HHS Public Access

Author manuscript

Statistics (Ber). Author manuscript; available in PMC 2024 October 13.

Published in final edited form as:

Statistics (Ber). 2023 ; 57(5): 987–1009. doi:10.1080/02331888.2023.2258429.

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

Lei Yang,

Yongzhao Shao

Department of Population Health New York University

Abstract

The design of multi-center study is increasingly used for borrowing strength from multiple research groups to obtain broadly applicable and reproducible study findings. Regression analysis is widely used for analyzing multi-group studies, however, some of the large number of regression predictors are nonlinear and/or often measured with batch effects in many large scale collaborative studies. Also, the group compositions of the nonlinear predictors are potentially heterogeneous across different centers. The conventional pooled data analysis ignores the interplay between nonlinearity and batch effect, group composition heterogeneity, measurement error and other data incoherence in multi-center setting that can cause biased regression estimates and misleading outcomes. In this paper, we propose an integrated partially linear regression model (IPLM) based analysis to account for the predictor's nonlinearity, general batch effect, group composition heterogeneity, high-dimensional covariates, potential measurement-error in covariates, and combinations of these complexities 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 that can be either homogeneous or heterogeneous across groups. In particular, when the effects of all predictors are homogeneous across the study centers, the proposed IPLM can automatically reduce to one single parsimonious partially linear model for all centers. The proposed method has asymptotic estimation and variable selection consistency including high-dimensional covariates. Moreover, it has a fast computing algorithm and its effectiveness is supported by numerical simulation studies. A multi-center Alzheimer's disease research project is provided to illustrate the proposed IPLM based analysis.

Keywords

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

1 Introduction

The design of multi-center study becomes increasingly used because it enables researchers to obtain more generalizable and reproducible study findings in many fields, including

Disclosure statement

No potential conflict of interest was reported by the authors

cancer study (Arslan et al. 2020, Cruz et al. 2018, He et al. 2021, Rahbar et al. 2017, Roach et al. 2018, Sturdza et al. 2016, Sun et al. 2019) and Alzheimer's disease (AD) research (Boada et al. 2017, Ewers et al. 2015, Khan et al. 2017, Lim et al. 2018, Niemantsverdriet et al. 2018, Pannee et al. 2021, Van Steenoven et al. 2016). Regression analysis is widely used to analyze multi-center 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-center 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 and many of them are nonlinear predictors. For example, in ovarian cancer, both the mean and variance effect of some DNA methylations are significant (Ahn & Wang 2013) in risk prediction. Thus, it is desired to develop flexible regression models that can incorporate both linear and nonlinear predictors, e.g. using the partially linear model (PLM) (Härdle et al. 2012), $Y = \boldsymbol{\beta}^{*\mathrm{T}}\boldsymbol{X} + f^{*}(W) + \epsilon$, 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 centers. Thus, for multi-center studies, a natural modeling choice is the following systems of partially linear models:

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

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

Another common challenge in multi-center 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 (Chen et al. 2011, Scherer 2009). 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 Lim et al. (2018), the levels of CSF Aβ_{42} protein have a nonlinear cyclic seasonal pattern over measurement dates. Similarly, we examined a recent multi-center 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 centers are heterogeneous. For example, in AD research, some study cohorts are younger while others are much older. According to Shoji et al. (2001), the levels of CSF $A\beta_{42}$ 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 $A\beta_{42}$. If a conventional linear model is applied in a younger versus an older centers independently to study the relationship between age and $A\beta_{42}$, 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-center analyses can potentially lead to contradictory study findings among different centers in the presence of heterogeneity in group compositions. Instead of single center 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-centers via suitable frequency matching or propensity score matching. In this context, the analysis using the simple pooling of multi-center data via the commonly-used z-score method can cause severe biases.

Additional common issues in multi-center 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 are 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 (Hessell et al. 2021). One popular choice to account for the measurement error in variable selection procedure is subtracting a bias correction term from the loss function, e.g. Liang & Li (2009), Zhao & Xue (2010), Zhou et al. (2011). There have been extensive research on both variable selection and measurement error in regression models. Therefore, regression analysis for multi-center 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-center study as demonstrated using the multi-center 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-center 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 (Fan & Gijbels 1996) 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 centers. If all the predictors' effects are homogeneous across centers, the proposed IPLM can automatically reduce to one parsimonious partially linear regression

model that is applicable to all centers while simultaneously account for nonlinearity, batch effect, heterogeneity, high-dimensionality, and measurement errors. The integrated analysis facilitates generalizable and reproducible outcomes in multi-center 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-center AD research project to illustrate the proposed procedures. A short summary is provided in Section 6.

2 Models

For the kth center in a multi-center 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. Aβ_{42} protein measured in different seasons in AD research (Lim et al. 2018) and $\boldsymbol{X}_k \in \mathscr{R}^{p_n}$ 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-center multi-center study:

$$Y_k = \boldsymbol{X}_k^{\mathrm{T}}\left(\boldsymbol{\beta}^* + \boldsymbol{\alpha}_k^*\right) + f_k^*(W_k) + \epsilon_k,$$

$$\boldsymbol{Z}_k = \boldsymbol{X}_k + \boldsymbol{U}_k,$$

$$V_k = W_k + g_k\left(m_k; \boldsymbol{\psi}_k^*\right) \quad \text{for} \quad 1 \le k \le K,$$

with the constraint $\Sigma_{k=1}^{K}\boldsymbol{\alpha}_k^* = \mathbf{0}$, 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. $g_k\left(\cdot\,; \boldsymbol{\psi}_k^*\right)$ is the general batch effect. Note that the function form of $g_k\left(\cdot\,; \boldsymbol{\psi}_k^*\right)$ is assumed known while the parameters $\boldsymbol{\psi}_k$ need to be estimated. The batch effect $g_k\left(\cdot\,; \boldsymbol{\psi}_k^*\right)$ 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-center 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 rigor 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 (Liang & Li 2009). 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 $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^{\mathrm{T}}, \cdots, \boldsymbol{\alpha}_K^{\mathrm{T}})^{\mathrm{T}}$. 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. In order 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 $\Sigma_{k=1}^{K}\boldsymbol{\alpha}_k = \mathbf{0}$. Then the regularized square loss function is defined follows

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

subject to the constraint $\Sigma_{k=1}^{K}\boldsymbol{\alpha}_k = \mathbf{0}$, where $p_{\lambda_\beta}(\boldsymbol{\beta}) = \lambda_\beta \Sigma_{j=1}^{p_n} \pi_j |\beta_j|$ is the penalty term to identify the informative mean effect, $p_{\lambda_\alpha}(\boldsymbol{\alpha}_k) = \lambda_\alpha \Sigma_{k=1}^{K}\Sigma_{j=1}^{p_n} \pi_{kj}|\alpha_{kj}|$ is the penalty term to identify the informative heterogeneous effect, $\Sigma_{k=1}^{K} n_k(\boldsymbol{\beta} + \boldsymbol{\alpha}_k)^{\mathrm{T}}\Sigma_k(\boldsymbol{\beta} + \boldsymbol{\alpha}_k)$ is the penalty term to correct the measurement error and π_j and π_{kj} are the adaptive Lasso weight (Zou 2006). Note that the adaptive Lasso weight can be achieved by setting $\pi_j = 1/|\tilde{\beta}_j|$ and $\pi_{kj} = 1/|\tilde{\alpha}_{kj}|$, where

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

subject to the constraint $\Sigma_{k=1}^{K}\boldsymbol{\alpha}_k = 0$. 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 (Liang & Li 2009), i.e. we observe $\boldsymbol{Z}_{kij} = \boldsymbol{X}_{ki} + \boldsymbol{U}_{kij}$ for $j = 1, \cdots, J_{ki}$. Let $\overline{\boldsymbol{Z}}_{ki} = J_{ki}^{-1}\Sigma_{i=1}^{J_{ki}}\boldsymbol{Z}_{kij}$ to be the sample mean of J_{ki} replicates for ith subject in kth group. Then a consistent moments estimate of Σ_k is

