^K \xi_{kj}/K \right|. \qquad \text{(A4)}$$

Next, for $\delta_l = l(\boldsymbol{\Theta}^* + r_n\boldsymbol{\theta}) - \sum_{k=1}^K n_k(\boldsymbol{\beta}^*_k + r_n\boldsymbol{\theta}_k)^{\mathrm{T}} \Sigma_k (\boldsymbol{\beta}^*_k + r_n\boldsymbol{\theta}_k) - l(\boldsymbol{\Theta}^*) + \sum_{k=1}^K n_k \boldsymbol{\beta}_k^{*\mathrm{T}} \Sigma_k \boldsymbol{\beta}^*_k$, we have

$$\delta_l = -2r_n \sum_{k=1}^K \sum_{i=1}^{n_k} (\widehat{y}_{ki}\widehat{\boldsymbol{z}}_{ki}^{\mathrm{T}} - \widehat{\boldsymbol{z}}_{ki}^{\mathrm{T}} \boldsymbol{\beta}^*_k \widehat{\boldsymbol{z}}_{ki}^{\mathrm{T}} + \boldsymbol{\beta}_k^{*\mathrm{T}} \Sigma_k) \boldsymbol{\xi}_k + r_n^2 \sum_{k=1}^K n_k \boldsymbol{\xi}_k^{\mathrm{T}} (n_k^{-1} \sum_{i=1}^{n_k} \widehat{\boldsymbol{z}}_{ki} \widehat{\boldsymbol{z}}_{ki}^{\mathrm{T}} - \Sigma_k) \boldsymbol{\xi}_k.$$

Now we calculate the order of the first term. Note that $\widehat{y}_{ki}$ and $\widehat{\boldsymbol{z}}_{ki}$ can be decomposed as

$$\begin{aligned}\widehat{y}_{ki} &= y_{ki} - \widehat{m}_{ky}(v'_{ki}) = y_{ki} - m_{ky}(w_{ki}) + m_{ky}(w_{ki}) - m_{ky}(v'_{ki}) + m_{ky}(v'_{ki}) - \widehat{m}_{ky}(v'_{ki}),\\ \widehat{\boldsymbol{z}}_{ki} &= \boldsymbol{z}_{ki} - \widehat{\boldsymbol{m}}_{kz}(v'_{ki}) = \boldsymbol{z}_{ki} - \boldsymbol{m}_{kz}(w_{ki}) + \boldsymbol{m}_{kz}(w_{ki}) - \boldsymbol{m}_{kz}(v'_{ki}) + \boldsymbol{m}_{kz}(v'_{ki}) - \widehat{\boldsymbol{m}}_{kz}(v'_{ki}).\end{aligned} \tag{A5}$$

