Supplementary materials for this article are available at https://doi.org/10.1007/s13253-023-00529-2.

Fusion Learning of Functional Linear Regression with Application to Genotype-by-Environment Interaction Studies

Shan YU, Aaron M. KUSMEC, Li WANG, and Dan NETTLETON

We propose a sparse multi-group functional linear regression model to simultaneously estimate multiple coefficient functions and identify groups, such that coefficient functions are identical within groups and distinct across groups. By borrowing information from relevant subgroups of subjects, our method enhances estimation efficiency while preserving heterogeneity in model parameters and coefficient functions. We use an adaptive fused lasso penalty to shrink coefficient estimates to a common value within each group. We also establish theoretical properties of the proposed estimators. To enhance computation efficiency and incorporate neighborhood information, we propose to use graph-constrained adaptive lasso with a computationally efficient algorithm. Two Monte Carlo simulation studies have been conducted to study the finite-sample performance of the proposed method. The proposed method is applied to sorghum flowering-time data and hybrid maize grain yields from the Genomes to Fields consortium.

Supplementary materials accompanying this paper appear online.

Key Words: Graph-constrained lasso; Heterogeneity; Spline approximation; Time-varying coefficient.

1. INTRODUCTION

Genotype-by-environment (G × E) interaction analysis can reveal interplay between genetic and environmental factors underlying observed phenotypes. The topic of G × E has a long history in genetics and statistics and has garnered much attention recently (Li et al. 2018). G × E interaction is essential to the development of breeding strategies, which seek to customize genotypes to the characteristics of particular environments (Nelsen 2002).

S. Yu (✉), Department of Statistics, University of Virginia, Halsey Hall, P.O. Box 400135, Charlottesville, VA 22904-4135, USA (E-mail: *sy5jx@virginia.edu*). A. M. Kusmec, Department of Agronomy, Iowa State University, 2035 Carver Co-Lab, 1111 WOI Rd, Ames, IA 50011-1085, USA. L. Wang, Department of Statistics, George Mason University, 1721 Nguyen Engineering Building, 4400 University Drive, MS 4A7 , Fairfax, VA 22030, USA. D. Nettleton, Department of Statistics, Iowa State University, 1121H Snedecor Hall, 2438 Osborn Drive , Ames, IA 50011-1090, USA.

Journal of Agricultural, Biological, and Environmental Statistics, Volume 28, Number 3, Pages 401–422
https://doi.org/10.1007/s13253-023-00529-2

Grouping genotypes according to their responses to environmental characteristics is of scientific interest.

Modern technologies enable data collection to be essentially continuous over time. In precision agriculture, researchers record weather data at small time increments from real-time local weather station and wireless sensors for environmental monitoring. The Genomes to Fields initiative aims to accurately predict of the phenotypes of corn plants in diverse environments (McFarland et al. 2020) by using daily recorded weather data from various devices. Such data, recorded daily or more frequently, are usually treated as time-indexed functional predictors. They are valuable because understanding how environmental characteristics influence plant growth during various stages of development provides farmers with a way of optimizing agricultural practices, e.g., the time of irrigation, fertilization, harvesting, etc.

This paper considers multi-group functional linear models (FLMs) to describe the relationship between functional predictors and a scalar response (Ogden et al. 2002). Define the subgroup k as a set of subjects (e.g., plants or plots) from the same analytic unit (e.g., genotype), $k = 1, \ldots, K$. For the ith subject within subgroup k, $i = 1, \ldots, n_k$, let Y_{ik} be a response variable, such as yield. Let $\{\mathbf{X}_{ik}(t), t \in \mathcal{T}\} = \{X_{ik\ell}(t), t \in \mathcal{T}\}_{\ell=1}^{L}$ be a vector of square-integrable random functions defined on a closed interval $\mathcal{T}$, for example, time-indexed weather data, such as temperature, soil moisture, etc. Let ℓ be the index of the ℓth functional predictor. The unknown coefficient functions $\alpha_{k\ell}$ characterize the associations between the scalar response variable and functional predictors by:

$$Y_{ik} = \beta_k + \sum_{\ell=1}^{L} \int_{\mathcal{T}} X_{ik\ell}(t)\alpha_{k\ell}(t)dt + \varepsilon_{ik}, \quad i = 1, \ldots, n_k, \; k = 1, \ldots, K, \tag{1}$$

where β_k's are unknown intercepts, $\alpha_{k\ell}$'s are unknown functions defined on $\mathcal{T}$, and the ε_{ik} terms are independent random errors with zero means and variances bounded above by a finite constant. This model assumes that the intercepts and coefficient functions for environmental variables are subgroup-dependent. For each unknown model component (the intercepts being one component, the coefficient functions on the first environmental variable being another, etc.), subgroups can be partitioned into unknown groups, where the values of the corresponding component are identical for subgroups in the same group but different across groups. The proposed model allows different model components to have different group structures. The G × E effect is described through genotype-dependent coefficient functions, representing different response patterns to the environmental changes for various genotypes. This paper aims to estimate the FLMs with heterogeneous coefficient functions and automatically identify the underlying group structure.

Functional linear models (FLMs) have been widely used to describe the relationship between functional predictors and a scalar response; see, for example, Cardot et al. (2003), Giraldo et al. (2010), Wang et al. (2016), Lin et al. (2017) and Guan et al. (2020). However, the standard approaches in the above FLMs do not explicitly consider multi-group functional data with subgroups that may have different coefficient functions for one or more functional predictors. In many applications, the available sample size in each subgroup greatly varies; therefore, even if one could fit an FLM for each subgroup separately, the estimator of

coefficient functions in an FLM based on a small number of subjects is often inefficient. To address this challenge, instead of fitting FLMs separately to each subgroup, we propose a novel fusion learning approach for a multi-group FLM, strategically combining similar subgroups to enhance estimation efficiency while capturing G × E interaction by allowing heterogeneous subgroup-specific coefficient functions. Several works have been conducted to apply fusion learning to a suite of statistical models, including linear regression (Ma and Huang 2017; Tang and Song 2016), longitudinal data analysis (Zhu and Qu 2018), and confidence distribution (Shen et al. 2020). In this paper, we consider fusion learning for functional data analysis. This approach is appealing for several reasons: (1) Fusion learning combines information to improve estimation efficiency; (2) It preserves parameter heterogeneity and can be used as a convenient tool to study G × E interaction; (3) By merging similar subgroups, it reduces the estimation variance and enhances the prediction performance.

Specifically, we utilize regularized likelihood methods to simultaneously identifying grouping structure and estimating intercepts and coefficient functions. First, we can use spline functions to approximate the coefficient functions and incorporate structural constraints on the coefficient functions through their spline coefficients; see Cardot et al. (2003), Wang et al. (2011) and Xue et al. (2020). Second, we adopt pairwise fused penalty as the regularization term to shrink similar estimators. In this paper, we consider the adaptive lasso (Zou 2006) as the penalty function, which facilities efficient numerical optimization using convex optimization methods. We use an Alternating Direction Method of Multipliers (ADMM) algorithm in our optimization procedure.

In addition to considering a pairwise fused penalty, we implement a graph-constrained adaptive lasso, which not only reduces the computation burden but also incorporates prior information about the underlying group structure. The graph-constrained lasso has wide applications in the literature. These include works on tree-based fused lasso in nonparametric regression (Tibshirani et al. 2005) and spatial data analysis (Li and Sang 2019), K-nearest-neighbors graphs in multivariate nonparametric regression (Madrid Padilla et al. 2020), and genetic networks in variable selection for genomic studies (Li and Li 2010). In this paper, we apply graph-constrained adaptive lasso in our model fitting, where each subgroup is treated as a node, graph edges are determined by genetic similarity, and the edges determine which pairwise differences in estimation are penalized. In addition, we propose a two-step penalization procedure that improves estimation and prediction performance if the chosen penalty graph is missing relevant edges.

The rest of the paper is structured as follows. Section 2 describes our model, then briefly reviews univariate splines and introduces the penalized adaptive fused lasso estimation method and the graph-constrained adaptive fused lasso estimator. Section 3 provides the implementation details of the proposed methods. Section 4 provides the theoretical properties of the proposed estimators. In Section 5, we conduct simulation studies to evaluate the finite-sample performance of the proposed method. Section 6 illustrates the newly proposed method using two real datasets. Some concluding remarks are given in Section 7. The proof of the asymptotic results and additional simulation studies are given in the online Supplementary Material.

2. METHODOLOGY

2.1. MULTI-GROUP FLMS

Consider the FLM (1) defined in Sect. 1. For $\ell = 0, \ldots, L$, let $\mathbb{G}_\ell = \{\mathcal{G}_{\ell,1}, \ldots, \mathcal{G}_{\ell,M_\ell}\}$ be a partition of $\{1, \ldots, K\}$, which represents the unknown underlying group structure. For group $\mathcal{G}_{0,m}$, if $k \in \mathcal{G}_{0,m}$, then $\beta_k = \beta_m^{\mathrm{G}}$, where β_m^{G} is the common value for all the β_k's from group $\mathcal{G}_{0,m}$. For group $\mathcal{G}_{\ell,m},\ \ell = 1, \ldots, L$, if $k \in \mathcal{G}_{\ell,m}$, then $\alpha_{k\ell} = \alpha_{m\ell}^{\mathrm{G}}$, where $\alpha_{m\ell}^{\mathrm{G}}$ is the common value for all the $\alpha_{k\ell}$'s from group $\mathcal{G}_{\ell,m}$.

As a simple example, suppose we have $K = 3$ genotypes and $L = 1$ time-indexed functional predictor. One possible group structure would be given by $\mathbb{G}_0 = \{\mathcal{G}_{0,1}, \mathcal{G}_{0,2}\}$ and $\mathbb{G}_1 = \{\mathcal{G}_{1,1}, \mathcal{G}_{1,2}\}$, where $\mathcal{G}_{0,1} = \{1, 2\}, \mathcal{G}_{0,2} = \{3\}, \mathcal{G}_{1,1} = \{1, 3\}$, and $\mathcal{G}_{1,2} = \{2\}$. This structure indicates that genotypes 1 and 2 have a common intercept that differs from the intercept for genotype 3 (i.e., $\beta_1 = \beta_2 \neq \beta_3$) and that genotypes 1 and 3 share a functional coefficient that differs from that of genotype 2 (i.e., $\alpha_{11} = \alpha_{31} \neq \alpha_{21}$).

Our goal is to estimate intercepts and coefficient functions and simultaneously identify distinct group patterns for intercepts and all coefficient functions. If, for a given time-indexed covariates, multiple genotypes are in the same group, we can infer that their response to the environmental factors measured by the covariate is the same (i.e., no G × E for those genotypes and that environmental factor). We can also borrow or share information across those genotypes to improve estimation of their shared coefficient function.

2.2. FLMS WITH ADAPTIVE FUSED LASSO PENALTY

In this project, we approximate the unknown coefficient functions via univariate polynomial splines for their simplicity in computation and automatically identify the homogeneous subgroups based on an adaptive pairwise fusion penalty.

