DOI: 10.1111/biom.12496

Fused Lasso with the Adaptation of Parameter Ordering in Combining Multiple Studies with Repeated Measurements

Fei Wang,[1]* Lu Wang,[2] and Peter X.-K. Song[2]*****

[1]Global Analytics, Ford Motor Credit, Dearborn, Michigan, U.S.A. 48126
[2]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, U.S.A. 48109
**email:* fwang55@ford.com
***email:* luwang@umich.edu
****email:* pxsong@umich.edu

Summary. Combining multiple studies is frequently undertaken in biomedical research to increase sample sizes for statistical power improvement. We consider the marginal model for the regression analysis of repeated measurements collected in several similar studies with potentially different variances and correlation structures. It is of great importance to examine whether there exist common parameters across study-specific marginal models so that simpler models, sensible interpretations, and meaningful efficiency gain can be obtained. Combining multiple studies via the classical means of hypothesis testing involves a large number of simultaneous tests for all possible subsets of common regression parameters, in which it results in unduly large degrees of freedom and low statistical power. We develop a new method of *fused lasso with the adaptation of parameter ordering* (FLAPO) to scrutinize only adjacent-pair parameter differences, leading to a substantial reduction for the number of involved constraints. Our method enjoys the oracle properties as does the full fused lasso based on all pairwise parameter differences. We show that FLAPO gives estimators with smaller error bounds and better finite sample performance than the full fused lasso. We also establish a regularized inference procedure based on bias-corrected FLAPO. We illustrate our method through both simulation studies and an analysis of HIV surveillance data collected over five geographic regions in China, in which the presence or absence of common covariate effects is reflective to relative effectiveness of regional policies on HIV control and prevention.

Key words: Data integration; Error bounds; Estimating equation; Inference; Regularization.

1. Introduction

This article concerns regression analysis of repeated measurements from multiple studies using the marginal model. When the sample size of a biomedical study is not large enough to achieve adequate statistical precision, it is a common practice to combine data from several similar studies (Zhang et al., 2007; Thase et al., 2009). For instance, in a study of prostate-specific antigen, Inoue et al. (2004) studied the pattern of the prostate-specific antigen growth by combining three longitudinal studies to obtain adequate sample sizes to reach satisfactory statistical power.

Arguably the increased sample size by combining data from similar studies cannot always lead to desirable improvement in estimation efficiency or testing power, especially when datasets are sampled from heterogeneous subpopulations. In meta analysis, a strong assumption of equal parameters from individual studies is routinely imposed in order to combine study-specific estimates. When data from different subpopulations are blindly assumed to have common regression parameters without any *a priori* data evidence or as such, it would be hard to interpret the estimated covariate effects. Thus, with the availability of subject-level data, one of the primary tasks before combining multiple datasets is to check parameter homogeneity across multiple studies. In this article, we are interested in developing a methodology that enables us to examine and identify sets of homogeneous (or common) regression coefficients across multiple studies. As a result, we may simplify the formation of the mean model, and consequently yield sensible interpretations and meaningful efficiency gain from combining multiple data sets.

Our methodology development was motivated by a national HIV surveillance project on injection drug users (IDUs) in a southwestern province of China. By the end of 2006, China had established 393 national and 370 provincial monitoring sites reporting HIV incidences to the national center for AIDS/sexually transmitted disease control and prevention (Sun et al., 2007). Provincial HIV sentinel surveillance program involved community health center, hospitals, and drug addiction treatment centers at which surveys were conducted among high-risk groups of IDUs.

The HIV surveillance data were collected between 2006 and 2009 using stratified sampling from 67 hospitals, community health center, and drug addiction treatment centers as primary sample units to monitor incidences of HIV infection among IDUs in the study area. All IDUs sampled in the surveys were tested for HIV and interviewed for their behavioral characteristics related to drug usage, e.g., if they inhale drugs, if they share needles with other IDUs, and if they are infected by syphilis virus. Cluster sizes of primary sample units varied greatly from 11 to 440 IDUs.

The study contains five regions termed as A, B, C, D, and E, which are very different in many aspects, such as population size, HIV prevalence, and socioeconomic status. For example, A is the largest metropolitan city in the province, whereas E is primarily dominated by minorities living in mountain villages. Thus, it is expected that highly diversified backgrounds and behaviors of IDUs across these regions possibly lead to different trends and covariate effects on HIV positive.

The focus of this study was on the association between behavioral activities and HIV positive, among which needle sharing is the central variable that has been proved as a critical factor for the infection of HIV. In particular, the provincial Center for Disease Control was interested in assessing the effectiveness of measures on needle sharing control across the five regions. This required to identify common effects of needle sharing so that similar effectiveness of policies on disease control and prevention may be clustered in the five regions.

Desirable properties for an approach used in combining multiple studies with repeated measurements include flexibility and robustness with respect to heterogeneous characteristics across study cohorts, such as discrepancies of within-cluster correlation, dispersion, or longitudinal follow-up schedule. Meta analysis (Hedges and Olkin, 1985; Hartung et al., 2008), e.g., Cocharn's test (Cochran, 1954), assumes all study-specific parameters are equal to a population parameter. Meta analysis utilizes individual estimates, instead of full datasets, to provide an overall combined estimator for the population parameter. This approach focuses more on providing inferential summary than identifying parameter structures existing in multiple studies. Also, Wang et al. (2012) showed that Cocharn's test is unable to control Type I error against heterogeneous covariances in multiple longitudinal studies. In regard to generalized estimating equations (Zeger and Liang, 1986), several versions of modified sandwich covariance estimators have been proposed to account for various types of heterogeneities; refer to Wang et al. (2012) and more references therein. However, all existing approaches are greatly challenged by the large number of simultaneous hypotheses to be checked for coefficient homogeneity when many studies and/or many covariates in individual studies are involved. In effect, the number of tests required in the case of K studies, each of which contains p covariates, is of order $C(K,2)^p$, where $C(K,2)=K(K-1)/2$ is the number of combinations of two studies out of K studies. When either K, or p, or both are large, the degrees of freedom of a test statistic increase rapidly, leading to low power. To deal with such issue of high-dimensionality, Ke et al. (2015) proposed a clustering algorithm to identify homogeneous parameter groups in a single regression model by taking the advantage of preliminary estimates obtained under full heterogeneity.

Alternatively, meta analysis may be tackled by Bayesian approaches in that random effects models are typically used to account for similarity and discrepancy among multiple studies (Smith et al., 1995). Müller et al. (2004) proposed a combined inference over several Bayesian models using a mixture of a common distribution and an idiosyncratic distribution specific to each study. Dunson (2006) considered a dynamic mixture of Dirichlet processes to account for heterogeneity of latent response distributions. Also see Dunson et al. (2008) concerning an approach of matrix stick-breaking processes for inter-study heterogeneity. In most of these Bayesian approaches, prior specification and computing based on the MCMC algorithm are not straightforward.

In contrast to Ke et al. (2015)'s method in a single study, we consider issues arising from combining multiple studies from a Frequentist point of view. We propose a new method by generalizing the fused lasso method (Tibshirani et al., 2005) in a system of parallel estimating functions, each formed for one study. We propose an objective function that automatically allocates balanced weighting on different studies, so that none of studies would dominate the resulting objective function. Another contribution in this article is rooted in an appealing adjustment on the penalty function through the adaptation of parameter ordering. This new adaptive approach is different from Zou's (2006) adaptive lasso, which incorporates the magnitudes of initial estimates to rescale the amounts of penalty on individual regression parameters. In the specification of contrasts in the fused lasso, we hope make a trade-off between sufficiency and conciseness, so that although only using a subset of adjacent parameter differences we can still sufficiently cover the spectrum of parameter structures in the regularized estimation. As a result, our proposed method, termed as *fused lasso with the adaptation of parameter ordering* (FLAPO), not only can identify common coefficients shared in multiple studies but also can reduce the uncertainty and complications pertinent to redundant constraints in pairwise comparisons. As shown in simulation studies and Theorem A in the Supplemental Materials, our proposed FLAPO exhibits smaller error bounds and better finite-sample performance than the full fused lasso that uses all possible pairwise constraints in the regularization. Following van de Geer et al. (2014), we provide an inference procedure in FLAPO, which is applied to analyze the HIV surveillance data with conclusion of statistical significance.

The rest of this article is organized as follows. Section 2 concerns both model formulation and FLAPO methodology. Section 3 presents an algorithm for algorithmic implementation. Section 4 present theoretical results for FLAPO. After simulation studies in Section 5, Section 6 presents the analysis of the HIV surveillance data. Section 7 provides concluding remarks. The Supplementary Web Materials include relevant technical details and extra numerical results.

2. Formulation and Method

We consider K studies, where study k, $k=1,\ldots,K$, collects n_k clusters, and cluster i contains $m_{k,i}$ repeated measurements, $i=1,\ldots,n_k$. Let $Y_{k,ij}$ denote the outcome and $\mathbf{X}_{k,ij}$ denote a p-dimensional covariate vector for the jth observation of cluster i in study k, where $j=1,\ldots,m_{k,i}$. For the ease of exposition, we let $n=\sum_{k=1}^{K} n_k$ and assume all studies have the same number of repeated measurements; that is, all $m_{k,i}=m$. For study k, the marginal model is specified as follows: the conditional mean of $Y_{k,ij}$ takes the form of $E(Y_{k,ij}\mid \mathbf{X}_{k,ij})=\mu_{k,ij}=h(\mathbf{X}_{k,ij}^T\boldsymbol{\beta}_k^0)$, and the conditional variance of $Y_{k,ij}$ is given by $\mathrm{var}(Y_{k,ij}\mid \mathbf{X}_{k,ij})=\sigma_k v(\mu_{k,ij})$, where σ_k is the dispersion parameter, $h(\cdot)$ and $v(\cdot)$ are the known link and variance functions, respectively, and $\boldsymbol{\beta}_k^0=(\beta_{k,1}^0,\ldots,\beta_{k,p}^0)^T$ is the vector of regression coefficients associated with $\mathbf{X}_{k,ij}$.