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

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

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

subject to the constraint $\Sigma_{k=1}^{K} \boldsymbol{\alpha}_k = \mathbf{0}$. 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 $f_k(V_k^{'}) = E(Y_k \mid V_k^{'}) - E(\boldsymbol{Z}_k \mid V_k^{'})^{\mathrm{T}}(\boldsymbol{\beta} + \boldsymbol{\alpha}_k)$. Denote $m_{ky}(V_k^{'}) = E(Y_k \mid V_k^{'})$ and $\boldsymbol{m}_{kz}(V_k^{'}) = E(\boldsymbol{Z}_k \mid V_k^{'})$. 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 $\Sigma_{k=1}^{K} \boldsymbol{\alpha}_k = \mathbf{0}$ 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 (Fan & Gijbels 1996) 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, $\hat{y}_{ki} = y_{ki} - \widehat{m}_{ky}(v_{ki}^{'})$ and $\hat{\boldsymbol{z}}_{ki} = \boldsymbol{z}_{ki} - \widehat{\boldsymbol{m}}_{ky}(v_{ki}^{'})$. 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 $l(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \Sigma_{k=1}^{K} \Sigma_{i=1}^{n_k} (\hat{y}_{ki} - \hat{\boldsymbol{z}}_{ki}^{\mathrm{T}}(\boldsymbol{\beta} + \boldsymbol{\alpha}_k))^2$. 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 (Boyd et al. 2011) to solve $\boldsymbol{\alpha}$ under linear constraint $\Sigma_{k=1}^{K} \boldsymbol{\alpha}_k = \mathbf{0}$. The details are showed in next subsection. Given estimated $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}$, the estimated nonlinear function is

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

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

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

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 $\boldsymbol{\Lambda}^{(t+1)} = \boldsymbol{\Lambda}^{(t)} - \delta^{-1}\left(A\boldsymbol{\alpha}^{(t+1)} + B\boldsymbol{\eta}^{(t+1)}\right)$. 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 centers, i.e. $\Sigma_{k=1}^{K} \Sigma_{j=1}^{p_n} \alpha_{kj}^2 = 0$ and $f_k(\cdot) = f_1(\cdot)$ for any $k = 1, \cdots, K$, and thus the IPLM automatically reduce 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

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

It is clear that we can build the unified model by minimizing (5) and solve $\boldsymbol{\beta}$ and $f(\cdot)$ as in section 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 & Li (2009). Without loss of generality, we assume that $\boldsymbol{\beta}_j^* = 0$ 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 1:

The support for W_k and m_k are bounded for $k = 1, \cdots, K$.

Assumption 2:

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

Assumption 3:

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

Assumption 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 5:

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

Assumption 6:

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

Assumption 7:

$n_k = O(n)$ for $k = 1, \cdots, K$.

Assumption 8:

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

Assumption 1 to 5 are included and discussed in Liang & Li (2009). Assumption 6 ensures that the batch effect adjustment using least square method is adequate. Assumption 7 ensures that the sample size across different centers are comparable. No center has a dominating or negligible sample size. Assumption 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 literature (Zou & Zhang 2009). To ensure consistency of the estimates of nonparametric component we need $p_n = p[n^a]$ with $0 \le a < 1/3$. 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 4.

Theorem 1.—*Under Assumptions 1-8, we have* $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\| = O_p\left((n/p_n)^{-\frac{1}{2}}\right)$, $\|\hat{\boldsymbol{\alpha}}_k - \boldsymbol{\alpha}_k^*\| = O_p\left((n/p_n)^{-\frac{1}{2}}\right)$ *and* $E\left|\hat{f}_k(\cdot) - f_k^*(\cdot)\right| = O_p\left(\max\left\{n^{-\frac{1}{4}}, (n/p_n^3)^{-\frac{1}{2}}\right\}\right)$ *for any* $k = 1, \cdots, K$ *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 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 2.—*Under Assumptions 1-8, we have* $\lim_{n\to\infty} P\left(\hat{\beta}_j = 0\right) = 1$ *and* $\lim_{n\to\infty} P(\hat{\alpha}_{kj} = 0) = 1$ *for* $k = 1, \cdots, K$ *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 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 centers, and high-dimensional predictors.

Similar to Theorem 1 and Theorem 2, both the estimation and variable selection consistency holds 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.—*Under Assumptions 1-7 with* $p_n = p$, *we have* $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\| = O_p\left(n^{-\frac{1}{2}}\right)$, $\|\hat{\boldsymbol{\alpha}}_k - \boldsymbol{\alpha}_k^*\| = O_p\left(n^{-\frac{1}{2}}\right)$ *and* $E\left|\hat{f}_k(\cdot) - f_k^*(\cdot)\right| = O_p\left(n^{-\frac{1}{4}}\right)$, $\lim_{n\to\infty} P\left(\hat{\beta}_j = 0\right) = 1$ *and* $\lim_{n\to\infty} P(\hat{\alpha}_{kj} = 0) = 1$ *for any* $k = 1, \cdots, K$ *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 center 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, $\Sigma^k_{X\,|\,W} = cov(\boldsymbol{X}_{kI} - E(\boldsymbol{X}_{kI} \mid W_k))$. As $n \to \infty$, $\hat{\boldsymbol{\beta}}_{kI}$ are asymptotic normally distributed.

Theorem 4.—*Under Assumptions 1-7 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* $k = 1, \cdots, K$, *we have* $\sqrt{n_k}\Sigma^k_{X\,|\,W}\left(\hat{\boldsymbol{\beta}}_{kI} - \boldsymbol{\beta}^*_{kI}\right) \to N(\mathbf{0}, \Gamma_k)$, where $\Gamma_k = E\left\{(\boldsymbol{X}_{kI} - E(\boldsymbol{X}_{kI} \mid W_k))\left(\epsilon_k - \boldsymbol{U}^{\mathrm{T}}_{kI}\boldsymbol{\beta}^*_{kI}\right) + \epsilon_k \boldsymbol{U}_{kI} + \left(\Sigma_{kI} - \boldsymbol{U}_{kI}\boldsymbol{U}^{\mathrm{T}}_{kI}\right)\boldsymbol{\beta}^*_{kI}\right\}^{\otimes 2}$

The proofs of the above theorems can be found in the Appendix while the proof of theorem 3 was essentially the same as proof of theorem 1 and 2, which was omitted.

4 Numerical studies

In this section, we report numerical studies conducted to demonstrate 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 centers with covariate dimensions comparable to the sample size. Also, the conventional pooled data analysis (e.g. using z-scores) or individual-center based analysis ignores the interplay between nonlinearity and group composition heterogeneity, batch effect, measurement error and other data incoherence in multicenter 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

$$Y_k = \boldsymbol{X}_k^{\mathrm{T}}\left(\boldsymbol{\beta}^* + \boldsymbol{\alpha}_k^*\right) + f_k^*(W_k) + \epsilon_k,$$

$$\boldsymbol{Z}_k = \boldsymbol{X}_k + \boldsymbol{U}_k,$$

$$V_k = W_k + g_k\left(m_k; \boldsymbol{\psi}_k^*\right) \text{ for } 1 \le k \le 2.$$

Here $\boldsymbol{X}$ is from a p_n dimensional normal distribution, i.e. $\boldsymbol{X}_{1i} \sim N(\mathbf{0}, \Sigma)$, $\boldsymbol{X}_{2i} \sim N(\mathbf{0}, \Sigma)$, 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 $\boldsymbol{U}_k \sim N(\mathbf{0}, \Sigma_k)$, 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 $\boldsymbol{Z}_k^R$, are generated. Linear coefficients $\boldsymbol{\beta}^*$, $\boldsymbol{\alpha}_1^*$ and $\boldsymbol{\alpha}_2^*$ are p_n dimensional vectors, i.e. $\boldsymbol{\beta}^* = (0.5, 0.5, 0.5, 0, \ldots, 0)$, $\boldsymbol{\alpha}_1^* = \boldsymbol{\alpha}_2^* = (0, 0, \ldots, 0)$, 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 centers are:

$$f_1^*(W) = f_2^*(W) = (W-1)^2.$$

Moreover, W_k has uniform distributions:

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

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 (Ahn & Wang 2013). Second, the nonlinear predictor V_k contains batch effects, e.g. the cyclic seasonal pattern of Aβ protein (Lim et al. 2018). 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 $\hat{b}_k$ and correct the batch effects by $v_{ki}' = v_{ki} - \hat{b}_k \sin(m_{ki})$ 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 (Schwarz et al. 1978) (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 are 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. $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$) were summarized in Table 2. Specifically, the average MSE of estimated group 1 effect $\hat{\boldsymbol{\beta}}_1$ and group 2 effect $\hat{\boldsymbol{\beta}}_2$ was defined as

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

where $\hat{\boldsymbol{\beta}}_1^b$ and $\hat{\boldsymbol{\beta}}_2^b$ 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 4 (denoted as SD2) of $\hat{\boldsymbol{\beta}}_{11}$ (estimated value of first element of $\boldsymbol{\beta}_1$) and $\hat{\beta}_{21}$ (estimated value of first element of $\boldsymbol{\beta}_2$), and the coverage of the 95% confidence interval estimated using Theorem 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 in order to properly use the adaptive Lasso. A 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 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%.

