Journal of Machine Learning Research 17 (2016) 1-23 Submitted 11/15; Revised 6/16; Published 7/16

Fused Lasso Approach in Regression Coefficients Clustering – Learning Parameter Heterogeneity in Data Integration

Lu Tang LUTANG@UMICH.EDU
Peter X.K. Song PXSONG@UMICH.EDU
Department of Biostatistics
University of Michigan
Ann Arbor, MI 48109, USA

Editor: Hui Zou

Abstract

As data sets of related studies become more easily accessible, combining data sets of similar studies is often undertaken in practice to achieve a larger sample size and higher power. A major challenge arising from data integration pertains to data heterogeneity in terms of study population, study design, or study coordination. Ignoring such heterogeneity in data analysis may result in biased estimation and misleading inference. Traditional techniques of remedy to data heterogeneity include the use of interactions and random effects, which are inferior to achieving desirable statistical power or providing a meaningful interpretation, especially when a large number of smaller data sets are combined. In this paper, we propose a regularized fusion method that allows us to identify and merge inter-study homogeneous parameter clusters in regression analysis, without the use of hypothesis testing approach. Using the fused lasso, we establish a computationally efficient procedure to deal with large-scale integrated data. Incorporating the estimated parameter ordering in the fused lasso facilitates computing speed with no loss of statistical power. We conduct extensive simulation studies and provide an application example to demonstrate the performance of the new method with a comparison to the conventional methods.

Keywords: Fused lasso, Data integration, Extended BIC, Generalized Linear Models

1. Introduction

Combining data sets collected from multiple studies is undertaken routinely in practice to achieve a larger sample size and higher statistical power. Such information integration is commonly seen in biomedical research, for example, the study of genetics or rare diseases where data repositories are available. The motivation of this paper arises from the consideration of data heterogeneity during data integration. Although data integration has different meanings, in here, we consider the concatenation of data sets of similar studies over different subjects, where the number of integrated data sets can be very large.

Inter-study heterogeneity can result from the differences in study environment, population, design and protocols (Leek and Storey, 2007; Sutton and Higgins, 2008; Liu et al., 2015). Data heterogeneity is likely attributed to population parameter heterogeneity, where the association of interest can differ across different study populations from which data sets are collected. Examples include multi-center clinical trials when participant data from different sites are combined (Shekelle et al., 2003) and genetics studies when genomic data

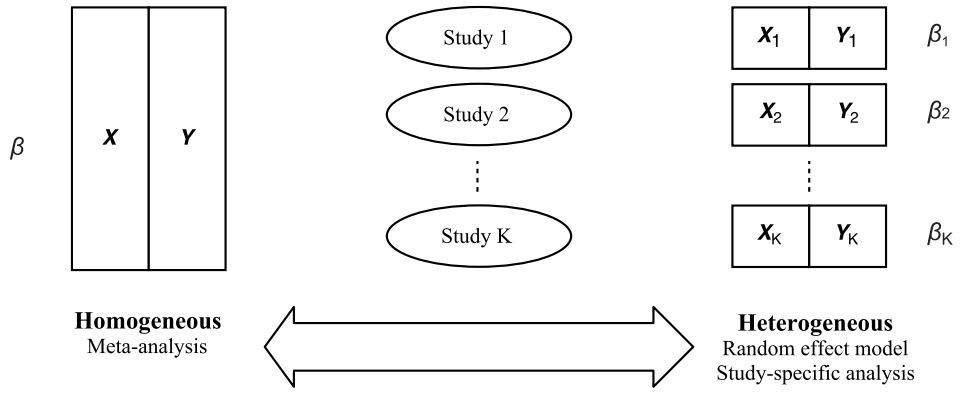

Figure 1: Homogeneous assumption (left) versus heterogeneous assumption (right).

from multiple similar studies are combined (Lohmueller et al., 2003; Sullivan et al., 2000). Discrepancies in treatment effect or trait-gene association may arise due to the differences in facilities, practices and patient characteristics across studies, albeit the adjustment of confounding (Leek and Storey, 2007). The parameter heterogeneity introduced in data integration compromises the power of the larger sample size and may even lead to biased results and misleading scientific conclusions. Thus, counterintuitively, the model obtained from the combined studies may not serve as a proper prediction model for each individual study in the case of heterogeneous study populations.

Traditional treatments of parameter heterogeneity are not optimal. Meta-analysis methods such as combining summary statistics (Glass, 1976), estimating functions (Hansen, 1982; Qin and Lawless, 1994) or p-values functions (Xie et al., 2012) are built upon the assumption of complete parameter homogeneity, as shown in the left panel of Figure 1. This assumption is hardly valid in practice. When individual participant data from multiple data sets are available, a retreat to the classical meta-analysis methods is necessary, because in this case assessing the assumption of inter-study homogeneity becomes possible. The two most common approaches to handling parameter heterogeneity include (i) specifying study-specific effects by including interaction terms between study indicator and covariates (e.g., Lin et al. (1998)), and (ii) utilizing random covariate effects by allowing variations across studies as random variables (e.g., DerSimonian and Kacker (2007)). Both approaches essentially assume fully heterogeneous covariate effects, namely, each study having its own set of regression coefficients, as shown in the right panel of Figure 1.

When study-specific effects are of interest, the interaction-based formulation may lead to over-parameterization, which impairs statistical power. The most straightforward way to reduce the number of parameters is to identify clusters of homogeneous parameters through exhaustive tests for the differences between every pair of study-specific coefficients. However, when the number of data sets is large, the use of hypothesis testing to determine parameter clusters becomes untrackable in addition to the multiple-testing problem. One may draw different or even conflicting conclusions due to different orders of hypotheses performed.

In reality, covariate effects from multiple studies are likely to form groups, a scenario falling in between the complete heterogeneity and the complete homogeneity. This leads

to the following two essential yet related analytic tasks: (i) to assess the inter-study heterogeneity, so to determine an appropriate form of parsimonious parameterization in model specification; and (ii) to identify and merge groups of homogeneous parameters for better statistical power for parameter estimation and inference based on a more parsimonious model. Along the idea of lasso shrinkage estimator (Tibshirani, 1996), fused lasso methods (Tibshirani et al., 2005; Friedman et al., 2007; Yang et al., 2012) have been introduced to achieve covariate grouping, where covariate adjacencies are naturally defined by a metric of time, location or network structure. In our problem of data integration, there does not exist a natural metric to define the ordering of regression coefficients from different studies. Shen and Huang (2010) proposed the grouping pursuit via penalization of all pairwise coefficient differences in a single study, where covariate orderings are not considered. To reduce the computational burden in the all-pairs based regularization, Wang et al. (2016) and Ke et al. (2015) used the initial coefficient estimates to establish certain ordering and then to define parameter adjacencies. However, most of these studies have been entirely focusing on a single cohort of subjects from a single study. For example, Shin et al. (2016) proposed to fuse regression coefficients of different loss functions obtained from a single study, such as coefficients from different quantile regression models. Limited publication of fusion learning and grouping pursuit has been available in the literature, except Wang et al. (2016), to assess the differences and similarities among regression coefficients across multiple studies in the scenario of data integration.

In this paper, we propose an agglomerative clustering method for regression coefficients in the context of data integration, named as the *Fused Lasso Approach in Regression Coefficients Clustering (FLARCC)*. FLARCC is proposed to identify heterogeneity patterns of regression coefficients across studies (or data sets) and to provide estimates of all regression coefficients simultaneously. It is interesting to draw a connection between our method and Pan et al. (2013) where they consider a classic clustering problem of individual responses by pairwise coefficient fusion via penalized regression. Their method aims at clustering subjects, while our method focuses on clustering regression coefficients across multiple data sets, and these two methods coincide only in a special case where each study is composed of only one subject. FLARCC achieves clustering of study-specific effects by penalizing the ℓ_1-norm differences of adjacent coefficients, with adjacency defined by the estimated ranks. Our method extends the bCARDS method in Ke et al. (2015) from one study to multiple studies as well as from the linear model to the generalized linear models, and focuses on simultaneous clustering of regression coefficients of individual covariates from multiple studies in data integration. An R package `metafuse` is created as part of our methodology development to perform the proposed integrated data analysis which can be downloaded from the Comprehensive R Archive Network (web link `https://cran.r-project.org/web/packages/metafuse`).

In the proposed method, tuning parameter is used to determine the clustering pattern of coefficients across data sets. Specifically, let λ be the tuning parameter of regularization. If $\lambda = 0$ (i.e., no penalty), FLARCC becomes a method under the setting of complete heterogeneity, so that study-specific regression coefficients for each covariate are assumed different across data sets. If λ is large enough that all differences of regression coefficients are shrunk to zero, FLARCC reduces to a homogeneous model in that a common regression coefficient for each covariate is assumed for all studies. In light of the hierarchical clustering

scheme, these two extreme cases above correspond to the start and end of an agglomerative clustering, respectively; however, the reality is believed to reside in between. Analogous to dendrograms in the hierarchical clustering, we propose a new tree-type graphic display, named as *fusogram*, which presents tree-based coefficient clusters according to solution paths obtained from FLARCC. The selection of optimal λ pertains to pruning of clustering trees, which can be based on certain model selection criterion. We use the extended Bayesian information criterion (EBIC) proposed by Chen and Chen (2008) as our model selection criterion and show that EBIC exhibits better performance than BIC when the number of studies (or data sets) is large. In addition, we propose a scaling strategy to "harmonize" solution paths by covariate-wise adaptive weights to allow flexible tuning, which further improves the clustering performance.

The rest of this paper is organized as follows. Section 2 describes FLARCC in detail under the generalized linear models (GLM) framework. Section 3 presents the theoretical properties of the proposed method (with technical proofs presented in the Appendix). Section 4 discusses the interpretation and selection of the tuning parameter. In Section 5, we use simulation studies to evaluate the performance of our method. A real data analysis is given in Section 6 with interpretation of coefficient estimates and illustration of *fusograms*. Discussion and concluding remarks are in Section 7.

2. Method of Parameter Fusion

In this section, we present the method and algorithm of FLARCC.