Denote $\mathbf{Y}_{k,i} = (Y_{k,i1}, \ldots, Y_{k,im})^T$, $\boldsymbol{\mu}_{k,i} = (\mu_{k,i1}, \ldots, \mu_{k,im})^T$, $\boldsymbol{\beta}^0 = ({\boldsymbol{\beta}_1^0}^T, \ldots, {\boldsymbol{\beta}_K^0}^T)^T$. To describe the underlying parameter configuration for each covariate x_l, we introduce a collection of study-index sets, $\mathcal{G}_0^l$, that constitutes, *say*, B distinct groups of parameters among K coefficients of x_l, $\boldsymbol{\beta}_{(l)}^0 = (\beta_{1,l}^0, \ldots, \beta_{K,l}^0)^T$. It takes the form of $\mathcal{G}_0^l = \uplus_{b=1}^B \mathcal{G}_0^{l,b}$, where $\mathcal{G}_0^{l,b} \subset \{1, \ldots, K\}$ contains indices of studies whose $\beta_{k,l}$'s equal to a common value. Operation $\uplus$ denotes a union of multiple subsets (not elements in subsets). Take an example of five studies in which parameters for the first covariate have two clusters given by $\beta_{1,1}^0 = \beta_{2,1}^0 = \beta_{3,1}^0 < 0 < \beta_{4,1}^0 = \beta_{5,1}^0$. Then, $\mathcal{G}_0^{1,1} = \{1, 2, 3\}, \mathcal{G}_0^{1,2} = \{4, 5\}$, and moreover $\mathcal{G}_0^1 = \{1, 2, 3\} \uplus \{4, 5\} = \{\{1, 2, 3\}, \{4, 5\}\}$. Since these parameters can be equivalently represented by parameter differences, we may instead use their pairwise differences to describe $\mathcal{G}_0^l$. As a convention, elements in each cluster $\{\beta_{k,l}^0, k \in \mathcal{G}_0^{l,b}\}$ are always listed by an order of their study indices. Thus, without loss of generality, we assume the true ordering is $\beta_{1,l}^0 \leq \cdots \leq \beta_{K,l}^0$. Then, $\boldsymbol{\beta}_{(l)}^0$ may be reparameterized by $\boldsymbol{\phi}_{(l)}^0 = (\phi_{1,l}^0, \phi_{2,l}^0, \ldots, \phi_{K,l}^0)^T$ where $\phi_{1,l}^0 = \beta_{1,l}^0$ and $\phi_{k,l}^0 = \beta_{k,l}^0 - \beta_{k-1,l}^0$ for $k = 2, \ldots, K$. Denote $\boldsymbol{\phi}^0 = ({\boldsymbol{\phi}^0}^T_{(1)}, \ldots, {\boldsymbol{\phi}^0}^T_{(p)})^T$. By the above convention, two sets $\mathcal{G}_0^l$ and $\mathcal{A}_0^l$ can be fully determined each other. Let $\mathcal{A}_0^l = \{\{k\} : \ \phi_{k,l} = 0, 1 \leq k \leq K\}$ is the set of study indices whose $\beta_{k,l}^0$'s are identical to the lower adjacent ones (or no jumps between pairs of adjacent coefficients). For the above example, $\mathcal{A}_0^1 = \{\{2\}, \{3\}, \{5\}\}$, and sets $\mathcal{G}_0^1$ and $\mathcal{A}_0^1$ can be uniquely converted into each other, because $\phi_{2,1}^0 = \phi_{3,1}^0 = 0$ is equivalent to $\beta_{1,1}^0 = \beta_{2,1}^0 = \beta_{3,1}^0$, and so is $\phi_{5,1}^0 = 0$ to $\beta_{4,1}^0 = \beta_{5,1}^0$. Thus, we can characterize the underlying parameter configuration by covariate-specific sets $\mathcal{A}_0^1, \ldots, \mathcal{A}_0^p$. Denote $\mathcal{A}_0 = \uplus_{l=1}^p \mathcal{A}_0^l$. Let the cardinality of $\mathcal{A}_0$ be $a_0 = \text{card}(\mathcal{A}_0) = \sum_{l=1}^p \text{card}(\mathcal{A}_0^l)$. Then, the cardinality of its complement, $\mathcal{A}_0^c$, is $b_0 = \text{card}(\mathcal{A}_0^c) = Kp - a_0$.

Our objective is twofold: to determine grouping structures in the set $\mathcal{A}_0$, and to estimate coefficients under the parameter configuration by $\mathcal{A}_0$. These two tasks can be achieved simultaneously by using the regularization technique proposed in the article.

As pointed out by Wang et al. (2012), the traditional estimating function approach is questionable to draw inference when the data are heterogeneous from one study to another. To account for such heterogeneity, we first establish a system of K study-specific estimating functions for $\boldsymbol{\beta}_k^0$, each for one study, and then combine them by the means of the generalized method of moments (Hansen, 1982), which is also referred to as the quadratic inference function (QIF) by Qu et al. (2000). This way of creating a meta estimating function enjoys the flexibility of accommodating different variance–covariance structures across different studies. Another advantage of this approach is that it allows data from multiple studies to contribute equally to the formation of the meta objective function, regardless of individual study sample size. A detailed discussion of this point is given at the end of this section. For each study, we first approximate the inverse of working correlation matrix $\mathbf{R}_k(\alpha_k)$ by $\mathbf{R}_k^{-1}(\alpha_k) \approx \sum_{s=1}^{s_k} \varrho_s \mathbf{M}_{k,s}$, where $\varrho_1, \ldots, \varrho_{s_k}$ are constants possibly dependent on α_k, and $\mathbf{M}_{k,1}, \ldots, \mathbf{M}_{k,s_k}$ are known basis matrices with elements 0 and 1. Refer to Qu et al. (2000) for more details concerning the basis matrices given in different working correlation structures, such as compound symmetry (CS) and first order autoregressive (AR-1). Also refer to Song et al. (2009) for the extension of QIF for data of unequal cluster sizes.

Using the above expansion of $\mathbf{R}_k^{-1}$, we can construct a system of study-specific estimating functions, $\bar{\mathbf{g}}_k(\boldsymbol{\beta}_k)$ for study $k = 1, \ldots, K$, given as follows:

$$\begin{aligned}\bar{\mathbf{g}}_k(\boldsymbol{\beta}_k) &= \frac{1}{n_k} \sum_{i=1}^{n_k} \mathbf{g}_{k,i}(\boldsymbol{\beta}_k) \\ &= \frac{1}{n_k} \sum_{i=1}^{n_k} \left\{ \begin{array}{c} \partial\boldsymbol{\mu}_{k,i}^T \mathbf{A}_{k,i}^{-1/2} \mathbf{M}_{k,1} \mathbf{A}_{k,i}^{-1/2} (\mathbf{Y}_{k,i} - \boldsymbol{\mu}_{k,i}) \\ \vdots \\ \partial\boldsymbol{\mu}_{k,i}^T \mathbf{A}_{k,i}^{-1/2} \mathbf{M}_{k,s_k} \mathbf{A}_{k,i}^{-1/2} (\mathbf{Y}_{k,i} - \boldsymbol{\mu}_{k,i}) \end{array} \right\},\end{aligned}$$

$\partial\boldsymbol{\mu}_{k,i} = \partial\boldsymbol{\mu}_{k,i}^T / \partial\boldsymbol{\beta}_k$ and $\mathbf{A}_{k,i} = \text{diag}\Big\{ v(\mu_{k,i1}), \ldots, v(\mu_{k,im}) \Big\}$. The dimension of $\bar{\mathbf{g}}_k(\boldsymbol{\beta}_k)$ is $s_k \text{dim}(\boldsymbol{\beta}_k)$. Instead of summing these study-specific $\bar{\mathbf{g}}_k(\boldsymbol{\beta}_k)$, we stack them to form an extended score function: $\bar{\mathbf{g}}(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n \mathbf{g}_i(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n \left\{ \delta_i(1) \mathbf{g}_{1,i}(\boldsymbol{\beta}_1)^T, \ldots, \delta_i(K) \mathbf{g}_{K,i}(\boldsymbol{\beta}_K)^T \right\}^T$, where $\delta_i(k)$=1 denotes that subject i belongs to study k, and $\delta_i(k) = 0$ otherwise. Because the dimension of $\bar{\mathbf{g}}(\boldsymbol{\beta})$ is much larger than that of $\boldsymbol{\beta}$, namely the case of over-identification, following Qu et al. (2000), we construct an objective function of the form: $Q(\boldsymbol{\beta}) = n \bar{\mathbf{g}}(\boldsymbol{\beta})^T \mathbf{C}^{-1}(\boldsymbol{\beta}) \bar{\mathbf{g}}(\boldsymbol{\beta})$, where $\mathbf{C}(\boldsymbol{\beta}) = \text{block-diag} \left\{ \frac{n_1}{n} \mathbf{C}_1(\boldsymbol{\beta}_1), \ldots, \frac{n_k}{n} \mathbf{C}_K(\boldsymbol{\beta}_k) \right\}$ and $\mathbf{C}^{-1}(\boldsymbol{\beta})$ is the inverse matrix of $\mathbf{C}(\boldsymbol{\beta})$. Thus, the classical QIF estimator is $\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} Q(\boldsymbol{\beta})$. Note that the objective function $Q(\boldsymbol{\beta})$ can also be written as $Q(\boldsymbol{\beta}) = \sum_{k=1}^K Q_k(\boldsymbol{\beta}_k) = \sum_{k=1}^K n_k \bar{\mathbf{g}}_k(\boldsymbol{\beta}_k)^T \mathbf{C}_k^{-1}(\boldsymbol{\beta}_k) \bar{\mathbf{g}}_k(\boldsymbol{\beta}_k)$, where $Q_k(\boldsymbol{\beta}_k)$ is a study-specific QIF. According to Qu et al. (2000), when the mean model in study k is correctly specified, $Q_k(\hat{\boldsymbol{\beta}}_k)$ converges in distribution to $\chi^2_{r_k - p}$ where r_k is the dimension of $\mathbf{C}_k(\boldsymbol{\beta}_k^0)$. It is worth pointing out that the asymptotic behavior of Q_k does not depend on the sample size n_k, nor on the dispersion σ_k. In other words, the sample size will not dictate the contribution of an individual QIF to the meta inference function.