Individual-group analysis can lead to contradictory findings: If we analyze data for each group (or center) separately, we find that the effect of V is negative in center 1 while positive in center 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 center/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 touse 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.

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 are quite good as reported in Table 6, which is consistent with the findings in scenario 4.1 when all effects are homogeneous.

5 Real data illustration

In this section, we illustrate how to apply the proposed IPLM-based analysis to rigorously analyze biomarker data in an multi-center 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 ($A\beta_{42}$) and tau proteins in brain underlie the plaques and NFT, respectively. Moreover, existence of within-neuron NFT indicates 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 (Hansson et al. 2006, Herukka et al. 2007). Additionally, among elderly persons who are at risk for AD, reductions in the CSF $A\beta_{42}$ are associated with brain $A\beta_{42}$ deposition and precede elevations in CSF tau levels. Therefore, there has been persistent interest to investigate the relationship between CSF $A\beta_{42}$ and CSF tau protein and CSF tau can be used as a surrogate outcome variable in these AD research. Moreover, the biological function of $A\beta_{42}$ 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 centers used in de Leon et al. (2018), 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 $A\beta_{42}$ 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 $A\beta_{42}$ first for data harmonization. In the NYU and ADNI databases, we fit the observed $A\beta_{42}$ 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 center 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 $A\beta_{42}$ values via:

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

If we investigated the relationship between CSF $A\beta_{42}$ and tau protein within each center alone separately, we will get seemingly contradictory findings as seen in Figure 3. From the upper left sub-figure of Figure 3, the NYU center data mainly shows an increasing trend while from the lower left sub-figure we can see that the ADNI center 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 center only, e.g. the ADNI cohort, one might easily reach the conclusion of an monotone relationship between CSF tau and CSF $A\beta_{42}$ 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-center 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 centers: NYU and ADNI. Also, the biological relationship between CSF tau and Aβ proteins and the

mechanism should be largely identical across different centers (de Leon et al. 2018). Thus it is natural to combine two study centers together and build a unified model. However, the two centers 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 $A\beta_{42}$ and tau protein between younger adults and older adults differ (Shoji et al. 2001). Due to the heterogeneous age distribution between the study centers, if we use simple pooled analysis, i.e. pooling z-scores from the two centers as commonly used in 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:

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

The fitted nonlinear curve $f^*(\cdot)$, i.e. the effect of $A\beta_{42}$ on tau protein, is shown in figure 3. From figure 3, we find that the effect of $A\beta_{42}$ on tau protein is clearly nonlinear. More specifically, as the value of $A\beta_{42}$ increases, the value of tau protein decreases first and then increases. Moreover, the study findings, i.e. the nonlinear relationship between $A\beta_{42}$ and tau protein, can be expected to be more robust and achieve higher generalizability and reproducibility because we used data from two independent study centers and the non-linear 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 exist between study centers. As extra examples, in AD research, Chen et al. (2021) showed that the effect of APoE4ϵ on 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 $A\beta_{42}$ values is more invasive than measuring the blood-based $A\beta_{42}$ level. Therefore, AD researchers across different centers have started using plasma $A\beta_{42}$ as non-invasive biomarkers. As reviewed in Figure 1 of Pannee et al. (2021), the plasma $A\beta_{42}$ values have very low between-center correlations indicating major batch effects across centers that need to be harmonized carefully in order to produce reproducible findings (Pannee et al. 2021). In short, this AD-research real-data example demonstrates that the commonly used single-center analysis can produce non-generalizable findings while our newly proposed IPLM-based approach can be used in multi-center studies to automatically account for nonlinear predictors and adjust for batch effect and heterogeneity in center compositions yielding generally valid findings.

6 Summary

Achieving generalizable and reproducible findings in multi-center 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-center 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-center 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-center 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 centers. 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 centers. 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 an multi-center 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-center based analysis can produce misleading findings while our newly proposed IPLM-based approach can be used in multi-center studies to automatically account for nonlinear predictors, heterogeneity in center 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 modeling procedure while, in the existing literature, neglected or casual batch-effect removal in data pooling before any careful statistical modeling often contributes to non-reproducible findings.

Acknowledgement

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 Alzheimer's disease data sets used in Section 5 for 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). Part of the data used in 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/datasamples/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 analysis or writing of this report. A complete list of ADNI investigators are at: http://adni.loni.usc.edu/wpcontent/uploads/howtoapply/ADNIAcknowledgementList.pdf.

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.

Appendix: technical proofs

Proof of Theorem 1:

First, we adjust the batch effect using linear regression model. By Assumption 6, we have

$$\left|w_{k(i)} - v_{k(i)}^{'}\right| = \left|g\left(m_{ki}; \widehat{\boldsymbol{\psi}}_k\right) - g\left(m_{ki}; \boldsymbol{\psi}_k^*\right)\right| \leq \left|g'\left(m_{ki}; \widetilde{\boldsymbol{\psi}}_k\right)\right|\left|\widehat{\boldsymbol{\psi}}_k - \boldsymbol{\psi}_k^*\right| = O_p\left(n^{-\frac{1}{2}}\right) \tag{6}$$

for any $i = 1, \cdots, n_k$ and $k = 1, \cdots, K$, where $\widetilde{\boldsymbol{\psi}}_k \in \left(\boldsymbol{\psi}_k^*, \widehat{\boldsymbol{\psi}}_k\right)$. Then by Assumption 1 to 5, the local polynomial estimates satisfy

$$\sup_{V_k^{'}}\left|\widehat{m}_{ky}\left(V_k^{'}\right) - m_{ky}\left(V_k^{'}\right)\right| = o_p\left(n^{-\frac{1}{4}}\right) \text{ and } \sup_{V_k^{'}}\left|\widehat{m}_{kzj}\left(V_k^{'}\right) - m_{kzj}\left(V_k^{'}\right)\right| = o_p\left(n^{-\frac{1}{4}}\right) \tag{7}$$

for $j = 1 \cdots, p_n$, 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 $\boldsymbol{\beta} = \Sigma_{k=1}^{K}\boldsymbol{\beta}/K$ and $\boldsymbol{\alpha}_k = \boldsymbol{\beta}_k - \boldsymbol{\beta}$.

