AIMS Mathematics, 10(8): 17518–17542.
DOI: 10.3934/math.2025783
Received: 29 April 2025
Revised: 10 July 2025
Accepted: 18 July 2025
Published: 01 August 2025

https://www.aimspress.com/journal/Math

Research article

Simultaneous clustering and optimization in function-on-scalar regression

Shan Sha[1,*] **and Yan Li**[2]

[1] School of Mathematics and Statistics, Beijing Jiaotong University, Beijing 100044, China
[2] School of Insurance and Economics, University of International Business and Economics, Beijing 100029, China

* **Correspondence:** Email: shansha@bjtu.edu.cn.

Abstract: Predicting a functional response from scalar predictors is challenging, especially with complex data structures. Traditional function-on-scalar regression (FOSR) methods emphasize smoothness or sparsity, but few address group structures in functional data. To address this gap, we introduce the network function-on-scalar Lasso (NFL), an innovative FOSR model that integrates simultaneous clustering and optimization (SCO) principles. The NFL model introduces a graph-structured sum-of-norms regularization to encourage similar functional responses for related observations (e.g., neighboring regions), while also performing sparse variable selection. An efficient semi-proximal alternating direction method of multipliers (ADMM) algorithm is developed for model estimation, scaling to high-dimensional functional data. We provide theoretical guarantees for the NFL estimator under regularity conditions, ensuring model accuracy and insight into its clustering consistency. Simulations and an environmental application predicting US county-level air quality trends demonstrate the NFL's superior prediction accuracy and ability to uncover meaningful group structures compared to existing methods.

Keywords: functional data analysis; function-on-scalar regression; sum of norms regularizer; convergence rate; semi-proximal ADMM
Mathematics Subject Classification: 62R10, 62J07, 65K10

1. Introduction

Functional data analysis (FDA) provides a unified framework for studying observations recorded over a continuum—curves, images, or surfaces—by treating them as elements in infinite-dimensional spaces. Within the FDA, functional regression lies at the core, encompassing scalar-on-function, function-on-scalar, and function-on-function models [1]. Foundational estimation theory was laid by, among others, [2–6], and subsequent surveys charted key methodological advances [7, 8]. More

recently, extensive reviews of non-parametric and semiparametric approaches have appeared [9–11].

Rapid progress in scalar-on-function [12–16] and function-on-function regression [17–19] has set the stage for intensive study of function-on-scalar models. Important themes include non-parametric estimation of mean and covariance functions [20], robust functional mixed modelling [21], and Bayesian discrimination [22]. Central to these developments is regularization, where ideas such as the Lasso [23], Elastic Net [24], and MCP [25] have been adapted to infinite-dimensional settings. Functional counterparts include basis expansion with MCP penalties [26], ℓ_1-sparse coefficient estimators [27], adaptive smooth-sparse procedures [28], and weighted group bridge methods [29]. Broader innovations blend sparsity with semiparametrics [30], deliver consistent variable selection [31], model discrete points of impact [16], extend to additive frameworks [32, 33], accommodate multi-functional predictors [34], and expand shrinkage to categorical covariates [35].

Despite these advances, most functional-on-scalar methods ignore known relational structure among subjects—information naturally represented by a graph and often crucial for interpretability and prediction. Simultaneous clustering and optimization (SCO) addresses this by augmenting a convex loss with a graph-based sum-of-norms penalty, an idea pioneered in the Network Lasso for scalar data [36, 37] and later analyzed for generic [38] and dynamic graphs [39], with inferential theory in [40]. Functional extensions remain limited: [41] introduces penalized clustering for large-scale data, [42] extends graphical lasso to functional setting, AFSSEN adapts SCO to scalar-on-function models [28], and the Smooth-Lasso fuses smoothness and sparsity in a common-surface function-on-function setting [43]. To date, no existing work explicitly fuses subject-specific coefficient functions along a pre-specified graph within a function-on-scalar model. Bridging this gap is the goal of the present study.

Motivated by this gap, we introduce a function-on-scalar linear regression model that directly captures similarities among functional responses. Such structure often comes from external knowledge—geographic proximity, network connections, or hierarchical links between observations. When these relationships are taken into account, the resulting coefficients become easier to interpret and tend to improve predictive accuracy because each scalar predictor can act in a structured way across observations.

Environmental science offers a clear example of why structural similarities matter in functional regression—consider forecasting air quality. Daily air quality index (AQI) curves for each county are functional data, and we often try to explain them with scalar covariates such as population density, industrial activity, or weather. Yet AQI shows strong spatial dependence: neighboring counties tend to share similar temporal patterns. Figure 1 highlights the year-long AQI profiles for several U.S. counties, underscoring temporal dynamics and the need for FDA tools, while Figure 2 maps AQI on a single day, making the spatial clustering obvious. Standard function-on-scalar models can include the covariates, but they rarely encode these spatial ties unless a separate—and often complex—spatial layer is added. Although existing air-quality studies do incorporate space in various ways [44, 45], few embed prior geographic information directly into a functional-regression framework through a principled tool like SCO. Our goal, therefore, is to build a function-on-scalar method that uses a graph of geographical proximity to impose similarity across related counties, yielding more interpretable coefficients and better predictions.

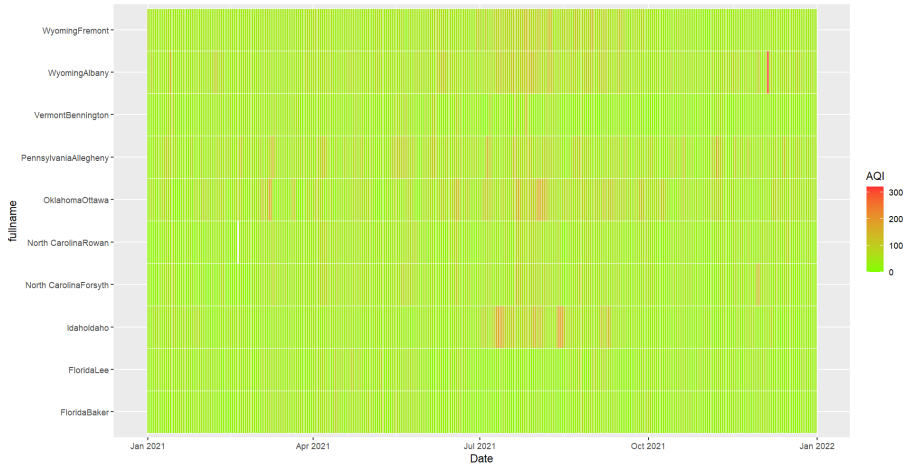

Figure 1. AQI index heatmap for 10 random counties.

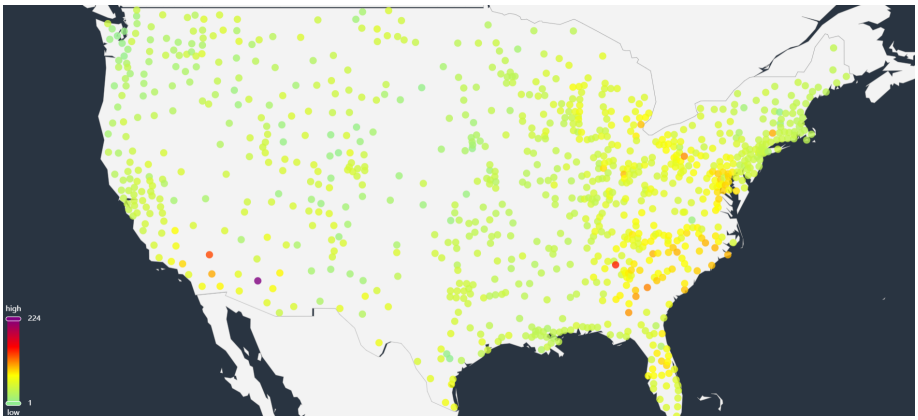

Figure 2. Geographical distribution of AQI.

Beyond the challenge of incorporating structural information, fitting functional regression models, especially with complex regularizers or large datasets, poses a significant computational challenge. The infinite-dimensional nature of functional data, coupled with potentially large numbers of observations and predictors and the structure-inducing penalties, can strain the efficiency of generic optimization solvers like standard implementations of the alternating direction method of multipliers (ADMM) within frameworks such as CVX [46]. Developing efficient and scalable algorithms specifically tailored to the proposed model is therefore essential to meet the need in modern practice.

In summary, this paper makes the following principal contributions:

(1) We extend the graph-fused penalty framework to linear function-on-scalar regression, proposing the network function-on-scalar Lasso (NFL) model. The model leverages prior structural information, represented by a graph, to jointly capture model similarities and improve prediction accuracy.

(2) We establish non-asymptotic error bounds for the proposed NFL estimator in a Hilbert-space setting. This analysis extends the existing error-bound theory for graph-fused lasso from the standard vector setting to the more complex functional coefficient setting.

(3) We implement an efficient semi-proximal alternating direction method of multipliers (sp-ADMM) algorithm specifically designed to solve the NFL optimization problem, addressing the computational challenges associated with large functional datasets.

(4) We demonstrate the model's effectiveness through simulation studies and showcase its practical utility by applying it to spatially linked functional data in a U.S. air quality analysis, highlighting its improved predictive power and interpretability in real-world scenarios.

The remainder of this paper is structured as follows. Section 2 details the formulation of the network function-on-scalar Lasso model. Section 3 presents its theoretical properties and describes the estimation procedure, including the proposed semi-proximal ADMM algorithm. Section 4 reports the findings from our simulation studies, evaluating the performance of NFL and comparing it to other existing methods under various simulated scenarios. Section 5 presents the results of the real-data application, demonstrating its practical application in analyzing air quality data. Concluding remarks and potential future work are discussed in Section 6. Theoretical proofs are provided in the Appendix.

2. Network function-on-scalar Lasso (NFL) model

2.1. Preliminaries

Function-on-scalar regression (FOSR) is an important branch of functional data analysis focused on modeling the relationship between a functional response and scalar covariates. In this setting, we consider n observations and p scalar covariates. Let $\mathcal{H} = L^2(\mathcal{T})$ denote the separable Hilbert space of real, square-integrable functions on a compact interval $\mathcal{T} \subset \mathcal{R}$ (without loss of generality, $\mathcal{T} = [0, 1]$), where $\mathcal{R}$ is real space. For $f, g \in \mathcal{H}$, we equip this space with the inner product $\langle f, g\rangle = \int_{\mathcal{T}} f(t)\, g(t)\, dt$, and induced norm $\|f\|_{\mathcal{H}} = \langle f, f\rangle^{1/2}$. A standard FOSR model for the i-th observation is typically expressed as:

$$y_i = \sum_{d=1}^{p} x_{id}\beta_d + \varepsilon_i, i = 1, 2, \ldots, n. \tag{2.1}$$

Here, $y_i \in \mathcal{H}$ is the functional response (e.g., a curve), $\beta_d \in \mathcal{H}$ are the functional coefficients associated with the d-th scalar covariate $x_{id} \in \mathcal{R}$, and $\varepsilon_i \in \mathcal{H}$ is the functional error term. This traditional formulation assumes that the same set of coefficient functions $\{\beta_1, \ldots, \beta_p\}$ applies to all n observations.

Our approach incorporates a network model to leverage potential group structures or similarities among observations, as inspired by the simultaneous clustering and optimization (SCO) framework. We represent the relationships among the n observations using a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{v_1, \cdots, v_n\}$ is the set of vertices, with each v_i corresponding to the i-th observation, and $\mathcal{E}$ is the set of edges connecting pairs of similar observations:

$$\mathcal{V} = \{v_1, \cdots, v_n\}, \mathcal{E} = \left\{e_{ij}, i, j \in N\right\}, N = \{1, 2, \cdots, n\}.$$

The edge set is typically constructed based on some prior information or data-driven proximity criterion. A common method is to connect each vertex v_i to its K nearest neighbors based on a relevant distance metric (e.g., ℓ_2 distance on sample features or external coordinates like geographical location). An edge e_{ij} exists between vertices i and j if they are deemed similar according to this criterion.