We start with a brief review of univariate splines. Let $\boldsymbol{\upsilon}$ be a partition of the interval $\mathcal{T} = [a, b]$ with N interior knots, where $\boldsymbol{\upsilon} = \{a = \upsilon_0 < \upsilon_1 < \cdots < \upsilon_N < \upsilon_{N+1} = b\}$. The polynomial splines of order $\varrho + 1$ are polynomial functions with ϱ-degree (or less) on intervals $[\upsilon_j, \upsilon_{j+1}),\ j = 0, \ldots, N$ and $\left[\upsilon_N, \upsilon_{N+1}\right]$, and have $\varrho - 1$ continuous derivatives globally. Suppose that $\alpha_{k\ell}$'s can be approximated well by a spline function: $\alpha_{k\ell}(t) \approx \sum_{j=1}^{q} \gamma_{k\ell j} U_j(t) = \mathbf{U}(t)^\top \boldsymbol{\gamma}_{k\ell}$, where $q = N+\varrho+1, \mathbf{U}(t) = \{U_1(t), \ldots, U_q(t)\}^\top$ is a vector of the B-spline basis functions with degree ϱ and partition $\boldsymbol{\upsilon}$, and $\boldsymbol{\gamma}_{k\ell} = \left(\gamma_{k\ell 1}, \ldots, \gamma_{k\ell q}\right)^\top$ is a vector of coefficients. Through spline approximation, the estimation of coefficient functions has been simplified to obtain spline coefficients. Then, the subgroup-specific least squares estimator (SSLS) is defined as

$$
\begin{aligned}
(\widehat{\beta}_k^{\mathrm{SSLS}}, \widehat{\boldsymbol{\gamma}}_{k\ell}^{\mathrm{SSLS}}) = & \underset{\substack{\beta_k \in R \\ \boldsymbol{\gamma}_{k\ell} \in R^q, 1 \le \ell \le L}}{\arg\min} \sum_{i=1}^{n_k} \left(Y_{ik} - \beta_k - \sum_{\ell=1}^{L} \mathbf{Z}_{ik\ell}^\top \boldsymbol{\gamma}_{k\ell} \right)^2 + \sum_{\ell=1}^{L} \lambda_\ell^{\mathrm{R}} \boldsymbol{\gamma}_{k\ell}^\top \mathbf{P} \boldsymbol{\gamma}_{k\ell}, \\
\widehat{\alpha}_{k\ell}^{\mathrm{SSLS}} = & \mathbf{U}(t)^\top \widehat{\boldsymbol{\gamma}}_{k\ell}^{\mathrm{SSLS}},
\end{aligned} \tag{2}
$$

where $\mathbf{Z}_{ik\ell} = (Z_{ik\ell 1}, \ldots, Z_{ik\ell q})^\top$, $Z_{ik\ell j} = \int_{\mathcal{T}} X_{ik\ell}(t)U_j(t)dt$, $\mathbf{P}$ is a roughness penalty matrix satisfying $\boldsymbol{\gamma}_{k\ell}^\top \mathbf{P}\boldsymbol{\gamma}_{k\ell} = \int_{\mathcal{T}}[\partial^2\{\mathbf{U}(t)^\top\boldsymbol{\gamma}_{k\ell}\}/\partial t^2]^2 dt$, and $\lambda_\ell^{\mathrm{R}}$ is the roughness tuning parameter. Note that when the number of subjects of each subgroup n_k is small, the SSLS estimator can be inefficient.

To combine subgroups with similar intercept terms and coefficient functions, we propose a pairwise fusion penalized least squares approach where the penalty terms tend to shrink the difference between subgroups. The penalty function is critical to the clustering results. Similar to the variable selection situation where the L_1 penalty gives biased estimates for large coefficients, the estimates based on the lasso penalty tend to underestimate between-group differences. This leads to biased estimates and unreliable clustering. Hence, penalties like the adaptive lasso or SCAD (Fan and Li 2001), which can produce less biased estimates, are more appealing.

In this paper, we focus on the adaptive lasso penalty as its convex property guarantees the existence of a global optimizer and a faster convergence rate of the ADMM algorithm. Then, the estimator of model (1) is given by minimizing the following regularized objective function:

$$\frac{1}{2}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\left(Y_{ik} - \beta_k - \sum_{\ell=1}^{L}\mathbf{Z}_{ik\ell}^\top\boldsymbol{\gamma}_{k\ell}\right)^2 + \sum_{k=1}^{K}\sum_{\ell=1}^{L}\lambda_\ell^{\mathrm{R}}\boldsymbol{\gamma}_{k\ell}^\top\mathbf{P}\boldsymbol{\gamma}_{k\ell}$$
$$+ \sum_{1\le k<k'\le K}\lambda_0^{\mathrm{F}}\omega_{0,kk'}|\beta_k - \beta_{k'}| + \sum_{\ell=1}^{L}\sum_{1\le k<k'\le K}\lambda_\ell^{\mathrm{F}}\omega_{\ell,kk'}\|\boldsymbol{\gamma}_{k\ell} - \boldsymbol{\gamma}_{k'\ell}\|, \tag{3}$$

where $\lambda_1^{\mathrm{R}}, \ldots, \lambda_L^{\mathrm{R}}$ are the tuning parameters for roughness penalty, $\lambda_0^{\mathrm{F}}, \ldots, \lambda_L^{\mathrm{F}}$ are the tuning parameters for fused lasso, and $\omega_{\ell,kk'}$, $0 \le \ell \le L$ are adaptive weights. The adaptive penalty tends to shrink some of the pairs $\beta_k - \beta_{k'}$ and $\boldsymbol{\gamma}_{k\ell} - \boldsymbol{\gamma}_{k'\ell}$ to zero. Combining similar subgroups can be regarded as clustering analysis of the subgroups based on their response patterns to changes in functional predictors. Based on this, we can arrange the subgroups into clusters. We consider $\omega_{0,kk'} \asymp |\bar{\beta}_k - \bar{\beta}_{k'}|^{-\tau}$, $w_{\ell,kk'} \asymp \|\bar{\boldsymbol{\gamma}}_{k\ell} - \bar{\boldsymbol{\gamma}}_{k'\ell}\|^{-\tau}$, $1 \le \ell \le L$, $\bar{\beta}_k$'s and $\bar{\boldsymbol{\gamma}}_{k\ell}$'s are some consistent initial estimators of β_k and $\boldsymbol{\gamma}_{k\ell}$, and τ is a tuning parameter. In Section 3.3, we discuss how to obtain the initial estimators $\bar{\beta}_k$ and $\bar{\boldsymbol{\gamma}}_{k\ell}$, the choice of τ.

Let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)^\top$, $\boldsymbol{\gamma}_\ell = (\boldsymbol{\gamma}_{1\ell}^\top, \ldots, \boldsymbol{\gamma}_{K\ell}^\top)^\top$, $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^\top, \ldots, \boldsymbol{\gamma}_L^\top)^\top$, and $\boldsymbol{\omega}_0 = (\omega_{0,kk'}, 1 \le k, k' \le K)^\top$, and $\boldsymbol{\omega}_\ell = (\omega_{\ell,kk'}, 1 \le k, k' \le K)^\top$. Next, let $\mathbf{Y} = (\mathbf{Y}_1^\top, \ldots, \mathbf{Y}_K^\top)^\top$, where $\mathbf{Y}_k = (Y_{1k}, Y_{2k}, \ldots, Y_{n_k k})^\top$. Let $\mathbf{Z}_0$ and $\mathbf{Z}_\ell$ be the block diagonal matrix with block matrices $\mathbf{1}_k$ and $\mathbf{Z}_{k\ell}$, $1 \le k \le K$, $1 \le \ell \le L$, respectively, where $\mathbf{1}_k$ is a vector of ones of length n_k and $\mathbf{Z}_{k\ell} = (\mathbf{Z}_{1k\ell}, \ldots, \mathbf{Z}_{n_k k\ell})^\top$. Let $\boldsymbol{h}_{0,kk'}$ be a vector of length K such that $\beta_k - \beta_{k'} = \boldsymbol{h}_{0,kk'}^\top\boldsymbol{\beta}$. The vector $\boldsymbol{h}_{0,kk'}$ contains two nonzero entries, 1 at the kth element and -1 at the k'th element. Denote $\mathbf{H}_0 = \{\boldsymbol{h}_{0,kk'}, k \ne k'\}^\top$ as a $(K-1)K/2 \times K$ matrix. Then, the objective function $L^P(\boldsymbol{\beta}, \boldsymbol{\gamma})$ is

$$\frac{1}{2}\Big\|\mathbf{Y}-\mathbf{Z}_0\boldsymbol{\beta}-\sum_{\ell=1}^{L}\mathbf{Z}_\ell\boldsymbol{\gamma}_\ell\Big\|^2+\sum_{\ell=1}^{L}\lambda_\ell^{\mathrm{R}}\boldsymbol{\gamma}_\ell^\top(\mathbf{I}_K\otimes\mathbf{P})\boldsymbol{\gamma}_\ell$$
$$+\lambda_0^{\mathrm{F}}\|\mathbf{H}_0\boldsymbol{\beta}\|_{1,\boldsymbol{\omega}_0}+\sum_{\ell=1}^{L}\lambda_\ell^{\mathrm{F}}\|(\mathbf{H}_0\otimes\mathbf{I}_q)\boldsymbol{\gamma}_\ell\|_{1,\boldsymbol{\omega}_\ell}^q, \tag{4}$$

where $\|\cdot\|$ is L_2 norm of a vector, $\|\cdot\|_{1,\boldsymbol{\omega}_0}$ is L_1 norm with adaptive weights $\boldsymbol{\omega}_0$, and $\|\cdot\|_{1,\boldsymbol{\omega}_\ell}^q$ be a mixed-norm with adaptive weights $\boldsymbol{\omega}_\ell$. Let a vector $\boldsymbol{a}$ be partitioned into the sub-vectors with length q and its mixed-norm $\|\boldsymbol{a}\|_{1,\boldsymbol{\omega}_\ell}^q$ is equal to the summation of L_2 norms of sub-vectors multiplied by the corresponding weights. Now, we denote the fusion function linear model estimator by $(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\gamma}}_1,\ldots,\widehat{\boldsymbol{\gamma}}_L)$, which is the minimizer of objective function (4). Subsequently, the estimators of coefficient functions are $\widehat{\alpha}_{k\ell}(t)=\mathbf{U}(t)^\top\widehat{\boldsymbol{\gamma}}_{k\ell}$. Subgroups with same intercepts or coefficients form one cluster. The estimated group structures are denoted as $\widehat{\mathbb{G}}_\ell=\{\widehat{\mathcal{G}}_{\ell,1}\ldots,\widehat{\mathcal{G}}_{\ell,\widehat{M}_\ell}\}$, $\ell=0,\ldots,L$.

2.3. GRAPH-CONSTRAINED ADAPTIVE FUSED LASSO (FGAFL)

Graphs and networks are common ways to describe biological information (Li and Li 2010). People use graphs to represent many different biological processes, such as regulatory networks, gene co-expression network and between-/within- species interaction networks. A prior use of such graphs can be a useful supplement to empirical data. With regard to the computation, in the objective function (3), we consider pairwise differences for each subgroups. There are $(L+1)\binom{K}{2}$ penalty terms, and the total number of terms increases by an order of $O(K^2)$, which imposes challenges in computation. In our application, the original maize yields data contain thousands of hybrids, which generates on the order of 10^6 penalty terms. To make use of the prior information and tackle computational challenges, we consider a graph-constrained adaptive fused lasso method.