Denote $\boldsymbol{\Theta} = \left(\boldsymbol{\beta}_1^{\mathrm{T}}, \cdots, \boldsymbol{\beta}_K^{\mathrm{T}}\right)^{\mathrm{T}}$. 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 $l(\boldsymbol{\Theta}) = \Sigma_{k=1}^{K}\Sigma_{i=1}^{n_k}\left(\widehat{y}_{ki} - \widehat{\boldsymbol{z}}_{ki}^{\mathrm{T}}\boldsymbol{\beta}_k\right)^2$, $p_{\lambda_\beta}(\boldsymbol{\Theta}) = \lambda_\beta \Sigma_{j=1}^{p_n}\left|K^{-1}\Sigma_{k=1}^{K}\beta_{kj}\right|$ and $p_{\lambda_\alpha}(\boldsymbol{\Theta}) = \lambda_\alpha \Sigma_{k=1}^{K}\Sigma_{j=1}^{p_n}\pi_{kj}\left|\beta_{kj} - K^{-1}\Sigma_{k=1}^{K}\beta_{kj}\right|$. Denote $\boldsymbol{\theta} = \left(\boldsymbol{\xi}_1^{\mathrm{T}}, \cdots, \boldsymbol{\xi}_K^{\mathrm{T}}\right)^{\mathrm{T}}$. 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

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

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

$$\begin{aligned} p_{\lambda_\beta}(\boldsymbol{\Theta}^* + r_n\boldsymbol{\theta}) - p_{\lambda_\beta}(\boldsymbol{\Theta}^*) &\geq \lambda_\beta \sum_{j=1}^{p_{n,0}} \pi_j \left|\sum_{k=1}^{K} \beta_{kj}^*/K + r_n \sum_{k=1}^{K} \xi_{kj}/K\right| - \lambda_\beta \sum_{j=1}^{p_{n,0}} \pi_j \left|\sum_{k=1}^{K} \beta_{kj}^*/K\right| \\ &\geq -\lambda_\beta r_n \sum_{j=1}^{p_{n,0}} \pi_j \left|\sum_{k=1}^{K} \xi_{kj}/K\right|. \end{aligned} \tag{8}$$

Similarly, we have

$$p_{\lambda_\alpha}(\boldsymbol{\Theta}^* + r_n\boldsymbol{\theta}) - p_{\lambda_\alpha}(\boldsymbol{\Theta}^*) \geq -\lambda_\alpha r_n \sum_{k=1}^{K}\sum_{j=1}^{p_{n,0}} \pi_{kj}\left|\xi_{kj} - \sum_{k=1}^{K}\xi_{kj}/K\right|. \tag{9}$$

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

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

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

$$\begin{aligned}\hat{y}_{ki} &= y_{ki} - \hat{m}_{ky}(v_{ki}') = y_{ki} - m_{ky}(w_{ki}) + m_{ky}(w_{ki}) - m_{ky}(v_{ki}') + m_{ky}(v_{ki}') - \hat{m}_{ky}(v_{ki}'),\\ \hat{\boldsymbol{z}}_{ki} &= \boldsymbol{z}_{ki} - \hat{\boldsymbol{m}}_{kz}(v_{ki}') = \boldsymbol{z}_{ki} - \boldsymbol{m}_{kz}(w_{ki}) + \boldsymbol{m}_{kz}(w_{ki}) - \boldsymbol{m}_{kz}(v_{ki}') + \boldsymbol{m}_{kz}(v_{ki}') - \hat{\boldsymbol{m}}_{kz}(v_{ki}').\end{aligned} \tag{10}$$

Denote $\tilde{y}_{ki} = y_{ki} - m_{ky}(w_{ki})$, $\bar{y}_{ki} = m_{ky}(w_{ki}) - m_{ky}(v_{ki}')$, $\tilde{\boldsymbol{z}}_{ki} = \boldsymbol{z}_{ki} - \boldsymbol{m}_{kz}(w_{ki})$ 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

$$\begin{aligned}&\Sigma_{k=1}^{K}\Sigma_{i=1}^{n_k}(\hat{y}_{ki} - \tilde{y}_{ki} - \bar{y}_{ki})(\hat{\boldsymbol{z}}_{ki} - \tilde{\boldsymbol{z}}_{ki} - \bar{\boldsymbol{z}}_{ki})^{\mathrm{T}} + (\hat{y}_{ki} - \tilde{y}_{ki} - \bar{y}_{ki})\tilde{\boldsymbol{z}}_{ki}^{\mathrm{T}} + (\hat{y}_{ki} - \tilde{y}_{ki} - \bar{y}_{ki})\bar{\boldsymbol{z}}_{ki}^{\mathrm{T}} + \tilde{y}_{ki}(\hat{\boldsymbol{z}}_{ki} - \tilde{\boldsymbol{z}}_{ki} - \bar{\boldsymbol{z}}_{ki})^{\mathrm{T}}.\\ &+ \bar{y}_{ki}(\hat{\boldsymbol{z}}_{ki} - \tilde{\boldsymbol{z}}_{ki} - \bar{\boldsymbol{z}}_{ki})^{\mathrm{T}} + \left(\tilde{y}_{ki}\tilde{\boldsymbol{z}}_{ki}^{\mathrm{T}} - \tilde{\boldsymbol{z}}_{ki}^{\mathrm{T}}\boldsymbol{\beta}_k^*\tilde{\boldsymbol{z}}_{ki}^{\mathrm{T}} + \boldsymbol{\beta}_k^{*\mathrm{T}}\Sigma_k\right) + \tilde{y}_{ki}\bar{\boldsymbol{z}}_{ki}^{\mathrm{T}} + \bar{y}_{ki}\tilde{\boldsymbol{z}}_{ki}^{\mathrm{T}} + \bar{y}_{ki}\bar{\boldsymbol{z}}_{ki}^{\mathrm{T}} - \tilde{\boldsymbol{z}}_{ki}^{\mathrm{T}}\boldsymbol{\beta}_k^*\bar{\boldsymbol{z}}_{ki}^{\mathrm{T}} - \bar{\boldsymbol{z}}_{ki}^{\mathrm{T}}\boldsymbol{\beta}_k^*\tilde{\boldsymbol{z}}_{ki}^{\mathrm{T}} - \bar{\boldsymbol{z}}_{ki}^{\mathrm{T}}\boldsymbol{\beta}_k^*\bar{\boldsymbol{z}}_{ki}^{\mathrm{T}}\\ &- (\hat{\boldsymbol{z}}_{ki} - \tilde{\boldsymbol{z}}_{ki} - \bar{\boldsymbol{z}}_{ki})^{\mathrm{T}}\boldsymbol{\beta}_k^*(\hat{\boldsymbol{z}}_{ki} - \tilde{\boldsymbol{z}}_{ki} - \bar{\boldsymbol{z}}_{ki})^{\mathrm{T}} - \tilde{\boldsymbol{z}}_{ki}^{\mathrm{T}}\boldsymbol{\beta}_k^*(\hat{\boldsymbol{z}}_{ki} - \tilde{\boldsymbol{z}}_{ki} - \bar{\boldsymbol{z}}_{ki})^{\mathrm{T}}\\ &- \bar{\boldsymbol{z}}_{ki}^{\mathrm{T}}\boldsymbol{\beta}_k^*\left(\hat{\boldsymbol{z}}_{ki} - \tilde{\boldsymbol{z}}_{ki} - \bar{\boldsymbol{z}}_{ki}\right)^{\mathrm{T}} - (\hat{\boldsymbol{z}}_{ki} - \tilde{\boldsymbol{z}}_{ki} - \bar{\boldsymbol{z}}_{ki})^{\mathrm{T}}\boldsymbol{\beta}_k^*\tilde{\boldsymbol{z}}_{ki}^{\mathrm{T}} - (\hat{\boldsymbol{z}}_{ki} - \tilde{\boldsymbol{z}}_{ki} - \bar{\boldsymbol{z}}_{ki})^{\mathrm{T}}\boldsymbol{\beta}_k^*\bar{\boldsymbol{z}}_{ki}\Big\}\end{aligned}$$

By Assumption 5 and (6), we have $\bar{y}_{ki} = O_p\left(n^{-\frac{1}{2}}\right)$ and $\|\bar{\boldsymbol{z}}_{ki}\| = O_p\left((n/p_n)^{-\frac{1}{2}}\right)$.