Let $|\mathcal{E}| = m$ be the number of edges in the network. To represent the graph structure and its use in promoting similarity between connected nodes, we define a matrix $Q_{\mathcal{E}} \in \mathcal{R}^{m\times n}$. Each row of $Q_{\mathcal{E}}$ corresponds to an edge $e_{ij} \in \mathcal{E}$. If the k-th row corresponds to the edge connecting vertices i and j, this row is constructed as $(Q_{\mathcal{E}})_k^\top = e_i - e_j$, where $e_i \in \mathcal{R}^n$ is the standard basis vector with a 1 at the i-th position and 0 elsewhere. This matrix effectively captures the differences between connected observations across the network.

2.2. The network function-on-scalar Lasso model

In contrast with the standard FOSR specification (2.1), which assigns the same set of coefficient functions to every subject, the proposed network function-on-scalar Lasso (NFL) allows the coefficients $\beta_d^i \in \mathcal{H}$ to vary by subject so that group-specific relationships can be captured. For subject i, we assume

$$y_i = \sum_{d=1}^{p} x_{id}\beta_d^i + \varepsilon_i, \qquad i = 1, \ldots, n. \tag{2.2}$$

Let $\beta^i = (\beta_1^i, \ldots, \beta_p^i)^\top \in \mathcal{H}^p$ denote the vector of coefficient functions for subject i. When $\beta^1 = \cdots = \beta^n$ the model coincides with (2.1); in general, $\beta^i = \beta^j$ whenever observations i and j belong to the same group.

The NFL model is formulated as an optimization problem designed to estimate the coefficient vectors $\beta^1, \ldots, \beta^n$. It incorporates three components: A squared loss term, a penalty for sparsity, and a sum-of-norms regularizer leveraging the network structure to encourage similarity between coefficients of connected observations. The objective function is

$$\min_{\beta^1,\ldots,\beta^n} \frac{1}{2}\sum_{i\in\mathcal{V}} \left\| y_i - \sum_{d=1}^{p} x_{id}\beta_d^i \right\|_{\mathcal{H}}^2 + \alpha_1 \sum_{i\in\mathcal{V}} \left\|\beta^i\right\|_{\mathcal{H},1} + \alpha_2 \sum_{(i,j)\in\mathcal{E}} \omega_{ij} \left\|\beta^i - \beta^j\right\|_{\mathcal{H},1}, \tag{2.3}$$

where $\left\|\beta^i\right\|_{\mathcal{H},1} = \sum_{d=1}^{p} \left\|\beta_d^i\right\|_{\mathcal{H}}$ is the ℓ_1 norm of the vector of function norms for subject i, $\alpha_1, \alpha_2 > 0$ are regularization parameters, and $\omega_{ij} > 0$ is a weight associated with the edge (i, j), typically inversely proportional to the dissimilarity between observations i and j. The first penalty term induces sparsity in the coefficient functions so that—within each subject's model—only those scalar predictors that meaningfully affect the functional response remain active. The second penalty term involves a sum over all edges in the network $\mathcal{E}$ and penalizes the functional difference between the coefficient vectors of connected subjects. A larger ω_{ij} strongly encourages β^i and β^j to be similar; this enables the model to estimate models customized for each subject while capturing inherent similarities dictated by the network structure.

3. Theoretical properties and computation for NFL

3.1. Statistical guarantees

To establish the theoretical properties of the NFL estimator, we rely on several assumptions:

Assumption 1. The functional responses y_i are independent random elements of a separable Hilbert space $\mathcal{H}$, generated according to the model (2.2).

Assumption 2. The true model is sparse, with S_0 representing the set of non-zero covariates, and the

cardinality of S_0 denoted as q.

Assumption 3. The design matrix $X = (x_{id}) \in \mathcal{R}^{n\times p}$ is deterministic with standardized columns.

Assumption 4. The errors ε_i are independent and identically distributed Gaussian elements with mean 0 and covariance operator C, denoted $\varepsilon_i \sim \mathcal{GP}(0, C)$.

Assumption 5. Let x_i be the i-th row of X, and let $Z = \mathrm{diag}(x_1, \ldots, x_n) \in \mathcal{R}^{n\times np}$ be a block diagonal matrix where each block is x_i viewed as a $1 \times p$ vector. Z satisfies the functional restricted eigenvalue condition $\mathrm{RE_F}(q, \alpha)$ with parameter $\alpha > 0$, such that for all subsets $S \subset \{1, \ldots, np\}$ with $|S| \le q$, we have $\|Zz\|^2_{\mathcal{H}} \ge \alpha n \|z\|^2_{\mathcal{H}}$ for all $z \in \mathcal{H}^{np}$ satisfying $\|z_{S^c}\|_{\mathcal{H},1} \le 3\|z_S\|_{\mathcal{H},1}$.

Remark 1. *Assumptions 1 and 2 are standard in linear regression under scalar settings and have also been explored in various functional settings. Assumptions 3 and 4 are standard in the context of functional regression, as demonstrated in the research of [26, 29].*

Remark 2. *The functional restricted eigenvalue condition* $\mathrm{RE_F}(q, \alpha)$*, as stated in Assumption 5, is an adaptation of the scalar restricted eigenvalue condition, which is a well-established concept in Lasso-related research. It has been proved that the functional restricted eigenvalue condition is no more restrictive than its scalar counterpart in the work of [27].*

Remark 3. *In practical applications, functional data y_i are often observed as discrete measurements $y^i = (y^{i1}, \ldots, y^{it_i}) \in \mathcal{R}^{t_i}$ at time points. When t_i is sufficiently large, these discrete data points can be used to reconstruct or represent the underlying continuous function $y_i \in \mathcal{H}$ through smoothing or basis expansion techniques [47]. Our subsequent computations operate directly with the reconstructed functional objects.*

Under these assumptions, we establish theoretical guarantees for the NFL model.

Theorem 1. *For the design matrix $X \in \mathcal{R}^{n\times p}$ and the block diagonal matrix $Z = \mathrm{diag}(x_1, \ldots, x_n) \in \mathcal{R}^{n\times np}$, if Assumptions 1–5 hold, and the regularization parameters satisfy $\alpha_1 - \alpha_0 - \alpha_2 r > 0$ (the explicit forms of α_0 and r are given in the Appendix), then for any minimizer $\widehat{\beta} = (\mathrm{vec}(\widehat{\beta}^1)^T, \ldots, \mathrm{vec}(\widehat{\beta}^n)^T)^T \in \mathcal{H}^{np}$ of problem* (2.3) *and the true coefficients $\beta^\star$, with high probability, we have the following error bounds:*

$$\left\| Z\left(\widehat{\beta} - \beta^\star\right)\right\|_{\mathcal{H}^n} \le \frac{4\alpha_1\sqrt{q}}{\sqrt{\alpha n}},$$

$$\|\widehat{\beta} - \beta^\star\|_{\mathcal{H},1} \le \frac{4\alpha_1^2 q}{\alpha n\left(\alpha_1 - \alpha_0 - \alpha_2 r\right)},$$

where $\mathrm{vec}(\beta^i) = (\beta^i_1, \ldots, \beta^i_p)^T \in \mathcal{H}^p$.

We emphasize that our proof handle tackles three challenges absent from earlier work: (i) an arbitrary weighted graph Laplacian Q, whose spectrum and heterogeneous degrees invalidate standard Lipschitz arguments; (ii) infinite-dimensional coefficient functions, approximated via FPCA, which require simultaneous control of truncation and fusion errors; and (iii) the joint presence of an ℓ_1 sparsity term and a graph-fusion term, which introduces distinct selection and fusion biases. We established convergence rates in this broader framework. Also, in practice we avoid overly dense graphs, because a large $r = \|Q\|_\infty$ forces α_1 to grow relative to α_2, over-sparsifying the model and potentially discarding important variables.

3.2. Computational implementation

Directly solving the optimization problem (2.3) involving infinite-dimensional functional objects is computationally challenging. Therefore, we employ a data preprocessing step using functional principal component analysis (FPCA) to represent the functional data in a finite-dimensional space.

3.2.1. Functional data representation via FPCA

FPCA is a standard technique for dimension reduction and representation of functional data. For a collection of sample curves $\{y_1, \ldots, y_n\}$ drawn from a process $y \in \mathcal{L}^2[0,1]$, we consider the covariance kernel $C_y(s,t) = \mathrm{Cov}(y(s), y(t))$. By Mercer's Theorem, $C_y(s,t)$ admits a spectral decomposition $C_y(s,t) = \sum_{k=1}^{\infty} \lambda_k e_k(s) e_k(t)$, where λ_k are non-negative eigenvalues ($\lambda_1 \geq \lambda_2 \geq \ldots$) and $e_k(t)$ are orthonormal eigenfunctions forming a complete orthonormal system (CONS) in $\mathcal{L}^2[0,1]$ [48]. Estimation of e_k and λ_k is based on estimating $C_y(s,t)$ [49, 50].

The Karhunen–Loève expansion allows each functional object y_i to be represented as a linear combination of the eigenfunctions:

$$y_i = \sum_{k=1}^{\infty} \langle y_i, e_k \rangle e_k = \sum_{k=1}^{\infty} s_{ik} e_k, \tag{3.1}$$

similarly, the functional coefficients β_d^i can be expanded:

$$\beta_d^i = \sum_{k=1}^{\infty} \langle \beta_d^i, e_k \rangle e_k = \sum_{k=1}^{\infty} b_{dk}^i e_k, \tag{3.2}$$

where $s_{ik} = \langle y_i, e_k \rangle$ and $b_{dk}^i = \langle \beta_d^i, e_k \rangle$ are the FPC scores. For practical computation, we truncate these infinite series expansions at a chosen number of components, K:

$$y_i \approx \sum_{k=1}^{K} s_{ik} e_k, \quad \beta_d^i \approx \sum_{k=1}^{K} b_{dk}^i e_k. \tag{3.3}$$

The truncation is justified by the property that the truncated expansion provides the best K-dimensional approximation [51]. The number of components K is typically selected to explain a large proportion of the total variance using the criterion:

$$\mathrm{PVE}(K) = \frac{\sum_{k=1}^{K} \lambda_k}{\sum_{k=1}^{\infty} \lambda_k}. \tag{3.4}$$

In most cases, a PVE of 95%-99% can be achieved with small K. This allows us to work with low-dimensional vectors of FPC scores.

Following the FPCA procedure and employing the representation in (3.3), each functional object in model (2.3) can be represented by its corresponding FPC score vector. Substituting these into (2.3), the functional ℓ_2 norms become Euclidean ℓ_2 norms of the score vectors, and the optimization problem transforms into a finite-dimensional problem in terms of the scores b_{dk}^i:

$$\min_{\{b_{dk}^i\}} \frac{1}{2} \sum_{i \in \mathcal{V}} \sum_{k=1}^{K} \left(s_{ik} - \sum_{d=1}^{p} x_{id} b_{dk}^i \right)^2 + \alpha_1 \sum_{i \in \mathcal{V}} \sum_{d=1}^{p} \left\| \mathbf{b}_d^i \right\|_2 + \alpha_2 \sum_{(i,j) \in \mathcal{E}} \omega_{ij} \left(\sum_{d=1}^{p} \left\| \mathbf{b}_d^i - \mathbf{b}_d^j \right\|_2 \right), \tag{3.5}$$

where $\mathbf{b}_d^i = (b_{d1}^i, \ldots, b_{dK}^i)^T \in \mathcal{R}^K$ is the vector of FPC scores for the d-th functional coefficient of sample i.

The notation that follows involves several block-structured matrices. For ease of reference we collect the most important objects, their sizes and, their roles in Table 1; symbols not listed are standard and explained in the text.

Table 1. Key matrices and their roles in the NFL optimization.