Consider an undirected graph G with vertices $\{1,\ldots,K\}$ representing the K subgroups in the data. Let $e(G)$ be the set of edges in the graph G, where each edge connects two vertices k and k'. For example, in our application, the graph represents genetic relationships among genotypes, where vertices are the genotypes, and an edge between genotypes k and k' on the graph indicates genetic similarity. We incorporate such information through graph-constrained fused lasso. When there is an edge connecting k and k', we add the corresponding penalty term. Otherwise, there is no penalty term between two subgroups. The basic intuition is that subgroups with an edge connecting them are more likely to be placed into the same cluster. In our application, this penalty implies that genetically similar subgroups are encouraged to have similar, or even identical, parameter estimates. The objective function in (3) can be generalized as

$$\frac{1}{2}\Big\|\mathbf{Y}-\mathbf{Z}_0\boldsymbol{\beta}-\sum_{\ell=1}^{L}\mathbf{Z}_\ell\boldsymbol{\gamma}_\ell\Big\|^2+\sum_{\ell=1}^{L}\lambda_\ell^{\mathrm{R}}\boldsymbol{\gamma}_\ell^\top(\mathbf{I}_K\otimes\mathbf{P})\boldsymbol{\gamma}_\ell$$
$$+\lambda_0^{\mathrm{F}}\|\mathbf{H}_G\boldsymbol{\beta}\|_{1,\boldsymbol{\omega}_0}+\sum_{\ell=1}^{L}\lambda_\ell^{\mathrm{F}}\|(\mathbf{H}_G\otimes\mathbf{I}_q)\boldsymbol{\gamma}_\ell\|_{1,\boldsymbol{\omega}_\ell}^q, \tag{5}$$

where $\mathbf{H}_G = \{\boldsymbol{h}_{0,kk'}, (k,k') \in e(G)\}^\top$ is a $|e(G)| \times K$ matrix constructed from the graph G. In this paper, we call the minimizer of (5) the *f*unctional *G*raph-constrained *A*daptive *F*used *L*asso (fGAFL) estimator. When the graph G is a minimal spanning tree, for example, there are at most $K-1$ edges, which reduces the number of penalty terms from $O(K^2)$ to $O(K)$ and dramatically reduces the computation burden. When the graph G is a complete graph, the objective function (5) is equivalent to the objective function (3). Thus, the estimator proposed in Sect. 2.2 is a special case of the fGAFL estimators, and we call it the fCGAFL estimators. The fGAFL and fCGAFL methods use the same approaches to solve objective functions and select proper hyperparameters.

3. IMPLEMENTATION

3.1. TUNING PARAMETERS FOR FGAFL

This section states the algorithm we use to obtain the fGAFL estimators. Selecting suitable values of tuning parameters is important to good model fitting. By selecting appropriate fusion tuning parameters $\lambda_\ell^{\mathrm{F}}$'s, the proper number of clusters for each functional predictor or intercept term can be determined. For fixed tuning parameters, we can identify pairs of subgroups with the same coefficient functions/intercepts. Then, the number of distinct coefficient functions/intercepts across subgroups determines the number of clusters. In this paper, we employ the Bayesian information criterion (BIC) to select appropriate tuning parameter values. Algorithm 1 in Sect. S.1.2 in the Supplementary Material introduces the overall model fitting structure. Given a set of candidate tuning parameters, we use the ADMM to fit the models and obtain the corresponding solution path. The selection criterion is defined as $\mathrm{BIC} = \log(\mathrm{RSS}) + n^{-1}\log(n)\widehat{df}$, where RSS is the residual sum of squares, and $\widehat{df}$ is the degrees of freedom of the fitted model; see Sect. S.1.1 in the Supplementary Materials for the estimation of the degrees of freedom.

3.2. ADMM ALGORITHM FOR FGAFL

Studies in Hallac et al. (2015) developed an algorithm based on ADMM to solve the network lasso optimization problem in a distributed and scalable manner, which allows for guaranteed global convergence even on large graphs. In this paper, we adopt the results in Hallac et al. (2015) and use ADMM to solve (5). In the following, we describe the details of the ADMM algorithm. First of all, we re-frame the objective function into a typical ADMM problem. Minimizing objective function (5) is equivalent to minimizing

$$
\begin{aligned}
S(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\eta}) &= \frac{1}{2}\Big\|\mathbf{Y} - \mathbf{Z}_0\boldsymbol{\beta} - \sum_{\ell=1}^{L}\mathbf{Z}_\ell\boldsymbol{\gamma}_\ell\Big\|^2 + \sum_{\ell=1}^{L}\lambda_\ell^{\mathrm{R}}\boldsymbol{\gamma}_\ell^\top((\mathbf{I}_K\otimes\mathbf{P})\boldsymbol{\gamma}_\ell \\
&+ \lambda_0^{\mathrm{F}}\|\boldsymbol{\eta}_0\|_{1,\boldsymbol{\omega}_0} + \sum_{\ell=1}^{L}\lambda_\ell^{\mathrm{F}}\|\boldsymbol{\eta}_\ell\|_{1,\boldsymbol{\omega}_\ell}^q, \\
&\text{subject to } \boldsymbol{\eta}_0 = \mathbf{H}_G\boldsymbol{\beta},\ \boldsymbol{\eta}_\ell = (\mathbf{H}_G\otimes\mathbf{I}_q)\boldsymbol{\gamma}_\ell, \text{ for } \ell = 1,\ldots,L, \qquad (6)
\end{aligned}
$$

where $\boldsymbol{\eta}_{\ell,kk'}$ is a q vector and $\boldsymbol{\eta} = (\boldsymbol{\eta}_0^\top, \ldots, \boldsymbol{\eta}_L^\top)^\top$ with $\boldsymbol{\eta}_0 = \{(\eta_{0,kk'}, (k,k') \in e(G)\}^\top$ and $\boldsymbol{\eta}_\ell = \{(\boldsymbol{\eta}_{\ell,kk'}^\top, (k,k') \in e(G)\}^\top, \ell = 1, \ldots, L$. By the augmented Lagrangian method, (6) can be solved by minimizing

$$L(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\nu}) = S(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\eta}) + \boldsymbol{\nu}_0^\top (\mathbf{H}_G \boldsymbol{\beta} - \boldsymbol{\eta}_0) + \sum_{\ell=1}^{L} \boldsymbol{\nu}_\ell^\top \{(\mathbf{H}_G \otimes \mathbf{I}_q)\boldsymbol{\gamma}_\ell - \boldsymbol{\eta}_\ell\} + \frac{\theta}{2} \left\{ \|\mathbf{H}_G \boldsymbol{\beta} - \boldsymbol{\eta}_0\|^2 + \sum_{\ell=1}^{L} \|(\mathbf{H}_G \otimes \mathbf{I}_q)\boldsymbol{\gamma}_\ell - \boldsymbol{\eta}_\ell\|^2 \right\}, \tag{7}$$

where $\boldsymbol{\nu} = (\boldsymbol{\nu}_0^\top, \boldsymbol{\nu}_1^\top, \ldots, \boldsymbol{\nu}_L^\top)^\top$ with $\boldsymbol{\nu}_0 = \{\nu_{0,kk'}, (k,k') \in e(G)\}^\top$ and $\boldsymbol{\nu}_\ell = \{\boldsymbol{\nu}_{\ell,kk'}^\top, (k,k') \in e(G)\}^\top$ and all the elements of $\boldsymbol{\nu}$ are larger than zero. We take the update step size $\theta = 1$. See Section S.1.3 in the Supplementary Material for details of the ADMM algorithm.

3.3. Weights and Penalty Graph G for fGAFL

The weights of adaptive lasso $\omega_{\ell,kk'}, \ell = 0, \ldots, L$ are critical to the performance of our proposed estimators. Denote initial estimators by $\bar{\beta}_k$ and $\bar{\boldsymbol{\gamma}}_{k\ell}$. Then, the weights of adaptive lasso are $\omega_{0,kk'} = |\bar{\beta}_k - \bar{\beta}_{k'}|^{-\tau}$ and $\omega_{\ell,kk'} = \|\bar{\boldsymbol{\gamma}}_{\ell k} - \bar{\boldsymbol{\gamma}}_{\ell k'}\|^{-\tau}, \ell = 1, \ldots, L$, where τ is some constant larger than zero. We can use the BIC to choose the best τ. In Sect. 5.1, we compare our proposed method based on different τ's with the value of τ selected by BIC. The results show that the BIC-selected τ provides the best estimation and clustering performance. In practice, there are several options for initial estimators: subgroup-specific least squares (SSLS), K-means, and fused lasso. Section S.1.4 in the Supplementary Materials provide details for these methods.

The choice of graph G is crucial to the performance of the fGAFL method. We first define a distance matrix measuring the pairwise dissimilarity among subgroups. Then, graph G is constructed by connecting subgroups to their k-nearest neighbors. The distance matrix can be created using one of the following two methods. The first approach is to use prior knowledge and gauge subgroup dissimilarity based on that knowledge. For example, we use the kinship matrix, which reflects the genetic covariance structure, to measure the dissimilarity between genotypes. The second approach is data-driven. We first obtain the subgroup-specific estimators. The distance matrix is based on the differences among these estimators. The second step of fGAFL–2 estimators (defined Sect. 3.4) is an example of a data-driven approach, where the distance matrix is defined as the L_2 norm of pairwise differences for subgroup-specific estimates from the first step.

3.4. Two-Step Graph-Constrained Adaptive Fused Lasso (fGAFL–2)

A sparse graph G can improve the computation efficiency with a smaller number of penalty terms. However, if the graph G does not perfectly match the underlying group structure, we may obtain more clusters than the true number of groups. Table 1 in Simulation 2 shows that the average number of the estimated clusters based on the sparse graph-constrained fused lasso is larger than the true number of underlying groups, espe-

cially when we increase the noise. One possible reason is that some subgroups within the same group are far away from each other in the graph G and therefore less likely to be shrunk into one cluster. This phenomenon motivates us to consider a second step of model fitting, in which we update penalty graphs in (5) with graphs $G'_\ell, \ell = 0, \ldots, L$. The new penalty graphs are constructed by the k-nearest neighbors method based on the pairwise distance matrix of the estimates for each subgroup from the first step. Note that k can be selected via BIC criterion. The objective function is formulated as

$$\begin{aligned}\tfrac{1}{2}\left\|\mathbf{Y} - \mathbf{Z}_0\boldsymbol{\beta} - \textstyle\sum_{\ell=1}^{L}\mathbf{Z}_\ell\boldsymbol{\gamma}_\ell\right\|^2 + \textstyle\sum_{\ell=1}^{L}\lambda_\ell^{\mathrm{R}}\boldsymbol{\gamma}_\ell^\top((\mathbf{I}_K\otimes\mathbf{P})\boldsymbol{\gamma}_\ell + \lambda_0^{\mathrm{F}}\|\mathbf{H}_{G'_0}\boldsymbol{\beta}\|_{1,\boldsymbol{\omega}_0}\\ + \textstyle\sum_{\ell=1}^{L}\lambda_\ell^{\mathrm{F}}\|(\mathbf{H}_{G'_\ell}\otimes\mathbf{I}_q)\boldsymbol{\gamma}_\ell\|_{1,\boldsymbol{\omega}_\ell}^q. \qquad (8)\end{aligned}$$

The minimizer of (8), which we call the fGAFL–2 estimator, is our final estimator.