Denote $\widetilde{y}_{ki} = y_{ki} - m_{ky}(w_{ki})$, $\bar{y}_{ki} = m_{ky}(w_{ki}) - m_{ky}(v'_{ki})$, $\widetilde{\boldsymbol{z}}_{ki} = \boldsymbol{z}_{ki} - \boldsymbol{m}_{kz}(w_{ki})$ and $\bar{\boldsymbol{z}}_{ki} = \boldsymbol{m}_{kz}(w_{ki}) - \boldsymbol{m}_{kz}(v'_{ki})$. Then we can decompose the first term of δ_l as $\sum_{k=1}^{K}\sum_{i=1}^{n_k}\{(\widehat{y}_{ki} - \widetilde{y}_{ki} - \bar{y}_{ki})(\widehat{\boldsymbol{z}}_{ki} - \widetilde{\boldsymbol{z}}_{ki} - \bar{\boldsymbol{z}}_{ki})^{\mathrm{T}} + (\widehat{y}_{ki} - \widetilde{y}_{ki} - \bar{y}_{ki})\widetilde{\boldsymbol{z}}_{ki}^{\mathrm{T}} + (\widehat{y}_{ki} - \widetilde{y}_{ki} - \bar{y}_{ki})\bar{\boldsymbol{z}}_{ki}^{\mathrm{T}} + \widetilde{y}_{ki}(\widehat{\boldsymbol{z}}_{ki} - \widetilde{\boldsymbol{z}}_{ki} - \bar{\boldsymbol{z}}_{ki})^{\mathrm{T}} + \bar{y}_{ki}(\widehat{\boldsymbol{z}}_{ki} - \widetilde{\boldsymbol{z}}_{ki} - \bar{\boldsymbol{z}}_{ki})^{\mathrm{T}} + (\widetilde{y}_{ki}\widetilde{\boldsymbol{z}}_{ki}^{\mathrm{T}} - \widetilde{\boldsymbol{z}}_{ki}^{\mathrm{T}}\boldsymbol{\beta}_k^*\widetilde{\boldsymbol{z}}_{ki}^{\mathrm{T}} + \boldsymbol{\beta}_k^{*\mathrm{T}}\Sigma_k) + \widetilde{y}_{ki}\bar{\boldsymbol{z}}_{ki}^{\mathrm{T}} + \bar{y}_{ki}\widetilde{\boldsymbol{z}}_{ki}^{\mathrm{T}} + \bar{y}_{ki}\bar{\boldsymbol{z}}_{ki}^{\mathrm{T}} - \widetilde{\boldsymbol{z}}_{ki}^{\mathrm{T}}\boldsymbol{\beta}_k^*\bar{\boldsymbol{z}}_{ki}^{\mathrm{T}} - \bar{\boldsymbol{z}}_{ki}^{\mathrm{T}}\boldsymbol{\beta}_k^*\widetilde{\boldsymbol{z}}_{ki}^{\mathrm{T}} - \bar{\boldsymbol{z}}_{ki}^{\mathrm{T}}\boldsymbol{\beta}_k^*\bar{\boldsymbol{z}}_{ki}^{\mathrm{T}} - (\widehat{\boldsymbol{z}}_{ki} - \widetilde{\boldsymbol{z}}_{ki} - \bar{\boldsymbol{z}}_{ki})^{\mathrm{T}}\boldsymbol{\beta}_k^*(\widehat{\boldsymbol{z}}_{ki} - \widetilde{\boldsymbol{z}}_{ki} - \bar{\boldsymbol{z}}_{ki})^{\mathrm{T}} - \widetilde{\boldsymbol{z}}_{ki}^{\mathrm{T}}\boldsymbol{\beta}_k^*(\widehat{\boldsymbol{z}}_{ki} - \widetilde{\boldsymbol{z}}_{ki} - \bar{\boldsymbol{z}}_{ki})^{\mathrm{T}} - \bar{\boldsymbol{z}}_{ki}^{\mathrm{T}}\boldsymbol{\beta}_k^*(\widehat{\boldsymbol{z}}_{ki} - \widetilde{\boldsymbol{z}}_{ki} - \bar{\boldsymbol{z}}_{ki})^{\mathrm{T}} - (\widehat{\boldsymbol{z}}_{ki} - \widetilde{\boldsymbol{z}}_{ki} - \bar{\boldsymbol{z}}_{ki})^{\mathrm{T}}\boldsymbol{\beta}_k^*\widetilde{\boldsymbol{z}}_{ki}^{\mathrm{T}} - (\widehat{\boldsymbol{z}}_{ki} - \widetilde{\boldsymbol{z}}_{ki} - \bar{\boldsymbol{z}}_{ki})^{\mathrm{T}}\boldsymbol{\beta}_k^*\bar{\boldsymbol{z}}_{ki}\}$. By Assumption 3.5 and (A1), we have $\bar{y}_{ki} = O_p(n^{-\frac{1}{2}})$ and $\|\bar{\boldsymbol{z}}_{ki}\| = O_p((n/p_n)^{-\frac{1}{2}})$. Combining $E(\widetilde{y}_{ki}) = 0$, $E(\widetilde{\boldsymbol{z}}_{ki}) = \boldsymbol{0}$, Equation (A2), Assumption 3.7, Lemma A.1 in Liang and Li [21] and only the first $p_{n,0}$ elements in $\boldsymbol{\beta}_k^*$ are nonzero, we have $\sum_{i=1}^{n_k}(\widehat{y}_{ki} - \widetilde{y}_{ki} - \bar{y}_{ki})(\widehat{\boldsymbol{z}}_{ki} - \widetilde{\boldsymbol{z}}_{ki} - \bar{\boldsymbol{z}}_{ki})^{\mathrm{T}} = o_p((np_n)^{\frac{1}{2}})$, $\sum_{i=1}^{n_k}(\widehat{y}_{ki} - \widetilde{y}_{ki} - \bar{y}_{ki})\widetilde{\boldsymbol{z}}_{ki}^{\mathrm{T}} = o_p((np_n)^{\frac{1}{2}})$, $\sum_{i=1}^{n_k}(\widehat{y}_{ki} - \widetilde{y}_{ki} - \bar{y}_{ki})\bar{\boldsymbol{z}}_{ki}^{\mathrm{T}} = o_p((np_n)^{\frac{1}{2}})$, $\sum_{k=1}^{K}\sum_{i=1}^{n_k}\widetilde{y}_{ki}(\widehat{\boldsymbol{z}}_{ki} - \widetilde{\boldsymbol{z}}_{ki} - \bar{\boldsymbol{z}}_{ki})^{\mathrm{T}} = o_p((np_n)^{\frac{1}{2}})$, $\sum_{i=1}^{n_k}\bar{y}_{ki}(\widehat{\boldsymbol{z}}_{ki} - \widetilde{\boldsymbol{z}}_{ki} - \bar{\boldsymbol{z}}_{ki})^{\mathrm{T}} = o_p((np_n)^{\frac{1}{2}})$, $\sum_{i=1}^{n_k}\widetilde{y}_{ki}\bar{\boldsymbol{z}}_{ki}^{\mathrm{T}} = o_p((np_n)^{\frac{1}{2}})$, $\sum_{i=1}^{n_k}\bar{y}_{ki}\widetilde{\boldsymbol{z}}_{ki}^{\mathrm{T}} = o_p((np_n)^{\frac{1}{2}})$, $\sum_{i=1}^{n_k}\bar{y}_{ki}\bar{\boldsymbol{z}}_{ki}^{\mathrm{T}} = o_p((np_n)^{\frac{1}{2}})$, $\sum_{i=1}^{n_k}\widetilde{\boldsymbol{z}}_{ki}^{\mathrm{T}}\boldsymbol{\beta}_k^*\bar{\boldsymbol{z}}_{ki}^{\mathrm{T}} = o_p((np_n)^{\frac{1}{2}})$, $\sum_{i=1}^{n_k}\bar{\boldsymbol{z}}_{ki}^{\mathrm{T}}\boldsymbol{\beta}_k^*\widetilde{\boldsymbol{z}}_{ki}^{\mathrm{T}} = o_p((np_n)^{\frac{1}{2}})$, $\sum_{i=1}^{n_k}\bar{\boldsymbol{z}}_{ki}^{\mathrm{T}}\boldsymbol{\beta}_k^*\bar{\boldsymbol{z}}_{ki}^{\mathrm{T}} = o_p((np_n)^{\frac{1}{2}})$, $\sum_{i=1}^{n_k}(\widehat{\boldsymbol{z}}_{ki} - \widetilde{\boldsymbol{z}}_{ki} - \bar{\boldsymbol{z}}_{ki})^{\mathrm{T}}\boldsymbol{\beta}_k^*(\widehat{\boldsymbol{z}}_{ki} - \widetilde{\boldsymbol{z}}_{ki} - \bar{\boldsymbol{z}}_{ki})^{\mathrm{T}} = o_p((np_n)^{\frac{1}{2}})$, $\sum_{i=1}^{n_k}\widetilde{\boldsymbol{z}}_{ki}^{\mathrm{T}}\boldsymbol{\beta}_k^*(\widehat{\boldsymbol{z}}_{ki} - \widetilde{\boldsymbol{z}}_{ki} - \bar{\boldsymbol{z}}_{ki})^{\mathrm{T}} = o_p((np_n)^{\frac{1}{2}})$, $\sum_{i=1}^{n_k}\bar{\boldsymbol{z}}_{ki}^{\mathrm{T}}\boldsymbol{\beta}_k^*(\widehat{\boldsymbol{z}}_{ki} - \widetilde{\boldsymbol{z}}_{ki} - \bar{\boldsymbol{z}}_{ki})^{\mathrm{T}} = o_p((np_n)^{\frac{1}{2}})$, $\sum_{i=1}^{n_k}(\widehat{\boldsymbol{z}}_{ki} - \widetilde{\boldsymbol{z}}_{ki} - \bar{\boldsymbol{z}}_{ki})^{\mathrm{T}}\boldsymbol{\beta}_k^*\widetilde{\boldsymbol{z}}_{ki}^{\mathrm{T}} = o_p((np_n)^{\frac{1}{2}})$, $\sum_{i=1}^{n_k}(\widehat{\boldsymbol{z}}_{ki} - \widetilde{\boldsymbol{z}}_{ki} - \bar{\boldsymbol{z}}_{ki})^{\mathrm{T}}\boldsymbol{\beta}_k^*\bar{\boldsymbol{z}}_{ki} = o_p((np_n)^{\frac{1}{2}})$. Moreover, by central limit theorem, we have $\sum_{k=1}^{K}\sum_{i=1}^{n_k}(\widetilde{y}_{ki}\widetilde{\boldsymbol{z}}_{ki}^{\mathrm{T}} - \widetilde{\boldsymbol{z}}_{ki}^{\mathrm{T}}\boldsymbol{\beta}_k^*\widetilde{\boldsymbol{z}}_{ki}^{\mathrm{T}} + \boldsymbol{\beta}_k^{*\mathrm{T}}\Sigma_k) = O_p((np_n)^{\frac{1}{2}})$. Thus $\sum_{i=1}^{n_k}(\widehat{y}_{ki}\widehat{\boldsymbol{z}}_{ki}^{\mathrm{T}} - \widehat{\boldsymbol{z}}_{ki}^{\mathrm{T}}\boldsymbol{\beta}_k^*\widehat{\boldsymbol{z}}_{ki}^{\mathrm{T}} + \boldsymbol{\beta}_k^{*\mathrm{T}}\Sigma_k) = o_p((np_n)^{\frac{1}{2}})$ for any $k = 1, \ldots, K$.