2.1 Notations and Method

We start by introducing necessary notations. Throughout this paper, i, j and k are used to index subject, covariate and study, respectively. For instance, $X_{j,k}^{(i)}$ denotes the measurement of the jth covariate from the ith individual from study k, and $Y_k^{(i)}$ is the measurement of a response variable from the ith individual from study k. The total number of studies is denoted as K and the number of covariates involved is p. The sample size for study k is n_k, $k = 1, \ldots, K$, and the combined sample size is $N = \sum_{k=1}^{K} n_k$. The collection of all coefficients (covariates-wise) is denoted as $\boldsymbol{\beta} = (\boldsymbol{\beta}_{1,\bullet}^\top, \boldsymbol{\beta}_{2,\bullet}^\top, \ldots, \boldsymbol{\beta}_{p,\bullet}^\top)^\top$ with $\boldsymbol{\beta}_{j,\bullet} = (\beta_{j,1}, \ldots, \beta_{j,K})^\top$ for $j = 1, \ldots, p$. An indicator vector $\boldsymbol{c} = (c_1, \ldots, c_p)^\top$ is used to flag heterogeneous covariates, namely if the jth covariate is treated as heterogeneous (i.e., all different coefficients across K studies) then $c_j = 1$ and as homogeneous (a common coefficient across K studies) otherwise. Thus $c_j = 0$ for some $j \in \{1, \ldots, p\}$ implies that coefficient vector $\boldsymbol{\beta}_{j,\bullet}$ reduces to a common scalar parameter β_j for all K studies.

For illustration, let us consider a simple scenario of $\boldsymbol{c} = (1, 1, 0, \ldots, 0)^\top$, in which the first two covariates are set as heterogeneous and the remaining $p - 2$ covariates are set as homogeneous. The resulting coefficient vector is $\boldsymbol{\beta} = (\boldsymbol{\beta}_{1,\bullet}^\top, \boldsymbol{\beta}_{2,\bullet}^\top, \beta_3, \ldots, \beta_p)^\top$. Then the corresponding design matrix $\boldsymbol{X}$ can be written as

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_{1,1} & & & \boldsymbol{X}_{2,1} & & & \boldsymbol{X}_{3,1} & \ldots & \boldsymbol{X}_{p,1} \\ & \ddots & & & \ddots & & \vdots & \ddots & \vdots \\ & & \boldsymbol{X}_{1,K} & & & \boldsymbol{X}_{2,K} & \boldsymbol{X}_{3,K} & \ldots & \boldsymbol{X}_{p,K} \end{pmatrix}_{N \times (2K+p-2)}$$

where $\boldsymbol{X}_{j,k} = (X_{j,k}^{(1)}, \ldots, X_{j,k}^{(n)})^\top$, $j = 1, \ldots, p$, $k = 1, \ldots, K$. The specification of $\boldsymbol{c}$ is can be dependent on the study interest. For example, in a multi-center clinical trial where we believe that the differences between the services provided across centers are non-negligible, but the study participants are similar, we can specify the clinic-related variables (e.g., treatment and cost) to be heterogeneous and the patient-related variables (e.g., age and gender) to be homogeneous. In addition, the specification of $\boldsymbol{c}$ can be dependent on preliminary marginal analysis of the homogeneousness of each variable, such as tests for random effects. When the homogeneousness of a covariate is unclear, we suggest specifying it as heterogeneous rather than homogeneous.

Under the assumption that both within-study and between-study samples are independent, for any $\boldsymbol{c} = (c_1, \ldots, c_p)^\top$ with $c_j \in \{0, 1\}$, $j = 1, \ldots, p$, the initial estimate of $\boldsymbol{\beta}$, which gives the starting level of clustering (i.e., $\lambda = 0$), can be consistently estimated by the maximum likelihood estimator

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{(K \times p)}}{\operatorname{argmax}} \frac{1}{K} \sum_{k=1}^{K} \frac{1}{n_k} \log L_k(\boldsymbol{\beta}), \tag{1}$$

where $L_k(\boldsymbol{\beta}) = \prod_{i=1}^{n_k} L_k^{(i)}(\boldsymbol{\beta}), k = 1, \ldots, K$ are the study-specific likelihoods from the given GLMs. For the purpose of parameter grouping and fusion, we propose the regularized maximum likelihood estimation for $\boldsymbol{\beta}$ by minimizing the following objective function:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^{(K \times p)}} \left(-\frac{1}{K} \sum_{k=1}^{K} \frac{1}{n_k} \log L_k(\boldsymbol{\beta}) + P(\boldsymbol{\beta}) \right), \tag{2}$$

where $P(\boldsymbol{\beta})$ is a penalty function of certain form. Here we adopt weighting $\frac{1}{n_k}$ to balance the contribution from each study so to avoid the dominance of large studies. Other types of weighting schemes may be considered to serve for different purposes, such as the inverse of estimated variances of initial estimates, which helps to achieve better estimation precision.

To achieve parameter fusion, Shen and Huang (2010) proposed the grouping pursuit algorithm, which specifies the sum of ℓ_1-norm differences of all study-specific coefficient pairs among individual heterogeneous coefficient vectors $\boldsymbol{\beta}_{j,\bullet}$, where $c_j = 1$, as the penalty:

$$P_\lambda(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{p} c_j \sum_{k=1}^{K-1} \sum_{k'>k}^{K} |\beta_{j,k} - \beta_{j,k'}|,$$

with $\lambda \geq 0$. In this penalty, there are $\binom{K}{2}$ terms of pairwise differences for each heterogeneous covariate and the total number of terms increases by an order of $O(K^2)$, given p fixed. This penalty contains many redundant constraints and imposes great computational challenges as pointed out in Shen and Huang (2010) and Ke et al. (2015).

Following arguments in Wang et al. (2016) and Ke et al. (2015), we develop the method of FLARCC by a simplified penalty function that uses the information on the ordering of coefficients. For the jth covariate, let $\boldsymbol{U}_j = (U_{j,1}, \ldots, U_{j,K})^\top$ be the ranking with no ties of $\boldsymbol{\beta}_{j,\bullet} = (\beta_{j,1}, \ldots, \beta_{j,K})^\top$, from the smallest to the largest. Specifically, $U_{j,k} = \sum_{k'=1}^{K} \mathbf{1}\{\beta_{j,k'} \leq \beta_{j,k}\}$ if there are no ties in $\boldsymbol{\beta}_{j,\bullet}$; otherwise, the ties in $\boldsymbol{U}_j$ are resolved by the first-occurrence-wins

rule according to k to ensure rank uniqueness. Then, the fusion penalty in FLARCC with parameter orderings $\boldsymbol{U}_j$, $j = 1, \ldots, p$, takes the form:

$$P_\lambda(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{p} c_j \nu_j \sum_{k=1}^{K-1} \sum_{k'>k}^{K} \mu_{j,k,k'} \mathbf{1}\{|U_{j,k} - U_{j,k'}| = 1\} |\beta_{j,k} - \beta_{j,k'}|, \tag{3}$$

where the constraints occur effectively only on adjacent ordered pairs. Clearly, the penalty in (3) only involves $K - 1$ terms for each case of $c_j = 1$, which is of an order $O(K)$, given p fixed. The ν_j's and $\mu_{j,k,k'}$'s in (3) are weights. Following Zou (2006), we choose adaptive weights $\hat{\mu}_{j,k,k'} = 1/|\hat{\beta}_{j,k} - \hat{\beta}_{j,k'}|^r$, $r > 0$, so that parameters with smaller difference will be penalized more than those with larger differences. Similarly, for a group of parameters $\boldsymbol{\beta}_{j,\bullet} = (\beta_{j,1}, \ldots, \beta_{j,K})^\top$, ν_j is an adaptive weight to characterize the degree of heterogeneousness of $\boldsymbol{\beta}_{j,\bullet}$. Specifically, in this paper we let $\hat{\nu}_j = 1/|\hat{\beta}_{j,(K)} - \hat{\beta}_{j,(1)}|^s$, the inverse of the range of the estimates, with $s \geq 0$; when a covariate is homogeneous, the differences of study-specific coefficients will be penalized more than those that are heterogeneous. In this way, we can "harmonize" solution paths so to greatly improve the performance by a single tuning parameter. We compare $s = 0$ and $s = 1$ in the simulation experiments and show in Section 5 that the introduction of such group-wise weights ν_j, $j = 1, \ldots, p$, gives rise to improvement on the performance of identifying homogeneous covariates when K and p are large.

A sparse version of FLARCC can also be achieved by including the traditional lasso penalty in (3) for covariate selection. In order to minimize the interference between fusion and sparsity penalties, we only encourage sparsity for the coefficient closest to zero in each $\boldsymbol{\beta}_{j,\bullet} = (\beta_{j,1}, \ldots, \beta_{j,K})^\top$, for $j = 1, \ldots, p$. Similar to the definition of $\boldsymbol{U}_j$, let $\boldsymbol{V}_j = (V_{j,1}, \ldots, V_{j,K})^\top$ be the ranking with no ties, from the smallest to the largest, of the absolute values of $\boldsymbol{\beta}_{j,\bullet}$, i.e., $(|\beta_{j,1}|, \ldots, |\beta_{j,K}|)^\top$. First we calculate $\boldsymbol{V}_j$ by $V_{j,k} = \sum_{k'=1}^{K} \mathbf{1}\{|\beta_{j,k'}| \leq |\beta_{j,k}|\}$, then we resolve the ties in $\boldsymbol{V}_j$ by the first-occurrence-wins rule according to k. Thus we can extend (3) to achieve variable selection by the following penalty function:

$$\begin{aligned} P_{\lambda,\alpha}(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{p} c_j \nu_j \sum_{k=1}^{K-1} \sum_{k'>k}^{K} \mu_{j,k,k'} \mathbf{1}\{|U_{j,k} - U_{j,k'}| = 1\} |\beta_{j,k} - \beta_{j,k'}| \\ + \alpha\lambda \sum_{j=1}^{p} \sum_{k=1}^{K} \mu_{j,k} \mathbf{1}\{V_{j,k} = 1\} |\beta_{j,k}|, \end{aligned} \tag{4}$$

where $\alpha \geq 0$ is another tuning parameter that controls the relative ratio between fusion and sparsity penalties, and $\hat{\mu}_{j,k} = 1/|\hat{\beta}_{j,k}|^r$. The sparsity penalty, although only enforced on the smallest coefficient in absolute value of $\boldsymbol{\beta}_{j,\bullet}$, is capable of shrinking a group of coefficients to zero when combined with the fusion penalty.

In practice, the weights (ν_j, $\mu_{j,k,k'}$ and $\mu_{j,k}$) and the parameter orderings ($\boldsymbol{U}_j$ and $\boldsymbol{V}_j$) are unknown, for $j = 1, \ldots, p$. We replace them with their estimates based on root-n consistent estimates $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_{1,\bullet}^\top, \ldots, \hat{\boldsymbol{\beta}}_{p,\bullet}^\top)^\top$, such as those from (1). In the simulation experiments and the real data application of this paper, we set $r = 1$ in $\hat{\mu}_{j,k,k'}$ and $\hat{\mu}_{j,k}$.