Now, we turn to the development of FLAPO methodology for parameter fusion. Denote all regression coefficients by $\boldsymbol{\beta} = (\boldsymbol{\beta}_{(1)}^T, \ldots, \boldsymbol{\beta}_{(p)}^T)^T$. To identify homogeneous parameter groups, we propose to regularize the above QIF objective function, $Q(\boldsymbol{\beta})$, using two new penalties with the adaption of parameter ordering. To proceed, let us begin with the adaptive fused lasso (Tibshirani et al., 2005; Zou, 2006), whose penalty takes the form: $P(\boldsymbol{\beta}) = \sum_{l=1}^p \sum_{k=1}^K \sum_{k'>k}^K w_{kk',l} |\beta_{k,l} - \beta_{k',l}| + \sum_{l=1}^p \sum_{k=1}^K w_{k,l} |\beta_{k,l}|$, where weights $w_{k,l} = 1/|\beta_{k,l}^*|^{\gamma_1}$ and $w_{kk',l} = 1/|\beta_{k,l}^* - \beta_{k',l}^*|^{\gamma_2}$ are typically specified by initial root-n consistent estimates $\beta_{k,l}^*$'s of $\beta_{k,l}$'s for some constants $\gamma_1, \gamma_2 > 0$ (Zou, 2006). In practice, often γ_1 and γ_2 are set equal to 1. Note that for each covariate, the total number of constraints is $s = K + C(K, 2)$. Ueki (2009) and Ueki and Kawasaki (2011) considered a similar problem of variable grouping in a much simpler setting of single cross-sectional study (i.e., $K = 1$, $m = 1$), where the ℓ_2-norm penalty for group lasso was used.

Equivalently, we may write the above fused lasso penalty $P(\boldsymbol{\beta})$ in a matrix notation: $P(\boldsymbol{\beta}) = \|\mathbf{D}\boldsymbol{\beta}\|_1 = \|\mathbf{W}\mathbf{B}\boldsymbol{\beta}\|_1$, where $\|\cdot\|_1$ is L_1-norm on R^{sp}, $\mathbf{B}$ is an $sp \times Kp$ matrix that defines sp constraints involving p covariates across K studies, and $\mathbf{W}$ is an $sp \times sp$ diagonal matrix containing all weights corresponding to the constraints in $\mathbf{B}$. Thus, to compare a pair $\beta_{k,l}$ and $\beta_{k',l}$, $k \neq k'$, the corresponding two entries in $\mathbf{B}$ are 1 and -1, and the corresponding diagonal entry in $\mathbf{W}$ is $w_{kk',l}$. For a single parameter $\beta_{k,l}$, the corresponding entry in $\mathbf{B}$ is 1 and the corresponding entry in $\mathbf{W}$ is $w_{k,l}$.

A potential caveat with the above fused lasso penalty $P(\boldsymbol{\beta})$ is that most of sp constraints in $\mathbf{B}$ are redundant, especially when the regression parameters of a covariate are ordered. For an example of $\beta^0_{1,1} \leq \beta^0_{2,1} \leq \beta^0_{3,1}$, the term $|\beta_{1,1} - \beta_{3,1}|$ may not be needed when two adjacent pairs $|\beta_{1,1} - \beta_{2,1}|$ and $|\beta_{2,1} - \beta_{3,1}|$ are used. Thus, when it is possible to arrange parameters in $\boldsymbol{\beta}^0_{(l)}$ in an increasing order as, *say*, $\beta^0_{1,l} \leq \beta^0_{2,l} \leq \cdots \leq \beta^0_{K,l}$, we can consider a simpler $K \times K$ constraint matrix $\widetilde{\mathbf{B}}_l$ for adjacent pairs in the $\boldsymbol{\beta}_{(l)}$ for covariate x_l. $\widetilde{\mathbf{B}}_l$ is a lower-triangular matrix of all zero entries, except the elements on the main diagonal being $(1, -1, \ldots, -1)$ and those on the subdiagonal (i.e., directly below the main diagonal) all equal to 1. Through row permutations in $\widetilde{\mathbf{B}}_l$, it is easy to accommodate different orderings of parameters in $\boldsymbol{\beta}^0_{(l)}$. Define a block-diagonal $Kp \times Kp$ matrix $\widetilde{\mathbf{B}} = \text{block-diag}\{\widetilde{\mathbf{B}}_1, \ldots, \widetilde{\mathbf{B}}_p\}$. In the fused lasso penalty $P(\boldsymbol{\beta})$ above, matrix $\mathbf{B}$ can be partitioned as $\mathbf{B} = (\widetilde{\mathbf{B}}^T, \overline{\mathbf{B}}^T)^T$ where $\overline{\mathbf{B}}$ is a $qp \times Kp$ matrix of the redundant pairs that are not included in $\widetilde{\mathbf{B}}$, $q = s - K$. Accordingly, matrix $\mathbf{W}$ may be partitioned as $\mathbf{W} = \text{block-diag}(\widetilde{\mathbf{W}}, \overline{\mathbf{W}})$, where $\widetilde{\mathbf{W}}$ is a $Kp \times Kp$ matrix consisting of weights corresponding to $\widetilde{\mathbf{B}}$, and $\overline{\mathbf{W}}$ is a $qp \times qp$ matrix of the weights associated with $\overline{\mathbf{B}}$. Unfortunately, such partition for matrix $\mathbf{B}$ is unknown in practice. However, if the parameter ordering were known and utilized, a new penalty (termed as the FLAPO penalty) would be specified by the form: $\widetilde{P}(\boldsymbol{\beta}) = \|\widetilde{\mathbf{D}}\boldsymbol{\beta}\|_1 = \|\widetilde{\mathbf{W}}\widetilde{\mathbf{B}}\boldsymbol{\beta}\|_1$. Thus, adequately estimating the parameter ordering is crucial to carry out the above strategy, and when such ordering is available from, *say*, certain initial root-n consistent estimates $\boldsymbol{\beta}^*$, we could construct a matrix $\widetilde{\mathbf{B}}_e$ to estimate $\widetilde{\mathbf{B}}$. Consequently, a new weight matrix $\widetilde{\mathbf{W}}_e$ replaces $\widetilde{\mathbf{W}}$, and moreover, an empirical counterpart of the FLAPO penalty $\widetilde{P}(\boldsymbol{\beta})$ is given by

$$\widetilde{P}_e(\boldsymbol{\beta}) = \sum_{l=1}^{p}\sum_{k=1}^{K}\sum_{k'>k}^{K} w_{kk',l}\,\delta\{|T^*_{k,l} - T^*_{k',l}| = 1\}|\beta_{k,l} - \beta_{k',l}| + \sum_{l=1}^{p} w_{k^*_l,l}|\beta_{k^*_l,l}|, \quad (1)$$

where $T^*_{k,l} = \sum_{k'=1}^{K} \delta\{\beta^*_{k',l} \geq \beta^*_{k,l}\}$ is the ranking of $\beta^*_{k,l}$ among the elements in $\boldsymbol{\beta}^*_{(l)}$, and k^*_l is the lowest position. A matrix form for (1) is now written as $\widetilde{P}_e(\boldsymbol{\beta}) = \|\widetilde{\mathbf{D}}_e\boldsymbol{\beta}\|_1 = \|\widetilde{\mathbf{W}}_e\widetilde{\mathbf{B}}_e\boldsymbol{\beta}\|_1$.

We consider three versions of the regularized estimators obtained, respectively, by minimizing the following penalized objective functions, $\widehat{\boldsymbol{\beta}}_D = \underset{\boldsymbol{\beta}\in R^{Kp}}{\arg\min}\big\{Q(\boldsymbol{\beta}) + \lambda P(\boldsymbol{\beta})\big\}$, and

$$\widehat{\boldsymbol{\beta}}_{\widetilde{D}} = \underset{\boldsymbol{\beta}\in R^{Kp}}{\arg\min}\big\{Q(\boldsymbol{\beta}) + \lambda \widetilde{P}(\boldsymbol{\beta})\big\},$$

$$\text{and } \widehat{\boldsymbol{\beta}}_{\widetilde{D}_e} = \underset{\boldsymbol{\beta}\in R^{Kp}}{\arg\min}\big\{Q(\boldsymbol{\beta}) + \lambda \widetilde{P}_e(\boldsymbol{\beta})\big\}, \quad (2)$$

where $\lambda > 0$ is a tuning parameter controlling the sparsity or cardinality of $\mathcal{A}_0$, which affects the search of common parameters. We refer to the proposed regularization method using penalty $\widetilde{P}(\boldsymbol{\beta})$ or $\widetilde{P}_e(\boldsymbol{\beta})$ as *the fused lasso with the adaptation of parameter ordering* (FLAPO). The second and third estimators $\widehat{\boldsymbol{\beta}}_{\widetilde{D}}$ and $\widehat{\boldsymbol{\beta}}_{\widetilde{D}_e}$ defined in (2) are our proposed estimators, using penalties with the true and estimated parameter orderings. The first estimator $\widehat{\boldsymbol{\beta}}_D$, which does not incorporate the ordering, is the traditional fused lasso with all possible pairwise differences in the penalty.

