Received: 7 November 2020 | Accepted: 22 April 2022

DOI: 10.1111/biom.13684

BIOMETRIC METHODOLOGY

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

Functional group bridge for simultaneous regression and support estimation

Zhengjia Wang[1] | **John Magnotti**[2] | **Michael S. Beauchamp**[2] | **Meng Li**[1]

[1]Department of Statistics, Rice University, Houston, Texas

[2]Department of Neurosurgery, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania

Correspondence
Meng Li, Department of Statistics, Rice University, Houston, TX.
Email: meng@rice.edu

Funding information
Division of Mathematical Sciences, Grant/Award Number: 2015569; National Institutes of Health, Grant/Award Number: 1R24MH117529

Abstract

This paper is motivated by studying differential brain activities to multiple experimental condition presentations in intracranial electroencephalography (iEEG) experiments. Contrasting effects of experimental conditions are often zero in most regions and nonzero in some local regions, yielding locally sparse functions. Such studies are essentially a function-on-scalar regression problem, with interest being focused not only on estimating nonparametric functions but also on recovering the function supports. We propose a weighted group bridge approach for simultaneous function estimation and support recovery in function-on-scalar mixed effect models, while accounting for heterogeneity present in functional data. We use B-splines to transform sparsity of functions to its sparse vector counterpart of increasing dimension, and propose a fast nonconvex optimization algorithm using nested alternative direction method of multipliers (ADMM) for estimation. Large sample properties are established. In particular, we show that the estimated coefficient functions are rate optimal in the minimax sense under the L_2 norm and resemble a phase transition phenomenon. For support estimation, we derive a convergence rate under the L_∞ norm that leads to a selection consistency property under δ-sparsity, and obtain a result under strict sparsity using a simple sufficient regularity condition. An adjusted extended Bayesian information criterion is proposed for parameter tuning. The developed method is illustrated through simulations and an application to a novel iEEG data set to study multisensory integration.

KEYWORDS

function-on-scalar regression, iEEG, locally sparse function, minimax rate, nonconvex optimization, selection consistency

1 | INTRODUCTION

Functional data analysis (FDA) is routinely encountered in modern applications due to the rapid advancement of new techniques to collect high-resolution data that can be viewed as curves; see Ramsay and Silverman (2005) for a comprehensive treatment. An overwhelming focus has been on nonparametric estimation of the underlying functions. However, shape constraints arise naturally in modern applications. One such example is *local sparsity*, that is, the function is exactly zero on subregions, in contrast to global sparsity that refers to a zero function. Local sparsity is a crucial characteristic for a nonparametric method to be interpretable in a variety of applications, and the estimation of the support as well as the function itself is of primary interest. This paper aims to develop a flexible method

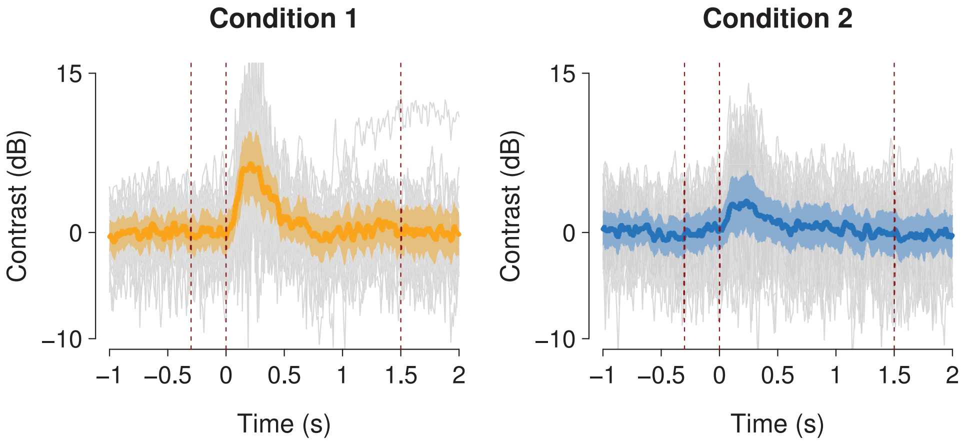

FIGURE 1 Baseline corrected response of iEEG high-gamma power under auditory-only (left) and audiovisual (right) conditions on one selected electrode. The data are obtained by decomposing the measured voltage signal into frequency space, converting to Decibel unit, calibrated to baseline (from −1 to −0.3 s), and then taking the average power in the 70-150 Hz range. Each individual gray line is one trial. The bold solid line is the mean responses, and the shaded area around the mean is a pointwise 95% confidence interval. This figure appears in color in the electronic version of this paper, and any mention of color refers to that version