2.2 Algorithm

Optimization problem (2) with $P(\boldsymbol{\beta}) = P_{\lambda,\alpha}(\boldsymbol{\beta})$ given in (4) can be carried out by a lasso regression through suitable reparameterization. Let the ordered coefficients of $\boldsymbol{\beta}_{j,\bullet}$ in an ascending order based on ranking $\boldsymbol{U}_j$ be $(\beta_{j,(1)}, \ldots, \beta_{j,(K)})^\top$, $j = 1, \ldots, p$. For the jth covariate, consider a set of transformed parameters $\boldsymbol{\theta}_{j,\bullet} = (\theta_{j,1}, \ldots, \theta_{j,K})^\top$ defined by

$$\begin{aligned} \theta_{j,1} &= \beta_{j,k}, && \text{for } k \text{ s.t. } V_{j,k} = 1; \\ \theta_{j,k} &= \beta_{j,(k)} - \beta_{j,(k-1)}, && \text{for } k = 2, \ldots, K. \end{aligned}$$

Then the $P_{\lambda,\alpha}(\boldsymbol{\beta})$ in (4) can be rewritten as

$$P_{\lambda,\alpha}(\boldsymbol{\theta}) = \lambda \sum_{j=1}^{p} \sum_{k=1}^{K} \omega_{j,k} |\theta_{j,k}|, \tag{5}$$

where

$$\hat{\omega}_{j,k} = \begin{cases} \alpha \frac{1}{|\hat{\theta}_{j,1}|^r}, & \text{if } k = 1 \\ c_j \frac{1}{|\sum_{k'=2}^{K} \hat{\theta}_{j,k'}|^s} \frac{1}{|\hat{\theta}_{j,k}|^r}, & \text{if } k = 2, \ldots, K, \end{cases} \tag{6}$$

for $j = 1, \ldots, p$. Since no ties are allowed in the parameter ordering of FLARCC, one-to-one transformation exists between $\boldsymbol{\beta} = (\boldsymbol{\beta}_{1,\bullet}^\top, \boldsymbol{\beta}_{2,\bullet}^\top, \ldots, \boldsymbol{\beta}_{p,\bullet}^\top)^\top$ and $\boldsymbol{\theta} = (\boldsymbol{\theta}_{1,\bullet}^\top, \boldsymbol{\theta}_{2,\bullet}^\top, \ldots, \boldsymbol{\theta}_{p,\bullet}^\top)^\top$ by suitable sorting matrix $\boldsymbol{S}$ and reparameterization matrix $\boldsymbol{R}$; that is, $\boldsymbol{\theta} = \boldsymbol{RS}\boldsymbol{\beta}$ and $\boldsymbol{\beta} = (\boldsymbol{RS})^{-1}\boldsymbol{\theta}$ with both $\boldsymbol{S}$ and $\boldsymbol{R}$ being full-rank square matrices. Thus, a solution to the fused lasso problem can be obtained equivalently by solving a routine lasso problem with respect to coefficient vector $\boldsymbol{\theta}$ and a transformed design matrix $\boldsymbol{X}(\boldsymbol{RS})^{-1}$. As aforementioned, the estimated parameter ordering is used to construct $\boldsymbol{S}$. It is obvious that the constraint in (5) is convex, thus FLARCC does not suffer from multiple local minimal issue. The optimization is done using R package `glmnet` (version 2.0-2) (Friedman et al., 2010), which accommodates GLMs with Gaussian, binomial and Poisson distributions.

3. Large-sample Properties

First we present the oracle property of our method when the parameter ordering is known, then we prove that the same large-sample properties are preserved when consistently estimated parameter ordering is used. Here we assume K is fixed. Theorems will be stated under the setting of all coefficients being heterogeneous, i.e., $\boldsymbol{c} = (1, \ldots, 1)^\top$. The large-sample theories for other specification of $\boldsymbol{c}$ can be established as a special case.

Denote the true parameter values as $\boldsymbol{\beta}^*$ and $\boldsymbol{\theta}^*$. Let the collection of true parameter orderings of all covariates and their absolute values be $\boldsymbol{W} = \{\boldsymbol{U}_j, \boldsymbol{V}_j\}_{j=1}^p$, and the estimated orderings based on the root-n consistent estimator $\hat{\boldsymbol{\beta}}$ from (1) as $\hat{\boldsymbol{W}} = \{\hat{\boldsymbol{U}}_j, \hat{\boldsymbol{V}}_j\}_{j=1}^p$. Denote the FLARCC estimator of $\boldsymbol{\theta}^*$ as $\hat{\boldsymbol{\theta}}^{\boldsymbol{W}}$ when $\boldsymbol{W}$ is known, and $\hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{W}}}$ when the estimated parameter ordering $\hat{\boldsymbol{W}}$ is used. Let $\mathcal{A} = \bigcup_{j=1}^p \{\mathcal{A}_j\}$ be the index set of nonzero values in $\boldsymbol{\theta}^*$, where $\mathcal{A}_j = \{(j,k) : \theta_{j,k}^* \neq 0\}$, and $\mathcal{A}^c$ be the complement of $\mathcal{A}$. Thus, $\boldsymbol{\theta}^*$ can be partitioned into two subsets, the true-zero set $\boldsymbol{\theta}_{\mathcal{A}^c}^*$ and the nonzero set $\boldsymbol{\theta}_{\mathcal{A}}^*$. Similarly, let $\hat{\mathcal{A}}^{\boldsymbol{W}}$ and $\hat{\mathcal{A}}^{\hat{\boldsymbol{W}}}$ be the index sets of nonzero elements in $\hat{\boldsymbol{\theta}}^{\boldsymbol{W}}$ and $\hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{W}}}$, respectively. Let $n = \min_{1 \le k \le K} n_k$, $N = \sum_{k=1}^K n_k$, and $\lambda_N = N\lambda$.

Theorem 1 *Suppose that tuning parameter λ_N satisfies $\lambda_N/\sqrt{N} \to 0$ and $\lambda_N N^{(r-1)/2} \to \infty$. Then under some mild regularity conditions (see Appendix A), the FLARCC estimator $\hat{\boldsymbol{\theta}}^{\boldsymbol{W}}$ based on the true parameter ordering $\boldsymbol{W}$ satisfies*

(i) (Selection Consistency) $\lim_n P(\hat{\mathcal{A}}^{\boldsymbol{W}} = \mathcal{A}) = 1$;

(ii) (Asymptotic Normality) $\sqrt{N}(\hat{\boldsymbol{\theta}}^{\boldsymbol{W}}_{\mathcal{A}} - \boldsymbol{\theta}^*_{\mathcal{A}}) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{I}_{11}^{-1})$ *as $n \to \infty$, where $\boldsymbol{I}_{11}$ is the submatrix of Fisher information matrix $\boldsymbol{I}$ corresponding to set $\mathcal{A}$.*

Theorem 1 states that when the coefficient orderings $\boldsymbol{W}$ of $\boldsymbol{\beta}$ is known, under mild regularity conditions, the FLARCC estimator $\hat{\boldsymbol{\theta}}^{\boldsymbol{W}}$ enjoys selection consistency and asymptotic normality. The proof of Theorem 1 follows Zou (2006) and is given in Appendix A. Now we present Theorem 3, which states that the same properties of Theorem 1 hold for $\hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{W}}}$, the FLARCC estimator of $\boldsymbol{\theta}^*$ based on the estimated parameter ordering $\hat{\boldsymbol{W}}$. In effect, Theorem 3 is a consequence of the following lemma.

Lemma 2 *If $\hat{\boldsymbol{\beta}}$ is a root-n consistent estimator of $\boldsymbol{\beta}$, then $\lim_n P(\hat{\boldsymbol{U}}_j = \boldsymbol{U}_j) = 1$ and $\lim_n P(\hat{\boldsymbol{V}}_j = \boldsymbol{V}_j) = 1$ for $j = 1, \ldots, p$.*

The proof of Lemma 2 is given in Appendix A. Lemma 2 implies that the parameter ordering can be consistently estimated. Using Lemma 2, we are able to extend the properties of $\hat{\boldsymbol{\theta}}^{\boldsymbol{W}}$ in Theorem 1 to the proposed FLARCC estimator $\hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{W}}}$.

Theorem 3 *Suppose that $\lambda_N/\sqrt{N} \to 0$ and $\lambda_N N^{(r-1)/2} \to \infty$. Let the estimated parameter ordering $\hat{\boldsymbol{W}}$ be the ranks from a root-n initial consistent estimator $\hat{\boldsymbol{\beta}}$. Under the same regularity conditions of Theorem 1, the FLARCC estimator $\hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{W}}}$ satisfies*

(i) (Selection Consistency) $\lim_n P(\hat{\mathcal{A}}^{\hat{\boldsymbol{W}}} = \mathcal{A}) = 1$;

(ii) (Asymptotic Normality) $\sqrt{N}(\hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{W}}}_{\mathcal{A}} - \boldsymbol{\theta}^*_{\mathcal{A}}) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{I}_{11}^{-1})$ *as $n \to \infty$, where $\boldsymbol{I}_{11}$ is the submatrix of Fisher information matrix $\boldsymbol{I}$ corresponding to set $\mathcal{A}$.*

The proof of Theorem 3 is given in Appendix A. The asymptotic normality for $\hat{\boldsymbol{\beta}}$ can also be derived by a simple linear transformation.

4. Tuning Parameter

In this section, we provide interpretation of the tuning parameter λ and discuss the selection criteria used for selecting λ.

4.1 Interpretation of ν_j's

Intuitively speaking, the study-specific coefficients of a homogeneous covariate tend to be fused at a small λ value, say λ_1, but the fusion of a heterogeneous covariate requires another λ value, λ_2, assuming $\lambda_2 > \lambda_1$. The region to draw correct clustering conclusion is $[\lambda_1, \lambda_2]$, that is, any λ within this region will produce the correct clustering result. However, when the number of covariates p is large, the region that λ can take value from to ensure the correct clustering of all p coefficient vectors simultaneously becomes narrower and may even

be empty. For example, when $\lambda_2 < \lambda_1$ in the above case, no single λ is able to correctly cluster both sets of parameters. The introduction of ν_j's in (4) creates larger separation between homogeneous and heterogeneous groups, so that the range for λ to identify the correct clustering pattern for all covariates is better established than the case with $s = 0$, namely no use of weighting ν_j's. When the number of covariates p is large, ν_j plays a more important role in harmonizing solution paths across covariates, and the performance will be greatly improved by simultaneous tuning via a single λ.

4.2 Model Selection

In the current literature, the tuning parameter λ may be selected by multiple model selection criteria, such as Bayesian information criterion (BIC) (Schwarz, 1978) and generalized cross-validation (GCV) (Golub et al., 1979). In this paper, we consider the widely used BIC and its modification, extended BIC, i.e., EBIC (Chen and Chen, 2008; Gao and Song, 2010), which has showed the benefit of achieving sparse solutions.