For the second term of δ_k, we have

$$n_k^{-1}\sum_{i=1}^{n_k}\widehat{\boldsymbol{z}}_{ki}\widehat{\boldsymbol{z}}_{ki}^{\mathrm{T}} - \Sigma_k \to E\Big[(\boldsymbol{x}_{ki} - \boldsymbol{m}_{kz}(w_{ki}))(\boldsymbol{x}_{ki} - \boldsymbol{m}_{kz}(w_{ki}))^{\mathrm{T}}\Big],$$

which is a positive definite matrix with constant eigenvalue by Assumption 3.3. Therefore, we conclude that there exists some constants c_1, c_2 and c_3 such that

$$\begin{aligned}\delta_l \geq{}& c_1 r_n^2 n\|\boldsymbol{\theta}\|_2^2 - c_2 r_n (np_n)^{\frac{1}{2}}\|\boldsymbol{\theta}\| - \lambda_\beta r_n p_n\|\boldsymbol{\theta}\| - \lambda_\alpha r_n p_n\|\boldsymbol{\theta}\|\\ \geq{}& p_n\big(c_1\|\boldsymbol{\theta}\|_2^2 - c_2\|\boldsymbol{\theta}\|_2 - \lambda_\beta(n/p_n)^{-\frac{1}{2}}\|\boldsymbol{\theta}\| - \lambda_\alpha(n/p_n)^{-\frac{1}{2}}\|\boldsymbol{\theta}\|\big).\end{aligned}$$