4. ASYMPTOTICS

We first introduce some technical notation. Denote the total number of subjects $n = \sum_{k=1}^{K} n_k$, the minimum number of subjects within a subgroup as $n_{\min} = \min_{1\le k\le K} n_k$, the maximum number of subjects within a subgroup as $n_{\max} = \max_{1\le k\le K} n_k$, the minimum number of subjects within a cluster as $n_{\min}(\mathbb{G}) = \min\{\sum_{k\in\mathcal{G}_{\ell,m}} n_k,\ 1 \le m \le M_\ell, \ell = 0, \ldots, L\}$, the maximum number of subjects within a cluster as $n_{\max}(\mathbb{G}) = \max\{\sum_{k\in\mathcal{G}_{\ell,m}} n_k,\ 1 \le m \le M_\ell, \ell = 0, \ldots, L\}$. Let $M_{\max} = \max_{0\le\ell\le L} M_\ell$ be the maximum number of clusters. For a scalar function $\phi(t)$, we define its L_2 norm by $\|\phi\| = \{\int_{\mathcal{T}}\phi(t)^2dt\}^{1/2}$. The covariance operator $\Gamma_{\mathbf{X}}$ is defined as $\Gamma_{\mathbf{X}}\boldsymbol{\phi}(t) = \int_{\mathcal{T}}\mathrm{E}\{\mathbf{X}(t)\mathbf{X}(s)^\top\}\boldsymbol{\phi}(s)ds$, and the corresponding induced norm is $\|\boldsymbol{\phi}\|_{\Gamma_{\mathbf{X}}} = \langle\boldsymbol{\phi}, \Gamma_{\mathbf{X}}\boldsymbol{\phi}\rangle^{1/2}$.

4.1. ORACLE ESTIMATORS

In this section, we study theoretical properties of oracle estimators, in which we assume that the underlying cluster structure is known. Denote the space of the oracle estimators as $\mathcal{M}_{\mathbb{G}_0} = \{\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)^\top \in R^K : \beta_k = \beta_{k'}, \text{ if } k, k' \in \mathcal{G}_{0,m} \text{ for some } m\}$, and the spline approximation space of the oracle estimators as $\mathcal{M}_{\mathbb{G}_\ell} = \{\boldsymbol{\alpha}_\ell = (\alpha_{1\ell}, \ldots, \alpha_{K\ell})^\top \in \mathcal{U}^K : \alpha_{k\ell} = \alpha_{k'\ell}, \text{ if } k, k' \in \mathcal{G}_{\ell,m} \text{ for some } m\}$. The oracle estimators $(\widehat{\boldsymbol{\beta}}^{o\top}, \widehat{\boldsymbol{\alpha}}_1^{o\top}, \ldots, \widehat{\boldsymbol{\alpha}}_L^{o\top})^\top$ can be obtained by solving:

$$\begin{aligned}\min_{\boldsymbol{\beta}\in\mathcal{M}_{\mathbb{G}_0},\boldsymbol{\alpha}_\ell\in\mathcal{M}_{\mathbb{G}_\ell}} \tfrac{1}{2}\textstyle\sum_{k=1}^{K}\sum_{i=1}^{n_k}\left\{Y_{ik} - \beta_k - \sum_{\ell=1}^{L}\int_{\mathcal{T}}X_{ik\ell}(t)\alpha_{k\ell}(t)dt\right\}^2\\ + \textstyle\sum_{k=1}^{K}\sum_{\ell=1}^{L}\lambda_\ell^R\int_{\mathcal{T}}\{\alpha''_{k\ell}(t)\}^2dt, \qquad (9)\end{aligned}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)^\top$, and $\boldsymbol{\alpha}_\ell(t) = \{\alpha_{1\ell}(t), \ldots, \alpha_{K\ell}(t)\}^\top$. Section S.4.1 in the Supplementary Material gives the solution of (9).

Below, we first introduce some technical assumptions. Let ν be a nonnegative integer, and $\delta \in (0, 1]$ such that $\varrho = \delta + \nu \ge 1$. Define $\mathcal{H}^{(\varrho)}(\mathcal{T})$ as the space of functions ψ on $\mathcal{T}$ whose ν-th derivative exists and satisfies a Lipschitz condition of order δ : $\left|\psi^{(\nu)}(x) - \psi^{(\nu)}\left(x'\right)\right| \le C_v\left|x - x'\right|^\delta$, for $x, x' \in \mathcal{T}$.

(A1) For $1 \le k \le K$, $\ell = 1, \ldots, L$, the coefficient functions $\alpha_{k\ell} \in \mathcal{H}^{(\varrho)}(\mathcal{T})$.

(A2) For $\ell = 1, \ldots, L$, $X_{ik\ell}(t)$ is a continuous function on $\mathcal{T}$ and its L_2 norm is finite, that is $\|X_{ik\ell}\| < \infty$, almost surely. For any $\ell, \ell' = 1, \ldots, L$, $\mathrm{E}|X_\ell(t)X_{\ell'}(s)|^3 < \infty$, for any $(s,t) \in \mathcal{T} \times \mathcal{T}$, and $\mathrm{E}X_\ell(t) = 0$, for $t \in \mathcal{T}$.

(A3) For $k = 1, \ldots, K$, $i = 1, \ldots, n_k$, $(Y_{ik}, \mathbf{X}_{ik}, \varepsilon_{ik})$ are independently distributed and $\mathrm{Var}(\varepsilon_{ik}) \le C < \infty$ holds.

(A4) For $1 \le k \le K$, the subgroup sample size n_k satisfies that $c_1 n K^{-1} \le n_k \le C_1 n K^{-1}$, $0 < c_1 < C_1 < \infty$. For $\ell = 0, 1, \ldots, L$, $1 \le m \le M_\ell$, the number of subgroups within each cluster satisfies $c_2 K^\zeta \le |\mathcal{G}_{\ell,m}| \le C_2 K^\zeta$, for some $0 \le \zeta < 1$ and $0 < c_2 < C_2 < \infty$.

(A5) As the total number of subjects n goes to infinity, the number of interior knots N and the roughness tuning parameter satisfy, for $\ell = 1, \ldots, L$, $\lambda_\ell^{\mathrm{R}} N^4 n^{-1} \to 0$, $\lambda_\ell^{\mathrm{R}} K n^{-1} M_{\max}^{1/2} \to 0$, $(\lambda_\ell^{\mathrm{R}} N^{7/2})^{-1} n_{\max}^{1/2}(\mathbb{G}) M_{\max}^{1/2} \to 0$, and $M_{\max}^{1/(2\varrho+2)} N^{-1} \to 0$.

The above assumptions are mild conditions that can be satisfied in many practical situations. Assumptions (A1–A3) are common assumptions in the literature of functional linear regression; see Assumptions (H.1–H.3) in Cardot et al. (2003). Assumption (A1) assumes each functional coefficient $\alpha_{k\ell}$ is a smooth function (Li et al. 2019; Yu et al. 2020). Assumption (A2) imposes requirements on the functional predictors. For model identification, we assume $\mathrm{E}X_\ell(t) = 0$, for $t \in \mathcal{T}$. Assumption (A3) assumes all the subjects are independent, and the error terms have variances bounded above by some finite constant. Assumption (A4) ensures that subgroup sample sizes are similar and clusters have similar numbers of subgroups. Assumption (A5) states the requirement on the number of interior knots and the roughness tuning parameter to ensure the consistency property of the spline estimator.

In the following, denote $d_{1,n} = \sum_{\ell=1}^{L} M_{\max}^{1/2}(N^{-(\varrho+1)} + \lambda_\ell^{\mathrm{R}} K n^{-1}) + (\lambda_\ell^{\mathrm{R}} N^4 K^\zeta)^{-1} n^{1/2}$, and $d_{2,n} = \sum_{\ell=1}^{L} M_{\max}^{1/2}(N^{-(\varrho+1)} + \lambda_\ell^{\mathrm{R}} K n^{-1}) + (\lambda_\ell^{\mathrm{R}} N^{7/2} K^\zeta)^{-1} n^{1/2}$. Theorem 1 shows the convergence rate of the oracle estimators. See Section S.4.3 in the Supplementary Material for its detailed proof.

Theorem 1. *Under Assumptions (A1)–(A5), the oracle estimators satisfy* $\sup_{1\le k\le K} |\widehat{\beta}_k^o - \beta_k| = O_P(d_{1,n})$, *and* $\sup_{1\le k\le K} \|\widehat{\boldsymbol{\alpha}}_k^o - \boldsymbol{\alpha}_k\|_{\Gamma_{\mathbf{X}}} = O_P(d_{2,n})$.

Remark 1. The convergence rate of the oracle estimators depends on the number of subjects within each cluster. If the number of clusters is finite, the convergence rate of the oracle estimator is consistent with Theorem 3.1 in Cardot et al. (2003). Results in Theorem 1 present a convergence rate for general cases where all the model components are allowed to have different cluster structures. When the cluster structures are the same for all the components, one can apply the conclusion in Cardot et al. (2003) and the convergence rate of functional-coefficient estimators of the mth cluster is $N^{-(\varrho+1)} + \sum_{\ell=1}^{L} \lambda_\ell^{\mathrm{R}} (\sum_{k\in\mathcal{G}_m} n_k)^{-1} + \sum_{\ell=1}^{L} (\lambda_\ell^{\mathrm{R}} N^{7/2})^{-1} (\sum_{k\in\mathcal{G}_m} n_k)^{1/2}$.

4.2. CONSISTENCY OF THE FCGAFL ESTIMATOR

Theorem 2 states the L_2 convergence rate of the adaptive fused lasso estimator. The proof details are given in Section S.4.4 in the Supplementary Material. Denote the difference between the initial estimators and the true values as $\delta_0 = \sup_{1\le k\le K} |\bar{\beta}_k - \beta_k|$ and $\delta_\ell = \sup_{1\le k\le K} \|\bar{\alpha}_{\ell k} - \alpha_{\ell k}\|$, for $\ell = 1, \ldots, L$. We introduce some technical assumptions.

(B1) The noise vector $\boldsymbol{\varepsilon} = (\varepsilon_{11}, \ldots, \varepsilon_{n_1 1}, \ldots, \varepsilon_{n_K K})^\top$ has sub-Gaussian tails such that $\mathrm{pr}(\boldsymbol{a}^\top \boldsymbol{\varepsilon} > \|\boldsymbol{a}\| x) \le 2\exp(-c_1 x^2)$ for any vector $\boldsymbol{a}$ and $x > 0$, where $0 < c_1 < \infty$.

(B2) The weights of adaptive lasso satisfy the conditions (i) $\delta_0^\tau K^{5/2-\zeta} n^{-1/2} d_{1,n}^{-1} \to 0$, and (ii) for $1 \le \ell \le L$, $\delta_\ell^\tau (\lambda_\ell^{\mathrm{R}})^{-1} N^{-3} K^{3/2-\zeta} n^{1/2} d_{2,n}^{-1} \to 0$.