Symbol	Dimension	Definition	Purpose
s_i	$1 \times K$	$(s_{i1}, \ldots, s_{iK})^T$	response FPC scores for sample i
Y	$n \times K$	$(s_1^\top, \ldots, s_n^\top)^\top$	response FPC scores
Z	$n \times np$	$\mathrm{diag}(x_1, \ldots, x_n)$	block-diagonal design
B^i	$p \times K$	$(\mathbf{b}_1^i, \ldots, \mathbf{b}_p^i)^T$	matrix of coefficient scores for sample i
B	$np \times K$	$((B^1)^T, \ldots, (B^n)^T)^T$	matrix of coefficient scores
$Q_\mathcal{E}$	$m \times n$	graph incidence	node differences
$Q'_\mathcal{E}$	$mp \times np$	$I_p \otimes Q_\mathcal{E}$	predictor-wise fusion
U	$np \times K$	auxiliary	sparsity penalty
V	$mp \times K$	$Q'_\mathcal{E} B$	fusion penalty
Λ	$np \times K$	dual variables	ADMM duals
M	$mp \times K$	dual variables	ADMM duals

For computational efficiency, we express (3.5) in matrix form. The term $\sum_{d=1}^p x_{id}\mathbf{b}_d^i$ is the i-th row of ZB; the penalty terms in (3.5) involve norms of rows of B. Specifically, $\mathbf{b}_d^i$ is the $(p(i-1)+d)$-th row of B, denoted $B_{p(i-1)+d,:}$. The term $\sum_{i\in\mathcal{V}}\sum_{d=1}^p \left\|\mathbf{b}_d^i\right\|_2$ is $\sum_{r=1}^{np}\|B_{r,:}\|_2$. The difference $\mathbf{b}_d^i - \mathbf{b}_d^j$ corresponds to the difference between row $p(i-1)+d$ and row $p(j-1)+d$ of B. Let $e_r \in \mathcal{R}^{np}$ be the r-th standard basis vector. The vector difference is $(e_{p(i-1)+d} - e_{p(j-1)+d})^T B$. The term $\sum_{(i,j)\in\mathcal{E}}\omega_{ij}\sum_{d=1}^p \left\|\mathbf{b}_d^i - \mathbf{b}_d^j\right\|_2$ is $\sum_{(i,j)\in\mathcal{E}}\omega_{ij}\sum_{d=1}^p \|(e_{p(i-1)+d} - e_{p(j-1)+d})^T B\|_2$. This leads to the equivalent matrix optimization problem:

$$\min_{B\in\mathcal{R}^{np\times K}} \frac{1}{2}\|Y - ZB\|_F^2 + \alpha_1 \sum_{r=1}^{np}\left\|B_{r,:}\right\|_2 + \alpha_2 \sum_{(i,j)\in\mathcal{E}}\omega_{ij}\sum_{d=1}^p \|(e_{p(i-1)+d} - e_{p(j-1)+d})^T B\|_2. \tag{3.6}$$

3.2.2. Optimization algorithm: Semi-proximal ADMM

Solving the optimization problem (3.6) is challenging because it contains two coupled $\ell_{2,1}$-norm penalties, particularly for large n and p. We implement an efficient scheme based on the semi-proximal alternating direction method of multipliers (sp-ADMM) to address this challenge.

We reformulate problem (3.6) into a constrained optimization problem suitable for ADMM. The penalties can be written using auxiliary variables U for the $\ell_{2,1}$ row norm penalty and V for the network penalty. Let $P_1(U) = \alpha_1 \sum_{r=1}^{np}\left\|U_{r,:}\right\|_2$, $P_2(V) = \alpha_2 \sum_{t=1}^{mp}\omega'_t\|V_{t,:}\|_2$, where ω'_t are the corresponding weights ω_{ij} associated with each stacked difference. The problem (3.6) becomes:

$$\begin{aligned} &\min_{B,U,V} \frac{1}{2}\|Y - ZB\|_F^2 + P_1(U) + P_2(V), \\ &\text{s.t. } B - U = 0, \end{aligned}$$

$$Q'_{\mathcal{E}}B - V = 0. \tag{3.7}$$

The augmented Lagrangian function for (3.7) with penalty parameter $\sigma > 0$ and dual variables Λ, M is:

$$\begin{aligned} L_\sigma(B, U, V; \Lambda, M) =& \frac{1}{2}\|Y - ZB\|_F^2 + P_1(U) + P_2(V) + \langle \Lambda, B - U\rangle + \langle M, Q'_{\mathcal{E}}B - V\rangle \\ &+ \frac{\sigma}{2}\|B - U\|_F^2 + \frac{\sigma}{2}\left\|Q'_{\mathcal{E}}B - V\right\|_F^2. \end{aligned}$$

The sp-ADMM algorithm iteratively updates B, (U, V), and the dual variables. Let $W = (U, V)$ and $\Psi = (\Lambda, M)$. With a step length $\tau > 0$ and positive semi-definite matrices $\mathcal{S}, \mathcal{T}$ for the proximal terms, the updates are (following [52, 53]):

$$\begin{cases} B^{k+1} = \text{argmin}_B \left\{ L_\sigma\left(B, W^k; \Psi^k\right) + \frac{\sigma}{2}\|B - B^k\|_{\mathcal{S}}^2 \right\}, \\ W^{k+1} = \text{argmin}_W \left\{ L_\sigma\left(B^{k+1}, W; \Psi^k\right) + \frac{\sigma}{2}\|W - W^k\|_{\mathcal{T}}^2 \right\}, \\ \Lambda^{k+1} = \Lambda^k + \tau\sigma\left(B^{k+1} - U^{k+1}\right), \\ M^{k+1} = M^k + \tau\sigma\left(Q'_{\mathcal{E}}B^{k+1} - V^{k+1}\right), \end{cases} \tag{3.8}$$

where $\|B - B^k\|_{\mathcal{S}}^2 := \text{trace}((B - B^k)^T\mathcal{S}(B - B^k))$. The $\mathcal{T}$ proximal term on $W = (U, V)$ decomposes due to the separability of $P_1(U)$ and $P_2(V)$.

The B-subproblem update in (3.8) is:

$$\begin{aligned} B^{k+1} &= \text{argmin}_B \left\{ L_\sigma\left(B, W^k; \Psi^k\right) + \frac{\sigma}{2}\|B - B^k\|_{\mathcal{S}}^2 \right\} \\ &= \text{argmin}_B \frac{1}{2}\|Y - ZB\|_F^2 + \frac{\sigma}{2}\|B - U^k + \Lambda^k/\sigma\|_F^2 + \frac{\sigma}{2}\|Q'_{\mathcal{E}}B - V^k + M^k/\sigma\|_F^2 + \frac{\sigma}{2}\|B - B^k\|_{\mathcal{S}}^2. \end{aligned} \tag{3.9}$$

This strongly convex problem has a unique solution

$$B^{k+1} = C^{-1}\{Z^TY + \sigma\left(U^k + (Q'_{\mathcal{E}})^TV^k + \mathcal{S}B^k\right) - \Lambda^k - (Q'_{\mathcal{E}})^TM^k\}, \tag{3.10}$$

where $C = Z^TZ + (np\sigma + \sigma)I_{np} - \sigma 1_{np}1_{np}^T$ and $\mathcal{S} = npI_{np} - 1_{np}1_{np}^T - ((Q'_{\mathcal{E}})^TQ'_{\mathcal{E}}) \succeq 0$.

The W-subproblem in (3.8) separates into independent updates for U and V. Similarly, the closed-form solution can be given by row-wise group soft-thresholding:

$$\begin{aligned} U_{r,:}^{k+1} &= \max\{1 - \frac{\alpha_1}{\sigma\|B_{r,:}^{k+1} + \Lambda_{r,:}^k/\sigma\|_2}, 0\}(B_{r,:}^{k+1} + \Lambda_{r,:}^k/\sigma); \\ V_{t,:}^{k+1} &= \max\{1 - \frac{\alpha_2\omega'_t}{\sigma\|(Q'_{\mathcal{E}}B^{k+1})_t + M_t^k/\sigma\|_2}, 0\}((Q'_{\mathcal{E}}B^{k+1})_t + M_t^k/\sigma). \end{aligned} \tag{3.11}$$

The dual variables Λ and M are updated via standard ADMM steps based on the residuals of the constraints, as shown in (3.8). Algorithm 1 summarizes the sp-ADMM procedure. We set $\tau = 1.618$ and a stopping tolerance of 10^{-3}.

Remark 4. *The convergence of Algorithm 1 to an optimal solution of problem* (3.7) *(and thus* (3.6)*) is guaranteed under standard conditions for sp-ADMM for convex problems with affine constraints and separable objectives [54].*

Algorithm 1 Semi-Proximal ADMM for solving Problem (3.7)

Choose parameters $\tau > 0$, $\sigma > 0$, tolerance $tol = 10^{-3}$, maximum iterations $k_{max} = 10000$. Initialize $k = 0$.
Initialize $B^0, U^0, \Lambda^0 \in \mathcal{R}^{np\times K}$, $V^0, M^0 \in \mathcal{R}^{mp\times K}$.
Precompute $C^{-1} = (Z^TZ + \sigma I + \sigma(Q'_{\mathcal{E}})^T Q'_{\mathcal{E}} + \sigma\mathcal{S})^{-1}$.
repeat
    $k \leftarrow k + 1$
    Update B: $B^{k+1} = C^{-1}\left\{Z^TY + \sigma\left(U^k + (Q'_{\mathcal{E}})^T V^k + \mathcal{S}B^k\right) - \Lambda^k - (Q'_{\mathcal{E}})^T M^k\right\}$
    for $r = 1$ to np **do**
        Update U: $U^{k+1}_{r,:} = \max\left\{1 - \frac{\alpha_1}{\sigma\|B^{k+1}_{r,:} + \Lambda^k_{r,:}/\sigma\|_2}, 0\right\}\left(B^{k+1}_{r,:} + \Lambda^k_{r,:}/\sigma\right)$
    end for
    for $t = 1$ to mp **do**
        Update V: $V^{k+1}_{t,:} = \max\left\{1 - \frac{\alpha_2\omega'_t}{\sigma\|(Q'_{\mathcal{E}}B^{k+1})_t + M^k_t/\sigma\|_2}, 0\right\}\left((Q'_{\mathcal{E}}B^{k+1})_t + M^k_t/\sigma\right)$
    end for
    Update dual variables:
    $\Lambda^{k+1} = \Lambda^k + \tau\sigma\left(B^{k+1} - U^{k+1}\right)$,
    $M^{k+1} = M^k + \tau\sigma\left(Q'_{\mathcal{E}}B^{k+1} - V^{k+1}\right)$.
    Compute residuals: $r^B_p = B^{k+1} - U^{k+1}$, $r^Q_p = Q'_{\mathcal{E}}B^{k+1} - V^{k+1}$, $r^B_d = B^{k+1} - B^k$.
until $\max(\|r^B_p\|_F, \|r^Q_p\|_F) \le tol$ and $\|r^B_d\|_F \le tol$, or $k \ge k_{max}$.

4. Simulation studies

In this section, we conduct extensive simulation studies to evaluate the finite-sample performance of the proposed network function-on-scalar Lasso (NFL) regression model and compare it comprehensively with several existing methods.

4.1. Experimental settings

The simulation study consists of four distinct scenarios designed to assess model performance under varying data characteristics. In each scenario, we generate synthetic data comprising predictor variables $X \in \mathcal{R}^{n\times p}$, true functional coefficients $\beta^{*i} \in \mathcal{H}^p$ for each observation i, functional error terms $\varepsilon_i \in \mathcal{H}$, and functional response variables $y_i \in \mathcal{H}$. Functional data are simulated by generating values at 50 equally spaced time points and then converting them to functional objects using basis expansion techniques available in the 'fdaM' package.