If $\lambda_\beta(n/p_n)^{-\frac{1}{2}} \to 0$ and $\lambda_\alpha(n/p_n)^{-\frac{1}{2}} \to 0$, we can find a large enough constant C such that $\delta_l > 0$ for $\|\boldsymbol{\theta}\| = C$. Thus $\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^* = O_p((n/p_n)^{-\frac{1}{2}})$, which implies that $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* = O_p((n/p_n)^{-\frac{1}{2}})$ and $\widehat{\boldsymbol{\alpha}}_k - \boldsymbol{\alpha}_k^* = O_p((n/p_n)^{-\frac{1}{2}})$ for any $k = 1, \ldots, K$.

For the nonlinear components estimation, we have $\widehat{m}_{ky}(v'_{ki}) - \widehat{\boldsymbol{m}}_{kz}(v'_{ki})^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_k - f_k^*(w_{ki}) = \widehat{m}_{ky}(v'_{ki}) - m_{ky}(v'_{ki}) + m_{ky}(v'_{ki}) - m_{ky}(w_{ki}) + m_{ky}(w_{ki}) - (\widehat{\boldsymbol{m}}_{kz}(v'_{ki}) - \boldsymbol{m}_{kz}(v'_{ki}) + \boldsymbol{m}_{kz}(v'_{ki}) - \boldsymbol{m}_{kz}(w_{ki}) + \boldsymbol{m}_{kz}(w_{ki}))^{\mathrm{T}}(\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^* + \boldsymbol{\beta}_k^*) - f_k^*(w_{ki})$. By $\widehat{m}_{ky}(v'_{ki}) - m_{ky}(v'_{ki}) = o_p(n^{-\frac{1}{4}})$, $m_{ky}(v'_{ki}) - m_{ky}(w_{ki}) = O_p(n^{-\frac{1}{2}})$, $\widehat{\boldsymbol{m}}_{kz}(v'_{ki}) - \boldsymbol{m}_{kz}(v'_{ki}) = o_p(n^{-\frac{1}{4}})$, $\boldsymbol{m}_{kz}(v'_{ki}) - \boldsymbol{m}_{kz}(w_{ki}) = O_p(n^{-\frac{1}{2}})$, $\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^* = O_p((n/p_n)^{-\frac{1}{2}})$ and $m_{ky}(w_{ki}) = m_{kz}(w_{ki}) + f_k^*(w_{ki})$, we get $\widehat{m}_{ky}(v'_{ki}) - \widehat{\boldsymbol{m}}_{kz}(v'_{ki})^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_k - f_k^*(w_{ki}) = O_p(\max\{n^{-\frac{1}{4}}, (n/p_n^3)^{-\frac{1}{2}}\})$. Then the desired result can be obtained. ■

Proof of Theorem 3.2: We prove Theorem 3.2 by contradiction. Suppose $|\widehat{\beta}_j| > 0$ for $j > p_{n,0}$. Take the derivative for β_j and get the KKT condition

$$\sum_{k=1}^{K}\sum_{i=1}^{n_k}(\widehat{y}_{ki} - \widehat{z}_{ki}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_k)\widehat{z}_{kij} + \sum_{k=1}^{K} n_k\widehat{\boldsymbol{\beta}}_k^{\mathrm{T}}\Sigma_k\mathbf{J}_j = \frac{1}{2}\lambda_\beta\pi_j\mathrm{sign}(\widehat{\beta}_j),$$

where $\boldsymbol{J}_j$ is the vector of all zeros except jth element. By Theorem 3.1 and the decomposition (A5), the left-hand side is $\sum_{k=1}^{K}\sum_{i=1}^{n_k}(\widetilde{y}_{ki}+o_p(n^{-\frac{1}{4}})-(\widetilde{\boldsymbol{z}}_{ki}+o_p(n^{-\frac{1}{4}}))^{\mathrm{T}}(\boldsymbol{\beta}_k^*+O_p((n/p_n)^{-\frac{1}{2}})))(\widetilde{\boldsymbol{z}}_{ki}+o_p(n^{-\frac{1}{4}}))_j-\sum_{k=1}^{K}n_k(\boldsymbol{\beta}_k^*+O_p((n/p_n)^{-\frac{1}{2}}))^{\mathrm{T}}\Sigma_k\mathbf{J}_j$. Then by Lemma A.1 in Liang and Li [21] and Assumption 3.3, the left-hand side can be simplified as