(B3) As the sample size n goes to infinity, the tuning parameters $\lambda_\ell^{\mathrm{F}}, \ell = 0, \ldots, L$ satisfy the conditions (i) $\lambda_0^{\mathrm{F}} K^{1+\zeta} n^{-1} d_{1,n}^{-1} \to 0$ and $\lambda_0^{\mathrm{F}} \delta_0^{-\tau} K^{\zeta+1/2} n^{-1/2} \to \infty$ (ii) for $1 \le \ell \le L$, $\lambda_\ell^{\mathrm{F}} K^2 N^{-7/2} d_{2,n}^{-1} \to 0$ and $\lambda_\ell^{\mathrm{F}} \delta_\ell^{-\tau} K^{\zeta+1/2} n^{-1/2} N^{-1/2} \to \infty$.

Assumption (B1) is widely used in the high dimensional settings; see Condition (C2) in Ma and Huang (2017). Assumption (B2) describes the requirements of the adaptive weights in the adaptive lasso. Assumption (B3) illustrates the conditions of tuning parameters. According to our theoretical studies, the convergence rate of the proposed estimators depends on the subgroup sample size, weights in the adaptive lasso, and the number of subgroups. If we use the SSLS estimators as the initial weights for the adaptive lasso and $\tau = 1$, by Theorem 1 and Assumption (B2), one can obtain that $K^{5/2-2\zeta} n^{-1/2} \to 0$ and $K^{2-2\zeta} n_k^{-1/2} \to 0$. To achieve the estimation consistency, the subgroup sample size should increase as the number of subgroups or the number of clusters increases.

Theorem 2. *Under Assumptions (A1) – (A5) and (B1) – (B3), the adaptive fused lasso estimators satisfy* $\sup_{1\le k\le K} |\widehat{\beta}_k - \beta_k| = O_P(d_{1,n})$ *and* $\sup_{1\le k\le K} \|\widehat{\boldsymbol{\alpha}}_k - \boldsymbol{\alpha}_k\|_{\Gamma_{\mathbf{X}}} = O_P(d_{2,n})$.

Define $\mathcal{I}_0 = \{(k,k') : \beta_k \ne \beta_{k'},\ k,k' = 1,\ldots,K\}$ as the edge set for nonidentical intercepts and $\mathcal{I}_\ell = \{(k,k') : \alpha_{k\ell} \ne \alpha_{k'\ell},\ k,k' = 1,\ldots,K\}$, $1 \le \ell \le L$ as the edge set for nonidentical coefficient functions. Let $b_0 = \min_{(k,k')\in\mathcal{I}_0} |\beta_k - \beta_{k'}|$, $b_1 = \min_{(k,k')\in\cup_{1\le\ell\le L}\mathcal{I}_\ell} \|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}_{k'}\|_{\Gamma_{\mathbf{X}}}$.

Remark 2. Note that $|\widehat{\beta}_k - \widehat{\beta}_{k'}| \ge |\beta_k - \beta_{k'}| - |\widehat{\beta}_k - \beta_k - \widehat{\beta}_{k'} + \beta_{k'}|$, and $\|\widehat{\boldsymbol{\alpha}}_k - \widehat{\boldsymbol{\alpha}}_{k'}\|_{\Gamma_{\mathbf{X}}} \ge \|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}_{k'}\|_{\Gamma_{\mathbf{X}}} - \|\widehat{\boldsymbol{\alpha}}_k - \boldsymbol{\alpha}_k - \widehat{\boldsymbol{\alpha}}_{k'} + \boldsymbol{\alpha}_{k'}\|_{\Gamma_{\mathbf{X}}}$. Therefore, if the conditions $b_0(d_{1,n})^{-1} \to \infty$ and $b_1(d_{2,n})^{-1} \to \infty$ hold, we can discover the underlying group structure with probability approaching one.

5. SIMULATION STUDY

5.1. SIMULATION STUDY 1

In this section, we conduct a Monte Carlo simulation study to examine the finite-sample performance of the proposed methodology based on a complete graph. We consider 40 subgroups and generate response variables from the following FLM: $Y_{ik} =$

$\beta_k + \int_0^1 X_{ik}(t)\alpha_k(t)dt + \varepsilon_{ik}$, $i = 1, \ldots, n_k$, $k = 1, \ldots, 40$, where $\beta_k = 5$, $k = 1, \ldots, 40$, $\alpha_k(t) = \cos(2\pi t)$, if $k = 1, \ldots, 20$ and $\alpha_k(t) = \sin(2\pi t)$, if $k = 21, \ldots, 40$. Therefore, the underlying group structures are $\mathbb{G}_0 = \{\mathcal{G}_{0,1}\}$ with $\mathcal{G}_{0,1} = \{1, \ldots, 40\}$ and $\mathbb{G}_1 = \{\mathcal{G}_{1,1}, \mathcal{G}_{1,2}\}$ with $\mathcal{G}_{1,1} = \{1, \ldots, 20\}$ and $\mathcal{G}_{1,2} = \{21, \ldots, 40\}$. We simulate the functional covariates by $X_{ik}(t) = \sum_{s=1}^{9} \xi_{iks}\phi_s(t)$, where $\phi_s(t)$ are orthonormal basis functions, and ξ_{iks}'s are independently generated from $N(0, 1)$ and truncated by $[-3, 3]$. Figure 1A shows true coefficient functions. The noise term ε_{ik} is generated from $N(0, \sigma^2)$. We consider $n_k = 50, 100$, $k = 1, \ldots, 40$, and the noise level $\sigma = 0.1, 0.5$ and 1.0.

We denote the proposed complete graph-constrained adaptive fused lasso estimator described in Sect. 2.2 as fCGAFL, and we compare it with the oracle estimator (ORACLE), the SSLS estimator (SSLS), and K-means estimator (K-Means), which are described in Sect. 3.3. K-Means is implemented by R function *kmeans* with the number of clusters selected via the Gap statistic using the R package *cluster*. We place knots on a grid of evenly spaced sample quantiles based on the observed time points for spline smoothing. In this simulation study, we use quadratic splines with four interior knots for all the estimators.

The fCGAFL estimator depends on the method used to obtain adaptive weights. By convention, we use fCGAFL to denote the proposed graph-constrained adaptive fused lasso with adaptive weights given by the SSLS estimators. We also evaluate the performance of the proposed fCGAFL method based on two other different adaptive weights: K-means and the lasso; and the corresponding estimators are referred to as fCGAFL(K) and fCGAFL(L). For the sake of saving space, we only report the simulation results of the fCGAFL estimators here, and the simulation results of the fCGAFL(K) and fCGAFL(L) estimators are presented in the Supplementary Material. Tables 1 and 2 present the simulation results of fCGAFL based on 100 Monte Carlo experiments. We find that fCGAFL, fCGAFL(K), and fCGAFL(L) all have better performance than K-Means in most of the simulation scenarios. Compared with fCGAFL and fCGAFL(K), fCGAFL(L) generates a more aggressive shrinkage among the subgroups. When the sample size is 50 and σ is 1.0 (low signal-to-noise ratio), fCGAFL(L) tends to shrink all the subgroups into one cluster, while fCGAFL and fCGAFL(K) give an average cluster size of 1.45 and 1.82, respectively. However, the clustering fails for all proposed methods in the low signal-to-noise ratio scenario.

Figures S.1 (A–B) in the Supplementary Material depict the solution path based on fCGAFL of a typical simulation example. When we increase the tuning penalty parameters, more subgroups can be shrunk. In Table 1, we report the clustering results based on K-Means and fCGAFL. The clustering performance is evaluated by the following measures (see Sect. S.2.1 in the Supplementary Material for the details of how these measures are calculated): the adjusted rand index (aRI), and the Jaccard index (Jaccard), which should be as close to one as possible. We also list the estimated number of groups (Size), which should be close to the true number of clusters (one for intercept and two for coefficient functions). We observe that the proposed fCGAFL outperforms K-Means. Table 2 presents the root mean squared error (RMSE), the root mean integrated squared error (RMISE), and the average computing time for K-Means and fCGAFL. The results of ORACLE serve as a benchmark for comparison. When clustering results are reasonable, the RMSE and RMISE for K-Means and fCGAFL are close to those of the ORACLE. Increasing the sample size in each subgroup enhances the clustering performance and estimation performance, which

supports our theoretical findings in Sect. 4. The clustering performance of fCGAFL(K) and fCGAFL(L) is given in Table S.1 in the Supplementary Material, which reveals that both fCGAFL(K) and fCGAFL(L) have better performance than K-Means and SSLS.

In addition, we study the performance of using BIC to choose the proper order τ in the weights and the number of interior knots N. The simulation study is conducted with subgroup sample size $n_k = 100$ at noise levels $\sigma = 0.1$, 0.5 and 1.0. According to our simulation results, we observe that the BIC criterion can provide a reliable choice of τ and N. Figures S.2 (A)–(B) in the Supplementary Material summarize the estimation performance. The BIC-based method has better performance among the methods with different N's and τ's. The value of aRI ranges from 0.997 to 1, and the average size ranges from 2.00 to 2.06.

To evaluate the performance of the fGAFL–2 method, in this simulation study, we take the fCGAL estimator in step 1 and obtain graph G for step 2 based on the distance among the estimated intercepts and coefficient functions for each subgroup. Tables 1 and 2 show that the fGAFL–2 estimator has better clustering and estimation performance than fCGAL. According to Table 2, the average computing time of fGAFL–2 is smaller than that of fCGAL, indicating a sparse penalty graph can help improve the computational efficiency.

5.2. SIMULATION STUDY 2

In this simulation study, we investigate the performance of the proposed fGAFL and fGAFL–2 methods describe in Sect. 3, in which a graph is used to provide some prior information about the underlying group structure. We generate our dataset from the following model with 165 different subgroups: $Y_{ik} = \beta_k + \int_0^1 X_{ik}(t)\alpha_k(t)dt + \varepsilon_{ik}, i = 1, \ldots, n_k, k = 1, \ldots, 165$, where $X_{ik}(t) = \sum_{s=1}^{9} \xi_{iks}\phi_s(t)$ and $\phi_s(t)$ are orthonormal basis functions, and ξ_{ik}'s are independently generated from $N(0, 1)$ and truncated by $[-3, 3]$. The noise term ε_{ik} is generated from $N(0, \sigma^2)$. The underlying group structures are denoted as $\mathbb{G}_0$ for intercepts and $\mathbb{G}_1$ for coefficient functions. There are two clusters for intercept $\mathbb{G}_0 = \{\mathcal{G}_{0,1}, \mathcal{G}_{0,2}\}$, and five clusters for coefficient functions $\mathbb{G}_1 = \{\mathcal{G}_{1,1}, \ldots, \mathcal{G}_{1,5}\}$ with $|\mathcal{G}_{0,1}| = 84$, $|\mathcal{G}_{0,2}| = 81$, and $|\mathcal{G}_{1,1}| = 37$, $|\mathcal{G}_{1,2}| = 42$, $|\mathcal{G}_{1,3}| = 12$, $|\mathcal{G}_{1,4}| = 27$, $|\mathcal{G}_{1,5}| = 47$. True intercepts and coefficient functions are $\boldsymbol{\beta}^{\mathrm{G}} = (\beta_1^{\mathrm{G}}, \beta_2^{\mathrm{G}}) = (5, 6)$ and $\alpha_1^{\mathrm{G}}(t) = \cos(2\pi t)$, $\alpha_2^{\mathrm{G}}(t) = \sin(2\pi t)$, $\alpha_3^{\mathrm{G}}(t) = 1.5(t - 0.5)$, $\alpha_4^{\mathrm{G}}(t) = 1 - 2\exp(-6t)$, and $\alpha_5^{\mathrm{G}}(t) = 2\exp(-6t) - 1$. Figure 1B presents plots of true coefficient functions.