Combining $E(\tilde{y}_{ki}) = 0$, $E(\tilde{\boldsymbol{z}}_{ki}) = \mathbf{0}$, equation (7), Assumption 7, Lemma A.1 in Liang & Li (2009) and only the first p_0 elements in $\boldsymbol{\beta}_k^*$ are nonzero,, we have

$$\begin{aligned}&\Sigma_{i=1}^{n_k}(\hat{y}_{ki} - \tilde{y}_{ki} - \bar{y}_{ki})(\hat{\boldsymbol{z}}_{ki} - \tilde{\boldsymbol{z}}_{ki} - \bar{\boldsymbol{z}}_{ki})^{\mathrm{T}} = o_p\left((np_n)^{\frac{1}{2}}\right), \Sigma_{i=1}^{n_k}(\hat{y}_{ki} - \tilde{y}_{ki} - \bar{y}_{ki})\tilde{\boldsymbol{z}}_{ki}^{\mathrm{T}} = o_p\left((np_n)^{\frac{1}{2}}\right),\\ &\Sigma_{i=1}^{n_k}(\hat{y}_{ki} - \tilde{y}_{ki} - \bar{y}_{ki})\bar{\boldsymbol{z}}_{ki}^{\mathrm{T}} = o_p\left((np_n)^{\frac{1}{2}}\right), \Sigma_{k=1}^{K}\Sigma_{i=1}^{n_k}\tilde{y}_{ki}(\hat{\boldsymbol{z}}_{ki} - \tilde{\boldsymbol{z}}_{ki} - \bar{\boldsymbol{z}}_{ki})^{\mathrm{T}} = o_p\left((np_n)^{\frac{1}{2}}\right), \Sigma_{i=1}^{n_k}\bar{y}_{ki}(\hat{\boldsymbol{z}}_{ki} - \tilde{\boldsymbol{z}}_{ki} - \bar{\boldsymbol{z}}_{ki})^{\mathrm{T}}\\ &= o_p\left((np_n)^{\frac{1}{2}}\right), \Sigma_{i=1}^{n_k}\tilde{y}_{ki}\bar{\boldsymbol{z}}_{ki}^{\mathrm{T}} = o_p\left((np_n)^{\frac{1}{2}}\right), \Sigma_{i=1}^{n_k}\bar{y}_{ki}\tilde{\boldsymbol{z}}_{ki}^{\mathrm{T}} = o_p\left((np_n)^{\frac{1}{2}}\right), \sum_{i=1}^{n_k}\bar{y}_{ki}\bar{\boldsymbol{z}}_{ki}^{\mathrm{T}} = o_p\left((np_n)^{\frac{1}{2}}\right), \Sigma_{i=1}^{n_k}\tilde{\boldsymbol{z}}_{ki}^{\mathrm{T}}\boldsymbol{\beta}_k^*\bar{\boldsymbol{z}}_{ki}^{\mathrm{T}}\\ &= o_p\left((np_n)^{\frac{1}{2}}\right), \Sigma_{i=1}^{n_k}\bar{\boldsymbol{z}}_{ki}^{\mathrm{T}}\boldsymbol{\beta}_k^*\tilde{\boldsymbol{z}}_{ki}^{\mathrm{T}} = o_p\left((np_n)^{\frac{1}{2}}\right), \Sigma_{i=1}^{n_k}\bar{\boldsymbol{z}}_{ki}^{\mathrm{T}}\boldsymbol{\beta}_k^*\bar{\boldsymbol{z}}_{ki}^{\mathrm{T}} = o_p\left((np_n)^{\frac{1}{2}}\right),\\ &\Sigma_{i=1}^{n_k}(\hat{\boldsymbol{z}}_{ki} - \tilde{\boldsymbol{z}}_{ki} - \bar{\boldsymbol{z}}_{ki})^{T}\boldsymbol{\beta}_k^*(\hat{\boldsymbol{z}}_{ki} - \tilde{\boldsymbol{z}}_{ki} - \bar{\boldsymbol{z}}_{ki})^{\mathrm{T}} = o_p\left((np_n)^{\frac{1}{2}}\right), \Sigma_{i=1}^{n_k}\bar{\boldsymbol{z}}_{ki}^{\mathrm{T}}\boldsymbol{\beta}_k^*(\hat{\boldsymbol{z}}_{ki} - \tilde{\boldsymbol{z}}_{ki} - \bar{\boldsymbol{z}}_{ki})^{\mathrm{T}} = o_p\left((np_n)^{\frac{1}{2}}\right),\\ &\Sigma_{i=1}^{n_k}\bar{\boldsymbol{z}}_{ki}^{\mathrm{T}}\boldsymbol{\beta}_k^*(\hat{\boldsymbol{z}}_{ki} - \tilde{\boldsymbol{z}}_{ki} - \bar{\boldsymbol{z}}_{ki})^{\mathrm{T}} = o_p\left((np_n)^{\frac{1}{2}}\right), \Sigma_{i=1}^{n_k}(\hat{\boldsymbol{z}}_{ki} - \tilde{\boldsymbol{z}}_{ki} - \bar{\boldsymbol{z}}_{ki})^{T}\boldsymbol{\beta}_k^*\tilde{\boldsymbol{z}}_{ki}^{\mathrm{T}} = o_p\left((np_n)^{\frac{1}{2}}\right),\\ &\Sigma_{i=1}^{n_k}(\hat{\boldsymbol{z}}_{ki} - \tilde{\boldsymbol{z}}_{ki} - \bar{\boldsymbol{z}}_{ki})^{T}\boldsymbol{\beta}_k^*\bar{\boldsymbol{z}}_{ki} = o_p\left((np_n)^{\frac{1}{2}}\right)\end{aligned}$$

Moreover, by central limit theorem, we have $\Sigma_{k=1}^{K}\Sigma_{i=1}^{n_k}\left(\tilde{y}_{ki}\tilde{\boldsymbol{z}}_{ki}^{\mathrm{T}} - \tilde{\boldsymbol{z}}_{ki}^{\mathrm{T}}\boldsymbol{\beta}_k^{*}\tilde{\boldsymbol{z}}_{ki}^{\mathrm{T}} + \boldsymbol{\beta}_k^{*\mathrm{T}}\Sigma_k\right) = O_p\left((np_n)^{\frac{1}{2}}\right)$.

Thus $\Sigma_{i=1}^{n_k}\left(\hat{y}_{ki}\hat{\boldsymbol{z}}_{ki}^{\mathrm{T}} - \hat{\boldsymbol{z}}_{ki}^{\mathrm{T}}\boldsymbol{\beta}_k^{*}\hat{\boldsymbol{z}}_{ki}^{\mathrm{T}} + \boldsymbol{\beta}_k^{*\mathrm{T}}\Sigma_k\right) = o_p\left((np_n)^{\frac{1}{2}}\right)$ for any $k = 1, \cdots, K$.

For the second term of δ_k, we have

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

which is a positive definite matrix with constant eigenvalue by Assumption 3. Therefore, we conclude that there exists some constants c_1, c_2 and c_3 such that

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

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 $\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^{*} = O_p\left((n/p_n)^{-\frac{1}{2}}\right)$, which implies that $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{*} = O_p\left((n/p_n)^{-\frac{1}{2}}\right)$ and $\hat{\boldsymbol{\alpha}}_k - \boldsymbol{\alpha}_k^{*} = O_p\left((n/p_n)^{-\frac{1}{2}}\right)$ for any $k = 1, \cdots, K$.