$$\sum_{k=1}^{K}\sum_{i=1}^{n_k}(\widetilde{y}_{ki}-\widetilde{\boldsymbol{z}}_{ki}^{\mathrm{T}}\boldsymbol{\beta}_k^*)(\widetilde{\boldsymbol{z}}_{ki})_j-\sum_{k=1}^{K}n_k(\boldsymbol{\beta}_k^*)^{\mathrm{T}}\Sigma_k\mathbf{J}_j+O_p((np_n)^{\frac{1}{2}}),$$

which is equal to

$$\sum_{k=1}^{K}\sum_{i=1}^{n_k}\epsilon_{ki}U_{kij}+\sum_{k=1}^{K}n_k(\boldsymbol{\beta}_k^*)^{\mathrm{T}}(n_k^{-1}\boldsymbol{U}_{ki}\boldsymbol{U}_{ki}^{\mathrm{T}}-\Sigma_k)\mathbf{J}_j+O_p((np_n)^{\frac{1}{2}}).$$

Because $n_k^{-1}\boldsymbol{U}_{ki}\boldsymbol{U}_{ki}^{\mathrm{T}}-\Sigma_k=O_p(n^{-\frac{1}{2}})$, we get the left-hand side is $O_p((np_n)^{\frac{1}{2}})$. The right-hand side is $\lambda_\beta O_p((n/p_n)^{\frac{1}{2}})$. We divide $(np_n)^{\frac{1}{2}}$ on both left- and right-hand sides and get $O_p(1)=\lambda_\beta/p_n$, which is contradicted to $\lambda_\beta/p_n\to\infty$. We conclude that $|\widehat{\beta}_j|=0$ for $j>p_{n,0}$.

Similarly, we can prove $\widehat{\alpha}_{kj}=0$ for any $j>p_{n,0}$ and $k=1,\dots,K$. Suppose $|\widehat{\alpha}_{kj}|>0$ for $j>p_{n,0}$. Take the derivative for α_{kj} and get the KKT condition

$$\sum_{i=1}^{n_k}(\widetilde{y}_{ki}-\widetilde{\boldsymbol{z}}_{ki}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_k)\widetilde{z}_{kij}+n_k\widehat{\boldsymbol{\beta}}_k^{\mathrm{T}}\Sigma_k\mathbf{J}_j=\frac{1}{2}\lambda_\alpha\pi_{kj}\mathrm{sign}(\widehat{\alpha}_{kj}).$$

Same as the proof for showing $|\widehat{\beta}_j|=0$ for $j>p_{n,0}$ above, we have the left-hand side is $O_p((np_n)^{\frac{1}{2}})$ and the right-hand side is $\lambda_\alpha O_p((n/p_n)^{\frac{1}{2}})$. We divide $(np_n)^{\frac{1}{2}}$ on both left- and right-hand sides and get $O_p(1)=\lambda_\alpha/p_n$, which is contradicted to $\lambda_\alpha/p_n\to\infty$. We conclude that $|\widehat{\alpha}_{kj}|=0$ for $j>p_{n,0}$. ■

Proof of Theorem 3.3: The proof of Theorem 3.3 was essentially the same as the proof of Theorem 3.1 and 3.2, which was omitted here. ■

Proof of Theorem 3.4: The key idea of the proof is the same as the proof of Theorem 1 in Liang and Li [21]. Take the derivative for $\boldsymbol{\beta}_{kI}$ in Equation (3) and get the KKT condition that

$$\sum_{i=1}^{n_k}(\boldsymbol{z}_{ki}-\widehat{\boldsymbol{m}}_{kz}(w_{ki}))_I(y_{ki}-\widehat{m}_{ky}(w_{ki})-(\boldsymbol{z}_{ki}-\widehat{\boldsymbol{m}}_{kz}(w_{ki}))_I^{\mathrm{T}}\boldsymbol{\beta}_{kI})-n_k\Sigma_{kI}\boldsymbol{\beta}_{kI}+o_p(n^{1/2})=\mathbf{0}.$$