Scenario 1: Low-dimensional data. Scenario 1 focuses on evaluating performance in low-dimensional settings (p not significantly larger than n), exploring different relative sizes of n and p. The predictor variables $X \in \mathcal{R}^{n\times p}$ are generated from a multivariate normal distribution with zero mean. The covariance structure between variables $X_{i,d}$ and $X_{i,d'}$ for a given observation i is set as $Cov(X_{i,d}, X_{i,d'}) = \rho^{|d-d'|}$, implementing an AR(1) dependence with $\rho = 0.5$. Predictor variables are independent across observations. We vary the sample size $n \in \{30, 60, 90, 120, 150\}$ with either a fixed $p = 20$ or $p = n/5$, the sparsity level is set to be 0.1. The true functional coefficients β^{*i} are set to be the same ($\beta^{*i} = \beta^*$ for all i) and generated using a Matérn process kernel with parameters $\nu = 2.5$ and

$z = 0.25$. The corresponding covariance function is $C_{\frac{5}{2}}(d) = \sigma^2\left(1 + \frac{\sqrt{5}d}{\rho} + \frac{5d^2}{3r^2}\right)\exp\left(-\frac{\sqrt{5}d}{z}\right)$, where d represents the distance in the time domain (typically 1). The functional error terms ε_i are generated independently from a Matérn process with parameters $\nu = 1.5$ and $z = 0.25$. The functional responses y_i are then computed based on the linear model $y_i = \sum_{d=1}^{p} x_{id}\beta_d^* + \varepsilon_i$.

Scenario 2: Low-dimensional data with group structure. Scenario 2 evaluates model performance when the data have a built-in group structure—i.e., the true coefficient functions differ across groups—under both homoskedastic and heteroskedastic noise. Observations are randomly partitioned into $G = 3$ equally sized groups. For each group $g = 1, \ldots, 3$, the predictor variables $X^{(g)} \in \mathcal{R}^{n_g \times p}$ are generated from a uniform distribution $U(g-1, g)$, where $n_g = n/3$ is the size of group g. The true coefficient functions β^{*i} are specific to each group g, i.e., $\beta^{*i} = \beta^{*(g)}$ for all i in group g. Each group-specific coefficient function set $\{\beta_d^{*(g)}\}_{d=1}^{p}$ is generated similarly to Scenario 1 (using a Matérn kernel with $\nu = 2.5, z = 0.25$) but independently across groups. To study noise robustness, two error schemes are considered: a homoskedastic design where all ε_i are independent Matérn functions, and a heteroskedastic design where each group has its own independent generation of Matérn functions. The response y_i for an observation i in group g is $y_i = \sum_{d=1}^{p} x_{id}\beta_d^{*(g)} + \varepsilon_i$. Sample sizes considered are $n \in \{30, 60, 90, 120, 150\}$, with $p = 20$.

Scenario 3: High-dimensional data with group structure. Scenario 3 extends the group design of Scenario 2 to a high-dimensional setting ($p \geq n$). We set the total number of predictors to $p = 200$. To introduce sparsity, only $p_0 = 20$ randomly selected predictors are set to have non-zero coefficients. For these active predictors we generate the data exactly as in Scenario 2—using the Matérn kernel and the same two noise schemes (homoskedastic vs. heteroskedastic). The response y_i is computed based on this sparse model. Sample sizes $n \in \{30, 60, 90, 120, 150\}$ mirror those in Scenario 2.

For Scenarios 2 and 3, the network graph $\mathcal{G}$ used in the NFL model is constructed based on the true group memberships: an edge exists between any two observations i and j if and only if they belong to the same group. The weights ω_{ij} are set to 1 for existing edges and 0 otherwise.

4.2. Competing methods and evaluation criteria

We compare the proposed NFL model with four benchmark methods:

(1) Group Lasso (GLasso): A standard multivariate regression method with an $\ell_{2,1}$ penalty on coefficients grouped by predictor. Solved using the groupwise majorization-descent (GMD) algorithm [55]. This method does not account for functional structure but considers grouping predictors.

(2) Network Lasso (NLasso): A standard multivariate regression method with an ℓ_2-norm penalty on differences of coefficients across a predefined network, solved via ADMM [36]. This method accounts for the network structure but not the functional structure.

(3) Functional Principal Component Analysis (FPCA): Applies standard linear regression on the estimated FPC scores of the response, treating scores as the response variables and using scalar predictors X. Solved using the 'PACE' package for FPCA and subsequent regression. This method accounts for the functional structure through dimension reduction but does not incorporate sparsity or the network structure explicitly in the regression coefficients.

(4) Function-on-Scalar Lasso (FSL): A functional regression method that uses an $\ell_{2,1}$ penalty on the vector of coefficient functions $\beta_d \in \mathcal{H}$, assuming a common β_d for all observations, as in (2.1). Implemented following the strategy of [27] and solved using the CVX framework. This method handles functional structure and sparsity across predictors but assumes homogeneous coefficients across observations.

(5) Network Function-on-Scalar Lasso (NFL): The proposed model incorporates functional structure, sparsity, and the network-induced grouping structure. Solved using the proposed semi-proximal ADMM algorithm introduced in Section 3.

Finite-sample performance is gauged by four criteria: root mean-squared prediction error (RMSP), the Adjusted Rand Index (ARI), the Normalized Mutual Information (NMI), and computation time. RMSP compares the predicted functional response with the truth via

$$\text{RMSP} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{T}}\left(\hat{y}_i(t) - \sum_{d=1}^{p} x_{id}\beta_d^{*i}(t)\right)^2 dt}. \tag{4.1}$$

ARI assesses pairwise label agreement with chance correction ($-1 \le \text{ARI} \le 1$), while NMI quantifies the mutual information shared by the true and estimated partitions on a $0-1$ scale; larger values for either indicate more accurate clustering. For methods that do not natively output cluster labels, we first apply k-means (with the true number of groups) to their predicted response curves and then compute ARI and NMI from the resulting labels.

4.3. Results

The simulation results are presented in Figures 3–7 and Tables 2–3, based on 50 replications for each setting.

Result 1: Scenario 1 Performance (Low-dimensional). In Scenario 1, we evaluate performance when the true coefficient functions are homogeneous across all observations (i.e., no group structure). The prediction accuracy results are shown in Figure 3, where shaded regions represent the standard error across replications. The results show that NFL, FSL, and NLasso achieve significantly better prediction accuracy than the FPCA and GLasso methods. The potential disadvantage for FPCA is that its response-driven dimension reduction can discard predictor-relevant signals, while GLasso's penalty may introduce bias by shrinking unrelated coefficients toward each other. As expected, prediction accuracy increases monotonically with larger sample size N; the accompanying reduction in standard error confirms greater stability at larger N.

Result 2: Scenario 2 Performance (Low-dimensional Group Structure). When the coefficient functions are group-specific, Figure 4 shows that NFL achieves the lowest RMSP for every sample size and under both homoskedastic and heteroskedastic noise. FSL ranks second, whereas FPCA records the highest errors. Heteroskedastic noise raises RMSP for all methods, confirming its added difficulty. The NFL's narrow shaded band further indicates the most stable performance. Table 2 reports clustering accuracy: NFL consistently attains high ARI and NMI, FSL produces weaker but still usable partitions, and FPCA, GLasso, and NLasso exhibit virtually no clustering ability.

Result 3: Scenario 3 Performance (High-dimensional Group Structure). Under the high-dimensional setting, as shown in Figure 5, NFL consistently delivers the lowest RMSP for every

sample size and for both noise schemes. FSL remains the runner-up, while FPCA performs worst; notably, FPCA's error increases with larger N, suggesting overfitting. Table 3 confirms NFL's strong clustering ability across all N and noise configurations, FSL ranks next, and the remaining methods provide substantially poorer clustering. Figure 6 reports computation time with standard-error bands. NFL and FPCA are the fastest across all settings, and NFL scales more favorably as N increases. The other methods require substantially more time, with heteroskedastic noise imposing a particularly large overhead on GLasso.

Result 4: Influence of FPCA Components K. Figure 7 shows how the number of FPCA components retained in NFL affects performance. An intermediate range of K is optimal: Very small K misses relevant signals and yields larger RMSP, whereas overly large K introduces noisy or irrelevant directions, degrading accuracy and raising computation time. Heteroskedastic noise shifts the curve upward and widens the error bands, indicating generally higher RMSP and greater variability.

Regarding the assumptions in Section 3, from the simulations with heteroskedastic errors, we can find that NFL still converges, albeit more slowly. When sparsity or the restricted-eigenvalue condition is violated, key covariates may be missed, which in turn slows convergence and can introduce bias. In such cases the loss can be replaced by a robust alternative (e.g., the Huber loss), or an adaptive weighting scheme can be introduced to mitigate the impact of these violations.

In summary, the simulations show that NFL delivers the best prediction accuracy across all scenarios—especially when coefficients are group-specific or the problem is high-dimensional and sparse—while also outpacing competing methods in clustering quality and computation time.

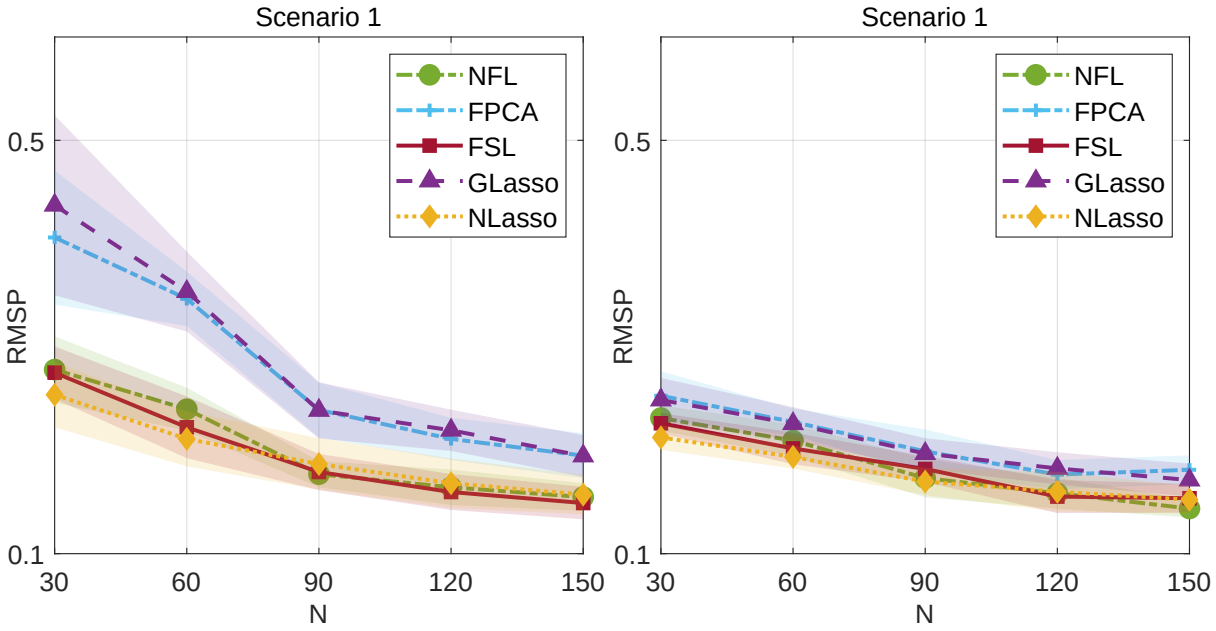

Figure 3. Scenario 1 RMSP: $p = 20$ (left) vs. $p = n/5$ (right).