3. Implementation

For convenience, here we focus on FLAPO with the empirical penalty $\widetilde{P}_e(\boldsymbol{\beta})$ in the algorithm; the entire procedure is applicable to the other two penalties $\widetilde{P}(\boldsymbol{\beta})$ and $P(\boldsymbol{\beta})$. We begin by approximating QIF $Q(\boldsymbol{\beta})$ by a second-order Taylor expansion at an initial consistent estimate $\boldsymbol{\beta}^*$. This initial estimate may be obtained by performing routine GEE analysis with one study at a time, where the estimation consistency holds when the mean models are correctly specified. The second-order approximation to the objective function $\Phi(\boldsymbol{\beta}) = Q(\boldsymbol{\beta}) + \lambda\widetilde{P}_e(\boldsymbol{\beta})$ around $\boldsymbol{\beta}^*$ is

$$\Phi(\boldsymbol{\beta}) \approx Q_* + \left(\partial\mathbf{Q}^T_*\right)(\boldsymbol{\beta} - \boldsymbol{\beta}^*) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T\left(\partial^2\mathbf{Q}_*\right)(\boldsymbol{\beta} - \boldsymbol{\beta}^*) + \lambda\|\widetilde{\mathbf{D}}_e\boldsymbol{\beta}\|_1, \quad (3)$$

where Q_*, $\partial\mathbf{Q}_*$, and $\partial^2\mathbf{Q}_*$ denote $Q(\boldsymbol{\beta}^*)$, the first- and second-order derivatives of $Q(\boldsymbol{\beta})$ evaluated at $\boldsymbol{\beta}^*$, respectively. Following Kim et al. (2009), we propose the following algorithm to minimize (3) for a fixed λ.

Step 1: Evaluate both first- and second-order approximations of $\Phi(\boldsymbol{\beta})$ at an update $\hat{\boldsymbol{\beta}}^{(r)}$ obtained at iteration r. Set $\hat{\boldsymbol{\beta}}^{(1)} = \boldsymbol{\beta}^*$.

Step 2: Obtain $\hat{\boldsymbol{\tau}}^{(r)}$ by the following minimization:

$$\min_{\boldsymbol{\tau}\in R_+^{(K-1)p}} - \boldsymbol{\tau}^T\widetilde{\mathbf{D}}_e\boldsymbol{\beta}^{(r)} + \frac{1}{2}(\partial\mathbf{Q}_* + \widetilde{\mathbf{D}}^T_e\boldsymbol{\tau})^T(\partial^2\mathbf{Q}_*)^{-1}(\partial\mathbf{Q}_* + \widetilde{\mathbf{D}}^T_e\boldsymbol{\tau})\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(r)}}$$

subject to $\|\boldsymbol{\tau}\|_\infty < \lambda$.

Step 3: Update $\hat{\boldsymbol{\beta}}^{(r+1)} = \hat{\boldsymbol{\beta}}^{(r)} - (\partial^2\mathbf{Q}_*)^{-1}(\partial\mathbf{Q}_* + \widetilde{\mathbf{D}}^T_e\hat{\boldsymbol{\tau}}^{(r)})\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(r)}}$.

Step 4: If $\|\hat{\boldsymbol{\beta}}^{(r)} - \hat{\boldsymbol{\beta}}^{(r+1)}\|_\infty < \epsilon$, then stop; otherwise, set $r = r + 1$ and go back to step 1.

In practice, ϵ is set at a small number, e.g., 10^{-5}, and an optimal λ may be chosen by the smallest BIC (Schwarz, 1978): $\mathrm{BIC}(\lambda) = Q(\hat{\boldsymbol{\beta}}_\lambda) + \mathrm{df}(\hat{\boldsymbol{\beta}}_\lambda)\log(n)$, where $\hat{\boldsymbol{\beta}}_\lambda$ is the final output given at the algorithm convergence, and $\mathrm{df}(\hat{\boldsymbol{\beta}}_\lambda)$ is the number of distinctive values in $\hat{\boldsymbol{\beta}}_\lambda$. This criterion has been widely used (e.g., Wang et al., 2007, 2009). See details in the Supplementary Materials.

4. Large Sample Properties

This section concerns the asymptotic properties of the three proposed estimators under certain regularity conditions listed in Section 1 of the Supplementary Materials. Given the parameter ordering, we consider reparametrize $\boldsymbol{\beta}$ by $\boldsymbol{\phi}$ as discussed in Section 2. Although this reparametrization is not required in the first estimator $\hat{\boldsymbol{\beta}}_D$, this formulation is still adopted for the ease of exposition. To present these three estimators in the setting of reparametrization, first note the relationship: block-diag$(\widetilde{\mathbf{W}}, \overline{\mathbf{W}})(\widetilde{\mathbf{B}}^T, \overline{\mathbf{B}}^T)^T\boldsymbol{\beta} = (\widetilde{\mathbf{W}}^T, (\overline{\mathbf{W}}\overline{\mathbf{B}}\widetilde{\mathbf{B}}^{-1})^T)^T\boldsymbol{\phi} \overset{\text{def}}{=} \mathbf{F}\boldsymbol{\phi}$, where $\mathbf{F} = \mathbf{D}\widetilde{\mathbf{B}}^{-1}$, $\widetilde{\mathbf{B}}^{-1}$ is the inverse of the full-rank square matrix $\widetilde{\mathbf{B}}$, and $\mathbf{F} = (\widetilde{\mathbf{F}}^T, \overline{\mathbf{F}}^T)^T$ with $\widetilde{\mathbf{F}} = \widetilde{\mathbf{W}}$ and $\overline{\mathbf{F}} = \overline{\mathbf{W}}\overline{\mathbf{B}}\widetilde{\mathbf{B}}^{-1}$. Thus, the fused lasso penalty, the FLAPO penalty, and the empirical FLAPO penalty become $P(\boldsymbol{\phi}) = \|\mathbf{F}\boldsymbol{\phi}\|_1$, $\widetilde{P}(\boldsymbol{\phi}) = \|\widetilde{\mathbf{F}}\boldsymbol{\phi}\|_1 = \|\widetilde{\mathbf{W}}\boldsymbol{\phi}\|_1$, $\widetilde{P}_e(\boldsymbol{\phi}) = \|\widetilde{\mathbf{F}}_e\boldsymbol{\phi}\|_1 = \|\widetilde{\mathbf{W}}_e\boldsymbol{\phi}\|_1$, respectively, and the expressions of both extended scores $\mathbf{g}(\cdot)$ and QIF objective function $Q(\cdot)$ remain the same. The three regularized estimators are equivalently obtained as follows:

$$\widehat{\boldsymbol{\phi}}_F = \underset{\boldsymbol{\phi} \in R^{Kp}}{\arg\min}\big\{ Q(\boldsymbol{\phi}) + \lambda P(\boldsymbol{\phi}) \big\}, \tag{4}$$

$$\widehat{\boldsymbol{\phi}}_{\widetilde{F}} = \underset{\boldsymbol{\phi} \in R^{Kp}}{\arg\min}\big\{ Q(\boldsymbol{\phi}) + \lambda \widetilde{P}(\boldsymbol{\phi}) \big\},$$

$$\widehat{\boldsymbol{\phi}}_{\widetilde{F}_e} = \underset{\boldsymbol{\phi} \in R^{Kp}}{\arg\min}\big\{ Q(\boldsymbol{\phi}) + \lambda \widetilde{P}_e(\boldsymbol{\phi}) \big\}. \tag{5}$$

Given an estimator $\widehat{\boldsymbol{\phi}}$, which may be $\widehat{\boldsymbol{\phi}}_F$, $\widehat{\boldsymbol{\phi}}_{\widetilde{F}}$, or $\widehat{\boldsymbol{\phi}}_{\widetilde{F}_e}$, the estimated set $\widehat{\mathcal{A}_0}$ is obtained by

$$\widehat{\mathcal{A}_0} = \uplus_{l=1}^{p} \widehat{\mathcal{A}_0}^l, \quad \text{with } \widehat{\mathcal{A}_0}^l = \{\{k\} : \widehat{\phi}_{k,l} = 0, 1 \le k \le K\}, \quad l = 1, \dots, p. \tag{6}$$

Using $\mathcal{A}_0$ and its complementary set, $\mathcal{A}_0^c$, we decompose $\boldsymbol{\phi}^0 = ({\boldsymbol{\phi}^0_{\mathcal{A}_0^c}}^T, {\boldsymbol{\phi}^0_{\mathcal{A}_0}}^T)^T = ({\boldsymbol{\phi}^0_{\mathcal{A}_0^c}}^T, \mathbf{0}^T)^T$, $\widehat{\boldsymbol{\phi}}_F = (\widehat{\boldsymbol{\phi}}^T_{F\mathcal{A}_0^c}, \widehat{\boldsymbol{\phi}}^T_{F\mathcal{A}_0})^T$, $\widehat{\boldsymbol{\phi}}_{\widetilde{F}} = (\widehat{\boldsymbol{\phi}}^T_{\widetilde{F}\mathcal{A}_0^c}, \widehat{\boldsymbol{\phi}}^T_{\widetilde{F}\mathcal{A}_0})^T$, $\widehat{\boldsymbol{\phi}}_{\widetilde{F}_e} = (\widehat{\boldsymbol{\phi}}^T_{\widetilde{F}_e\mathcal{A}_0^c}, \widehat{\boldsymbol{\phi}}^T_{\widetilde{F}_e\mathcal{A}_0})^T$, $\mathbf{D} = (\mathbf{D}^T_{\mathcal{A}_0^c}, \mathbf{D}^T_{\mathcal{A}_0})^T$, $\widetilde{\mathbf{F}} = (\widetilde{\mathbf{F}}^T_{\mathcal{A}_0^c}, \widetilde{\mathbf{F}}^T_{\mathcal{A}_0})^T$, and $\overline{\mathbf{F}} = (\overline{\mathbf{F}}^T_{\mathcal{A}_0^c}, \overline{\mathbf{F}}^T_{\mathcal{A}_0})^T$.