Following the derivation of BIC for weighted likelihoods in Lumley and Scott (2015), the conventional BIC for FLARCC is defined as follows:

$$BIC_\lambda = -2\sum_{k=1}^{K} \frac{\bar{n}}{n_k} \log L_k(\hat{\boldsymbol{\beta}}(\lambda)) + \mathrm{df}(\hat{\boldsymbol{\beta}}(\lambda)) \log(N), \tag{7}$$

where $\bar{n} = N/K$ is the average sample size per study, $L_k(\boldsymbol{\beta})$ is the study-specific likelihood, $\hat{\boldsymbol{\beta}}(\lambda)$ is the estimation of $\boldsymbol{\beta}$ at tuning parameter value λ, and $\mathrm{df}(\hat{\boldsymbol{\beta}}(\lambda)) = \sum_{j=1}^{p} \mathrm{df}(\hat{\boldsymbol{\beta}}_{j,\bullet}(\lambda))$ is the total number of distinct parameters in $\hat{\boldsymbol{\beta}}(\lambda)$. The study-specific log-likelihoods for three most common models are listed below:

$$\text{Normal: } \log L_k(\hat{\boldsymbol{\beta}}(\lambda)) \propto -\frac{n_k}{2} \log \left\{ \sum_{i=1}^{n_k} \left(Y_k^{(i)} - \boldsymbol{X}_k^{(i)\top} \hat{\boldsymbol{\beta}}(\lambda) \right)^2 / n_k \right\};$$

$$\text{Logistic: } \log L_k(\hat{\boldsymbol{\beta}}(\lambda)) \propto \sum_{i=1}^{n_k} \left\{ Y_k^{(i)} \boldsymbol{X}_k^{(i)\top} \hat{\boldsymbol{\beta}}(\lambda) - \log \left(1 + e^{\boldsymbol{X}_k^{(i)\top} \hat{\boldsymbol{\beta}}(\lambda)} \right) \right\};$$

$$\text{Poisson: } \log L_k(\hat{\boldsymbol{\beta}}(\lambda)) \propto \sum_{i=1}^{n_k} \left\{ Y_k^{(i)} \boldsymbol{X}_k^{(i)\top} \hat{\boldsymbol{\beta}}(\lambda) - e^{\boldsymbol{X}_k^{(i)\top} \hat{\boldsymbol{\beta}}(\lambda)} \right\}.$$

To improve the BIC by further controlling model size and encouraging sparer models, we adapt the EBIC for FLARCC, which takes the following form:

$$EBIC_\lambda = -2\sum_{k=1}^{K} \frac{\bar{n}}{n_k} \log L_k(\hat{\boldsymbol{\beta}}(\lambda)) + \mathrm{df}(\hat{\boldsymbol{\beta}}(\lambda)) \log(N) + 2\gamma \log \sum_{j=1}^{p} \binom{K}{\mathrm{df}(\hat{\boldsymbol{\beta}}_{j,\bullet}(\lambda))}, \tag{8}$$

where $\gamma \in [0, 1]$ is a tuning parameter that is typically fixed at 1 as done in our numerical experiments. Note that EBIC reduces to BIC when $\gamma = 0$. The last term in (8) encourages a sparser solution in comparison to the conventional BIC. Simulation studies in Section 5 provide numerical evidence to elucidate the difference between BIC and EBIC in terms of their performance on achieving sparsity.

In a view of hierarchical clustering, the solution path of each covariate can be thought of as a hierarchical clustering tree. For the jth covariate, $\lambda = 0$ corresponds to the bottom of the clustering tree; and $\lambda = \lambda_{Fuse,j}$, the smallest λ value to achieve complete parameter fusion, corresponds to the top of the clustering tree. The completely heterogeneous model corresponds to the position on the solution path at $\lambda = 0$ and the completely homogeneous model corresponds to the model at $\lambda = \lambda_{Fuse} := \max_{1 \le j \le p} \lambda_{Fuse,j}$.

5. Simulation Studies

This section presents results from two simulation experiments. The first simulation compares the performance of FLARCC under different GLM regression models. The second simulation is a more complicated scenario with large K and more non-important covariates, where covariate selection is also of interest.

5.1 Simulation Experiment 1

The first simulation study aims to assess the performance of our method for different GLM regression models. For this, we consider combining data sets from $K = 10$ different studies with, for simplicity, equal sample size $n_1 = \cdots = n_{10} = 100$. Data are simulated from the following mean regression model:

$$h\{E(Y_k^{(i)})\} = \beta_{1,k} X_{1,k}^{(i)} + \beta_{2,k} X_{2,k}^{(i)} + \beta_{3,k} X_{3,k}^{(i)},\ i = 1, \ldots, 100,\ k = 1, \ldots, 10,$$

where the true coefficient vectors have the following clustering structures:

$$\begin{aligned}
\boldsymbol{\beta}_{1,\bullet} &= (\underbrace{0, \ldots, 0}_{10})^\top; \\
\boldsymbol{\beta}_{2,\bullet} &= (\underbrace{0, \ldots, 0}_{5}, \underbrace{1, \ldots, 1}_{5})^\top; \\
\boldsymbol{\beta}_{3,\bullet} &= (\underbrace{-1, \ldots, -1}_{3}, \underbrace{0, \ldots, 0}_{4}, \underbrace{1, \ldots, 1}_{3})^\top.
\end{aligned}$$

The true values in $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_3$ are heterogeneous, while the true values in $\boldsymbol{\beta}_1$ are homogeneous across studies. The three covariates are correlated with exchangeable correlation of 0.3 and marginally distributed according to the standard normal distributions, $\mathcal{N}(0, 1)$. Three types of GLM regression models are considered: linear model for continuous normal outcomes (with errors simulated from $\mathcal{N}(0, 1)$), logistic model for binary outcomes and Poisson model for count outcomes.

To evaluate the performance of FLARCC to correctly detect patterns of all covariates, we assume all covariates are heterogeneous across studies with no prior knowledge on clustering structure of any covariate. Intercept is fitted and assumed to be homogeneous. No sparsity penalty is applied on the covariates (i.e., $\alpha = 0$) in this simulation experiment. Coefficients of all three covariates are fused simultaneously, and the optimal tuning parameter λ_{opt} is selected by EBIC. We report sensitivity and specificity as metrics of the performance of FLARCC to identify similar and distinct coefficient pairs. Sensitivity measures the proportion of equal coefficient pairs that are correctly identified. Similarly, specificity measures

Method ($\tilde{\lambda}_{opt}$)	$\boldsymbol{\beta}$	$\hat{\boldsymbol{\beta}}$ size	$\tilde{\lambda}_{Fuse,j}$	Sensitivity	Specificity	MSE when $\lambda =$		
						λ_{opt}	λ_{Fuse}	0
Linear: continuous response								
$s=0$ (0.154)	$\boldsymbol{\beta}_1$	1.067	0.111	0.974	–	0.001	0.002	0.012
	$\boldsymbol{\beta}_2$	2.075	1.275	0.982	1.000	0.003	0.253	0.012
	$\boldsymbol{\beta}_3$	3.081	1.368	0.982	1.000	0.004	0.603	0.012
$s=1$ (0.349)	$\boldsymbol{\beta}_1$	1.006	0.080	0.998	–	0.001	0.002	0.012
	$\boldsymbol{\beta}_2$	2.058	1.584	0.986	1.000	0.003	0.253	0.012
	$\boldsymbol{\beta}_3$	3.123	1.972	0.974	1.000	0.004	0.603	0.012
Logistic: binary response								
$s=0$ (0.066)	$\boldsymbol{\beta}_1$	1.270	0.064	0.898	–	0.010	0.005	0.070
	$\boldsymbol{\beta}_2$	2.572	0.318	0.819	0.963	0.047	0.268	0.087
	$\boldsymbol{\beta}_3$	3.682	0.437	0.784	0.964	0.069	0.607	0.091
$s=1$ (0.112)	$\boldsymbol{\beta}_1$	1.075	0.050	0.972	–	0.007	0.005	0.069
	$\boldsymbol{\beta}_2$	2.478	0.414	0.837	0.952	0.052	0.268	0.088
	$\boldsymbol{\beta}_3$	3.912	0.711	0.749	0.971	0.064	0.607	0.091
Poisson: count response								
$s=0$ (0.187)	$\boldsymbol{\beta}_1$	1.087	0.129	0.976	–	0.001	0.005	0.008
	$\boldsymbol{\beta}_2$	2.084	1.751	0.984	1.000	0.001	0.271	0.008
	$\boldsymbol{\beta}_3$	3.076	1.885	0.986	1.000	0.002	0.659	0.008
$s=1$ (0.433)	$\boldsymbol{\beta}_1$	1.047	0.087	0.992	–	0.001	0.005	0.008
	$\boldsymbol{\beta}_2$	2.088	2.060	0.984	1.000	0.002	0.271	0.008
	$\boldsymbol{\beta}_3$	3.111	2.536	0.978	1.000	0.002	0.657	0.008

Table 1: Results of simulation experiment 1 for FLARCC when scaling weight parameter $s = 0$ and $s = 1$ with λ selected by EBIC, for the linear, logistic and Poisson models. Tuning parameters are reported in log scale, i.e., $\tilde{\lambda} = \log_{10}(\lambda + 1)$. Results are summarized from 1,000 replications.

the proportion of unequal coefficients pairs that are correctly identified; however, specificity is not defined for homogeneous covariates which have no unequal coefficient pairs. In addition, we calculate the mean squared error (MSE) for each $\hat{\boldsymbol{\beta}}_{j,\boldsymbol{\cdot}}$ across all K studies, defined as $\text{MSE}_j = \sum_{k=1}^{K}(\hat{\beta}_{j,k} - \beta_{j,k})^2/K, j = 1, \ldots, p$, and compare with the MSE of each estimate based on homogeneous model ($\lambda = \lambda_{Fuse}$) and heterogeneous model ($\lambda = 0$).

Table 1 shows the results of simulation experiment 1 from 1,000 simulation replicates. The MSE of all estimated covariates based on FLARCC ($\lambda = \lambda_{opt}$) are consistently and significantly smaller than those based on the homogeneous ($\lambda = \lambda_{Fuse}$) and heterogeneous ($\lambda = 0$) models, regardless of the model type. FLARCC performs very well in the linear and Poisson regressions in terms of identifying the correct clustering, with the sensitivity and specificity both above 95% for all covariates (specificity is not reported for $\boldsymbol{\beta}_1$ since there is no unequal pair within $\boldsymbol{\beta}_{1,\boldsymbol{\cdot}}$). Sensitivity and specificity of FLARCC drop in the logistic regression, especially as the level of heterogeneity increases. One reason for the reduced performance of FLARCC in the logistic regression is that the estimated variances of

regression coefficients in the logistic model are larger than in the linear and Poisson models, given the same coefficient setting. Therefore, the estimated parameter ordering for which our method is based on may be less accurate. For the logistic regression, increasing sample sizes is one of the possible ways to improve the performance. The performance difference between scaling weight parameter $s = 0$ and $s = 1$ in (4) is small in this case because of the relatively small number of covariates $p = 3$. Additionally, since K is small in this case, the optimal λ selected by BIC and EBIC are very close, thus we only display results based on EBIC. As p and K become larger, FLARCC will increasingly benefit from the additional weights ν_j (i.e., $s = 1$) and EBIC, as will be shown in Section 5.2. A sensitivity analysis to investigate how the initial ordering affect the performance of FLARCC is conducted, with results shown in Appendix B. We show that when the initial parameter ordering is slightly distorted, our method still achieves satisfactory performance.