For the nonlinear components estimation, we have $\hat{m}_{ky}(v_{ki}') - \hat{\boldsymbol{m}}_{kz}(v_{ki}')^{\mathrm{T}}\hat{\boldsymbol{\beta}}_k - f_k^{*}(w_{ki}) = \hat{m}_{ky}(v_{ki}') - m_{ky}(v_{ki}') + m_{ky}(v_{ki}') - m_{ky}(w_{ki}) + m_{ky}(w_{ki})$. By $-\left(\hat{\boldsymbol{m}}_{kz}(v_{ki}') - \boldsymbol{m}_{kz}(v_{ki}') + \boldsymbol{m}_{kz}(v_{ki}') - \boldsymbol{m}_{kz}(w_{ki}) + \boldsymbol{m}_{kz}(w_{ki})\right)^{\mathrm{T}}\left(\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^{*} + \boldsymbol{\beta}_k^{*}\right) - f_k^{*}(w_{ki})$ $\hat{m}_{ky}(v_{ki}') - m_{ky}(v_{ki}') = o_p\left(n^{-\frac{1}{4}}\right)$, $m_{ky}(v_{ki}') - m_{ky}(w_{ki}) = O_p\left(n^{-\frac{1}{2}}\right)$, $\hat{\boldsymbol{m}}_{kz}(v_{ki}') - \boldsymbol{m}_{kz}(v_{ki}') = o_p\left(n^{-\frac{1}{4}}\right)$, $\boldsymbol{m}_{kz}(v_{ki}')$ and $-\boldsymbol{m}_{kz}(w_{ki}) = O_p\left(n^{-\frac{1}{2}}\right)$, $\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^{*} = O_p\left((n/p_n)^{-\frac{1}{2}}\right)$

$m_{ky}(w_{ki}) = m_{kz}(w_{ki}) + f_k^{*}(w_{ki})$, we get $\hat{m}_{ky}(v_{ki}') - \hat{\boldsymbol{m}}_{kz}(v_{ki}')^{\mathrm{T}}\hat{\boldsymbol{\beta}}_k - f_k^{*}(w_{ki}) = O_p\left(\max\left\{n^{-\frac{1}{4}}, (n/p_n^3)^{-\frac{1}{2}}\right\}\right)$.

Then the desired result can be obtained.

Proof of Theorem 2:

We prove theorem 2 by contradiction. Suppose $\left|\hat{\beta}_j\right| > 0$ for $j > p_{n,0}$. Take the derivative for β_j and get the KKT condition

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

where $\boldsymbol{J}_j$ is the vector of all zeros except jth element is 1.3

By theorem 1 and the decomposition (10), the left hand side is $\Sigma_{k=1}^{K}\Sigma_{i=1}^{n_k}\left(\tilde{y}_{ki}+o_p\left(n^{-\frac{1}{4}}\right)-\left(\tilde{\boldsymbol{z}}_{ki}+o_p\left(n^{-\frac{1}{4}}\right)\right)^{\mathrm{T}}\left(\boldsymbol{\beta}_k^*+O_p\left((n/p_n)^{-\frac{1}{2}}\right)\right)\right)\left(\tilde{\boldsymbol{z}}_{ki}+o_p\left(n^{-\frac{1}{4}}\right)\right)_j$. Then by Lemma A.1 $-\Sigma_{k=1}^{K}n_k\left(\boldsymbol{\beta}_k^*+O_p\left((n/p_n)^{-\frac{1}{2}}\right)\right)^{\mathrm{T}}\Sigma_k\mathbf{J}_j$

in Liang & Li (2009) and Assumption 3, the left hand side can be simplified as

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

which is equal to

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

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

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

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

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

Proof of Theorem 3

The proof of theorem 3 was essentially the same as proof of theorem 1 and 2, which was omitted here.

Proof of Theorem 4:

The key idea of the proof is the same as proof of theorem 1 in Liang & Li (2009). Take the derivative for $\boldsymbol{\beta}_{kl}$ in equation (3) and get the KKT condition that

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

Same as the proof of theorem 1 in Liang & Li (2009), because

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

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

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

By $\boldsymbol{z}_{ki} - \boldsymbol{m}_{kz}(w_{ki}) = \boldsymbol{x}_{ki} - E(\boldsymbol{x}_{ki} \mid w_{ki}) + \boldsymbol{U}_{ki}$ 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

$$\frac{1}{\sqrt{n_k}} \sum_{i=1}^{n_k} \left\{ [\boldsymbol{x}_{ki} - E(\boldsymbol{x}_{ki} \mid w_{ki}) + \boldsymbol{U}_{ki}]_I^{\otimes 2} - \Sigma_{kI} \right\} \left(\hat{\boldsymbol{\beta}}_{kI} - \boldsymbol{\beta}_{kI}^*\right)$$

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

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

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

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

References

Ahn S. & Wang T. (2013), A powerful statistical method for identifying differentially methylated markers in complex diseases, in 'Biocomputing 2013', World Scientific, pp. 69–79.

Arslan A, Tuminello S, Yang L, Zhang Y, Durmus N, Snuderl M, Heguy A, Zeleniuch-Jacquotte A, Shao Y. & Reibman J. (2020), 'Genome-wide dna methylation profiles in community members exposed to the world trade center disaster', International journal of environmental research and public health 17(1), 5493. [PubMed: 32751422]

Boada M, Anaya F, Ortiz P, Olazarán J, Shua-Haim JR, Obisesan TO, Hernández I, Muñoz J, Buendia M, Alegret M. et al. (2017), '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', Journal of Alzheimer's Disease 56(1), 129–143.

Boyd S, Parikh N, Chu E, Peleato B, Eckstein J. et al. (2011), 'Distributed optimization and statistical learning via the alternating direction method of multipliers', Foundations and Trends® in Machine learning 3(1), 1–122.

Chen C, Grennan K, Badner J, Zhang D, Gershon E, Jin L. & Liu C. (2011), 'Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods', PloS one 6(2), e17238.

Chen X, Shao Y, & Sadowski M. (2021), 'Segmented linear mixed model analysis reveals association of the apoe e4 allele with faster rate of alzheimer dementia progression', Journal of Alzheimer's Disease 82(3), 921–937.

Cruz C, Llop-Guevara A, Garber JE, Arun BK, Fidalgo JAP, Lluch A, Telli ML, Fernández C, Kahatt C, Galmarini CM et al. (2018), 'Multicenter phase ii study of lur-binectedin in brca-mutated and unselected metastatic advanced breast cancer and biomarker assessment substudy', Journal of Clinical Oncology 36(31), 3134. [PubMed: 30240327]

de Leon MJ, Pirraglia E, Osorio RS, Glodzik L, Saint-Louis L, Kim H-J, Fortea J, Fossati S, Laska E, Siegel C. et al. (2018), 'The nonlinear relationship between cerebrospinal fluid ab42 and tau in preclinical alzheimer's disease', PloS one 13(2), e0191240.

Ewers M, Mattsson N, Minthon L, Molinuevo JL, Antonell A, Popp J, Jessen F, Herukka S-K, Soininen H, Maetzler W. et al. (2015), 'Csf biomarkers for the differential diagnosis of alzheimer's disease: a large-scale international multicenter study', Alzheimer's & Dementia 11(11), 1306–1315.

Fan J. & Gijbels I. (1996), 'Local polynomial modelling and its applications'.

Hansson O, Zetterberg H, Buchhave P, Londos E, Blennow K. & Minthon L. (2006), 'Association between csf biomarkers and incipient alzheimer's disease in patients with mild cognitive impairment: a follow-up study', The Lancet Neurology 5(3), 228–234. [PubMed: 16488378]

Hardle W, Liang H. & Gao J. (2012),¨ Partially linear models, Springer Science & Business Media.

He X, Sun X. & Shao Y. (2021), 'Network-based survival model to discover target genes for developing cancer immunotherapies and predicting patient survival', Journal of Applied Statistics 48, 1352–1373. [PubMed: 35444359]

Herukka S-K, Helisalmi S, Hallikainen M, Tervo S, Soininen H. & Pirttilä T. (2007), 'Csf¨ ab42, tau and phosphorylated tau, apoe4 allele and mci type in progressive mci', Neurobiology of aging 28(4), 507–514. [PubMed: 16546302]

Hessell A, Li L, Malherbe D, Barnette P, Pandey e. a., Haigwood N. & Gorny M. (2021), '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', Journal of Immunology 206, 1266–1283.

Khan W, Giampietro V, Banaschewski T, Barker GJ, Bokde AL, Buchel C, Conrod P,¨ Flor H, Frouin V, Garavan H. et al. (2017), 'A multi-cohort study of apoe4 and amyloid-β effects on the hippocampus in alzheimer's disease', Journal of Alzheimer's Disease 56(3), 1159–1174.