The regularity conditions listed in Section 1 of the Supplementary Materials are required to establish Proposition 1 and Theorem 1. Proposition 1 presents the oracle property for the estimator $\widehat{\boldsymbol{\phi}}_F$ in the sense given by Fan and Li (2001), including selection consistency and asymptotic normality. Theorem 1 establishes these results for the estimator $\widehat{\boldsymbol{\phi}}_{\widetilde{F}}$ with known parameter ordering. Consequently, $\widehat{\boldsymbol{\phi}}_F$ and $\widehat{\boldsymbol{\phi}}_{\widetilde{F}}$ have the same asymptotic distribution despite different penalties.

PROPOSITION 1. *Suppose that $\lambda \to \infty$, $\lambda n^{-1/2} \to 0$, and the initial estimator $\boldsymbol{\phi}^*$ is root-n consistent. Under Assumptions 1–5 in the Supplementary Materials, the estimator $\widehat{\boldsymbol{\phi}}_F$ in (4) satisfies: (a) $\widehat{\boldsymbol{\phi}}_F$ is root-n consistent, namely $\widehat{\boldsymbol{\phi}}_F - \boldsymbol{\phi}^0 = O_p(n^{-1/2})$; (b) (selection consistency) $\widehat{\mathcal{A}_0} \to \mathcal{A}_0$ in probability, where the estimator $\widehat{\mathcal{A}_0}$ is given in* (6) *based on the estimator $\widehat{\boldsymbol{\phi}}_F$; (c) (asymptotic normality) $n^{1/2}(\widehat{\boldsymbol{\phi}}_{F\mathcal{A}_0^c} - \boldsymbol{\phi}^0_{\mathcal{A}_0^c}) {=} {-}(\mathbf{G}_{\mathcal{A}_0^c}\Sigma^{-1}\mathbf{G}^T_{\mathcal{A}_0^c})^{-1}\mathbf{G}_{\mathcal{A}_0^c}\Sigma^{-1}\Psi{+}o_p(1)$, where $\mathbf{G} = (\mathbf{G}^T_{\mathcal{A}_0^c}, \mathbf{G}^T_{\mathcal{A}_0})^T$ and $\Psi = (\Psi^T_{\mathcal{A}_0^c}, \Psi^T_{\mathcal{A}_0})^T$ with $n^{1/2}\bar{\mathbf{g}}(\boldsymbol{\phi}^0) \to \Psi \sim N(\mathbf{0}, \Sigma)$ in distribution.*

The proof of Proposition 1 is provided in Section 3.1 of the Supplementary Materials. Proposition 1 implies that the nonzero parameter $\boldsymbol{\phi}^0_{\mathcal{A}_0^c}$ can be consistently estimated at root-n rate, and that the estimator of the zero parameter $\widehat{\boldsymbol{\phi}}_{F\mathcal{A}_0}$ can be asymptotically shrunk to $\mathbf{0}$. The penalty used in $\widehat{\boldsymbol{\phi}}_F$ contains many redundant constraints, giving rise of unnecessary extra noise to the regularization procedure. The following theorem (its proof is given in Section 3.3 of the Supplementary Materials) shows that our proposed estimator $\widehat{\boldsymbol{\phi}}_{\widetilde{F}}$ based only on adjacent-pair contrasts in the penalty can achieves the same asymptotic results as those of $\widehat{\boldsymbol{\phi}}_F$.

THEOREM 1. *Suppose that $\lambda \to \infty$, $\lambda n^{-1/2} \to 0$, the initial estimator $\boldsymbol{\phi}^*$ is root-n consistent and the ordering of regression coefficients is known. Under Assumptions 1–4 in the Supplementary Materials, all results in parts (a), (b), and (c) stated for $\widehat{\boldsymbol{\phi}}_F$ in Proposition 1 hold for $\widehat{\boldsymbol{\phi}}_{\widetilde{F}}$, where $\widehat{\boldsymbol{\phi}}_{\widetilde{F}} = (\widehat{\boldsymbol{\phi}}^T_{\widetilde{F}\mathcal{A}_0^c}, \widehat{\boldsymbol{\phi}}^T_{\widetilde{F}\mathcal{A}_0})^T$, and estimator $\widehat{\mathcal{A}_0}$ is given in* (6) *based on $\widehat{\boldsymbol{\phi}}_{\widetilde{F}}$.*

In practice the ordering of parameters is unknown. For each covariate x_l, we use $T^*_{k,l}$ defined in (1) based on initial root-n consistent estimates $\boldsymbol{\beta}^*_{(l)}$ to estimate the true position $T_{k,l}$ of $\beta^0_{k,l}$, i.e., $T^*_{k,l} = \sum_{k'=1}^{K} \delta\{\beta^*_{k',l} \ge \beta^*_{k,l}\}$. Let sets $\mathbf{T}_l = \{T_{1,l}, \dots, T_{K,l}\}$ and $\mathbf{T}^*_l = \{T^*_{1,l}, \dots, T^*_{K,l}\}$, in which the elements are arranged in the same order of the $\boldsymbol{\beta}^*$. Consider an event $\{\mathbf{T}^*_l = \mathbf{T}_l\}$ that represents the coincidence of the estimated ordering with the true ordering of the parameters in $\boldsymbol{\beta}_{(l)}$. Take an example of four studies where the first covariate has two distinct parameter groups $\{\beta^0_{1,1}, \beta^0_{2,1}\}$ and $\{\beta^0_{3,1}, \beta^0_{4,1}\}$ listed as, *say*, $\beta^0_{1,1} = \beta^0_{2,1} < \beta^0_{3,1} = \beta^0_{4,1}$. Then, event $\{\mathbf{T}^*_1 = \mathbf{T}_1\}$ occurs if one of these four scenarios occurs: (i) $\beta^*_{1,1} \le \beta^*_{2,1} \le \beta^*_{3,1} \le \beta^*_{4,1}$; (ii) $\beta^*_{2,1} \le \beta^*_{1,1} \le \beta^*_{3,1} \le \beta^*_{4,1}$; (iii) $\beta^*_{2,1} \le \beta^*_{1,1} \le \beta^*_{4,1} \le \beta^*_{3,1}$; and (iv) $\beta^*_{1,1} \le \beta^*_{2,1} \le \beta^*_{4,1} \le \beta^*_{3,1}$.

LEMMA 1. *Assume estimator $\boldsymbol{\beta}^*$ is root-n consistent for $\boldsymbol{\beta}^0$. Then, pr$(\{\mathbf{T}^*_l = \mathbf{T}_l\}) \to 1$ as $n \to \infty$, for $l = 1, \dots, p$.*

The proof of Lemma 1 is given in Section 3.1 of the Supplementary Materials. This lemma means that we can estimate the parameter ordering correctly with probability tending to 1 as the sample size n increases to infinity. Therefore, we can extend the results of Theorem 1 to the proposed third estimator $\widehat{\boldsymbol{\phi}}_{\widetilde{F}_e}$, as stated in Theorem 2.

Theorem 2. *When the parameter ordering $\mathbf{T}_l$ is estimated by the $\mathbf{T}_l^*$ using an initial root-n consistent estimator $\boldsymbol{\beta}^*$, under Assumptions 1–4, the results given in Theorem 1 hold for the estimator $\widehat{\boldsymbol{\phi}}_{\widetilde{F}_e}$ defined in* (5).

The proof of Theorem 2 is given in Section 3.4 of the Supplementary Materials.

To close this section, we make an important remark on the inference. All the above results are not applicable to conduct statistical inference for parameter $\boldsymbol{\phi}$ or $\boldsymbol{\beta}^0$. Following van de Geer et al. (2014), we managed to establish the needed asymptotic distributions for bias-corrected PLAPO estimator $\widehat{\boldsymbol{\phi}}^{c}_{\widetilde{F}_e}$. To do so, we first construct a bias-corrected estimator, $\widehat{\boldsymbol{\phi}}^{c}_{\widetilde{F}_e}$, for $\widehat{\boldsymbol{\phi}}_{\widetilde{F}_e}$, given by the following form: $\widehat{\boldsymbol{\phi}}^{c}_{\widetilde{F}_e} = \widehat{\boldsymbol{\phi}}_{\widetilde{F}_e} + n^{-1}\lambda\{\partial\bar{\mathbf{g}}(\widehat{\boldsymbol{\phi}}_{\widetilde{F}_e})\mathbf{C}(\widehat{\boldsymbol{\phi}}_{\widetilde{F}_e})^{-1}\partial\mathbf{g}(\widehat{\boldsymbol{\phi}}_{\widetilde{F}_e})^T\}^{-1}\widetilde{\mathbf{F}}_e^T\kappa$, where κ is the subdifferential of $\|\widetilde{\mathbf{F}}_e\boldsymbol{\phi}\|_1$ and λ is the tuning parameter selected by the BIC. Applying similar arguments given in van de Geer et al. (2014), we obtained $\sqrt{n}(\widehat{\boldsymbol{\phi}}^{c}_{\widetilde{F}_e} - \boldsymbol{\phi}^0) \xrightarrow{d} N(\mathbf{0}, \{\mathbf{G}\boldsymbol{\Sigma}^{-1}\mathbf{G}^T\}^{-1})$, where $\mathbf{G}$ and $\boldsymbol{\Sigma}$ are defined in Assumptions 3 and 4 in the Supplementary Materials. Moreover, a bias-corrected estimator of $\widehat{\boldsymbol{\beta}}_{\widetilde{D}_e}$ of $\boldsymbol{\beta}^0$ is $\widehat{\boldsymbol{\beta}}^{c}_{\widetilde{D}_e} = \widehat{\boldsymbol{\beta}}_{\widetilde{D}_e} + \frac{1}{n}\lambda\{\widetilde{\mathbf{B}}_e^T\partial\bar{\mathbf{g}}(\widehat{\boldsymbol{\phi}}_{\widetilde{F}_e})\mathbf{C}(\widehat{\boldsymbol{\phi}}_{\widetilde{F}_e})^{-1}\partial\mathbf{g}(\widehat{\boldsymbol{\phi}}_{\widetilde{F}_e})^T\widetilde{\mathbf{B}}_e\}^{-1}\widetilde{\mathbf{B}}_e^T\widetilde{\mathbf{W}}_e^T\kappa$, which leads to $\sqrt{n}(\widehat{\boldsymbol{\beta}}^{c}_{\widetilde{D}_e} - \boldsymbol{\beta}^0) \xrightarrow{d} N\left(\mathbf{0}, (\widetilde{\mathbf{B}}_e^T\ \mathbf{G}\boldsymbol{\Sigma}^{-1}\mathbf{G}^T\widetilde{\mathbf{B}}_e)^{-1}\right)$.