5.2 Simulation Experiment 2

The second simulation study aims to evaluate the performance of FLARCC in a more challenging setting. More specifically, we consider data sets from $K = 100$ studies, each with a sample size 100, totaling 10,000 subject-level observations. Comparing to the previous setting, we increase the number of covariates and reduce the gaps between heterogeneous coefficients. For each study, we simulate data from the following linear regression model:

$$E(Y_k^{(i)}) = \sum_{j=1}^{8} \beta_{j,k} X_{j,k}^{(i)},\ i = 1, \ldots, 100,\ k = 1, \ldots, 100.$$

The signals are set sparse; only the first four covariates with coefficient vectors, $\boldsymbol{\beta}_1$ to $\boldsymbol{\beta}_4$, are influential to Y with the true clustered effect patterns given as follows:

$$\begin{aligned}
\boldsymbol{\beta}_{1,\bullet} &= (\underbrace{0, \ldots, 0}_{50}, \underbrace{0.5, \ldots, 0.5}_{50})^\top, \\
\boldsymbol{\beta}_{2,\bullet} &= (\underbrace{-0.5, \ldots, -0.5}_{30}, \underbrace{0, \ldots, 0}_{40}, \underbrace{0.5, \ldots, 0.5}_{30})^\top, \\
\boldsymbol{\beta}_{3,\bullet} &= (\underbrace{-0.5, \ldots, -0.5}_{25}, \underbrace{0, \ldots, 0}_{25}, \underbrace{0.5, \ldots, 0.5}_{25}, \underbrace{1, \ldots, 1}_{25})^\top, \\
\boldsymbol{\beta}_{4,\bullet} &= (\underbrace{-1, \ldots, -1}_{20}, \underbrace{-0.5, \ldots, -0.5}_{20}, \underbrace{0, \ldots, 0}_{20}, \underbrace{0.5, \ldots, 0.5}_{20}, \underbrace{1, \ldots, 1}_{20})^\top,
\end{aligned}$$

whereas $\boldsymbol{\beta}_5$ to $\boldsymbol{\beta}_8$ are all zero, i.e., $\boldsymbol{\beta}_{j,\bullet} = (0, \ldots, 0)^\top$, for $j = 5, 6, 7, 8$. All covariates are equally correlated with an exchangeable correlation of 0.3 and marginally distributed according to $\mathcal{N}(0, 1)$. We set $\boldsymbol{\beta}_1$ to $\boldsymbol{\beta}_8$ as being heterogeneous from the start and fuse all of them simultaneously. We apply the additional sparsity penalty to all covariates by setting $\alpha = 1$. The intercept is assumed to be homogeneous in the analysis.

Since K is large, we also present results from individual covariate K-means clustering. This is a two-step method where we first estimate regression coefficients within each study, and then separately for each covariate, we perform the K-means clustering on the estimated study-specific coefficients of each covariate. The number of clusters is selected by the generalized cross-validation criterion $\sum_{k=1}^{K} (\hat{\beta}_k - \hat{\beta}_{c(k)})^2 / (K - \text{GDF})^2$, with $\hat{\beta}_{c(k)}$ being the

Method ($\tilde{\lambda}_{opt}$)	$\boldsymbol{\beta}$	$\hat{\boldsymbol{\beta}}$ size	$\tilde{\lambda}_{Fuse,j}$	Sensitivity	Specificity	Sparsity	MSE when $\lambda =$ λ_{opt}	λ_{Fuse}	0
$s=0$ BIC (0.143)	$\boldsymbol{\beta}_1$	8.115	1.517	0.373	0.995	0.199	0.006	0.063	0.014
	$\boldsymbol{\beta}_2$	10.689	1.551	0.401	0.996	0.166	0.008	0.150	0.014
	$\boldsymbol{\beta}_3$	13.107	1.962	0.443	0.997	0.113	0.008	0.313	0.014
	$\boldsymbol{\beta}_4$	15.178	1.984	0.461	0.997	0.091	0.009	0.500	0.014
	$\boldsymbol{\beta}_5$	4.818	0.301	0.322	–	0.338	0.003	0.000	0.014
	$\boldsymbol{\beta}_6$	4.860	0.305	0.322	–	0.339	0.003	0.000	0.014
	$\boldsymbol{\beta}_7$	4.860	0.302	0.321	–	0.330	0.003	0.000	0.014
	$\boldsymbol{\beta}_8$	4.820	0.301	0.319	–	0.336	0.003	0.000	0.014
$s=0$ EBIC (0.159)	$\boldsymbol{\beta}_1$	7.538	1.509	0.417	0.994	0.224	0.006	0.063	0.014
	$\boldsymbol{\beta}_2$	9.975	1.546	0.441	0.995	0.182	0.007	0.150	0.014
	$\boldsymbol{\beta}_3$	12.212	1.953	0.483	0.996	0.124	0.008	0.313	0.014
	$\boldsymbol{\beta}_4$	14.096	1.978	0.503	0.997	0.101	0.008	0.500	0.014
	$\boldsymbol{\beta}_5$	4.388	0.298	0.377	–	0.394	0.003	0.000	0.014
	$\boldsymbol{\beta}_6$	4.413	0.303	0.379	–	0.397	0.003	0.000	0.014
	$\boldsymbol{\beta}_7$	4.408	0.299	0.374	–	0.385	0.003	0.000	0.014
	$\boldsymbol{\beta}_8$	4.385	0.298	0.374	–	0.392	0.003	0.000	0.014
$s=1$ EBIC (0.492)	$\boldsymbol{\beta}_1$	3.563	1.589	0.800	0.981	0.422	0.006	0.063	0.014
	$\boldsymbol{\beta}_2$	6.111	1.810	0.708	0.989	0.291	0.007	0.150	0.014
	$\boldsymbol{\beta}_3$	8.843	2.388	0.667	0.994	0.168	0.007	0.313	0.014
	$\boldsymbol{\beta}_4$	11.358	2.521	0.635	0.995	0.128	0.008	0.500	0.014
	$\boldsymbol{\beta}_5$	1.329	0.275	0.928	–	0.932	0.000	0.000	0.014
	$\boldsymbol{\beta}_6$	1.321	0.280	0.933	–	0.937	0.000	0.000	0.014
	$\boldsymbol{\beta}_7$	1.311	0.274	0.934	–	0.935	0.000	0.000	0.014
	$\boldsymbol{\beta}_8$	1.321	0.273	0.929	–	0.936	0.000	0.000	0.014
									MSE from K-means
K-means GCV (GDF)	$\boldsymbol{\beta}_1$	7.196	–	0.753	0.971	0.000			0.008
	$\boldsymbol{\beta}_2$	11.236	–	0.671	0.983	0.000			0.009
	$\boldsymbol{\beta}_3$	13.721	–	0.639	0.984	0.000			0.011
	$\boldsymbol{\beta}_4$	18.308	–	0.527	0.985	0.000			0.014
	$\boldsymbol{\beta}_5$	6.415	–	0.759	–	0.000			0.004
	$\boldsymbol{\beta}_6$	5.271	–	0.769	–	0.000			0.004
	$\boldsymbol{\beta}_7$	5.629	–	0.767	–	0.000			0.004
	$\boldsymbol{\beta}_8$	5.080	–	0.794	–	0.000			0.004

Table 2: Result of simulation experiment 2 under the linear model. Scaling weight parameter is set at $s=0$ and $s=1$. Tuning parameters are reported in log scale, i.e., $\tilde{\lambda} = \log_{10}(\lambda+1)$. Sparsity denotes the proportion of zero in estimation. Results are summarized from 1,000 replications.

cluster center of $\hat{\beta}_k$ and GDF is the generalized degrees of freedom estimated according to Ye (1998), where purturbations are generated independently from $\mathcal{N}(0, 0.01)$. The cluster centroids are then used as the estimates of the group-level parameters.

Table 2 summarizes the simulation results for linear model where the errors are generated independently from $\mathcal{N}(0,1)$. Similar to simulation 1, FLARCC gives the smallest MSE for heterogeneous covariates, $\boldsymbol{\beta}_1$ to $\boldsymbol{\beta}_4$, among all three models, and has comparable MSE as the homogeneous model for homogeneous covariates, $\boldsymbol{\beta}_5$ to $\boldsymbol{\beta}_8$. More interestingly, when K is large, BIC does not provide satisfactory model selection, erring on the lack of parsimony, while EBIC encourages stronger fusion and improves the ability to detect equal coefficient pairs in all eight covariates, regardless of their levels of heterogeneity. In addition, EBIC improves the sparsity detection among both the important and nonimportant covariates. It is interesting to note that the choice between BIC and EBIC does not alter solution paths, but only model selection. FLARCC with scaling weight parameter $s = 1$ has the best clustering performance among all compared methods. The difference between the choices of $s = 0$ and $s = 1$ is substantial in simulation 2, in contrast to the results from simulation 1. This indicates that the covariate-specific weights for heterogeneity $\{\nu_j\}_{j=1}^{p}$ are very effective to improve the performance of the proposed fusion learning, especially when K and p are large. Sensitivity and specificity of the two-step K-means clustering method are higher than those of FLARCC with $s = 0$, but lower than those of FLARCC with $s = 1$. The two-step K-means has larger MSE than FLARCC because it does not consider the correlation between covariates. More importantly, the K-means clustering is a model-free method, so the results obtained from this method cannot be plugged in back to the model for prediction. As suggested from the empirical results of both simulation experiments, EBIC tends to provide better model selection for FLARCC than the conventional BIC.

6. Application: Clustering of Cohort Effects

In this data analysis example, we like to demonstrate the use of our method to derive clusters of cohort effects. Here we consider the Panel Study of Income Dynamics (PSID), which is a household survey study following thousands of families across different states in the US. PSID collects information of employment, income, health, and so on. In this data analysis, we focus on the association of household income with body mass index (BMI) on school-aged children between age of 11 and 19, adjusted for age, gender and birth weight. Data of 1880 children were gathered from four census regions (1-Northeast, 2-Midwest, 3-South and 4-West), as defined by U.S. Census Bureau (2015). All variables are standardized before model fitting. We are interested in investigating if regional heterogeneity exists and if the effects of interest differ across regions with region-dependent patterns.