Liang H. & Li R. (2009), 'Variable selection for partially linear models with measurement errors', Journal of the American Statistical Association 104(485), 234–248. [PubMed: 20046976]

Lim AS, Gaiteri C, Yu L, Sohail S, Swardfager W, Tasaki S, Schneider JA, Paquet C,Stuss DT, Masellis M. et al. (2018), 'Seasonal plasticity of cognition and related biological measures in adults with and without alzheimer disease: Analysis of multiple cohorts', PLoS medicine 15(9), e1002647. [PubMed: 30180184]

Niemantsverdriet E, Ribbens A, Bastin C, Benoit F, Bergmans B, Bier J-C, Bladt R, Claes L, De Deyn PP, Deryck O. et al. (2018), 'A retrospective belgian multi-center mri biomarker study in alzheimer's disease (remember)', Journal of Alzheimer's disease 63(4), 1509–1522.

Pannee J, Shaw L, Korecka e. a., Blennow K. & Zetterberg H. (2021), 'The global alzheimer's association round robin study on plasma amyloid beta methods', Alzheimer's Dement. 13, e12242.

Rahbar K, Ahmadzadehfar H, Kratochwil C, Haberkorn U, Schafers M, Essler M, Baum RP, Kulkarni HR, Schmidt M, Drzezga A. et al. (2017), 'German multicenter study investigating 177lu-psma-617 radioligand therapy in advanced prostate cancer patients', Journal of Nuclear Medicine 58(1), 85–90. [PubMed: 27765862]

Roach PJ, Francis R, Emmett L, Hsiao E, Kneebone A, Hruby G, Eade T, Nguyen QA, Thompson BD, Cusick T. et al. (2018), 'The impact of 68ga-psma pet/ct on management intent in prostate cancer:

results of an australian prospective multicenter study', Journal of Nuclear Medicine 59(1), 82–88. [PubMed: 28646014]

Scherer A. (2009), Batch effects and noise in microarray experiments: sources and solutions, Vol. 868, John Wiley & Sons.

Schwarz G. et al. (1978), 'Estimating the dimension of a model', The annals of statistics 6(2), 461–464.

Shoji M, Kanai M, Matsubara E, Tomidokoro Y, Shizuka M, Ikeda Y, Ikeda M, Harigaya Y, Okamoto K. & Hirai S. (2001), 'The levels of cerebrospinal fluid aβ40 and aβ42 (43) are regulated age-dependently', Neurobiology of aging 22(2), 209–215. [PubMed: 11182470]

Sturdza A, Pötter R, Fokdal LU, Haie-Meder C, Tan LT, Mazeron R, Petric P, Šegedin B, Jurgenliemk-Schulz IM, Nomden C. et al. (2016), 'Image guided brachytherapy in locally advanced cervical cancer: improved pelvic control and survival in retroembrace, a multicenter cohort study', Radiotherapy and Oncology 120(3), 428–433. [PubMed: 27134181]

Sun X, Liu X, Xia M, Shao Y. & Zhang XD (2019), 'Multicellular gene network analysis identifies a macrophage-related gene signature predictive of therapeutic response and prognosis of gliomas', Journal of translational medicine 17(1), 159. [PubMed: 31097021]

Van Steenoven I, Aarsland D, Weintraub D, Londos E, Blanc F, Van der Flier WM, Teunissen CE, Mollenhauer B, Fladby T, Kramberger MG et al. (2016), 'Cerebrospinal fluid alzheimer's disease biomarkers across the spectrum of lewy body diseases: results from a large multicenter cohort', Journal of Alzheimer's Disease 54(1), 287–295.

Zhao P. & Xue L. (2010), 'Variable selection for semiparametric varying coefficient partially linear errors-in-variables models', Journal of Multivariate Analysis 101(8), 1872–1883.

Zhou Z, Jiang R. & Qian W. (2011), 'Variable selection for additive partially linear models with measurement error', Metrika 74(2), 185–202.

Zou H. (2006), 'The adaptive lasso and its oracle properties', Journal of the American statistical association 101(476), 1418–1429.

Zou H. & Zhang HH (2009), 'On the adaptive elastic-net with a diverging number of parameters', Annals of statistics 37(4), 1733–1751. [PubMed: 20445770]

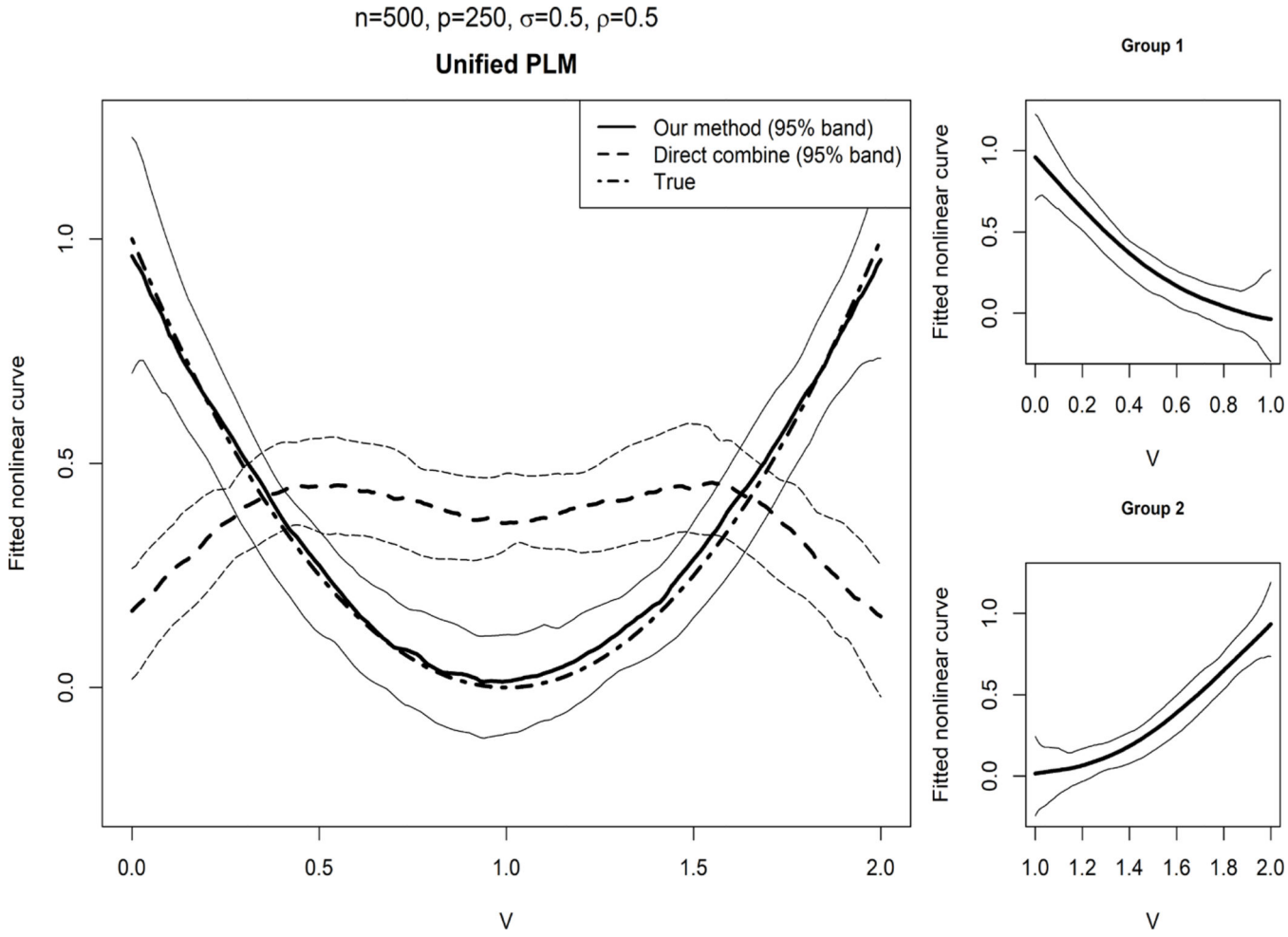

Figure 1:
Nonlinear curve estimation in simulation study.