5. Simulation Experiments

We conduct two simulation studies in the article. The first one, presented in this section, aims to examine the performance of the methods to identify the underlying homogeneity of parameters for continuous outcomes. The second one, presented in the Supplementary Materials (Section 4) due to the space limitations, considers binary longitudinal outcomes.

We simulate eight longitudinal studies with four repeated measurements through the following linear models: $Y_{k,ij} = \beta^0_{k,0} + \beta^0_{k,1}X_{k,i} + \beta^0_{k,2}Z_{k,ij} + \epsilon_{k,ij}$, $j = 1, \ldots, 4$, $k = 1, \ldots, 8$, $i = 1, \ldots, n_k$, where $\boldsymbol{\beta}^0_k = (\beta^0_{k,0}, \beta^0_{k,1}, \beta^0_{k,2})^T$ is the vector of true regression parameters and the error term $\boldsymbol{\epsilon}_{k,i} = (\epsilon_{k,i1}, \ldots, \epsilon_{k,i4})^T$ follows $N\{0, \sigma_k\mathbf{R}_k(\alpha_k)\}$. Covariate $X_{k,i}$ is a baseline covariate generated from $N(0, 0.5^2)$. Covariate $Z_{k,i} = (Z_{k,i1}, \ldots, Z_{k,i4})^T$ is time-dependent and simulated from $N(0, 0.5^2)$. Covariance structures are set to mimic a situation where these eight studies recruit subjects from different subpopulations; set $\mathbf{R}_k(\cdot)$ for $k = 1, 4, 6, 7, 8$ as AR-1 and for $k = 2, 3, 5$ as compound symmetry (CS), with equal correlation $\alpha_k = 0.5, 1 \le k \le 8$. Set the variances $(\sigma_1, \ldots, \sigma_8)^T = (0.6, 1.5, 1.5, 0.6, 1.5, 0.6, 0.6, 0.6)^T$. We consider two cases of the underlying homogeneous parameters: all intercepts are always set at -1; the slope parameters for X are also set the same in both cases at $\boldsymbol{\beta}^0_1 = (2, 2, 3, 3, 3, 3, 3.3, 3.3)^T$; and the parameters for Z are set different as $\boldsymbol{\beta}^0_2 = (2.3, 2.3, 2, 2, 2, 2, 3, 3)^T$ for case I and $\boldsymbol{\beta}^0_2 = (2, 2, 2, 2, 2, 2, 3, 3)^T$ for case II. The former is slightly harder than the latter because case I contains more distinct parameter groups with smaller magnitude of pairwise differences.

Clearly, an exhaustive search requires to check a total of 56 hypotheses to determine the homogeneity clusters for both slope parameters. The intercepts are ignored here as they may be removed by centralizing the response variables. Here, we mainly focus on identifying homogenous parameter clusters across different studies, so no penalty is imposed on individual coefficients. Figure 1 displays two BIC curves computed from one randomly chosen simulated dataset, and their shapes appear to be quite representative to those obtained in our entire simulation study.

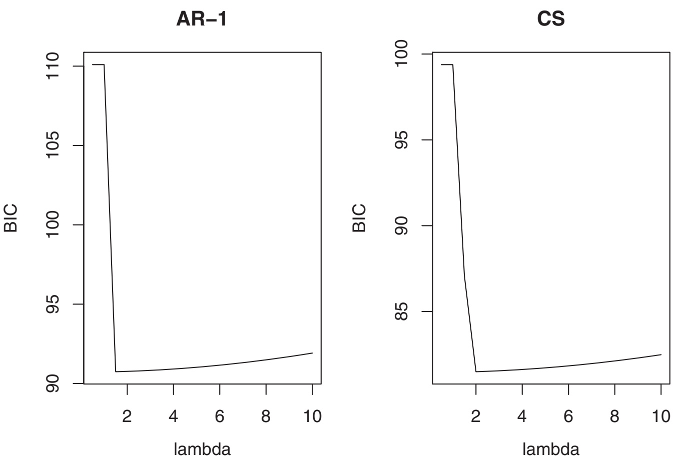

Figure 1. Bayesian information criterion (BIC) curves for the selection of tuning parameters in FLAPO under two different working correlations (AR-1 and Compound Symmetry (CS)).

To summarize the simulation results, we report results based on three criteria of sensitivity, specificity, and model size in Table 1 under two working correlation structures (i.e., AR-1 and CS) from 200 rounds of simulation. Sensitivity refers to the proportion of correctly identified equal coefficient pairs, while specificity refers to the proportion of correctly identified unequal coefficient pairs. Model size is the number of distinctive estimates in $\widehat{\boldsymbol{\beta}}_{(1)}$ and $\widehat{\boldsymbol{\beta}}_{(2)}$. The true numbers of parameter clusters are six for case I and five for case II, respectively.

Results in Table 1 provide us numerical evidence to compare $\widehat{\boldsymbol{\beta}}_{\widetilde{D}}$, $\widehat{\boldsymbol{\beta}}_{\widetilde{D}_e}$, and $\widehat{\boldsymbol{\beta}}_D$. Clearly, $\widehat{\boldsymbol{\beta}}_{\widetilde{D}}$ gives a better performance in terms of sensitivity and specificity than both $\widehat{\boldsymbol{\beta}}_{\widetilde{D}_e}$ and $\widehat{\boldsymbol{\beta}}_D$. But $\widehat{\boldsymbol{\beta}}_{\widetilde{D}_e}$, not $\widehat{\boldsymbol{\beta}}_{\widetilde{D}}$, is actually the method that is used in practice because the true parameter ordering is unknown. Focusing on the comparison between $\widehat{\boldsymbol{\beta}}_{\widetilde{D}_e}$ and $\widehat{\boldsymbol{\beta}}_D$, the former clearly outperforms the latter in both cases in terms of sensitivity, specificity, and model size. This example suggests that including redundant constraints in the regularization approach actually worsens the finite sample performance. This also provides the supporting evidence to Theorem A in the Supplementary Materials, which shows theoretically the FLAPO estimator has a smaller error bound than the full fused lasso estimator. The performance of $\widehat{\boldsymbol{\beta}}_{\widetilde{D}_e}$ is greatly improved if the size of between-pair differences is larger than 0.3, and a small difference of 0.3 or less considered in our simulation study presents a challenge for grouping. Also, Table 1 reveals that both sensitivity and specificity get improved along the increase of the sample size. In regard to the choice of working correlation structures, there is little effect observed on the performance of the methods. From the view of model

Table 1

Sensitivity (se100, se90), specificity (sp100, sp90), model size (size), and standard deviation of model size for the Case I and Case II in the simulation study I using different penalty matrices. Se100 and se90 represent the sensitivities computed based on 100% and 90% correct identification of all equal parameter pairs, respectively. Sp100 and sp90 are defined in the similar way but for unequal parameter pairs.

Case		AR-1			CS		
Penalty	n_k	Se100(Se90)	Sp100(Sp90)	Size(Std)	Se100(Se90)	Sp100(Sp90)	Size(Std)
I, $\widetilde{\mathbf{D}}$	100	0.57 (0.67)	0.65 (0.73)	6.19 (0.82)	0.63 (0.71)	0.59 (0.67)	6.01 (0.79)
	200	0.62 (0.68)	0.90 (0.94)	6.44 (0.76)	0.69 (0.75)	0.87 (0.91)	6.26 (0.67)
	400	0.79 (0.83)	0.99 (0.99)	6.25 (0.51)	0.83 (0.86)	0.99 (0.99)	6.19 (0.46)
I, $\widetilde{\mathbf{D}}_e$	100	0.42 (0.54)	0.26 (0.44)	6.00 (0.93)	0.43 (0.57)	0.26 (0.42)	5.95 (0.86)
	200	0.51 (0.62)	0.57 (0.74)	6.22 (0.77)	0.55 (0.64)	0.55 (0.72)	6.14 (0.76)
	400	0.60 (0.68)	0.84 (0.98)	6.36 (0.62)	0.62 (0.69)	0.83 (0.98)	6.30 (0.55)
I, $\mathbf{D}$	100	0.22 (0.57)	0.19 (0.28)	6.52 (1.12)	0.19 (0.57)	0.21 (0.32)	6.62 (1.17)
	200	0.27 (0.65)	0.54 (0.63)	6.89 (1.26)	0.31 (0.67)	0.50 (0.59)	6.76 (1.26)
	400	0.35 (0.71)	0.81 (0.90)	6.93 (1.14)	0.38 (0.74)	0.82 (0.90)	6.84 (1.05)
II, $\widetilde{\mathbf{D}}$	100	0.55 (0.62)	0.63 (0.71)	5.19 (0.80)	0.66 (0.71)	0.59 (0.67)	5.00 (0.78)
	200	0.62 (0.66)	0.90 (0.93)	5.41 (0.74)	0.72 (0.77)	0.86 (0.90)	5.21 (0.66)
	400	0.72 (0.77)	0.99 (0.99)	5.37 (0.68)	0.81 (0.83)	0.99 (0.99)	5.20 (0.47)
II, $\widetilde{\mathbf{D}}_e$	100	0.55 (0.64)	0.24 (0.40)	4.79 (0.79)	0.54 (0.66)	0.26 (0.42)	4.82 (0.79)
	200	0.56 (0.66)	0.56 (0.73)	5.14 (0.73)	0.58 (0.67)	0.55 (0.71)	5.08 (0.72)
	400	0.62 (0.68)	0.83 (0.98)	5.33 (0.61)	0.64 (0.69)	0.83 (0.98)	5.29 (0.56)
II, $\mathbf{D}$	100	0.37 (0.74)	0.18 (0.25)	5.02 (0.96)	0.37 (0.72)	0.19 (0.29)	5.06 (0.93)
	200	0.43 (0.78)	0.51 (0.59)	5.35 (1.06)	0.44 (0.78)	0.50 (0.58)	5.30 (1.02)
	400	0.46 (0.81)	0.78 (0.87)	5.56 (0.94)	0.44 (0.80)	0.80 (0.90)	5.59 (0.88)

size comparison, in general $\widehat{\boldsymbol{\beta}}_{\widetilde{D}}$ and $\widehat{\boldsymbol{\beta}}_{\widetilde{D}_e}$ can achieve better results in both smaller estimation bias and standard deviation. With no surprise, all three methods uniformly perform better in case II than in case I, due to the fact that case I has more complex parameter structures than case II.