Table 3 shows the results of coefficient estimates obtained from three different models: (A) homogeneous model ($\lambda = \lambda_{Fuse}$), coefficients estimated by combining data sets from four regions, (B) heterogeneous model ($\lambda = 0$), coefficients estimated separately by region-specific data, and (C) FLARCC ($\lambda = \lambda_{EBIC}$). Model A suggests that age and birth weight are positively associated with BMI for the subjects, but income was negatively associated with BMI. The estimates from Model B suggest that heterogeneous coefficient patterns exist among these associations since conclusions differ between regions. Model C appears more sensible when regression coefficients are heterogeneous across these regions. Since K and p are small in this data application, we apply FLARCC with $s = 0$ on the PSID data, assuming effects of income, age, gender and birth weight are heterogeneous across regions, and set sparsity parameter $\alpha = 1$ for variable selection.

Region	n	Intercept $\boldsymbol{\beta}_0$	Age $\boldsymbol{\beta}_1$	Sex $\boldsymbol{\beta}_2$	Birth Wt. $\boldsymbol{\beta}_3$	Income $\boldsymbol{\beta}_4$
(A) Homogeneous model – combine all regions						
All regions	1880	0.000	0.206	0.016	0.063	-0.096
(B) Heterogeneous model – region specific estimates						
1-Northeast	239	-0.133	0.228	-0.079	-0.003	0.004
2-Midwest	493	-0.054	0.229	0.017	0.124	-0.132
3-South	805	0.128	0.158	0.095	0.068	-0.071
4-West	343	-0.155	0.236	-0.083	0.057	-0.074
(C) Fused model using FLARCC						
1-Northeast	239	-0.093	0.201	-0.036	0.000	0.000
2-Midwest	493	-0.093	0.201	0.000	0.021	-0.047
3-South	805	0.075	0.201	0.000	0.021	-0.047
4-West	343	-0.093	0.201	-0.036	0.021	-0.047

Table 3: Coefficient estimates of the homogeneous model ($\lambda = \lambda_{Fuse}$), the heterogeneous model ($\lambda = 0$) and the fused model using FLARCC with λ selected by EBIC, respectively.

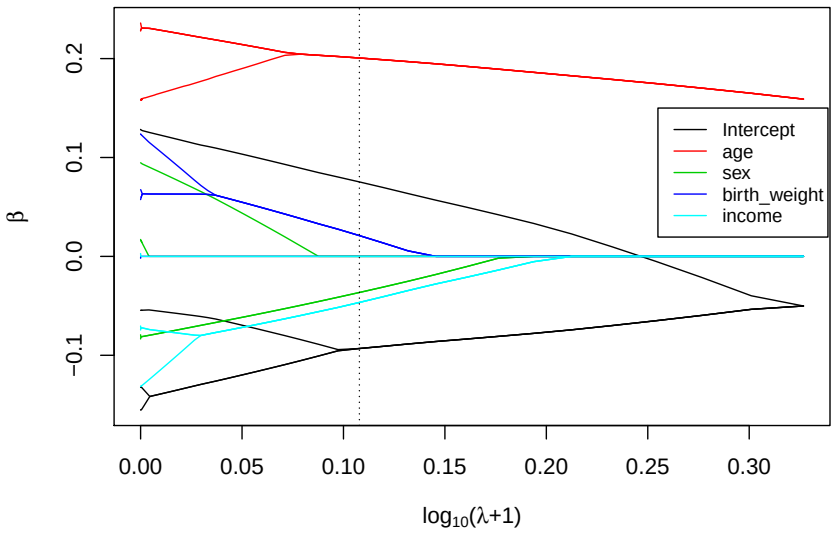

Figure 2: FLARCC solution paths of all covariates over the transformed tuning parameter $\tilde{\lambda} = \log_{10}(\lambda+1)$, with $s = 0$. The vertical dotted line denotes the optimal tuning parameter value $\tilde{\lambda}_{EBIC}$.

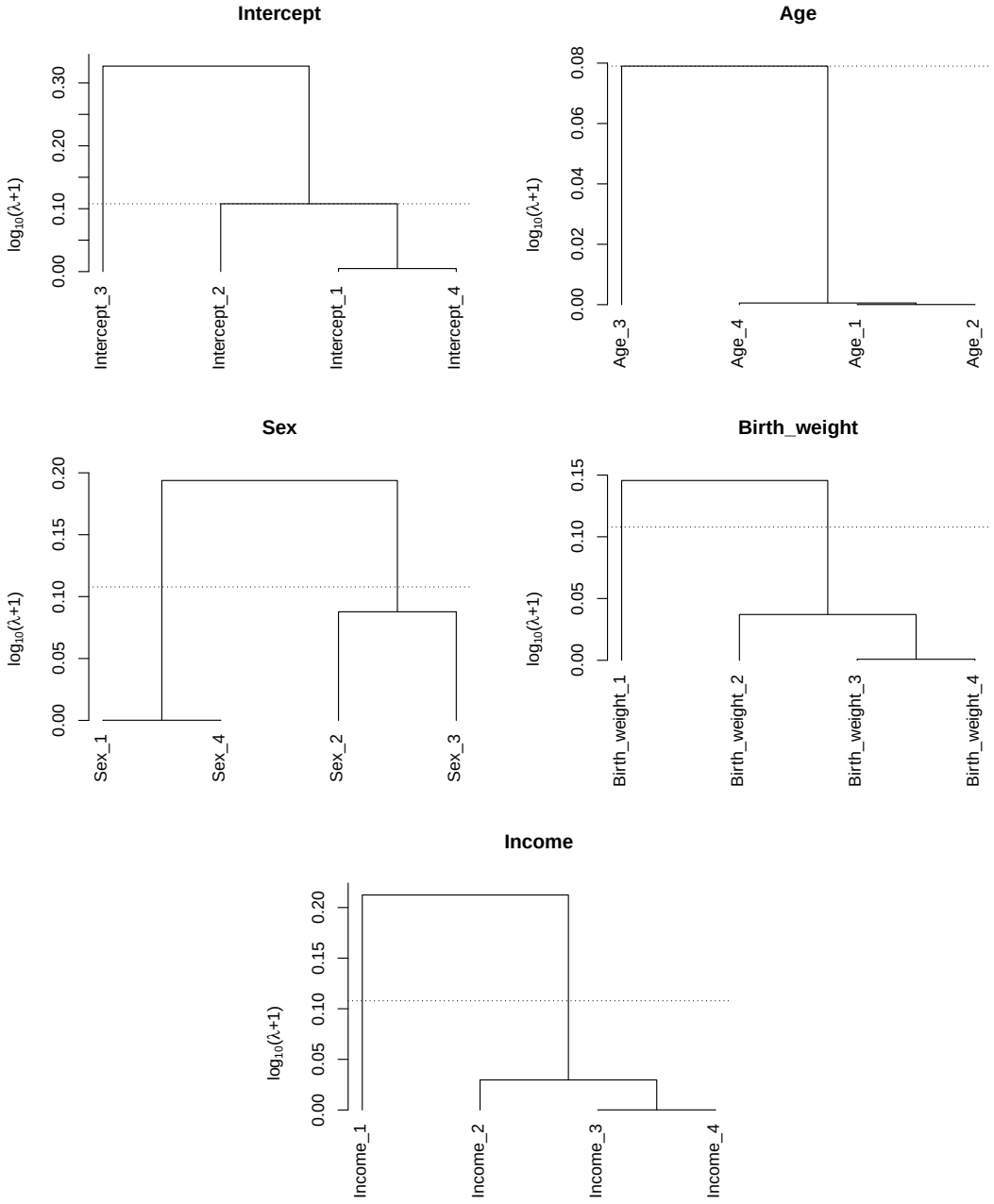

Figure 3: *Fusograms* of all covariates based on FLARCC solution paths. The horizontal dotted lines denote the optimal regression coefficient clustering determined by EBIC.

Based on the results from FLARCC, the estimated mean of standardized BMI in the South is 0.168 higher (or 0.97 higher in original scale of BMI) than that of the other three regions, which share the same mean. The effects of age are consistent across four regions. The effects of gender are classified into two clusters. The mean of standardized BMI of females is 0.036 lower (or 0.42 lower in original scale) than that of males in the Northeast and the West, but males and females have the same mean BMI in the Midwest and the South. Standardized BMI increases by 0.021 for every standard deviation increase of birth weight (or BMI increases by 0.19 for every unit increase of birth weight) in all regions except the Northeast. Similarly, standardized BMI decreases by 0.047 for every standard deviation increase of log income (or BMI decreases by 0.27 for every unit increase of income) in all regions except the Northeast where BMI is not affected by income. The leave-one-out mean squared prediction errors for model A, B and C are 0.953, 0.945 and 0.950, respectively. The differences between the prediction errors are small because of the relatively small effect sizes of the heterogeneous covariates identified by FLARCC, i.e., sex, birth weight and income. The most significant covariate, age, is homogeneous thus it does not differentiate the prediction power among the three models. Solution paths and *fusograms* of all covariates are shown in Figure 2 and Figure 3, respectively, for illustration. In summary, FLARCC ensures parsimony where necessary to maximize the prediction power of the final model; and it provides more informative interpretation and better visualization than the other two traditional models.

7. Concluding Remarks

The proposed method brings a new perspective to model fitting when combining multiple data sets from different sources is of primary interest. As data volumes and data sources grow fast, more and more opportunities and demands emerge in practice to borrow strengths of combined data sets. In such case, traditional methods are challenged by the complex data structures and do not provide desirable treatments and meaningful interpretations to data heterogeneity, especially when the number of data sets is very large. FLARCC allows the flexibility to explore the heterogeneity pattern of parameters among large number of data sets by tuning the shrinkage parameter.

When K and p are small, weights $\{\nu_j\}_{j=1}^{p}$ do not contribute to much difference in terms of clustering and estimation. However, since only one tuning parameter is used to regularize the fusion of all covariates, when both K and p are large, we suggest letting $s > 0$ to allow covariate-specific weights adapting to the heterogeneousness of coefficients from individual covariates to achieve better results. In addition, the estimation consistency of rank estimator is a critical component needed to determine adjacent pairs. The current consistency is established under the case of K being fixed, and the validity of its property is unknown when K increases along the total sample size.

FLARCC can be applied to various scientific problems, such as the detection of outlying studies by singling out outlying coefficients; it can also be applied to the clustering of patient trajectories by viewing the time series data of patients as individual studies. Essentially, all study that are interested in the group-specific effects may be analyzed from the perspective of parameter fusion using the proposed method.