with efficient implementation and theoretical guarantees for simultaneously estimating and recovering the support of locally sparse functions. Our motivation stems from neuroscientific studies using human intracranial electroencephalography (iEEG) data. iEEG is an emerging invasive method that offers anatomically precise measurements of human brain activity with electrodes placed on or implanted in the human brain, leading to excellent temporal resolution data and high signal-to-noise ratios (Lachaux et al., 2012; Kaiju et al., 2017). In most iEEG experiments, participants are presented with multiple experimental conditions. The brain response to each condition is recorded, and the experimenter wishes to know whether and how they differ. The contrast of brain activities is expected to be locally sparse (zero at certain period of time), and detecting nonsparse regions is of substantial interest to neuroscientists in addition to estimating the coefficient functions. For example, Figure 1 shows iEEG data from an audiovisual (AV) speech perception task (Ozker et al., 2018) under two experimental conditions, "auditory-only" and "audiovisual." In this study, the goal is to understand how the brain responds to auditory and visual stimuli through analyzing differential brain activities to these two conditions. Large trial-to-trial variation necessitates the use of statistical inference to automate the extraction of both population trajectories and supports of underlying brain activities.

Over the past several decades, there has been an extensive literature on sparsity. This leads to a rich menu of methods in the context of regularization-based variable selection with the parameter space being *sparse vectors*, including the Lasso (Tibshirani, 1996), minimax concave penalty (MCP) (Zhang, 2010), and bridge regression (Frank and Friedman, 1993), to name just a few. Such concepts have been extended to grouped variable selection and nonparametric *sparse functions*. Indeed, coupled with basis expansions, sparse coefficient vectors regularize the estimated function and lead to globally sparse coefficient functions via grouped sparsity. Along this line, Barber et al. (2017) extended group Lasso to functional data, and Chen et al. (2016) adopted group MCP, both achieving variable selection and parameter estimation.

However, comparatively little work has been done for functions with local sparsity. There is a closely related literature on utilizing various penalties to improve interpretability, for example, James et al. (2009), Zhou et al. (2013), Wang and Kai (2015), and Lin et al. (2017). These methods, however, rely on scalar responses and are not suitable for the motivating iEEG studies, where the response is functional and covariates are scalar, referred to as *function-on-scalar regression* hereafter. In function-on-scalar regression, the intrafunctional dependence of responses is vital, and large sample properties may be more intricately determined by the sample size (number of subjects) as well as the sampling frequency of individual trajectories of each subject. For concreteness, we focus on longitudinal data over time for the functional covariate that our motivating example corresponds to.

In this paper, we propose a method for simultaneous regression and support estimation for function-on-scalar regression, where the underlying functions are locally sparse. We adopt group bridge estimators coupled with B-splines to recover the sparse pattern of functional regression coefficients. Unlike the group Lasso and Lasso, group bridge penalty provably achieves variable selection in linear models at the group and individual levels simultaneously (Huang et al., 2009). This is particularly well suited for locally sparse functions by inducing exactly zero regions through grouping basis coefficients that contribute to the function at each time point, while maintaining parsimony and regularization of basis coefficients via selection at the individual level. The group bridge penalty has been used in other settings with a recent example being Ma and Kundu (2022) for scalar-on-image regression.

The proposed method does not require Gaussian assumptions and allows flexible heterogeneous correlation structures through random effects that possibly depend on various experimental phases. This leads to a novel *weighted functional group bridge* approach for function-on-scalar mixed effect models. On the algorithmic front, we introduce a nested alternating direction method of multipliers (ADMM) algorithm with "warm-start" and "early stopping" for fast computation. On the theoretical front, we establish a range of large sample properties, including minimax rate optimality for regression and selection consistency for support detection. Although under different models, our theory relies on substantially simplified assumptions than the existing literature to regularize regression functions, notably Assumption 2, facilitating interpretability. We allow flexible sampling designs including the case when the number of time points grows faster than the sample size, which is better suited for iEEG studies. In an application to the aforementioned iEEG experiment, our results complement previous studies by showing that multisensory interactions are a powerful modulator of activity throughout the speech perception network.

The rest of the paper is organized as follows. Section 2 introduces our model, the proposed estimator, and the optimization algorithm. Section 3 provides asymptotic properties. Section 4 contains simulations, followed by an application to iEEG analysis in Section 5.

2 | METHODOLOGY

2.1 | Model

Suppose a sample of n functional signals $\{y_i(t)\}_{i=1}^n$ is observed on a compact time set $\mathcal{T}$. We assume without loss of generality $\mathcal{T} = [0, 1]$. The linear function-on-scalar mixed effect model assumes

$$y_i(t) = \sum_{j=1}^{p} x_{ij}\beta_j(t) + \theta_i(t) + \epsilon_i(t), \qquad (1)$$

where x_{ij} is a scalar predictor, $\beta_j(t)$ is a fixed effect function, $\theta_i(t)$ is a zero mean random effect with covariance kernel $\Sigma_\theta(t,t')$ that captures the within-curve dependence, and $\epsilon_i(t)$ is the measurement error process independent of $\theta_i(t)$ with zero means and covariance kernel $\Sigma_\epsilon(t,t') = \sigma^2\delta_{t,t'}$; here $\delta_{t,t'}$ is the Kronecker delta. We assume $\beta_j(t)$ is smooth and locally sparse on the time domain. We propose to use phase-dependent random effects to account for heterogeneous dependence structures in various stages of an experiment, such as resting phase, trial onset, and stimuli offset. In particular, we partition $\mathcal{T}$ into a union of disjoint intervals $\{\mathcal{T}^{\text{pa}}\}_{\text{pa}=1}^P$, where each $\mathcal{T}^{\text{pa}}$ corresponds to a stage of the experiments, and random effects $\theta_i(t)$ are smooth within each phase. Note this generalizes traditional mixed effect models where P is typically set to one.