Same as the proof of Theorem 1 in Liang and Li [21], because

$$\sup_{W_k}|\widehat{m}_{ky}(W_k)-m_{ky}(W_k)|=o_p(n^{-\frac{1}{4}})\quad\text{and}\quad\sup_{W_k}|\widehat{m}_{kzj}(W_k)-m_{kzj}(W_k)|=o_p(n^{-\frac{1}{4}}),$$

$\widehat{\boldsymbol{\beta}}_{kI}$ has the same asymptotic distribution as the solution of

$$-\sum_{i=1}^{n_k}(\boldsymbol{z}_{ki}-\boldsymbol{m}_{kz}(w_{ki}))_I(y_{ki}-m_{ky}(w_{ki})-(\boldsymbol{z}_{ki}-\boldsymbol{m}_{kz}(w_{ki}))_I^{\mathrm{T}}\boldsymbol{\beta}_{kI})-n_k\Sigma_{kI}\boldsymbol{\beta}_{kI}+o_p(n^{1/2})=\mathbf{0}.$$

By $\boldsymbol{z}_{ki}-\boldsymbol{m}_{kz}(w_{ki})=\boldsymbol{x}_{ki}-E(\boldsymbol{x}_{ki}|w_{ki})+\boldsymbol{U}_{ki}$ and $y_{ki}=(\boldsymbol{x}_{ki}-\boldsymbol{m}_{kz}(w_{ki}))_I^{\mathrm{T}}\boldsymbol{\beta}_{kI}^*+m_{ky}(w_{ki})+\epsilon_{ki}$, a direct simplification yields that

$$\begin{aligned}&\frac{1}{\sqrt{n_k}}\sum_{i=1}^{n_k}\left\{\left[\boldsymbol{x}_{ki}-E(\boldsymbol{x}_{ki}|w_{ki})+\boldsymbol{U}_{ki}\right]_I^{\otimes 2}-\Sigma_{kI}\right\}(\widehat{\boldsymbol{\beta}}_{kI}-\boldsymbol{\beta}_{kI}^*)\\&=\frac{1}{\sqrt{n_k}}\sum_{i=1}^{n_k}\left\{[\boldsymbol{x}_{ki}-E(\boldsymbol{x}_{ki}|w_{ki})+\boldsymbol{U}_{ki}]_I(\epsilon_{ki}-\boldsymbol{U}_{kI,i}^{\mathrm{T}}\boldsymbol{\beta}_{kI}^*)+\Sigma_{kI}\boldsymbol{\beta}_{kI}^*\right\}+o_p(1).\end{aligned}$$

As $n \to \infty$, because $n_k^{-1} \sum_{i=1}^{n_k} [\boldsymbol{x}_{ki} - E(\boldsymbol{x}_{ki}|w_{ki}) + \boldsymbol{U}_{ki}]_I^{\otimes 2} \to \Sigma_{X|W}^k + \Sigma_{kI}$, we get

$$\sqrt{n_k}\Sigma_{X|W}^k(\widehat{\boldsymbol{\beta}}_{kI} - \boldsymbol{\beta}_{kI}^*) \to N(\mathbf{0}, \Gamma_k),$$

where $\Gamma_k = E\{(\boldsymbol{X}_{kI} - E(\boldsymbol{X}_{kI}|W_k))(\epsilon_k - \boldsymbol{U}_{kI}^{\mathrm{T}}\boldsymbol{\beta}_{kI}^*) + \epsilon_k\boldsymbol{U}_{kI} + (\Sigma_{kI} - \boldsymbol{U}_{kI}\boldsymbol{U}_{kI}^{\mathrm{T}})\boldsymbol{\beta}_{kI}^*\}^{\otimes 2}$. The desired result is obtained. ■