Acknowledgments

We thank Drs. Fei Wang and Ling Zhou for helpful discussion. We are grateful to the action editor and two anonymous reviewers for their constructive comments that have led to an improvement of this paper. This research is supported by an NIH grant R01 ES024732.

Appendix A. Theorem Proofs

Proof of Theorem 1: The proof of Theorem 1 closely follows arguments given in Zou (2006). Without loss of generality, we assume $n_1 = \cdots = n_K = n$ and $N = Kn$. As K is fixed, $n \to \infty$ implies $N \to \infty$ in the same order. We assume the following regularity conditions:

(i) The Fisher information matrix is finite and positive definite,

$$\boldsymbol{I}(\boldsymbol{\theta}^*) = E\left[\phi''\left(\boldsymbol{X}^\top\boldsymbol{\theta}^*\right)\boldsymbol{X}\boldsymbol{X}^\top\right].$$

Here, $\boldsymbol{\theta}^*_{(Kp\times 1)}$ is the true parameters, $\boldsymbol{X}_{(N\times Kp)}$ is the design matrix corresponding to $\boldsymbol{\theta}$ and ϕ is the link function (i.e., $\phi' = h^{-1}$) defined in the following optimization problem

$$\hat{\boldsymbol{\theta}}^{\boldsymbol{W}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\left\{-\frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_k}\sum_{i=1}^{n_k}\left(Y_k^{(i)}\boldsymbol{X}_k^{(i)\top}\boldsymbol{\theta}(\lambda) - \phi\left(\boldsymbol{X}_k^{(i)\top}\boldsymbol{\theta}(\lambda)\right)\right) + P_{\lambda,\alpha}(\boldsymbol{\theta})\right\}$$

with $P_{\lambda,\alpha}(\boldsymbol{\theta})$ as defined in (4), and $\hat{\boldsymbol{\theta}}^{\boldsymbol{W}}$ is the estimator with true ordering $\boldsymbol{W}$ given.

(ii) There is a sufficiently large open set $\mathcal{O}$ that contains $\boldsymbol{\theta}^*$ such that $\forall\boldsymbol{\theta} \in \mathcal{O}$,

$$|\phi'''(\boldsymbol{X}^\top\boldsymbol{\theta})| \le M(\boldsymbol{X}^\top) < \infty, \text{ and}$$
$$E\left[M(\boldsymbol{X})|x_j x_k x_l|\right] < \infty$$

for a suitable function M and all $1 \le j, k, l \le Kp$.

First we prove asymptotic normality. For $\forall s \ge 0$ and $r > 0$, let $\boldsymbol{\theta} = \boldsymbol{\theta}^* + \boldsymbol{u}/\sqrt{N}$. Define

$$\Gamma_N(\boldsymbol{u}) = -\sum_{k=1}^{K}\sum_{i=1}^{n}\left(Y_k^{(i)}\boldsymbol{X}_k^{(i)\top}\left(\boldsymbol{\theta}^* + \frac{\boldsymbol{u}}{\sqrt{N}}\right) - \phi\left(\boldsymbol{X}_k^{(i)\top}\left(\boldsymbol{\theta}^* + \frac{\boldsymbol{u}}{\sqrt{N}}\right)\right)\right)$$
$$+\lambda_N\sum_{j=1}^{p}\sum_{k=1}^{K}\hat{\omega}_{j,k}\left|\theta^*_{j,k} + \frac{u_{j,k}}{\sqrt{N}}\right|$$

where $\hat{\omega}_{j,k}$ is specified in (6). Let $\hat{\boldsymbol{u}}^{(N)} = \arg\min_{\boldsymbol{u}}\Gamma_N(\boldsymbol{u})$; then $\hat{\boldsymbol{u}}^{(N)} = \sqrt{N}(\hat{\boldsymbol{\theta}}^{\boldsymbol{W}} - \boldsymbol{\theta}^*)$. By Taylor expansion, we have $\Gamma_N(\boldsymbol{u}) - \Gamma_N(\boldsymbol{0}) = H^{(N)}(\boldsymbol{u})$, where

$$H^{(N)}(\boldsymbol{u}) \equiv A_1^{(N)} + A_2^{(N)} + A_3^{(N)} + A_4^{(N)},$$

with

$$A_1^{(N)} = -\sum_{k=1}^{K}\sum_{i=1}^{n}\left[Y_k^{(i)} - \phi'(\boldsymbol{X}_k^{(i)\top}\boldsymbol{\theta}^*)\right]\frac{\boldsymbol{X}_k^{(i)\top}\boldsymbol{u}}{\sqrt{N}},$$
$$A_2^{(N)} = \sum_{k=1}^{K}\sum_{i=1}^{n}\frac{1}{2}\phi''(\boldsymbol{X}_k^{(i)\top}\boldsymbol{\theta}^*)\boldsymbol{u}^\top\frac{\boldsymbol{X}_k^{(i)}\boldsymbol{X}_k^{(i)\top}}{\sqrt{N}}\boldsymbol{u},$$
$$A_3^{(N)} = \frac{\lambda_N}{\sqrt{N}}\sum_{j=1}^{p}\sum_{k=1}^{K}\hat{\omega}_{j,k}\sqrt{N}\left(\left|\theta_{j,k}^* + \frac{u_{j,k}}{\sqrt{N}}\right| - \left|\theta_{j,k}^*\right|\right),$$
$$\text{and } A_4^{(N)} = N^{-3/2}\sum_{k=1}^{K}\sum_{i=1}^{n}\frac{1}{6}\phi'''\left(\boldsymbol{X}_k^{(i)\top}\tilde{\boldsymbol{\theta}}_*\right)\left(\boldsymbol{X}_k^{(i)\top}\boldsymbol{u}\right)^3,$$

where $\tilde{\boldsymbol{\theta}}_*$ is between $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}^* + \boldsymbol{u}/\sqrt{N}$. The asymptotic limits of $A_1^{(N)}$, $A_2^{(N)}$ and $A_4^{(N)}$ is exactly the same as those in the proof of Theorem 4 in Zou (2006). It suffice to show that $A_3^{(N)}$ has the same asymptotic limit. If $\theta_{j,k}^* \neq 0$, $\hat{\omega}_{j,1} \to_p \alpha|\theta_{j,1}^*|^{-r}$, $\hat{\omega}_{j,k} \to_p |\sum_{k'=2}^{K}\theta_{j,k'}^*|^{-s}|\theta_{j,k}^*|^{-r}$ for $k = 2,\ldots,K$, and $\sqrt{N}\left(\left|\theta_{j,k}^* + \frac{u_{j,k}}{\sqrt{N}}\right| - \left|\theta_{j,k}^*\right|\right) \to u_{j,k}\,\mathrm{sgn}(\theta_{j,k}^*)$. Thus by Slutsky's theorem, $A_3^{(N)} \to 0$. If $\theta_{j,k}^* = 0$, for $k = 1$, since $\sqrt{N}\hat{\theta}_{j,1} = O_p(1)$, $\frac{\lambda_N}{\sqrt{N}}N^{r-2}\alpha(|\sqrt{N}\hat{\theta}_{j,1}|)^{-r} \to \infty$; for $k = 2,\ldots,K$, if $\sum_{k'=2}^{K}\theta_{j,k'}^* = 0$ (i.e., homogeneous), $\sqrt{N}\sum_{k'=2}^{K}\hat{\theta}_{j,k'} = O_p(1)$, thus $\frac{\lambda_N}{\sqrt{N}}N^{\frac{s+r}{2}}(|\sqrt{N}\sum_{k'=2}^{K}\hat{\theta}_{j,k'}|)^{-s}(|\sqrt{N}\hat{\theta}_{j,k}|)^{-r} \to \infty$; similarly, if $\sum_{k'=2}^{K}\theta_{j,k'}^* \neq 0$ (i.e., heterogeneous), $\sum_{k'=2}^{K}\hat{\theta}_{j,k'} \to_p \sum_{k'=2}^{K}\theta_{j,k'}^*$, $\frac{\lambda_N}{\sqrt{N}}\hat{\omega}_{j,k} \to \infty$ still holds. And since $\sqrt{N}\left(\left|\theta_{j,k}^* + \frac{u_{j,k}}{\sqrt{N}}\right| - \left|\theta_{j,k}^*\right|\right) \to |u_{j,k}|$, we have the following result summary:

$$\frac{\lambda_N}{\sqrt{N}}\hat{\omega}_{j,k}\sqrt{N}\left(\left|\theta_{j,k}^* + \frac{u_{j,k}}{\sqrt{N}}\right| - \left|\theta_{j,k}^*\right|\right) \to_p \begin{cases} 0 & \text{if } \theta_{j,k}^* \neq 0 \\ 0 & \text{if } \theta_{j,k}^* = 0 \text{ and } u_{j,k} = 0 \\ \infty & \text{if } \theta_{j,k}^* = 0 \text{ and } u_{j,k} \neq 0. \end{cases}$$

Following same arguments in Zou (2006)'s proof of Theorem 4, we have $\hat{\boldsymbol{u}}_{\mathcal{A}}^{(N)} \to_d \mathcal{N}(0, \boldsymbol{I}_{11}^{-1})$ and $\hat{\boldsymbol{u}}_{\mathcal{A}^c}^{(N)} \to_d \boldsymbol{0}$. The proof of the consistency part is similar and thus omitted. ■

Proof of Lemma 2: The estimated ordering $\hat{\boldsymbol{U}}_j$ of $\boldsymbol{\beta}_{j,\bullet}^*$ is only determined by the differences between distinct parameter groups within $\boldsymbol{\beta}_{j,\bullet}^*$. First note that for any $0 < \epsilon < 1$, if two parameters $\beta_{j,k}^*$ and $\beta_{j,k'}^*$ are in the same parameter group (i.e., $\beta_{j,k}^* = \beta_{j,k'}^*$), assigning arbitrary ordering between them will not affect the estimated ordering of the parameters between groups, because the ordering within the same parameter group is exchangeable. On the other hand, when two parameters $\beta_{j,k}^*$ and $\beta_{j,k'}^*$ are from different parameter groups,

without loss of generality, let $\beta^*_{j,k} > \beta^*_{j,k'}$, the probability of estimating a wrong ordering

$$\begin{aligned} P\left(\mathbf{1}\{\hat{\beta}_{j,k'} \geq \hat{\beta}_{j,k}\} > \epsilon\right) &= P\left(\hat{\beta}_{j,k'} \geq \hat{\beta}_{j,k}\right) \\ &= P\left(\hat{\beta}_{j,k'} - \hat{\beta}_{j,k} + \beta^*_{j,k} - \beta^*_{j,k'} \geq \beta^*_{j,k} - \beta^*_{j,k'}\right) \\ &\leq P\left(|\hat{\beta}_{j,k'} - \beta^*_{j,k'}| + |\hat{\beta}_{j,k} - \beta^*_{j,k}| > 0\right) \\ &= 1 - P\left(\hat{\beta}_{j,k'} = \beta^*_{j,k'}\right) P\left(\hat{\beta}_{j,k} = \beta^*_{j,k}\right) \to 0 \end{aligned}$$

as $n \to \infty$ since $\hat{\beta}_{j,k'}$ and $\hat{\beta}_{j,k}$ are independent and consistent estimators. Similarly, the consistency of the estimated ordering $\hat{\boldsymbol{V}}_j$ of the absolute values in vector $\boldsymbol{\beta}^*_{j,\bullet}$ can be derived by taking the square of the absolute values and following the same argument as for $\hat{\boldsymbol{U}}_j$. ■