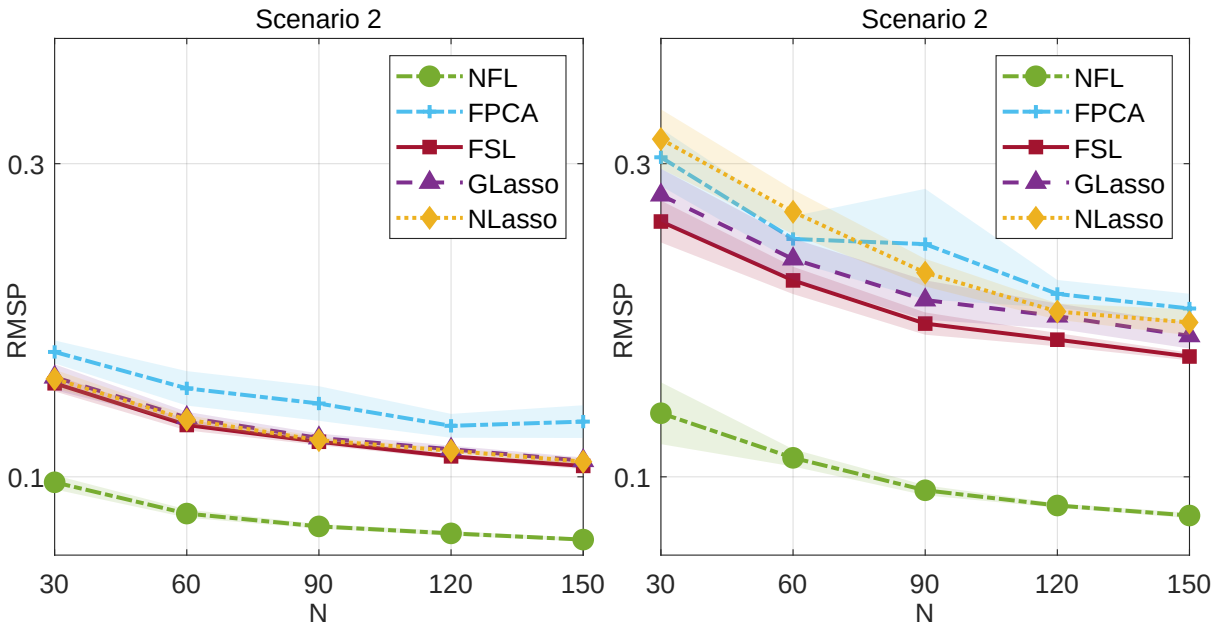

Figure 4. Scenario 2 RMSP: homoskedastic noise (left) vs. heteroskedastic noise (right).

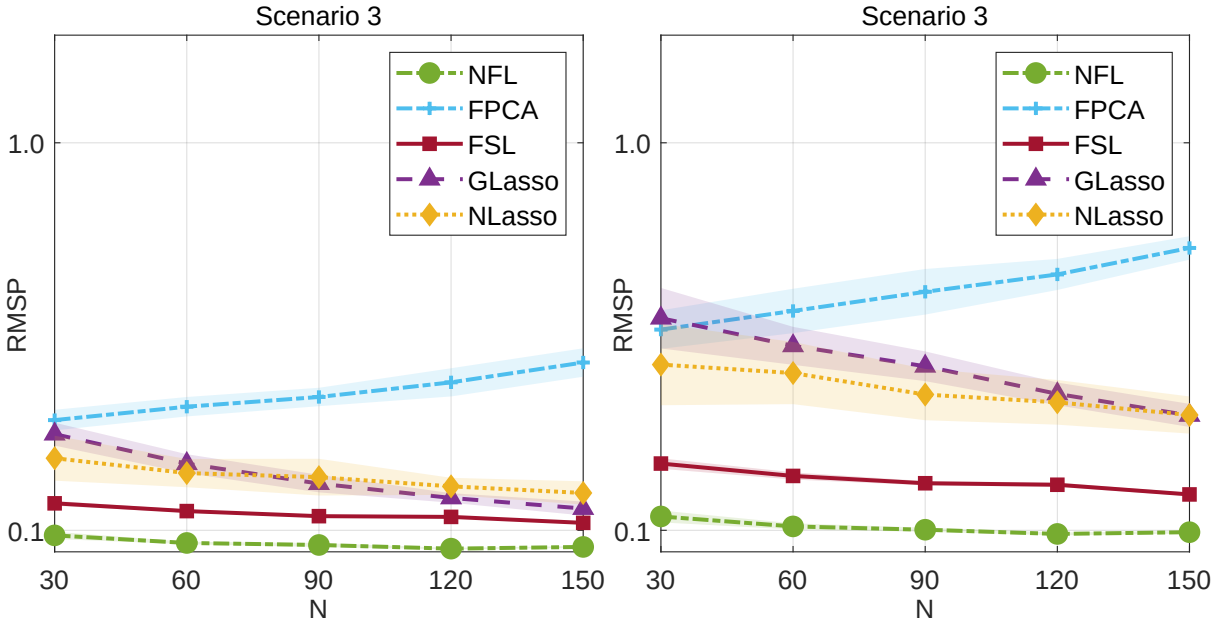

Figure 5. Scenario 3 RMSP: homoskedastic noise (left) vs. heteroskedastic noise (right).

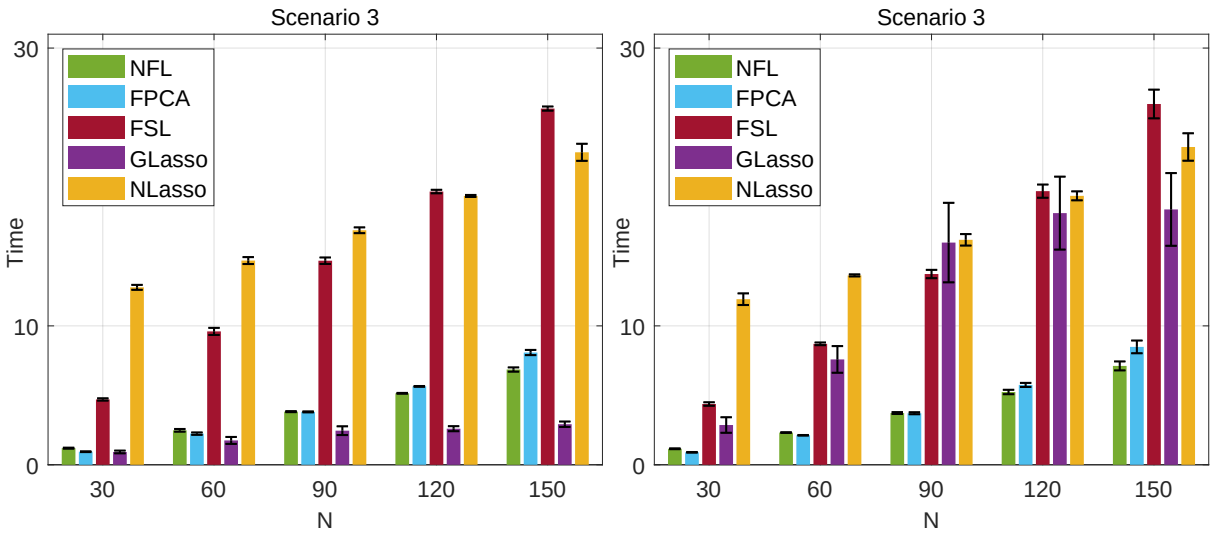

Figure 6. Scenario 3 runtime: homoskedastic noise (left) vs. heteroskedastic noise (right).

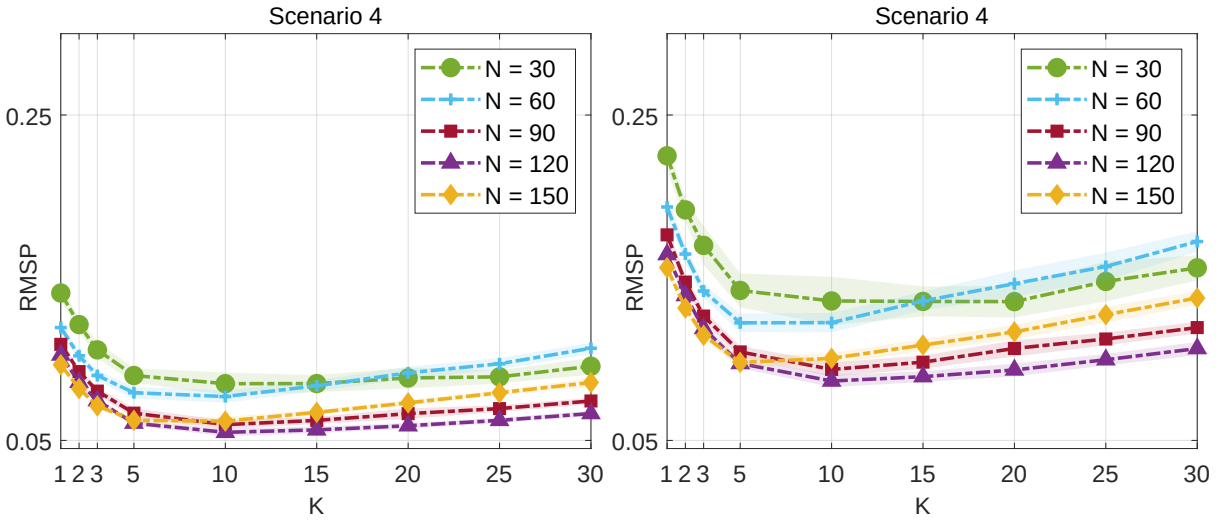

Figure 7. NFL RMSP: homoskedastic noise (left) vs. heteroskedastic noise (right).

Table 2. Clustering performance in Scenario 2.

		$N = 30$		$N = 60$		$N = 90$		$N = 120$		$N = 150$	
		ARI	NMI	ARI	NMI	ARI	NMI	ARI	NMI	ARI	NMI
Hom.	NFL	0.166	0.271	0.305	0.350	0.576	0.599	0.713	0.699	0.949	0.934
	FPCA	0.003	0.079	0.003	0.044	0.002	0.026	0.011	0.030	0.004	0.017
	FSL	0.160	0.260	0.427	0.445	0.611	0.631	0.744	0.715	0.611	0.716
	GLasso	0	0.109	0	0.063	0	0.043	0	0.039	0	0.031
	NLasso	0.051	0.152	0.173	0.205	0.346	0.396	0.425	0.449	0.468	0.520
Het.	NFL	0.050	0.222	0.227	0.297	0.383	0.458	0.380	0.452	0.402	0.455
	FPCA	0.070	0.200	0.071	0.169	0.083	0.167	0.094	0.174	0.089	0.093
	FSL	0.009	0.082	0	0.003	0.127	0.162	0.054	0.083	0.123	0.151
	GLasso	0.038	0.142	0.012	0.078	0.006	0.050	0.008	0.046	0.015	0.050
	NLasso	0.134	0.330	0.121	0.263	0.117	0.225	0.129	0.229	0.125	0.215

Table 3. Clustering performance in Scenario 3.

		$N = 30$		$N = 60$		$N = 90$		$N = 120$		$N = 150$	
		ARI	NMI	ARI	NMI	ARI	NMI	ARI	NMI	ARI	NMI
Homo.	NFL	0.445	0.563	0.520	0.633	0.508	0.630	0.568	0.673	0.691	0.739
	FPCA	0.014	0.089	0.001	0.033	0.002	0.022	0	0.015	0.002	0.015
	FSL	0.529	0.697	0.514	0.686	0.543	0.706	0.562	0.721	0.544	0.701
	GLasso	0	0.077	0.002	0.065	0.002	0.052	0.002	0.048	0.001	0.039
	NLasso	0.012	0.090	0.002	0.032	0.009	0.030	0.000	0.017	0.011	0.013
Hetero.	NFL	0.520	0.642	0.529	0.637	0.535	0.651	0.578	0.676	0.667	0.732
	FPCA	0.085	0.210	0.076	0.153	0.078	0.147	0.081	0.154	0.105	0.177
	FSL	0.483	0.676	0.509	0.684	0.535	0.701	0.562	0.721	0.550	0.704
	GLasso	0.056	0.164	0.002	0.059	0.011	0.056	0.010	0.045	0.010	0.039
	NLasso	0.097	0.244	0.082	0.172	0.104	0.186	0.108	0.188	0.103	0.179

5. Real data analysis

In this section, we apply the NFL model to analyze Air Quality Index (AQI) data for counties in the conterminous United States. The dataset consists of daily AQI records for various counties during 2021, obtained from the U.S. Environmental Protection Agency (US EPA)*. We incorporate several scalar predictors, including geographical coordinates (longitude and latitude) and socio-economic indicators from ICPSR†, which may influence AQI levels. The objective is to explore the relationship between the daily AQI curves (functional response) and these scalar predictors.

*https://www.epa.gov/outdoor-air-quality-data
†https://www.icpsr.umich.edu/web/pages/

Importantly, we utilize the geographical information (longitude and latitude) to construct the network graph among observations for the sum-of-norms regularizer in the NFL model, thereby integrating spatial considerations directly into the analysis framework.