Figure 1C, D shows the graph G used in this simulation. Each circle represents one subgroup, and the gray lines connecting circles represent the edges in the graph G. We use different colors to indicate different clusters in $\mathbb{G}_0$ and $\mathbb{G}_1$. To mimic real data in the second application example, we consider $n_k = 30,\ 50$, $k = 1, \ldots, 165$, and noise level $\sigma = 0.1$, 0.5 and 1.0.

In this simulation example, we compare the following methods: ORACLE, SSLS, K-Means, fGAFL, and fGAFL–2. Figure 1E, G depicts the solution path based on fGAFL of a typical simulation example. Tables 1 and 3 present the clustering performance, estimation accuracy, and the average computing time based on 100 Monte Carlo experiments using different methods. According to Table 1, fGAFL and fGAFL–2 have better clustering performance than K-Means. Table 3 shows that our proposed methods have better estimation accuracy than SSLS and K-Means. As expected, the estimation accuracy for all the

Table 1. Clustering results in simulation studies

Simu	n_k	σ	Method	β			α		
				aRI	Jaccard	Size	aRI	Jaccard	Size
1	50	0.1	K-Means	0.85	0.87	2.05	1.00	1.00	2.00
			fCGAFL	1.00	1.00	1.00	1.00	1.00	2.00
			fGAFL–2	1.00	1.00	1.00	1.00	1.00	2.00
		0.5	K-Means	0.82	0.85	2.28	0.70	0.70	4.40
			fCGAFL	1.00	1.00	1.00	0.81	0.81	5.74
			fGAFL–2	1.00	1.00	1.00	0.99	0.99	2.02
		1	K-Means	0.82	0.85	2.30	0.27	0.44	4.51
			fCGAFL	1.00	1.00	1.00	0.02	0.49	1.45
			fGAFL–2	1.00	1.00	1.00	0.16	0.42	11.73
	100	0.1	K-Means	0.74	0.78	2.90	1.00	1.00	2.00
			fCGAFL	1.00	1.00	1.00	1.00	1.00	2.00
			fGAFL–2	1.00	1.00	1.00	1.00	1.00	2.00
		0.5	K-Means	0.73	0.77	2.89	0.77	0.76	3.90
			fCGAFL	1.00	1.00	1.00	1.00	1.00	2.04
			fGAFL–2	1.00	1.00	1.00	1.00	1.00	2.00
		1	K-Means	0.74	0.78	2.82	0.62	0.62	4.94
			fCGAFL	1.00	1.00	1.00	0.85	0.84	5.10
			fGAFL–2	1.00	1.00	1.00	1.00	1.00	2.03
2	30	0.1	K-Means	1.00	1.00	2.00	0.79	0.72	6.68
			fGAFL	1.00	1.00	2.00	1.00	1.00	5.00
			fGAFL–2	1.00	1.00	2.00	1.00	1.00	5.00
		0.5	K-Means	1.00	1.00	2.00	0.81	0.74	6.48
			fGAFL	1.00	1.00	2.02	1.00	1.00	5.29
			fGAFL–2	1.00	1.00	2.00	1.00	1.00	5.00
		1	K-Means	0.98	0.98	2.00	0.77	0.70	4.36
			fGAFL	0.99	0.99	2.80	0.98	0.96	6.92
			fGAFL–2	1.00	1.00	2.00	0.99	0.98	5.04
	50	0.1	K-Means	1.00	1.00	2.00	0.76	0.69	6.89
			fGAFL	1.00	1.00	2.00	1.00	1.00	5.00
			fGAFL–2	1.00	1.00	2.00	1.00	1.00	5.00
		0.5	K-Means	1.00	1.00	2.00	0.78	0.71	6.80
			fGAFL	1.00	1.00	2.02	1.00	1.00	5.08
			fGAFL–2	1.00	1.00	2.00	1.00	1.00	5.00
		1	K-Means	1.00	1.00	2.00	0.83	0.78	4.81
			fGAFL	1.00	1.00	2.20	0.99	0.99	5.69
			fGAFL–2	1.00	1.00	2.00	1.00	1.00	5.00

Simu, simulation setting; aRI, average of the adjusted rand index; Jaccard, average of the Jaccard index; Size, average of the estimated number of groups

methods improves as the sample size increases or the noise level decreases. We observe the average computing time varies with sample size and noise level. The fGAFL–2 is more time-consuming than the fGAFL with its two-step fitting procedure. As one can observe from Figure 1 (G), there are clusters with similar estimated coefficient functions. To avoid the influence of the graph G and combine the similar clusters, we implement the fGAFL–2. Figure 1 (F) and (H) presents the solution path based on fGAFL–2. From Table 1 and 3, we can tell fGAFL–2 can improve the estimation performance of fGAFL.

Compared with Simulation 1, sparse graph-constrained adaptive lasso allows us to handle datasets with a large number of subgroups and improves the clustering and estimation accuracy.

Table 2. RMISEs (RMSEs) of the coefficient functions (intercepts) in Simulation 1

n_k	σ	Method	RMSE ($\times 10$)	RMISE ($\times 10$)		Time
			β	α_1	α_2	
50	0.1	ORACLE	0.00	0.15	0.19	1.28
		SSLS	0.15	4.76	4.71	4.61
		K-Means	0.02	0.15	0.17	2.56
		fCGAFL	0.00	0.15	0.18	481.30
		fGAFL–2	0.00	0.15	0.18	52.76
	0.5	ORACLE	0.01	0.47	0.49	1.25
		SSLS	0.76	4.83	4.79	4.57
		K-Means	0.13	0.88	0.95	2.52
		fCGAFL	0.02	0.92	0.92	135.80
		fGAFL–2	0.01	0.48	0.45	70.29
	1.0	ORACLE	0.01	0.91	0.93	1.47
		SSLS	1.51	5.05	5.00	5.61
		K-Means	0.29	3.68	3.64	3.11
		fCGAFL	0.09	4.93	4.99	168.85
		fGAFL–2	0.08	4.23	4.71	17.55
100	0.1	ORACLE	0.00	0.13	0.17	2.70
		SSLS	0.10	4.76	4.71	7.97
		K-Means	0.03	0.13	0.16	4.26
		fCGAFL	0.00	0.13	0.16	206.11
		fGAFL–2	0.00	0.13	0.16	10.27
	0.5	ORACLE	0.00	0.37	0.38	2.44
		SSLS	0.50	4.79	4.74	7.19
		K-Means	0.13	0.61	0.61	3.85
		fCGAFL	0.01	0.66	0.66	101.17
		fGAFL–2	0.00	0.35	0.35	64.36
	1.0	ORACLE	0.01	0.69	0.71	2.76
		SSLS	1.00	4.90	4.84	8.17
		K-Means	0.26	1.59	1.58	4.37
		fCGAFL	0.02	1.63	1.61	185.82
		fGAFL–2	0.01	0.64	0.68	29.39

RMSE, average of the root mean squared error of intercept terms; RMISE, average of the root mean integrated squared error of coefficient functions. Time, the average computing time (in seconds)

6. APPLICATION

Characterization of a plant variety's response to the environment is a key challenge for plant breeders (Nicotra et al. 2010). Traits, such as flowering time and grain yield, are the primary targets of breeding efforts but are responsive to environmental conditions. These conditions are integrated across the growing season, and in certain developmental periods, plants are often more sensitive to some environmental factors than others (Van de Pol et al. 2016). To make breeding decisions, plant breeders require flexible methods (1) to estimate the responses of different varieties to the environment to identify target response patterns and (2) to distinguish the responses of varieties from each other to choose the parents of the next generation. The proposed multi-FLM is useful to study the above problems. By combining the genotypes with similar response patterns, we obtain more efficient estimators. Meanwhile, the proposed multi-FLM automatically identifies the different response patterns,

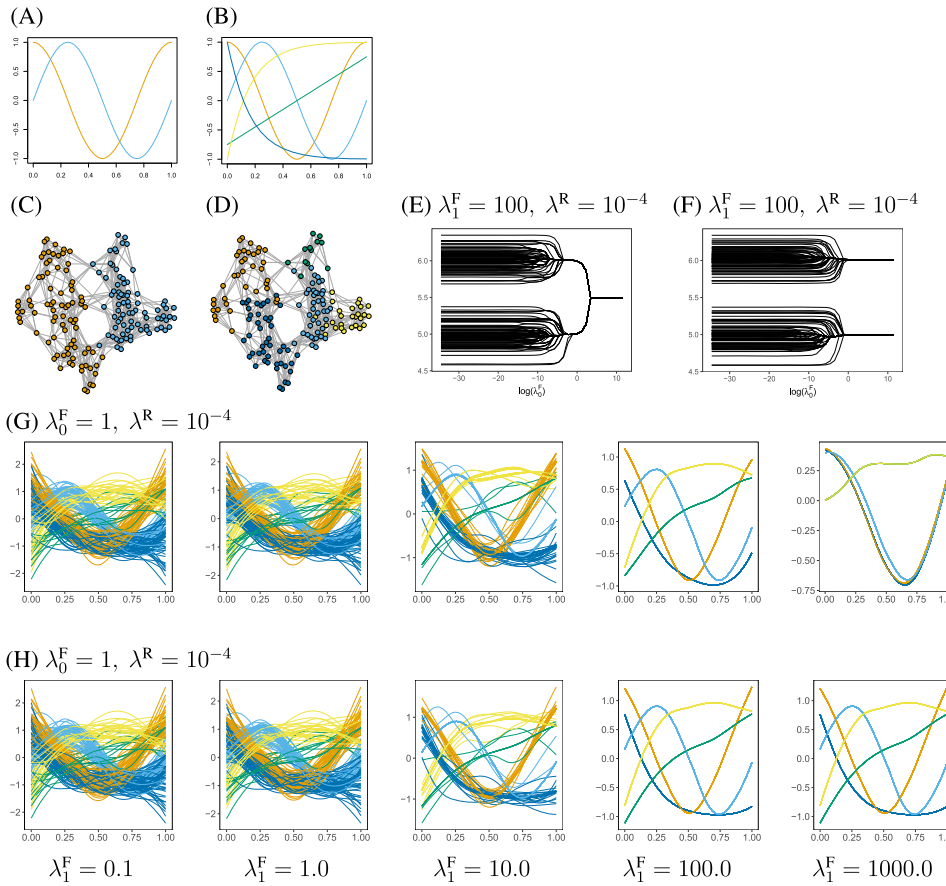

Figure 1. Plot of the true coefficient functions in Simulation 1 (**A**) and Simulation 2 (**B**). The penalty group G used in Simulation 2 for the intercept term (**C**), and the coefficient function (**D**). Each circle represents one subgroup, and gray lines connecting circles represent the edges in the graph G, and different colors indicate the true group structure. Solution paths of fGAFL (**E**) and (**G**) and fGAFL–2 (**F**) and (**H**) in Simulation 2 with $n_k = 50$ and $\sigma = 1.0$. (E) and (F) are solution paths for intercept terms for a given λ_1^F and λ^R. (G) and (H) are solution paths for coefficient functions for a given λ_0^F and λ^R. The color for each curve indicates its true group. The selected tuning parameters are $\lambda_0^F = 0.32$ (E), $\lambda_0^F = 3.16$ (F), $\lambda_1^F = 100$ (G), and $\lambda_1^F = 100$ (H) .

which serves as a tool to study genotype-by-environmental (G × E) effects. In the following, we consider two plant datasets in plant science to illustrate the proposed method.