Proof of Theorem 3: Here we assume the same regularity condition as in Theorem 1. To complete this proof, we first define the event $\mathcal{W}$ when the orderings of all p covariates are correctly assigned as

$$\mathcal{W} = \bigcap_{j=1}^{p} \left(\{\hat{\boldsymbol{U}}_j = \boldsymbol{U}_j\} \cap \{\hat{\boldsymbol{V}}_j = \boldsymbol{V}_j\}\right).$$

Let $\hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{W}}}$ be $\hat{\boldsymbol{\theta}}_{\mathcal{W}}$ when $\mathcal{W}$ occurs; otherwise, denote it as $\hat{\boldsymbol{\theta}}_{\mathcal{W}^c}$. Then, the estimator can be rewritten as

$$\hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{W}}} = \hat{\boldsymbol{\theta}}_{\mathcal{W}} \mathbf{1}\{\mathcal{W}\} + \hat{\boldsymbol{\theta}}_{\mathcal{W}^c} \mathbf{1}\{\mathcal{W}^c\}$$

and therefore

$$\sqrt{N}\left(\hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{W}}} - \boldsymbol{\theta}^*\right) = \sqrt{N}\left(\hat{\boldsymbol{\theta}}_{\mathcal{W}} - \boldsymbol{\theta}^*\right) \mathbf{1}\{\mathcal{W}\} + \sqrt{N}\left(\hat{\boldsymbol{\theta}}_{\mathcal{W}^c} - \boldsymbol{\theta}^*\right) \mathbf{1}\{\mathcal{W}^c\}. \tag{9}$$

By Theorem 1, we have $\sqrt{N}\left(\hat{\boldsymbol{\theta}}_{\mathcal{W}} - \boldsymbol{\theta}^*\right) = O(1)$ and $\sqrt{N}\left(\hat{\boldsymbol{\theta}}_{\mathcal{W}^c} - \boldsymbol{\theta}^*\right) = O(1)$ as $n \to \infty$. By Lemma 2, we have $P(\mathcal{W}) \to 1$ and $P(\mathcal{W}^c) \to 0$ as $n \to \infty$. Therefore, by Slutsky's Theorem, (9) converge to the same distribution as $\sqrt{N}\left(\hat{\boldsymbol{\theta}}_{\mathcal{W}} - \boldsymbol{\theta}^*\right)$. Similarly, by results from Theorem 1 and Lemma 2, we have selection consistency

$$P(\hat{\mathcal{A}}^{\hat{\boldsymbol{W}}} = \mathcal{A}) = P(\hat{\mathcal{A}}^{\hat{\boldsymbol{W}}} = \mathcal{A}|\mathcal{W})P(\mathcal{W}) + P(\hat{\mathcal{A}}^{\hat{\boldsymbol{W}}} = \mathcal{A}|\mathcal{W}^c)P(\mathcal{W}^c) \to 1$$

as $n \to \infty$. This completes the proof of the Theorem 3. ■

Appendix B. Performance with Distorted Parameter Ordering

Under the same setting as simulation experiment 1 in Section 5.1 with $\alpha = 0$ and $s = 0$, we conduct a sensitivity analysis to evaluate the performance of FLARCC when parameter ordering is incorrectly specified. Specifically, we report results of sensitivity, specificity and MSE for the linear regression model when the coefficient ordering is determined from the initial estimate with distortion through an added disturbance ϵ, $\hat{\boldsymbol{\beta}} + \epsilon$, where $\hat{\boldsymbol{\beta}}$ from (1) and $\epsilon \sim \mathcal{N}(0, v^2)$. As v^2 increases, the percent of order switching in initial estimates increases. Sensitivity, specificity and MSE in relation to the percentage of wrongly ordered

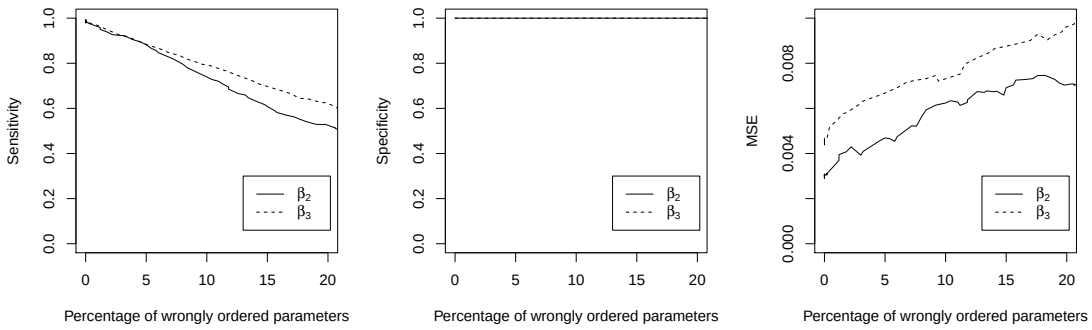

Figure 4: Clustering sensitivity and mean squared error of two heterogeneous slope parameters $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_3$ based on FLARCC with λ selected by EBIC, as the percent of distorted ordering increases. Results are summarized from 100 replications.

parameters are displayed in Figure 4 for the two heterogeneous effects $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_3$, and the homogeneous parameter $\boldsymbol{\beta}_1$ is not included in the comparison because of no effect from the distortion on its performance. As the percentage of wrongly ordered parameters increases, as expected, sensitivity becomes lower and MSE becomes larger. However, specificity remains unaffected. When the distortion of ordering is mild ($\leq 10\%$), the performance of FLARCC appears satisfactory in this simulation setting.

References

Jiahua Chen and Zehua Chen. Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771, 2008.

Rebecca DerSimonian and Raghu Kacker. Random-effects model for meta-analysis of clinical trials: an update. *Contemporary Clinical Trials*, 28(2):105–114, 2007.

Jerome Friedman, Trevor Hastie, Holger Höfling, and Robert Tibshirani. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332, 2007.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1, 2010.

Xin Gao and Peter X-K Song. Composite likelihood bayesian information criteria for model selection in high-dimensional data. *Journal of the American Statistical Association*, 105 (492):1531–1540, 2010.

Gene V Glass. Primary, secondary, and meta-analysis of research. *Educational Researcher*, 5(10):3–8, 1976.

Gene H Golub, Michael Heath, and Grace Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979.

Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054, 1982.

Zheng Tracy Ke, Jianqing Fan, and Yichao Wu. Homogeneity pursuit. *Journal of the American Statistical Association*, 110(509):175–194, 2015.

Jeffrey T Leek and John D Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, 3(9):1724–1735, 2007.

Hung-Mo Lin, H Myron Kauffman, Maureen A McBride, Darcy B Davies, John D Rosendale, Carol M Smith, Erick B Edwards, O Patrick Daily, James Kirklin, Charles F Shield, and Lawrence G Hunsicker. Center-specific graft and patient survival rates: 1997 united network for organ sharing (unos) report. *Journal of the American Medical Association*, 280(13):1153–1160, 1998.

Dungang Liu, Regina Y Liu, and Minge Xie. Multivariate meta-analysis of heterogeneous studies using only summary statistics: efficiency and robustness. *Journal of the American Statistical Association*, 110(509):326–340, 2015.

Kirk E Lohmueller, Celeste L Pearce, Malcolm Pike, Eric S Lander, and Joel N Hirschhorn. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nature Genetics*, 33(2):177–182, 2003.

Thomas Lumley and Alastair Scott. AIC and BIC for modeling with complex survey data. *Journal of Survey Statistics and Methodology*, 3(1):1–18, 2015.

Wei Pan, Xiaotong Shen, and Binghui Liu. Cluster analysis: unsupervised learning via supervised learning with a non-convex penalty. *The Journal of Machine Learning Research*, 14(1):1865–1889, 2013.

Jin Qin and Jerry Lawless. Empirical likelihood and general estimating equations. *The Annals of Statistics*, 22(1):300–325, 1994.

Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2): 461–464, 1978.

Paul G Shekelle, Mary L Hardy, Sally C Morton, Margaret Maglione, Walter A Mojica, Marika J Suttorp, Shannon L Rhodes, Lara Jungvig, and James Gagné. Efficacy and safety of ephedra and ephedrine for weight loss and athletic performance: a meta-analysis. *Journal of the American Medical Association*, 289(12):1537–1545, 2003.

Xiaotong Shen and Hsin-Cheng Huang. Grouping pursuit through a regularization solution surface. *Journal of the American Statistical Association*, 105(490):727–739, 2010.

Sunyoung Shin, Jason Fine, and Yufeng Liu. Adaptive estimation with partially overlapping models. *Statistica Sinica*, 26(1):235–253, 2016.

Patrick F Sullivan, Michael C Neale, and Kenneth S Kendler. Genetic epidemiology of major depression: review and meta-analysis. *American Journal of Psychiatry*, 157(10): 1552–1562, 2000.

Alexander J Sutton and Julian Higgins. Recent developments in meta-analysis. *Statistics in Medicine*, 27(5):625–650, 2008.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58(1):267–288, 1996.

Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

U.S. Census Bureau. Regions and divisions. `http://www.census.gov/econ/census/help/geography/regions_and_divisions.html`, 2015. [Online; accessed 10-31-2015].

Fei Wang, Lu Wang, and Peter X-K Song. Fused lasso with the adaptation of parameter ordering in combining multiple studies with repeated measurements. *Biometrics*, 2016. doi: 10.1111/biom.12496. Advance online publication.

Minge Xie, Kesar Singh, and William E Strawderman. Confidence distributions and a unifying framework for meta-analysis. *Journal of the American Statistical Association*, 106(493):320–333, 2012.

Sen Yang, Lei Yuan, Ying-Cheng Lai, Xiaotong Shen, Peter Wonka, and Jieping Ye. Feature grouping and selection over an undirected graph. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 922–930. ACM, 2012.

Jianming Ye. On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, 93(441):120–131, 1998.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.