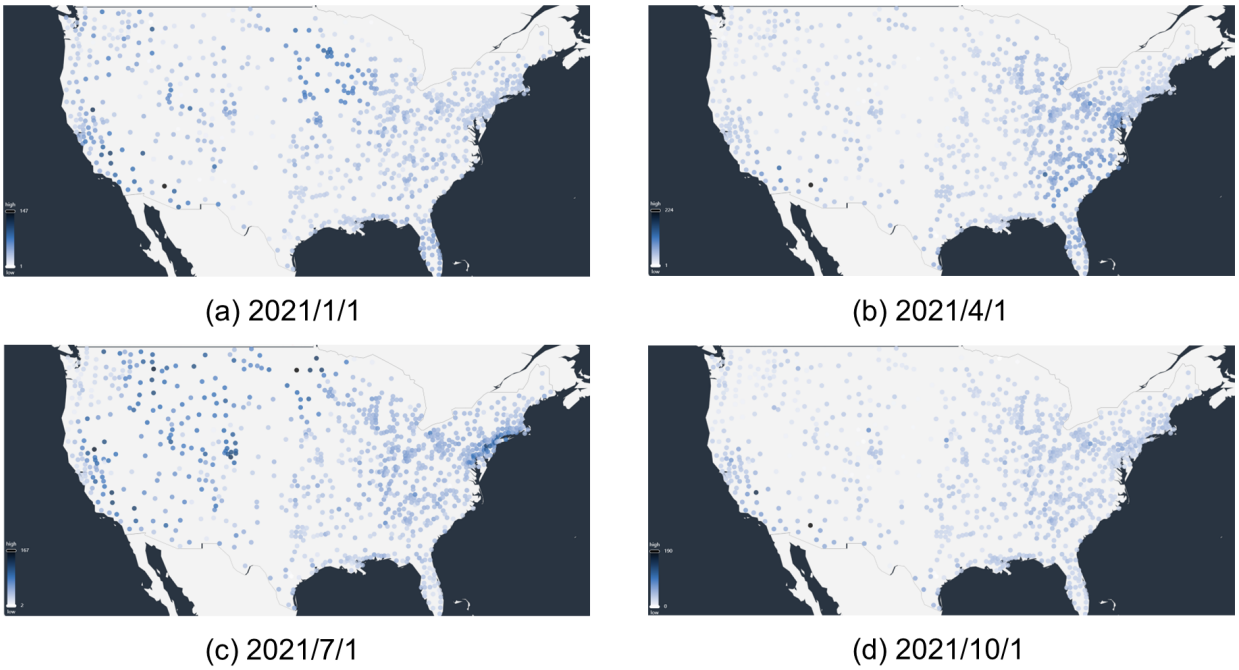

Figure 8. County-level AQI across the Conterminous United States on four selected day.

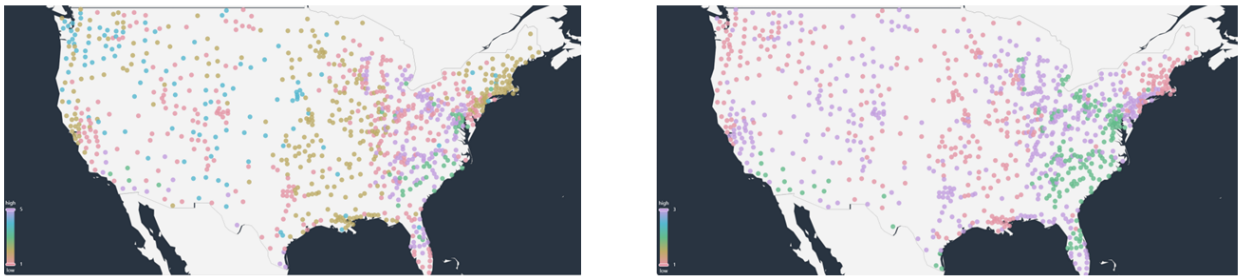

Figure 9. County clusters from NFL: $\alpha_2 = 3$ (left) vs $\alpha_2 = 5$ (right).

Figure 8 visualizes the spatial distribution of AQI values on four specific days throughout 2021. Each point represents a county, colored by its AQI value intensity. This visualization supports two key observations: Firstly, there is a discernible geographic pattern, where nearby counties tend to exhibit similar AQI values, suggesting spatial dependency and potential shared underlying models. Secondly, air quality patterns evolve temporally, with noticeable changes across January, April, July, and October, indicating non-stationarity or seasonal effects in the functional response.

To analyze the patterns, we first construct a K-nearest-neighbors graph from the latitude-longitude coordinates, yielding the network. To respect spatial autocorrelation, we adopt spatial leave-one-region-out cross-validation [56]: The data are partitioned into geographic blocks, and each fold omits

one entire block for testing. Within every training split, the tuning parameters α_1, α_2 and k are selected by five-fold cross-validation; all remaining settings follow Algorithm 1. The scalar predictors X are used to predict the daily AQI curves y_i. We apply the NFL model to the training data and evaluate performance on the testing data. Figure 9 shows the estimated clustering of counties (those assigned the same coefficient model $\hat{\beta}^i$) under different values of the network penalty parameter α_2. A larger α_2 imposes a stronger penalty on differences between coefficient functions of connected counties, leading to more aggressive merging of models and thus fewer estimated clusters. As shown, increasing α_2 from 3 to 5 reduces the number of estimated clusters, illustrating the parameter's role in controlling the granularity of spatial grouping.

Figure 10 displays the estimated functional coefficient curves for two selected scalar predictors, population and computer ownership rate, within the four clusters (M1–M4) identified when $\alpha_2 = 10$. For population, the effect differs by region: Northeast clusters M1 and M2 have a mainly negative link with AQI, whereas the Northwest cluster M4 is positive, likely reflecting regional industry, urban form, or climate. For computer ownership, clusters M1–M3 again give a negative effect—higher ownership, and thus greater urbanization, coincides with poorer air quality—while M4 is positive. The contrasting curves underscore NFL's ability to uncover region-specific relationships that a single global model would miss.

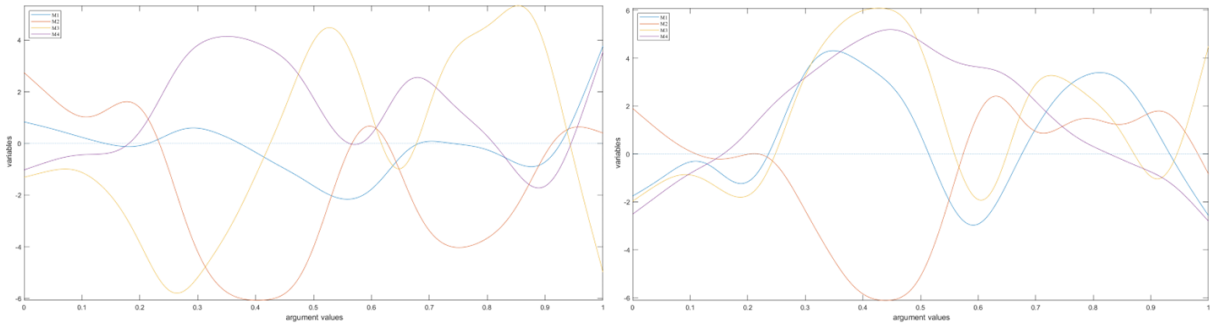

Figure 10. Cluster coefficients: Population (left) vs. Computer ownership (right).

Finally, we evaluated prediction accuracy on the real data using 100 replications. As Table 4 shows, NFL yields the lowest RMSP, the smallest standard error, and the shortest run-time of all methods.

Table 4. RMSP and Runtime (standard error) for five methods.

	NFL	FPCA	FSL	GLASSO	NLASSO
RMSP	0.330 (0.004)	1.982 (0.181)	0.493 (0.007)	0.448 (0.006)	0.644 (0.017)
Time (s)	0.110 (0.005)	1.781 (0.118)	1.821 (0.047)	14.558 (0.674)	1.367 (0.095)

6. Conclusions

We propose the network function-on-scalar Lasso (NFL), which extends existing graph-guided penalization techniques to function-valued regression coefficients. NFL links entire coefficient curves through a known network so that estimation and clustering are carried out simultaneously. We outline its finite-dimensional representation, develop an efficient FPCA-based semi-proximal ADMM

algorithm, and provide accompanying risk bounds. Simulated studies and an application to county-level AQI data show that NFL achieves the lowest prediction error, scales well to high-dimensional predictors, and reveals clear region-specific effects.

Despite these promising results, several avenues for future research remain. Establishing variable selection consistency is challenging because NFL estimates unit-specific coefficient functions rather than a single sparse vector; precise selection guarantees therefore require further investigation. Extending the framework to non-Gaussian errors, longitudinal functional responses, or alternative network penalties would also broaden its scope.

Author contributions

Shan Sha: Conceptualization, methodology, theoretical analysis, algorithm development, software, formal analysis, investigation, visualisation, writing – original draft; Yan Li: Data analysis, validation, writing – review. All authors have read and approved the final version of the manuscript for publication.

Acknowledgments

The authors thank the anonymous referees and the editor for their helpful comments and constructive suggestions, which greatly improved the quality and clarity of this manuscript. This work is supported in part by the project funded by BGP under Grant YA24L00020, and in part by the National Natural Science Foundation of China under Grant 12371322.

Use of Generative-AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Conflict of interest

The authors declare no conflict of interest.

Appendix

Proof. We provide the proof for Theorem 1, following the proof structure outlined in Theorem 4 by [27]. We begin by stating an equivalent formulation of Model (2.3) for convenience in computation:

$$\min_{\beta\in\mathcal{H}^{np}} \frac{1}{2}\left\|y - Z\beta\right\|_{\mathcal{H}}^2 + \alpha_1 \left\|\beta\right\|_{\mathcal{H},1} + \alpha_2 \left\|Q\beta\right\|_{\mathcal{H},1}, \tag{6.1}$$

where $y = (y_1^T, \ldots, y_n^T)^T \in \mathcal{H}^n$ (concatenated functional responses), denote $\|\cdot\|_{\mathcal{H}^n} = \|\cdot\|_{\mathcal{H}}$ for simplicity, $Z = \mathrm{diag}(x_1, x_2, \cdots, x_n) \in \mathcal{R}^{n\times np}$ is the block diagonal design matrix with x_i being the i-th row of the design matrix $X \in \mathcal{R}^{n\times p}$ as a $1 \times p$ vector, $\beta = (\mathrm{vec}(\beta^1)^T, \ldots, \mathrm{vec}(\beta^n)^T)^T \in \mathcal{H}^{np}$ is the concatenated vector of all functional coefficients, with $\mathrm{vec}(\beta^i) = (\beta_1^i, \ldots, \beta_p^i)^T$, and $\|\beta\|_{\mathcal{H},1} = \sum_{k=1}^{np} \|\beta_k\|_{\mathcal{H}}$. The operator $Q \in \mathcal{R}^{mp\times np}$ represents the differences between coefficients of connected nodes in the network as constructed in Section 2, such that $\|Q\beta\|_{\mathcal{H},1}$ represents the weighted sum of $\mathcal{H}$-norms of differences

between coefficients of connected observations, corresponding to the third term in (2.3) when weights ω_{ij} are incorporated. For clarity, we use $\|Q\beta\|_{\mathcal{H},1}$ to denote $\sum_{(i,j)\in\mathcal{E}} \omega_{ij} \sum_{d=1}^{p} \|\beta_d^i - \beta_d^j\|_{\mathcal{H}}$.