6.1. EXAMPLE 1: SORGHUM FLOWERING TIME

We first consider a study focusing on the G × E effects on flowering time, which has a significant bearing on evolution and adaptation of plants to different environments (Li et al. 2018). The dataset contains 206 recombinant inbred lines (RILs) from two parental inbreds, evaluated in nine environments with two replications in each environment. Our study uses flowering time in Growing Degree Days (GDD) as our response variable. Photothermal time (PT) within a growth period is a major environmental determinant for flowering time and

Table 3. RMISEs (RMSEs) of the coefficient functions (intercepts) in Simulation 2

n_k	σ	Method	RMSE ($\times 10$)		RMISE ($\times 10$)					Time
			β_1	β_2	α_1	α_2	α_3	α_4	α_5	
30	0.1	ORACLE	0.01	0.01	0.15	0.18	0.17	0.13	0.10	2.93
		SSLS	0.21	0.2	4.76	4.73	2.93	5.51	5.51	7.39
		K-Means	0.02	0.03	0.41	0.58	2.17	1.21	0.33	11.56
		fGAFL	0.01	0.01	0.15	0.16	0.16	0.14	0.12	741.38
		fGAFL–2	0.01	0.01	0.15	0.16	0.17	0.13	0.11	882.81
	0.5	ORACLE	0.06	0.06	0.47	0.44	0.74	0.53	0.42	2.97
		SSLS	0.98	0.98	4.88	4.91	3.02	5.58	5.58	7.34
		K-Means	0.06	0.07	0.84	0.98	2.74	1.44	0.85	11.06
		fGAFL	0.06	0.07	0.54	0.54	0.68	0.66	0.7	416.16
		fGAFL–2	0.06	0.06	0.52	0.46	0.62	0.54	0.46	650.55
	1.0	ORACLE	0.11	0.12	0.83	0.79	1.37	0.96	0.73	2.96
		SSLS	1.92	1.95	5.36	5.43	3.16	5.73	5.74	7.29
		K-Means	0.54	0.55	2.44	2.87	5.18	2.44	2.45	11.37
		fGAFL	0.29	0.27	1.85	1.64	2.14	1.49	1.93	334.34
		fGAFL–2	0.17	0.15	1.15	1.11	1.21	0.95	0.92	463.46
50	0.1	ORACLE	0.01	0.01	0.13	0.17	0.13	0.1	0.09	5.79
		SSLS	0.15	0.15	4.76	4.72	2.93	5.50	5.50	11.28
		K-Means	0.01	0.02	0.3	0.66	2.54	1.31	0.49	18.59
		fGAFL	0.01	0.01	0.13	0.15	0.13	0.10	0.10	656.68
		fGAFL–2	0.01	0.01	0.13	0.16	0.13	0.10	0.09	960.09
	0.5	ORACLE	0.04	0.04	0.38	0.34	0.59	0.39	0.34	5.72
		SSLS	0.73	0.73	4.83	4.83	2.97	5.56	5.56	10.99
		K-Means	0.04	0.05	0.59	0.93	2.54	1.31	0.93	18.21
		fGAFL	0.04	0.05	0.4	0.39	0.52	0.47	0.54	402.41
		fGAFL–2	0.04	0.04	0.4	0.35	0.52	0.39	0.37	765.65
	1.0	ORACLE	0.08	0.09	0.68	0.63	1.11	0.72	0.61	5.75
		SSLS	1.45	1.45	5.14	5.23	3.05	5.67	5.68	10.95
		K-Means	0.14	0.13	1.24	1.64	4.07	2.02	1.59	18.2
		fGAFL	0.14	0.14	1.05	1.07	1.25	1.01	1.32	271.39
		fGAFL–2	0.09	0.11	0.80	0.77	0.91	0.72	0.71	499.28

RMSE, average of the root mean squared error of intercept terms; RMISE, average of the root mean integrated squared error of coefficient functions; Time, the average computing time (in seconds)

can be used as an environmental index. Figure S.3A in the Supplementary Material shows the observed daily PT within each environmental setting. We are interested in (1) identifying how daily PT influences flowering time and (2) detecting whether the influence of PT varies across different RILs.

For each RIL, we first conduct a simple linear regression of flowering time on the average of daily photothermal time from 18 to 43 days after planting. Figure S.3B in the Supplementary Materials presents the histogram of estimated slopes. There are two peaks in the histogram indicating for some of the RILs, the increase in PT leads to a larger increase in flowering time.

Motivated by our findings in Figure S.3B, we consider the following multi-group FLM:

$$Y_{ik} = \beta_k + \int_{\mathcal{T}} \mathrm{PT}_{ik}(t)\alpha_k(t)dt + \varepsilon_{ik}, \tag{10}$$

where Y_{ik} is the logarithm of flowering time for plot i and RIL k, $\mathcal{T}$ is from one to eighty days after planting, and $\mathrm{PT}_{ik}(t)$ is the measurements of photothermal time at time t for plot i, RIL k. We use our proposed method to fit the model and consider quadratic splines with one interior knot. As an example of how prior information can be incorporated in our method, we consider a ten-nearest neighbor graph, based on genetic similarity at loci Ma_6 and FT among RILs, as the graph G in our fused lasso penalty function. More specifically, we use 40 single nucleotide polymorphisms (SNPs) located near loci Ma_6 and FT studied by Li et al. (2018), to compute a kinship matrix using the method of VanRaden (2008). The kinship matrix describes the genetic covariance structure among RILs with respect to these SNPs. Each RIL was connected with an edge in the graph to the ten most genetically similar RILs based on kinship matrix entries. We use BIC to select the number of neighbors in the penalty graph. See Figure 2A for plot of the graph G.

Figure S.4A in the Supplementary Material shows the fGAFL estimated clusters for coefficient functions, in which we use different colors to represent different clusters. Figure S.4B in the Supplementary Material presents the fGAFL estimate of the coefficient functions of the clusters containing more than one subgroup. Note that there are some clusters whose subgroups are sparsely connected in the graph G while with similar estimates. To combine RILs with similar estimates, we construct G_0' and G_1' based on the similarity among the estimated intercepts and coefficient functions, respectively, and then, fit model (10) with the graph G_0' and G_1'. Through the second step of model fitting, we shrink the coefficient functions of RILs into two major clusters. According to Figure 2 (D), the values of both coefficient functions are positive from 1 to 60 days after planting. The value of the coefficient function in the blue cluster (61 RILs) is higher than that of the red cluster (145 RILs), indicating a stronger association between the flowering time and PT for the RILs in blue cluster. Both coefficient functions reach their peak around 30 days after planting. Figure 2E, F presents the estimated coefficient functions for two RILs with 95% pointwise confidence intervals. We use the wild bootstrap to construct pointwise intervals; see Sect. S.3.3 in the Supplementary Material for detailed descriptions. Table S.3 in the Supplementary Material shows the number of RILs in each intercept cluster using fGAFL and fGAFL–2. To compare differences between two clusters, for each SNP and cluster, we calculate the relative frequency of the SNP genotype coded as zero. Figure 2C displays a boxplot of the genotype relative frequency differences between the two identified clusters, based on 1,462 SNPs. There are 16 SNPs with differences larger than 0.6, indicating a discrepancy between two clusters. See Sect. S.3.1 in the Supplementary Material for a list of these 16 SNPs. One thing to note is that all 16 SNPs are located at the same chromosome.

Finally, we use leave-one-field-out cross-validation to evaluate the prediction performance. Table 4 lists the mean squared prediction error for each field and total average. Figure S.5 in the Supplementary Materials presents the boxplots of the absolute value of prediction errors. Both fGAFL and fGAFL–2 outperform SSLS and K-Means. By creating graphs based on the fGAFL estimate, fGAFL–2 improves the prediction performance of fGAFL. With only 18 subjects per subgroup, SSLS has bad predictive performance. The fGAFL and fGAFL–2, which fuse the coefficient functions from a same group into one function, reduce the large variance of SSLS estimators.

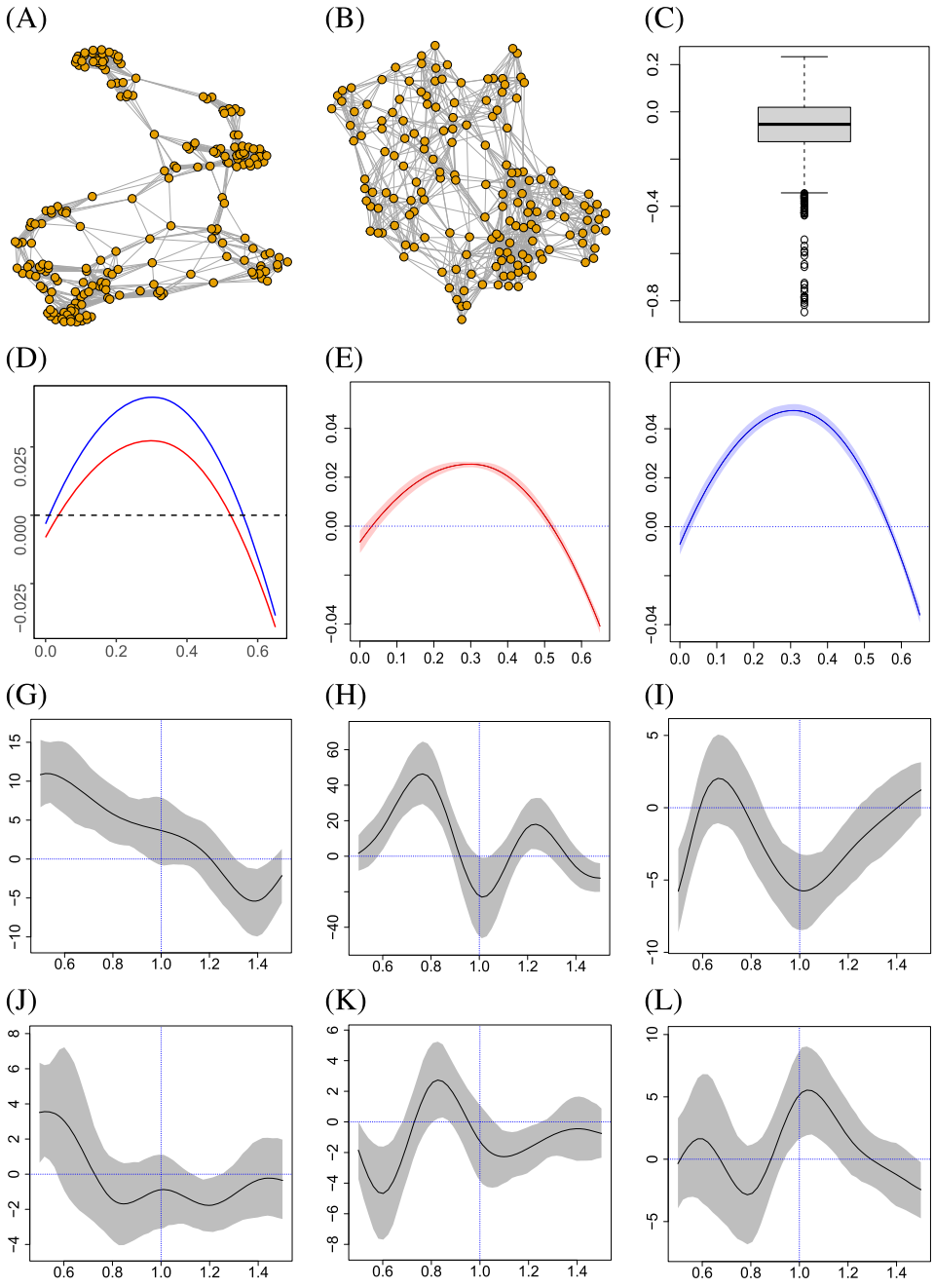

Figure 2. Plots of the graph G used in fGAFL in Application Example 1 **A** and Application Example 2 **B**. **C** boxplot of genotype relative frequency differences between clusters for 1,462 SNPs. **D** two identified clusters based on fGAFL–2, and the 95% pointwise confidence intervals of coefficient functions for RIL5 **E** and RIL10 **F**. The estimated coefficient functions with 95% pointwise confidence intervals in Application Example 2: **G** GDD, **H** CDD, **I** EDD, **J** PPT, **K** SR, and **L** VPD. The dashed blue horizontal and vertical lines represent the zero line and the anthesis date .