6. Analysis of HIV Surveillance Cohort Data

We now apply the proposed regularization method to analyze the clustered dataset of the motivating example from the HIV surveillance project on injection drug users (IDUs). Refer to Section 1 for more details of the study background. To reduce heterogeneity within each primary sample unit (e.g., spousal correlation), we further divide IDUs within each sample unit into three groups according to martial status (single, marriage, and divorce). This division enables to simplify the analysis, and does not affect the estimation for the effect of needle sharing according to the finding of insignificant association between marital status and needle sharing (Wu et al., 1996). This results in 194 smaller but more homogeneous clusters of IDUs. The primary aim is to identify homogeneous groups of association parameters between behavioral activities and HIV positive across five regions. We fit the following marginal logistic model: $\text{logit}\{E(Y_{k,ij} \mid X_{k,i1}, X_{k,i2}, X_{k,i3}, X_{k,i4})\} = \beta_{k,0} + \beta_{k,1}X_{k,i1} + \beta_{k,2}X_{k,i2} + \beta_{k,3}X_{k,i3} + \beta_{k,4}X_{k,i4}$, where $Y_{k,ij}$ is a binary outcome of HIV positive for the jth subject in the ith cluster from region k, and covariates $X_{k,i1}$ to $X_{k,i4}$ are gender (1 for male, 0 for female), time (0–4 years), needle sharing (1 for yes, 0 otherwise), and syphilis (1 for yes, 0 otherwise). Region index k is coded as $1 = A, 2 = B, 3 = C, 4 = D, 5 = D$. All covariates are standardized.

First, the data are analyzed separately by region using the existing QIF method (Qu et al., 2000) under the compound symmetry correlation. These initial estimates are reported in the upper panel of Table 2, which are used to estimate the parameter ordering required in FLAPO. Second, we apply FLAPO to identify groups of common effects of needle sharing and syphilis reflective to relative effectiveness of regional policies for disease control and prevention, while the other parameters are treated as confounding and not considered for fusion. In particular, using different intercepts in the model allows to account for unequal regional HIV prevalence. Both BIC curves and solution paths of needle sharing and syphilis are showed in Figure 2. FLAPO estimates and confidence intervals for all four covariates are shown in the lower panel of Table 2 and in Figure 3. These estimates are yielded at the minimum BIC, $\lambda = 1.00625$. This chosen tuning parameter is used to construct the 95% confidence intervals in Tables 2 and 3 according to the inference described in Section 4.

The solution paths concerning the effects of needle sharing in Figure 2 indicate regions A and B share a common effect of needle sharing, which is slightly higher than that in region C and much higher than those in regions D and E. Using the p-values in Table 3, at the significance level by the Bonferroni correction for multiplicity $0.05/4 = 0.0125$, we detect three clusters of needle sharing effects on HIV positive, $\{A, B, C\}, \{D\}, \{E\}$. We fail to conclude any significant differential effects among three regions A, B, and C. Furthermore, we apply the standard meta analysis approach to combining

Table 2

Parameter estimates obtained by QIF from initial individual analyses and FLAPO and bias-corrected (BC) FLAPO using HIV data from five regions A–E. All covariates are standardized with mean 0 and variance 1.

Region	Estimate	A	B	C	D	E
Intercept	QIF	−3.072	−3.863	−2.660	−1.911	−0.741
Gender	QIF	0.128	−0.232	−0.058	0.164	0.325
Time	QIF	−0.237	0.337	0.172	−0.378	−0.132
Needle	QIF	0.788	0.927	0.628	0.023	0.308
Syphilis	QIF	0.186	0.041	0.157	−0.059	0.088
Intercept	FLAPO	−3.0858	−3.7881	−3.0469	−2.0400	−0.7607
	BC FLAPO	−3.0859	−3.7878	−3.0469	−2.0400	−0.7607
	95% CI	(−3.2472, −2.9246)	(−4.0790, −3.4966)	(−3.3784, −2.7154)	(−2.1313, −1.9488)	(−0.8871, −0.6343)
Gender	FLAPO	0.1578	−0.1921	−0.1184	0.2416	0.3492
	BC FLAPO	0.1577	−0.1920	−0.1184	0.2416	0.3492
	95% CI	(0.0291, 0.2863)	(−0.3697, −0.0143)	(−0.2136, −0.0233)	(0.0896, 0.3935)	(0.2775, 0.4210)
Time	FLAPO	−0.2408	0.1723	0.3280	−0.6222	−0.2120
	BC FLAPO	−0.2407	0.1724	0.3279	−0.6223	−0.2120
	95% CI	(−0.3795, −0.1019)	(−0.0801, 0.4250)	(0.0176, 0.6382)	(−0.7327, −0.5119)	(−0.3343, −0.0896)
Needle	FLAPO	0.7832	0.7832	0.6237	−0.0544	0.3247
	BC FLAPO	0.7834	0.7829	0.6237	−0.0545	0.3247
	95% CI	(0.6388, 0.9281)	(0.5505, 1.0153)	(0.4957, 0.7517)	(−0.1662, 0.0572)	(0.2433, 0.4061)
Syphilis	FLAPO	0.1707	0.0076	0.1707	−0.0319	0.0964
	BC FLAPO	0.1708	0.0071	0.1707	−0.0322	0.0965
	95% CI	(0.1151, 0.2264)	(−0.1644, 0.1785)	(0.0830, 0.2584)	(−0.1653, 0.1009)	(0.0222, 0.1708)

both these three estimated effects of regions A, B, and C, and their confidence interval listed in Table 2. We obtain the weighted estimate equal to 0.7067 and 95% confidence interval (0.6181, 0.7953) for $\{A, B, C\}$. Clearly, D is the only region at which we do not find a significant effect of needle sharing on HIV positive as its confidence interval contains 0.

A similar procedure is applied to examine the effects of syphilis. The p-values from Table 3 indicate an equal effect of syphilis on HIV positive across the five regions. By the standard meta analysis approach, we obtain a weighted estimate of the common effect as 0.1286, with a 95% CI (0.0915, 0.1658), which does not cover 0. So, there is a significant effect of syphilis on HIV positive, and this effect has a smaller magnitude than that of needle sharing in regions A, B, and C.

In summary, one interesting finding from this analysis is that in region D there was no significant effect of needle sharing on HIV positive, while risk of HIV positive among IDUs in regions A, B, and C was significantly associated with needle sharing. This provides useful information to the provincial CDC for a further investigation. Because of potential

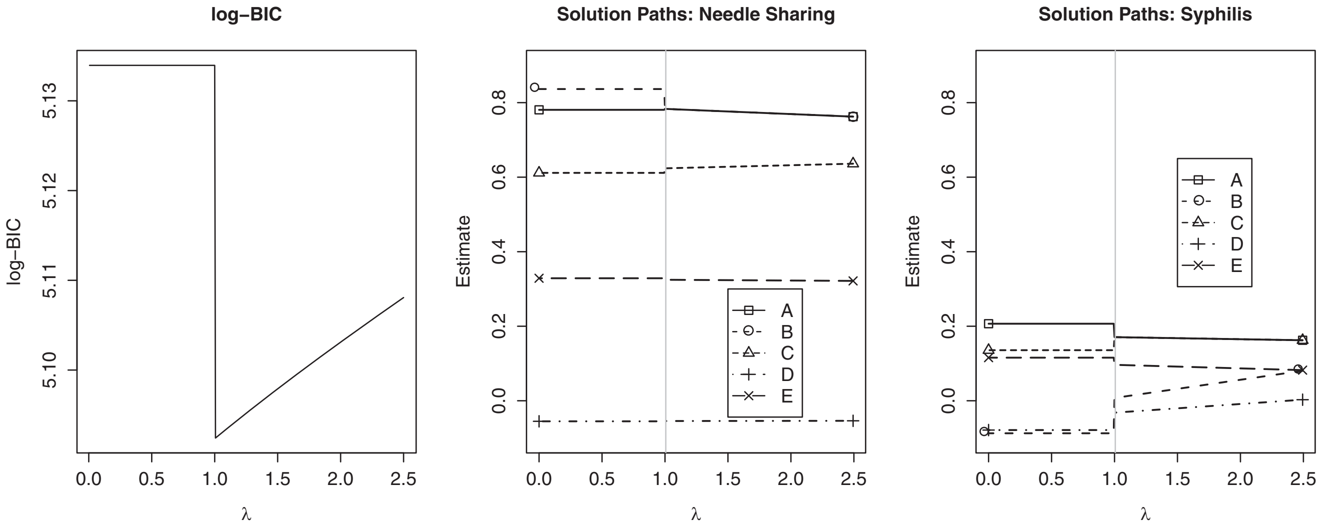

Figure 2. Bayesian information criterion and solution paths for the effects of needle sharing and syphilis on HIV positive in five regions A (square), B (circle), C(triangle), D(plus), and E(cross).