Part 1. Let $\widehat{\beta}$ be any minimizer of problem (6.1) and $\beta^\star$ be the true coefficients. We aim to derive the error bound $\|\widehat{\beta} - \beta^\star\|_{\mathcal{H},1}$. From the optimality condition of problem (6.1), we have:

$$0 \in -Z^T(y - Z\widehat{\beta}) + \alpha_1 \partial \|\widehat{\beta}\|_{\mathcal{H},1} + \alpha_2 \partial \|Q\widehat{\beta}\|_{\mathcal{H},1}.$$

Since $\widehat{\beta}$ is a minimizer, we have,

$$\frac{1}{2}\left\|y - Z\widehat{\beta}\right\|_{\mathcal{H}}^2 + \alpha_1 \left\|\widehat{\beta}\right\|_{\mathcal{H},1} + \alpha_2 \left\|Q\widehat{\beta}\right\|_{\mathcal{H},1} \leq \frac{1}{2}\left\|y - Z\beta^\star\right\|_{\mathcal{H}}^2 + \alpha_1 \left\|\beta^\star\right\|_{\mathcal{H},1} + \alpha_2 \left\|Q\beta^\star\right\|_{\mathcal{H},1}.$$

Let $y = Z\beta^\star + \varepsilon$, where $\varepsilon = (\varepsilon_1^T, \ldots, \varepsilon_n^T)^T \in \mathcal{H}$. Substituting this into the inequality:

$$\begin{aligned}&\frac{1}{2}\|Z\beta^\star + \varepsilon - Z\widehat{\beta}\|^2 + \alpha_1\|\widehat{\beta}\|_1 + \alpha_2\|Q\widehat{\beta}\|_1\\ &\leq \frac{1}{2}\|\varepsilon\|^2 + \alpha_1\|\beta^\star\|_1 + \alpha_2\|Q\beta^\star\|_1.\end{aligned}$$

Expanding the term on the left:

$$\begin{aligned}\frac{1}{2}\|Z(\beta^\star - \widehat{\beta}) + \varepsilon\|^2 &= \frac{1}{2}\|Z(\beta^\star - \widehat{\beta})\|^2 + \langle Z(\beta^\star - \widehat{\beta}), \varepsilon\rangle + \frac{1}{2}\|\varepsilon\|^2\\ &= \frac{1}{2}\|Z(\widehat{\beta} - \beta^\star)\|^2 + \langle Z(\beta^\star - \widehat{\beta}), \varepsilon\rangle + \frac{1}{2}\|\varepsilon\|^2.\end{aligned}$$

So,

$$\begin{aligned}&\frac{1}{2}\|Z(\widehat{\beta} - \beta^\star)\|^2 + \langle Z(\beta^\star - \widehat{\beta}), \varepsilon\rangle + \frac{1}{2}\|\varepsilon\|^2 + \alpha_1\|\widehat{\beta}\|_1 + \alpha_2\|Q\widehat{\beta}\|_1\\ &\leq \frac{1}{2}\|\varepsilon\|^2 + \alpha_1\|\beta^\star\|_1 + \alpha_2\|Q\beta^\star\|_1.\end{aligned}$$

Canceling $\frac{1}{2}\|\varepsilon\|^2$ and rearranging:

$$\begin{aligned}&\frac{1}{2}\|Z(\widehat{\beta} - \beta^\star)\|^2 + \alpha_1\|\widehat{\beta}\|_1 + \alpha_2\|Q\widehat{\beta}\|_1\\ &\leq \langle \varepsilon, Z(\widehat{\beta} - \beta^\star)\rangle + \alpha_1\|\beta^\star\|_1 + \alpha_2\|Q\beta^\star\|_1\\ &\leq \|\varepsilon^T Z\|_{\mathcal{H},\infty}\|\widehat{\beta} - \beta^\star\|_{\mathcal{H},1} + \alpha_1\|\beta^\star\|_{\mathcal{H},1} + \alpha_2\|Q\beta^\star\|_{\mathcal{H},1},\end{aligned} \tag{6.2}$$

where $\|\varepsilon^T Z\|_{\mathcal{H},\infty} = \max_{k=1,\ldots,np} \|\varepsilon^T Z^{(k)}\|_{\mathcal{H}}$, and $Z^{(k)}$ is the k-th column of Z.

Part 2. By Lemma 9 in [27], if $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ are iid Gaussian elements in $\mathcal{H}$ and Assumption 4 holds, and if Z has standardized columns, for $\delta > 0$, we have

$$P\left(\max_{1\leq k\leq np} \left\|\varepsilon^T Z^{(k)}\right\|_{\mathcal{H}} \leq \alpha_0\right) \geq 1 - \delta, \tag{6.3}$$

where $\alpha_0 = \sqrt{n}\max_{\substack{1\leq i\leq p\\ 1\leq j\leq n}} x_{ji}c$, $c = \sqrt{\|\Lambda\|_1 + 2\|\Lambda\|_2\sqrt{-\log\delta} + 2\|\Lambda\|_\infty(-\log\delta)}$. Here, $\Lambda = (\lambda_1, \lambda_2, \cdots)$ is the vector of the eigenvalues of the covariance operator of $\varepsilon = (\varepsilon_1^T, \ldots, \varepsilon_n^T)^T \in \mathcal{H}$.

Part 3. Using (6.3), with probability $1-\delta$, we have $\|\varepsilon^T Z\|_{\mathcal{H},\infty} \leq \alpha_0$. Substituting this into (6.2) and multiplying by 2:

$$\begin{aligned}&\|Z\left(\widehat{\beta}-\beta^\star\right)\|_{\mathcal{H}}^2 + 2\alpha_1\|\widehat{\beta}\|_{\mathcal{H},1} + 2\alpha_2\|Q\widehat{\beta}\|_{\mathcal{H},1}\\ &\leq 2\alpha_0\|\widehat{\beta}-\beta^\star\|_{\mathcal{H},1} + 2\alpha_1\|\beta^\star\|_{\mathcal{H},1} + 2\alpha_2\|Q\beta^\star\|_{\mathcal{H},1}.\end{aligned}$$

Rearranging the terms involving α_1 and α_2:

$$\|Z\left(\widehat{\beta}-\beta^\star\right)\|_{\mathcal{H}}^2 \leq 2\alpha_0\|\widehat{\beta}-\beta^\star\|_{\mathcal{H},1} + 2\alpha_1(\|\beta^\star\|_{\mathcal{H},1} - \|\widehat{\beta}\|_{\mathcal{H},1} + 2\alpha_2(\|Q\beta^\star\|_{\mathcal{H},1} - \|Q\widehat{\beta}\|_{\mathcal{H},1}), \tag{6.4}$$

with probability $1-\delta$.

We analyze the terms on the RHS. Let S_0 be the set of indices $\{k|\beta_k^\star \neq 0\}$, and S_0^c its complement. By the triangle inequality for the ℓ_1 norm in $\mathcal{H}^{np}$:

$$\|\widehat{\beta}\|_{\mathcal{H},1} = \|\widehat{\beta}_{S_0}\|_{\mathcal{H},1} + \|\widehat{\beta}_{S_0^c}\|_{\mathcal{H},1}.$$

Also, $\|\beta^\star\|_{\mathcal{H},1} = \|\beta^\star_{S_0}\|_{\mathcal{H},1}$ since $\beta^\star_{S_0^c} = 0$,

$$\|\beta^\star_{S_0}\|_{\mathcal{H},1} \leq \|\widehat{\beta}_{S_0}\|_{\mathcal{H},1} + \|\beta^\star_{S_0} - \widehat{\beta}_{S_0}\|_{\mathcal{H},1}.$$

So, $\|\beta^\star\|_{\mathcal{H},1} - \|\widehat{\beta}\|_{\mathcal{H},1} \leq \|\beta^\star_{S_0} - \widehat{\beta}_{S_0}\|_{\mathcal{H},1} - \|\widehat{\beta}_{S_0^c}\|_{\mathcal{H},1}$. The second term in (6.4) is bounded as:

$$2\alpha_1(\|\beta^\star\|_{\mathcal{H},1} - \|\widehat{\beta}\|_{\mathcal{H},1}) \leq 2\alpha_1\|\beta^\star_{S_0} - \widehat{\beta}_{S_0}\|_{\mathcal{H},1} - 2\alpha_1\|\widehat{\beta}_{S_0^c}\|_{\mathcal{H},1}.$$

For the network penalty term difference, by the triangle inequality for norms:

$$\begin{aligned}2\alpha_2\|Q\beta^\star\|_{\mathcal{H},1} - 2\alpha_2\|Q\widehat{\beta}\|_{\mathcal{H},1} &\leq 2\alpha_2\|Q\left(\beta^\star - \widehat{\beta}\right)\|_{\mathcal{H},1}\\ &\leq 2\alpha_2\|Q\|_\infty\|\beta^\star - \widehat{\beta}\|_{\mathcal{H},1}\\ &= 2\alpha_2 r\|\beta^\star - \widehat{\beta}\|_{\mathcal{H},1},\end{aligned}$$

where $r = \|Q\|_\infty$. Thus we have

$$\begin{aligned}\left\|Z\left(\widehat{\beta}-\beta^\star\right)\right\|_{\mathcal{H}}^2 \leq\ & 2\alpha_0\left(\|\widehat{\beta}_{S_0} - \beta^\star_{S_0}\|_{\mathcal{H},1} + \|\widehat{\beta}_{S_0^c}\|_{\mathcal{H},1}\right)\\ &- 2\alpha_1\left(\|\beta^\star_{S_0}\|_{\mathcal{H},1} - \|\widehat{\beta}_{S_0} - \beta^\star_{S_0}\|_{\mathcal{H},1} + \|\widehat{\beta}_{S_0^c}\|_{\mathcal{H},1}\right)\\ &+ 2\alpha_1\|\beta^\star_{S_0}\|_{\mathcal{H},1} + 2\alpha_2 r\left(\|\widehat{\beta}_{S_0} - \beta^\star_{S_0}\|_{\mathcal{H},1} + \|\widehat{\beta}_{S_0^c}\|_{\mathcal{H},1}\right).\end{aligned}$$

Which means

$$\left\|Z\left(\widehat{\beta}-\beta^\star\right)\right\|_{\mathcal{H}}^2 + (2\alpha_1 - 2\alpha_0 - 2\alpha_2 r)\|\widehat{\beta}_{S_0^c}\|_{\mathcal{H},1} \leq (2\alpha_1 + 2\alpha_0 + 2\alpha_2 r)\|\widehat{\beta}_{S_0} - \beta^\star_{S_0}\|_{\mathcal{H},1}.$$

If $\alpha_1 \geq 2\alpha_0 + 2\alpha_2 r$, then $RE_F(q,\alpha)$ property can be used to conclude

$$\begin{aligned}\left\|Z\left(\widehat{\beta}-\beta^\star\right)\right\|_{\mathcal{H}}^2 + (2\alpha_1 - 2\alpha_0 - 2\alpha_2 r)\|\widehat{\beta} - \beta^\star\|_{\mathcal{H},1} &\leq 4\alpha_1\|\widehat{\beta}_{S_0} - \beta^\star_{S_0}\|_{\mathcal{H},1}\\ &\leq 4\alpha_1\sqrt{q}\|\widehat{\beta}_{S_0} - \beta^\star_{S_0}\|_{\mathcal{H},1}\end{aligned}$$

$$\leq \frac{4\alpha_1\sqrt{q}}{\sqrt{\alpha n}}\|Z(\beta^\star - \widehat{\beta})\|_{\mathcal{H}}$$
$$\leq \frac{1}{2}\left\|Z\left(\beta^\star - \widehat{\beta}\right)\right\|_{\mathcal{H}}^2 + \frac{8\alpha_1^2 q}{\alpha n}.$$

We can conclude that

$$\frac{1}{2}\left\|Z\left(\widehat{\beta} - \beta^\star\right)\right\|_{\mathcal{H}}^2 + (2\alpha_1 - 2\alpha_0 - 2\alpha_2 r)\,\|\widehat{\beta} - \beta^\star\|_{\mathcal{H},1} \leq \frac{8\alpha_1^2 q}{\alpha n},$$
$$\left\|Z\left(\widehat{\beta} - \beta^\star\right)\right\|_{\mathcal{H}} \leq \frac{4\alpha_1\sqrt{q}}{\sqrt{\alpha n}},$$
$$\|\widehat{\beta} - \beta^\star\|_{\mathcal{H},1} \leq \frac{8\alpha_1^2 q}{\alpha n\,(2\alpha_1 - 2\alpha_0 - 2\alpha_2 r)}.$$