Table 4. The mean squared prediction error based on leave-one-field-out cross-validation

Field (Sample size)	IA13 (396)	IA14 (410)	KS11 (412)	KS12 (404)	PR11 (406)
SSLS	0.055	0.764	1.568	2.308	1.011
K-Means	0.310	0.114	1.054	0.781	0.135
fGAFL	0.189	0.089	0.516	0.197	0.260
fGAFL–2	0.059	0.051	0.405	0.250	0.145

Field (Sample size)	PR12 (412)	PR14S (398)	IA15 (409)	IA16 (407)	Average (SD)
SSLS	1.235	0.305	0.131	0.080	0.829 (0.779)
K-Means	0.085	0.639	0.227	0.289	0.404 (0.341)
fGAFL	0.189	0.214	0.142	0.119	0.213 (0.125)
fGAFL–2	0.111	0.199	0.080	0.078	0.153 (0.116)

6.2. EXAMPLE 2: HYBRID MAIZE GRAIN YIELD

In this application example, we illustrate our proposed method for a multi-dimensional functional linear model. We consider a dataset of hybrid maize grain yield from the Genomes to Fields consortium (AlKhalifah et al. 2018; McFarland et al. 2020). These hybrids were grown in 44 environments (combinations of location and year) where data on daily temperatures, rainfall, and incident solar radiation were collected by in-field weather stations.

We use a subset of these hybrids that were grown in at least 20 environments, yielding a final dataset of 165 hybrids with 5448 field plots total. We consider daily growing degree days (GDD), cold degree days (CDD; time exposed to temperatures below 10 °C), extreme degree days (EDD; time exposed to temperatures above 30 °C), total daily precipitation (PPT), incident solar radiation (SR) and vapor pressure deficit (VPD) as the environmental variable of interest due to their effects on maize grain yield (Tollenaar et al. 2017). We consider the following multi-group FLM:

$$\begin{aligned} Y_{ik} = \beta_k &+ \int_{\mathcal{T}} \mathrm{GDD}_{ik}(t)\alpha_{1k}(t)dt + \int_{\mathcal{T}} \mathrm{CDD}_{ik}(t)\alpha_{2k}(t)dt + \int_{\mathcal{T}} \mathrm{EDD}_{ik}(t)\alpha_{3k}(t)dt \\ &+ \int_{\mathcal{T}} \mathrm{PPT}_{ik}(t)\alpha_{4k}(t)dt + \int_{\mathcal{T}} \mathrm{SR}_{ik}(t)\alpha_{5k}(t)dt + \int_{\mathcal{T}} \mathrm{VPD}_{ik}(t)\alpha_{6k}(t)dt + \varepsilon_{ik}, \end{aligned}$$

where Y_{ik} is yield for plot i and hybrid k, $\mathcal{T} = [0, 2]$ is the range of relative GDD time from zero to two, and $\mathrm{GDD}_{ik}(t)$, $\mathrm{CDD}_{ik}(t)$, $\mathrm{EDD}_{ik}(t)$, $\mathrm{PPT}_{ik}(t)$, $\mathrm{SR}_{ik}(t)$ and $\mathrm{VPD}_{ik}(t)$ are the measurements of the environmental variables at the relative GDD time t with respect to plot i for hybrid k. Section S.3.2 in the Supplementary Material shows the construction of relative GDD time index.

Similar to the flowering-time example, we consider a ten-nearest neighbor graph, based on genetic similarity among hybrids, as the graph G in our fused lasso penalty function. See Figure 2 (B) for the plot of the graph G. We consider natural cubic splines with nine interior knots (selected by BIC criterion). One cluster is identified for each model component,

indicating that all the genotypes have similar response patterns to environmental changes. Figure 2 G–L shows the estimated coefficient functions with 95% pointwise confidence intervals. Sect. S.3.2 in the Supplementary Materials discusses the estimated coefficient curves.

7. DISCUSSION

In plant breeding studies, understanding heterogeneity among subgroups is essential to study the G × E interaction. This article proposes a fusion learning approach for multi-group functional linear regression, which simultaneously estimates FLMs and identifies the underlying group structure. By detecting homogeneous groups and merging them together, we can build a parsimonious model with great explanatory power, estimation efficiency and better interpretation. To reduce the computation burden and include prior information, we propose a graph-constrained adaptive fused lasso (fGAFL) estimator, in which the penalty graph can be constructed by measuring similarities among subgroups based on genetic information. In addition, a two-step method (fGAFL–2) is proposed to improve performance when a penalty graph lacks relevant edges. The merits of the proposed method have been illustrated by simulation studies and application examples.

Many issues still merit further research. The basic setup of FLMs can be extended to more flexible parametric and semiparametric models. For instance, we may develop a functional single index model. It is also feasible to extend our proposed methodology to many other types of functional data, such as next-generation functional data including 2D or 3D neuroimaging data.

Funding The project is funded by National Science Foundation Grant Nos. HDR:TRIPODS 19-34884 and DMS-2203207, National Institute of Food and Agriculture Hatch project IOW03617, Iowa State University Plant Sciences Institute Scholars Program, National Institute of Food and Agriculture Grant No. 2017-67007-26175.

[Received March 2022. Revised November 2022. Accepted January 2023. Published Online February 2023.]

REFERENCES

AlKhalifah N, Campbell DA, Falcon CM, Gardiner JM, Miller ND, Romay MC, Walls R, Walton R, Yeh C-T, Bohn M et al (2018) Maize genomes to fields: 2014 and 2015 field season genotype, phenotype, environment, and inbred ear image datasets. BMC Res Notes 11:452

Cardot H, Ferraty F, Sarda P (2003) Spline estimators for the functional linear model. Stat Sin 13:571–591

Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc 96:1348–1360

Giraldo R, Delicado P, Mateu J (2010) Continuous time-varying kriging for spatial prediction of functional data: an environmental application. J Agric Biol Environ Stat 15:66–82

Guan T, Lin Z, Cao J (2020) Estimating truncated functional linear models with a nested group bridge approach. J Comput Graph Stat 29:620–628

Hallac, D., Leskovec, J., and Boyd, S. (2015). Network lasso: Clustering and optimization in large graphs. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 387–396

Li C, Li H (2010) Variable selection and regression analysis for graph-structured covariates with an application to genomics. Ann Appl Stat 4:1498

Li F, Sang H (2019) Spatial homogeneity pursuit of regression coefficients for large datasets. J Am Stat Assoc 114:1050–1062

Li X, Guo T, Mu Q, Li X, Yu J (2018) Genomic and environmental determinants and their interplay underlying phenotypic plasticity. Proc Nat Acad Sci 115:6679–6684

Li X, Wang L, Nettleton D (2019) Additive partially linear models for ultra-high-dimensional regression. Stat 8:e223

Lin Z, Cao J, Wang L, Wang H (2017) Locally sparse estimator for functional linear regression models. J Comput Graph Stat 26:306–318

Ma S, Huang J (2017) A concave pairwise fusion approach to subgroup analysis. J Am Stat Assoc 112:410–423

Madrid Padilla OH, Sharpnack J, Chen Y, Witten DM (2020) Adaptive nonparametric regression with the k-nearest neighbour fused lasso. Biometrika 107:293–310

McFarland BA, AlKhalifah N, Bohn M, Bubert J, Buckler ES, Ciampitti I, Edwards J, Ertl D, Gage JL, Falcon CM et al (2020) Maize genomes to fields (g2f): 2014–2017 field seasons: genotype, phenotype, climatic, soil, and inbred ear image datasets. BMC Res Notes 13:1–6

Nelsen TC (2002) The state of statistics in agricultural science. J Agric Biol Environ stat 7:313–319

Nicotra AB, Atkin OK, Bonser SP, Davidson AM, Finnegan EJ, Mathesius U, Poot P, Purugganan MD, Richards CL, Valladares F et al (2010) Plant phenotypic plasticity in a changing climate. Trends Plant Sci 15:684–692

Ogden RT, Miller CE, Takezawa K, Ninomiya S (2002) Functional regression in crop lodging assessment with digital images. J Agric Biol Environ Stat 7:389–402

Shen J, Liu RY, Xie M-G (2020) ifusion: individualized fusion learning. J Am Stat Assoc 115:1251–1267

Tang L, Song PX (2016) Fused lasso approach in regression coefficients clustering: learning parameter heterogeneity in data integration. J Mach Learning Res 17:3915–3937

Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K (2005) Sparsity and smoothness via the fused lasso. J Royal Stat Soc Series B (Stat Methodology) 67:91–108

Tollenaar M, Fridgen J, Tyagi P, Stackhouse PW Jr, Kumudini S (2017) The contribution of solar brightening to the us maize yield trend. Nat Clim Chang 7:275–278

Van de Pol M, Bailey LD, McLean N, Rijsdijk L, Lawson CR, Brouwer L (2016) Identifying the best climatic predictors in ecology and evolution. Methods Ecol Evol 7:1246–1257

VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91:4414–4423

Wang J-L, Chiou J-M, Müller H-G (2016) Functional data analysis. Annu Rev Stat Appl 3:257–295

Wang L, Liu X, Liang H, Carroll RJ (2011) Estimation and variable selection for generalized additive partial linear models. Ann Stat 39:1827

Xue L, Shu X, Qu A (2020) Time-varying estimation and dynamic model selection with an application of network data. Stat Sin 30:251–284

Yu S, Wang G, Wang L, Liu C, Yang L (2020) Estimation and inference for generalized geoadditive models. J Am Stat Assoc 115:761–774

Zhu X, Qu A (2018) Cluster analysis of longitudinal profiles with subgroups. Electron J Stat 12:171–193

Zou H (2006) The adaptive lasso and its oracle properties. J Am Stat Assoc 101:1418–1429