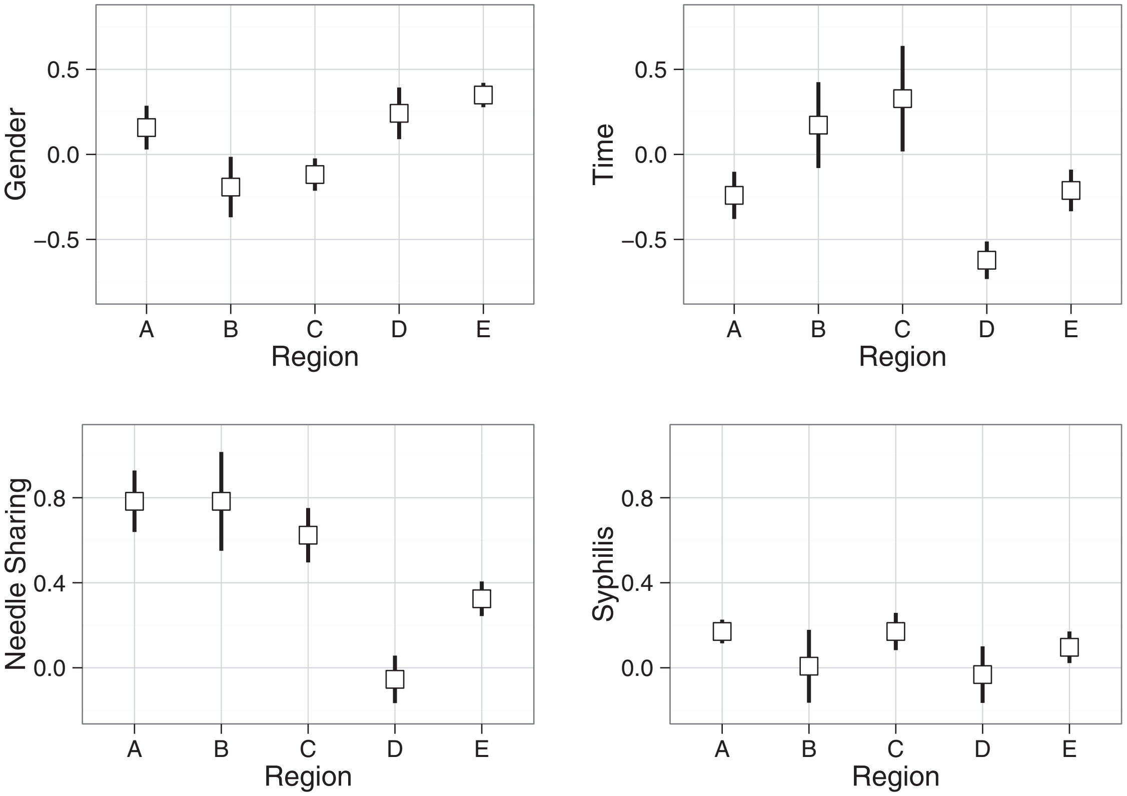

Figure 3. 95% confidence intervals of regression parameters in five regions A, B, C, D, and E, obtained by the bias-corrected FLAPO.

confounding in the data, the above findings should be interpreted with great caution and may not be generalizable to represent a general mechanism of risk prediction for HIV.

7. Concluding Remarks

In this article, we propose a new methodology of regularized estimation and inference to conduct statistical analysis for combined datasets of repeated measurements, including longitudinal data and clustered data as special cases. This method is developed to address the situation where the underlying parameter homogeneity cannot be analyzed by using the classical hypothesis testing procedures due to excessively high computing burden. The proposed method of *fused lasso with the adaptation of parameter ordering* (FLAPO) incorporates parameter ordering in the regularization procedures, so that the number of parameter constraints can be greatly reduced, leading to both improved computing speed and better finite-sample performance. The numerical examples further verify that the approach works well in terms of sensitivity and specificity as well as model size. However, the proposed method may be challenged by the increased computational complexity

Table 3

Estimates of parameter differences obtained by FLAPO and bias-corrected (BC) FLAPO, 95% *confidence intervals, and p-values obtained from the regularized inference*

		$\beta_{D,3}-\beta_{E,3}$	$\beta_{E,3}-\beta_{C,3}$	$\beta_{C,3}-\beta_{A,3}$	$\beta_{A,3}-\beta_{B,3}$
Needle	FLAPO	−0.3790	−0.2991	−0.1595	0.0000
	BC FLAPO	−0.3792	−0.2990	−0.1598	0.0005
	95% CI	(−0.5174, −0.2409)	(−0.4507, −0.1473)	(−0.3529, 0.0334)	(−0.2732, 0.2743)
	p-value	0.0000	0.0001	0.1050	0.9969
		$\beta_{D,4}-\beta_{B,4}$	$\beta_{B,4}-\beta_{E,4}$	$\beta_{E,4}-\beta_{C,4}$	$\beta_{C,4}-\beta_{A,4}$
Syphilis	FLAPO	−0.0395	−0.0889	−0.0743	0.0000
	BC FLAPO	−0.0393	−0.0894	−0.0742	−0.0001
	95% CI	(−0.2563, 0.1778)	(−0.2763, 0.0974)	(−0.1891, 0.0407)	(−0.1039, 0.1038)
	p-value	0.7230	0.3482	0.2058	0.9991

concerning the underlying pattern of homogeneous parameters. One of the limitations might be attributive to the inflexibility of bearing the variable selection only on a single tuning parameter in our method. In addition, when the number of parameters goes to infinity, it is not clear to us if it is possible to recover the true parameter ordering with probability 1. This problem deserves further improvement for handling data integration involving a large number of similar studies.

8. Supplementary Materials

Web supplementary sections 1–4 referenced in Sections 3–5 are available with this article at the *Biometrics* website on Wiley Online Library. It provides all technical detail, including regularity conditions, large-sample properties and finite-sample error bounds, as well as the proofs. Also, it includes some extra information of the algorithm and the results of the second simulation study. A file of the R code used in the simulation study is available online with the article.

Acknowledgements

The authors are grateful to the Co-Editor, the Associate Editor, and two anonymous reviewers for their very constructive comments that helped us improve the article greatly. The third author's research is supported in part by an NIH grant (R01ES024732) and an NSF(DMS1513595).

References

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics* **10**, 101–129.

Dunson, D. B. (2006). Bayesian dynamic modeling of latent trait distributions. *Biostatistics* **7**, 551–568.

Dunson, D. B., Xue, Y., and Carin, L. (2008). The matrix stick-breaking process: Flexible bayes meta-analysis. *Journal of the American Statistical Association* **103**, 317–327.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50**, 1029–1054.

Hartung, J., Knapp, G., and Sinha, B. K. (2008). *Statistical Meta-Analysis with Applications.* John Wiley and Sons, New Jersey.

Hedges, L. V. and Olkin, I. (1985). *Statistical Methods for Meta-Analysis.* New York: Academic Press.

Inoue, L. Y. T., Etzioni, R., Slate, E. H., Morrell, C., and Penson, D. F. (2004). Combining longitudinal studies of psa. *Biostatistics* **5**, 483–500.

Ke, T., Fan, J., and Wu, Y. (2015). Homogeneity pursuit. *Journal of the American Statistical Association* **110**, 175–194.

Kim, S.-J., Koh, K., Boyd, S. P., and Gorinevsky, D. M. (2009). l_1 trend filtering. *SIAM Review* **51**, 339–360.

Müller, P., Quintana, F., and Rosner, G. (2004). A method for combining inference across related nonparametric Bayesian models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **66**, 735–749.

Qu, A., Lindsay, B. G., and Li, B. (2000). Improving generalised estimating equations using quadratic inference functions. *Biometrika* **87**, 823–836.

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

Smith, T. C., Spiegelhalter, D. J., and Thomas, A. (1995). Bayesian approaches to random-effects meta-analysis: A comparative study. *Statistics in Medicine* **14**, 2685–2699.

Song, P. X.-K. X., Jiang, Z., Park, E., and Qu, A. (2009). Quadratic inference functions in marginal models for longitudinal data. *Statistics in Medicine* **28**, 3683–3696.

Sun, X., Wang, N., Li, D., Zheng, X., Qu, S., Wang, L., et al. (2007). The development of HIV/AIDS surveillance in China. *AIDS* **21**, 33–38.

Thase, M. E., Kornstein, S. G., Germain, J.-M., Jiang, Q., Guico-Pabia, C., and Ninan, P. T. (2009). An integrated analysis of the efficacy of desvenlafaxine compared with placebo in patients with major depressive disorder. *CNS Spectrums* **14**, 144–154.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society Series B* **67**, 91–108.

Ueki, M. (2009). A note on automatic variable selection using smooth-threshold estimating equations. *Biometrika* **96**, 1005–1011.

Ueki, M. and Kawasaki, Y. (2011). Automatic grouping using smooth-threshold estimating equations. *Electronic Journal of Statistics* **5**, 309–328.

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics* **42**, 1166–1202.

Wang, F., Wang, L., and Song, P. X. K. (2012). Quadratic inference function approach to merging longitudinal studies: Validation and joint estimation. *Biometrika* **99**, 755–762.

Wang, H., Li, B., and Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society Series B* **71**, 671–683.

Wang, H., Li, R., and Tsai, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94**, 553–568.

Wu, Z., Detels, R., Zhang, J., Duan, S., Cheng, H., Li, Z., et al. (1996). Risk factors for intravenous drug use and sharing equipment among young male drug users in longchuan county, south-west China. *AIDS* **10**, 1017–1024.

Zeger, S. L. and Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42**, 121–130.

Zhang, Z., Chen, D., and Fenstermacher, D. A. (2007). Integrated analysis of independent gene expression microarray datasets improves the predictability of breast cancer outcome. *BMC Genomics* **8**, 331.

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

Received November 2014. *Revised November* 2015.
Accepted December 2015.