□

References

1. J. S. Morris, Functional regression, *Annu. Rev. Stat. Appl.*, **2** (2015), 321–359. https://doi.org/10.1146/annurev-statistics-010814-020413
2. H. Cardot, F. Ferraty, P. Sarda, Functional linear model, *Stat. Probabil. Lett.*, **45** (1999), 11–22. https://doi.org/10.1016/S0167-7152(99)00036-X
3. F. Yao, H. G. Müller, J. L. Wang, Functional linear regression analysis for longitudinal data, *Ann. Stat.*, **33** (2005), 2873–2903. https://doi.org/10.1214/009053605000000660
4. G. Wahba, *Spline models for observational data*, SIAM, 1990.
5. C. Gu, C. Gu, *Smoothing spline ANOVA models*, Springer, **297** (2013).
6. G. M. James, J. Wang, J. Zhu, Functional linear regression that's interpretable, *Ann. Stat.*, **37** (2009), 2083–2108. https://doi.org/10.1109/TPS.2009.2033220
7. A. Goia, P. Vieu, An introduction to recent advances in high/infinite dimensional statistics, *J. Multivariate Anal.*, **146** (2016), 1–6. https://doi.org/10.1016/j.jmva.2015.12.001
8. G. Aneiros, R. Cao, R. Fraiman, C. Genest, P. Vieu, Recent advances in functional data analysis and high-dimensional statistics, *J. Multivariate Anal.*, **170** (2019), 3–9. https://doi.org/10.1016/j.jmva.2018.11.007
9. N. Ling, P. Vieu, Nonparametric modelling for functional data: Selected survey and tracks for future, *Statistics*, **52** (2018), 934–949. https://doi.org/10.1080/02331888.2018.1487120
10. N. Ling, P. Vieu, On semiparametric regression in functional data analysis, *WIREs Comput. Stat.*, **13** (2021), e1538. https://doi.org/10.1002/wics.1538
11. G. Aneiros, S. Novo, P. Vieu, Variable selection in functional regression models: A review, *J. Multivariate Anal.*, **188** (2022), 104871. https://doi.org/10.1016/j.jmva.2021.104871
12. P. T. Reiss, R. T. Ogden, Functional principal component regression and functional partial least squares, *J. Am. Stat. Assoc.*, **102** (2007), 984–996. https://doi.org/10.1198/016214507000000527

13. C. M. Crainiceanu, A. M. Staicu, C. Z. Di, Generalized multilevel functional regression, *J. Am. Stat. Assoc.*, **104** (2009), 1550–1561. https://doi.org/10.1198/jasa.2009.tm08564

14. R. A. Maronna, V. J. Yohai, Robust functional linear regression based on splines, *Comput. Stat. Data Anal.*, **65** (2013), 46–55. https://doi.org/10.1016/j.csda.2011.11.014

15. H. Zhu, F. Yao, H. H. Zhang, Structured functional additive regression in reproducing kernel Hilbert spaces, *J. Roy. Stat. Soc. B*, **76** (2014), 581–603. https://doi.org/10.1111/rssb.12036

16. A. Kneip, D. Poß, P. Sarda, Functional linear regression with points of impact, *Ann. Stat.*, **44** (2016), 1–30. https://doi.org/10.1214/15-AOS1323

17. J. Fan, J. T. Zhang, Two-step estimation of functional linear models with applications to longitudinal data, *J. Roy. Stat. Soc. B*, **62** (2000), 303–322. https://doi.org/10.1111/1467-9868.00233

18. J. Goldsmith, J. Bobb, C. M. Crainiceanu, B. Caffo, D. Reich, Penalized functional regression, *J. Comput. Graph. Stat.*, **20** (2011), 830–851. https://doi.org/10.1198/jcgs.2010.10007

19. M. W. McLean, G. Hooker, A.-M. Staicu, F. Scheipl, D. Ruppert, Functional generalized additive models, *J. Comput. Graph. Stat.*, **23** (2014), 249–269. https://doi.org/10.1080/10618600.2012.729985

20. Y. Li, T. Hsing, Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data, *Ann. Stat.*, **38** (2010), 3321–3351, https://doi.org/10.1214/10-AOS813

21. H. Zhu, P. J. Brown, J. S. Morris, Robust, adaptive functional regression in functional mixed model framework, *J. Am. Stat. Assoc.*, **106** (2011), 1167–1179. https://doi.org/10.1198/jasa.2011.tm10370

22. F. C. Stingo, M. Vannucci, G. Downey, Bayesian wavelet-based curve classification via discriminant analysis with markov random tree priors, *Stat. Sin.*, **22** (2012), 465. https://doi.org/10.1177/0962280213506395

23. R. Tibshirani, Regression shrinkage and selection via the Lasso, *J. Roy. Stat. Soc. B*, **58** (1996), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

24. H. Zou, T. Hastie, Regularization and variable selection via the elastic net, *J. Roy. Stat. Soc. B*, **67** (2005), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

25. C. H. Zhang, Nearly unbiased variable selection under minimax concave penalty, *Ann. Stat.*, **38** (2010), 894–942. https://doi.org/10.1214/09-AOS729

26. Y. Chen, J. Goldsmith, R. T. Ogden, Variable selection in function-on-scalar regression, *Stat*, **5** (2016), 88–101. https://doi.org/10.1002/sta4.106

27. R. F. Barber, M. Reimherr, T. Schill, The function-on-scalar LASSO with applications to longitudinal GWAS, *Electron. J. Stat.*, **11** (2017), 1351–1389. https://doi.org/10.1214/17-EJS1260

28. A. Mirshani, M. Reimherr, Adaptive function-on-scalar regression with a smoothing elastic net, *J. Multivariate Anal.*, **185** (2021), 104765. https://doi.org/10.1016/j.jmva.2021.104765

29. Z. Wang, J. Magnotti, M. S. Beauchamp, M. Li, Functional group bridge for simultaneous regression and support estimation, *Biometrics*, **79** (2023), 1226–1238. https://doi.org/10.1111/biom.13684

30. S. Novo, G. Aneiros, P. Vieu, Sparse semiparametric regression when predictors are mixture of functional and high-dimensional variables, *Test*, **30** (2021), 481–504. https://doi.org/10.3765/salt.v30i0.4830

31. J. A. Collazos, R. Dias, A. Z. Zambom, Consistent variable selection for functional regression models, *J. Multivariate Anal.*, **146** (2016), 63–71. https://doi.org/10.1016/j.jmva.2015.06.007

32. M. Febrero-Bande, W. González-Manteiga, Variable selection in functional additive regression models, *Comput. Stat.*, **34** (2019), 469–487. https://doi.org/10.1007/s00180-018-0844-5

33. G. Aneiros, P. Vieu, Sparse nonparametric model for regression with functional covariate, *J. Nonparametr. Stat.*, **28** (2016), 839–859. https://doi.org/10.1080/10485252.2016.1234050

34. G. Aneiros, P. Vieu, Partial linear modelling with multi-functional covariates, *Comput. Stat.*, **30** (2015), 647–671. https://doi.org/10.1007/s00180-015-0568-8

35. J. Gertheiss, G. Tutz, Sparse modeling of categorial explanatory variables, *Ann. Appl. Stat.*, **60** (2009), 2150–2180. https://doi.org/10.1214/10-AOAS355

36. D. Hallac, J. Leskovec, S. Boyd, Network Lasso: Clustering and optimization in large graphs, In: *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015, 387–396. https://doi.org/10.1145/2783258.2783313

37. A. Jung, N. Tran, A. Mara, When is network lasso accurate? *Front. Appl. Math. Stat.*, **3** (2018), 28.

38. J. Leskovec, R. Sosič, Snap: A general-purpose network analysis and graph-mining library, *ACM T. Intel. Syst. Tec.*, **8** (2016), 1–20. https://doi.org/10.1145/2898361

39. Y. Zhao, E. Zhu, X. Liu, C. Tang, D. Guo, J. Yin, Simultaneous clustering and optimization for evolving datasets, *IEEE T. Knowl. Data Eng.*, **33** (2019), 259–270. https://doi.org/10.1590/s0103-4014.2019.3395.0017

40. Y. Chen, S. Jewell, D. Witten, More powerful selective inference for the graph fused Lasso, *J. Comput. Graph. Stat.*, **32** (2023), 577–587. https://doi.org/10.1080/10618600.2022.2097246

41. P. Ma, W. Zhong, Penalized clustering of large-scale functional data with multiple covariates, *J. Am. Stat. Assoc.*, **103** (2008), 625–636. https://doi.org/10.1198/016214508000000247

42. X. Qiao, S. Guo, G. M. James, Functional graphical models, *J. Am. Stat. Assoc.*, **114** (2019), 211–222. https://doi.org/10.1080/01621459.2017.1390466

43. F. Centofanti, M. Fontana, A. Lepore, S. Vantini, Smooth Lasso estimator for the function-on-function linear regression model, *Comput. Stat. Data Anal.*, **176** (2022), 107556. https://doi.org/10.1016/j.csda.2022.107556

44. Z. Ma, X. Hu, A. M. Sayer, R. Levy, Q. Zhang, Y. Xue, et al., Satellite-based spatiotemporal trends in PM2.5 concentrations: China, 2004–2013, *Environ. Health Persp.*, **124** (2016), 184–192. https://doi.org/10.1289/ehp.1409481

45. S. Zhang, B. Guo, A. Dong, J. He, Z. Xu, S. X. Chen, Cautionary tales on air-quality improvement in Beijing, *P. Roy. Soc. A-Math. Phys.*, **473** (2017), 20170457. https://doi.org/10.1098/rspa.2017.0457

46. M. Grant, S. Boyd, Cvx: Matlab software for disciplined convex programming, version 2.1, 2014.

47. S. Fremdt, L. Horváth, P. Kokoszka, J. G. Steinebach, Functional data analysis with increasing number of projections, *J. Multivariate Anal.*, **124** (2014), 313–332. https://doi.org/10.1016/j.jmva.2013.11.009

48. T. Hsing, R. Eubank, *Theoretical foundations of functional data analysis, with an introduction to linear operators*, John Wiley & Sons, Chichester, **997** (2015). https://doi.org/10.1002/9781118762547

49. A. Cuevas, A partial overview of the theory of statistics with functional data, *J. Stat. Plan. Infer.*, **147** (2014), 1–23. https://doi.org/10.1016/j.jspi.2013.04.002

50. X. Zhang, J. L. Wang, From sparse to dense functional data and beyond, *Ann. Stat.*, **44** (2016), 2281–2321. https://doi.org/10.1214/16-AOS1446

51. P. Hall, M. Hosseini-Nasab, On properties of functional principal components analysis, *J. Roy. Stat. Soc. B*, **68** (2006), 109–126. https://doi.org/10.1111/j.1467-9868.2005.00535.x

52. Y. Xiao, L. Chen, D. Li, A generalized alternating direction method of multipliers with semi-proximal terms for convex composite conic programming, *Math. Program. Comput.*, **10** (2018), 533–555. https://doi.org/10.1007/s12532-018-0134-9

53. H. Chen, L. Kong, Y. Li, A novel convex clustering method for high-dimensional data using semiproximal ADMM, *Math. Probl. Eng.*, **2020** (2020), 9216351. https://doi.org/10.1155/2020/9216351

54. M. Fazel, T. K. Pong, D. Sun, P. Tseng, Hankel matrix rank minimization with applications to system identification and realization, *SIAM J. Matrix Anal. Appl.*, **34** (2013), 946–977. https://doi.org/10.1137/110853996

55. Y. Yang, H. Zou, A fast unified algorithm for solving group-lasso penalize learning problems, *Stat. Comput.*, **25** (2015), 1129–1141. https://doi.org/10.1007/s11222-014-9498-5

56. D. R. Roberts, V. Bahn, S. Ciuti, M. S. Boyce, J. Elith, G. Guillera-Arroita, et al., Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure, *Ecography*, **40** (2017), 913–929. https://doi.org/10.1111/ecog.02881

AIMS Press