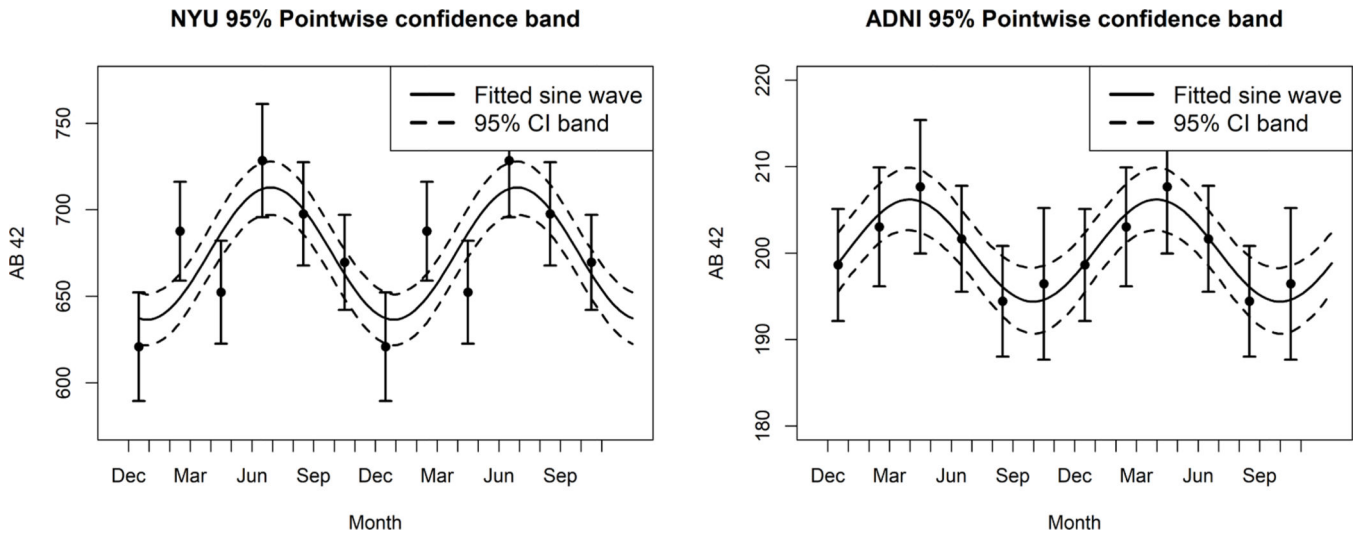

Figure 2:
Batch effect of $A\beta_{42}$.

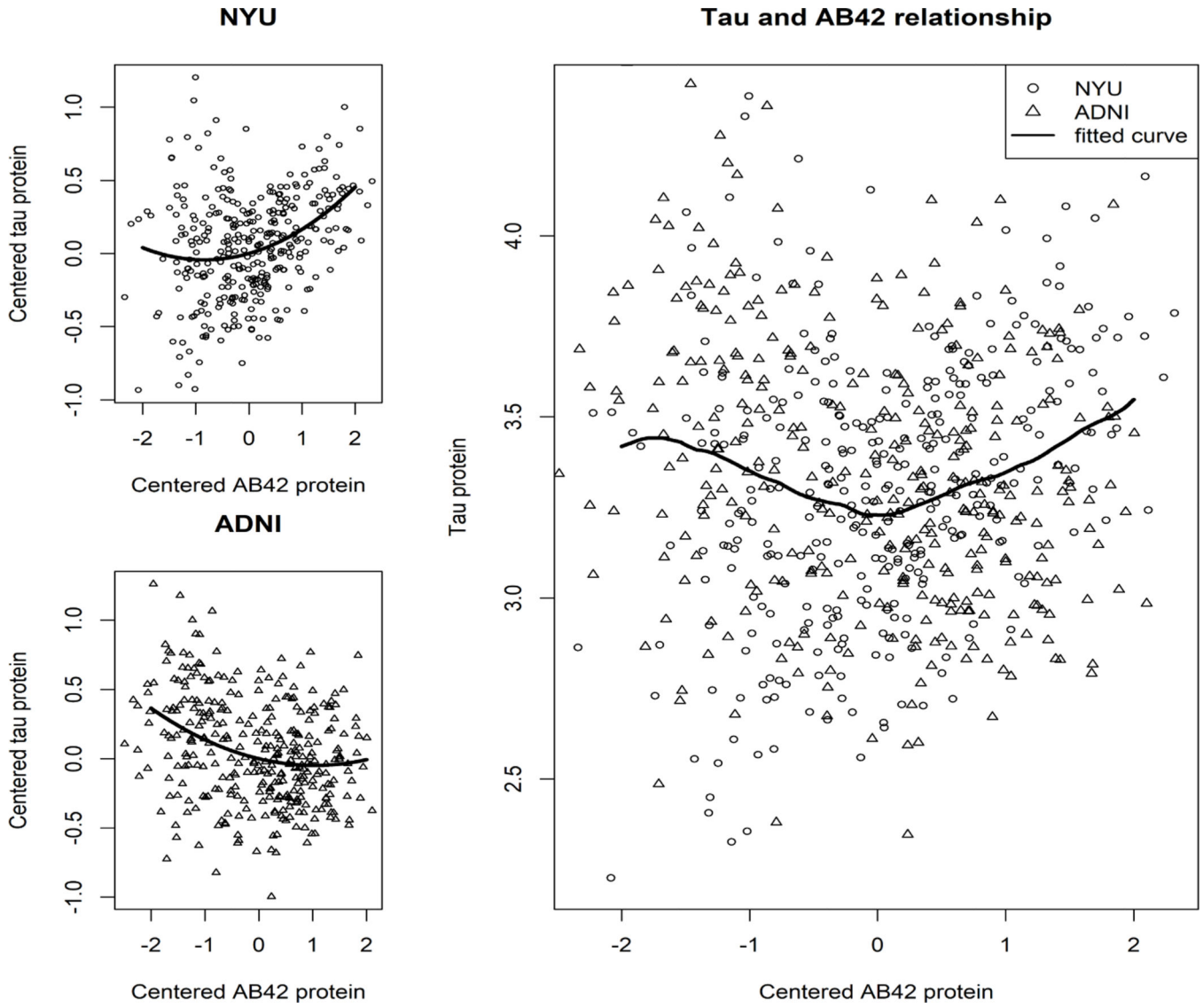

Figure 3:
Individual group and combined group analysis.

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

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

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)	$\hat{\beta}_{11}$	Ada Lasso	0.505	0.048	0.040	94.7%
		Lasso	0.503	0.047	0.040	94.7%
	$\hat{\beta}_{21}$	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)	$\hat{\beta}_{11}$	Ada Lasso	0.496	0.044	0.051	97.3%
		Lasso	0.496	0.044	0.051	97.3%
	$\hat{\beta}_{21}$	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)	$\hat{\beta}_{11}$	Ada Lasso	0.503	0.046	0.040	95.2%
		Lasso	0.503	0.047	0.040	95.2%
	$\hat{\beta}_{21}$	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)	$\hat{\beta}_{11}$	Ada Lasso	0.498	0.055	0.050	96.8%
		Lasso	0.498	0.055	0.050	96.8%
	$\hat{\beta}_{21}$	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.

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

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)	$\hat{\beta}_{11}$	Ada Lasso	5.057	0.302	0.327	96.3%
		Lasso	5.057	0.302	0.327	96.3%
	$\hat{\beta}_{21}$	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)	$\hat{\beta}_{11}$	Ada Lasso	5.060	0.375	0.370	93.8%
		Lasso	5.060	0.375	0.370	93.8%
	$\hat{\beta}_{21}$	Lasso	0.982	0.164	0.143	92.6%
		Lasso	0.982	0.164	0.143	92.6%
(500, 250, 0.25, 0.25)	$\hat{\beta}_{11}$	Ada Lasso	4.984	0.349	0.324	94.7%
		Lasso	4.984	0.349	0.324	94.7%
	$\hat{\beta}_{21}$	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)	$\hat{\beta}_{11}$	Ada Lasso	5.014	0.416	0.359	94.5%
		Lasso	5.014	0.416	0.359	94.5%
	$\hat{\beta}_{21}$	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.