We use B-splines $\{\phi_k(t)\}_{k=1}^K$ to approximate each fixed effect function, that is, $\beta_j(t) = \sum_{k=1}^K \gamma_{jk}\phi_k(t) + R_j(t)$, where $R_j(t)$ is the approximation error. In addition to sharp approximation bounds to smooth functions, B-splines are particularly well suited for sparse functions as they are locally supported and thus transfer sparsity in $\beta(t)$ to a sparse p by K matrix $\boldsymbol{\gamma} = \{\gamma_{jk}\}_{p\times K}$. Let a nonnegative integer d be the degree of B-splines and define the knots of length $K+d+1$ to be $t_{m_1} = \cdots = t_{m_d} = 0 = t_{m_{d+1}} < t_{m_{d+2}} < \cdots < t_{m_K} = 1 = t_{m_{K+1}} = \cdots = t_{m_{K+d+1}}$. Then B-splines are defined recursively (De Boor, 1978): $\phi_{k,1}(t) = \mathbf{1}_{[t_{m_k}, t_{m_{k+1}})}(t)$, and $\phi_{k,d+1}(t) = \frac{t-t_{m_k}}{t_{m_{k+d}}-t_{m_k}}\phi_{k,d}(t) + \frac{t_{m_{k+1}}-t}{t_{m_{k+q-1}}-t_{m_{k+1}}}\phi_{k+1,d}(t)$, where $k = 1, \ldots, K$. B-splines of order $q = d+1$ are $\phi_k(t) = \phi_{k,q}(t)$, and we typically choose $q \geq 4$.

In practice, functional data are observed at discrete time points. Let $\mathcal{T}_0 = \{t_m\}_{m=1}^T \subseteq \mathcal{T}$ be the set of time points at which $y_i(t)$ is observed. Each partition set $\mathcal{T}_0^{\text{pa}}$ is defined as $\mathcal{T}_0 \cap \mathcal{T}^{\text{pa}}$. Let $\boldsymbol{Y} = \{y_{im}\}_{n\times T}, \boldsymbol{\theta} = \{\theta_{im}\}_{n\times T}, \boldsymbol{E} = \{e_{it}\}_{n\times T}$ be the discretized responses, random effects, and random noise observed at $\mathcal{T}_0$, respectively, and $\boldsymbol{X} = \{x_{ij}\}_{n\times p}$ the design matrix, $\boldsymbol{B} = \{B_k^m\}_{K\times T}$ the basis functions $\{\phi_k(t)\}_{k=1}^K$ evaluated on $\mathcal{T}_0$, and $\boldsymbol{R} = \{R_j(t_m)\}_{p\times T}$ the corresponding approximation error. Then Model (1) can be written as $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\gamma}\boldsymbol{B} + \boldsymbol{X}\boldsymbol{R} + \boldsymbol{\theta} + \boldsymbol{E}$, where $\boldsymbol{\theta}$ and $\boldsymbol{E}$ have covariance $\boldsymbol{\Sigma}_\theta$ and $\boldsymbol{\Sigma}_\epsilon = \sigma^2\boldsymbol{I}$ that are discretized $\Sigma_\theta(t,t')$ and $\Sigma_\epsilon(t,t')$ on $\mathcal{T}_0$.

In what follows, we use subscript to index rows; for example, $\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{\gamma}_j, \boldsymbol{B}_k$ are the corresponding rows of $\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\gamma}, \boldsymbol{B}$, respectively. We use $\boldsymbol{B}^{(m)}$ to denote the m^{th} column of $\boldsymbol{B}$, and $\boldsymbol{X}^{(j)}$ to denote the jth column of $\boldsymbol{X}$. All vectors are column vectors.

2.2 | Estimation: Weighted functional group bridge

We propose a weighted functional group bridge approach to estimate $\boldsymbol{\gamma}$:

$$\begin{aligned}\hat{\boldsymbol{\gamma}} &= \underset{\boldsymbol{\gamma}}{\operatorname{argmin}}\ L(\boldsymbol{\gamma}; \lambda, \alpha, \boldsymbol{W}) \\ &= \underset{\boldsymbol{\gamma}}{\operatorname{argmin}}\ \{f(\boldsymbol{\gamma}; \boldsymbol{W}) + \lambda g(\boldsymbol{\gamma}; \alpha)\}, \qquad (2)\end{aligned}$$

where $f(\boldsymbol{\gamma};\boldsymbol{W}) = \frac{1}{2}\sum_{i=1}^{n} \|(\boldsymbol{y}_i^T - \boldsymbol{X}_i^T\boldsymbol{\gamma}\boldsymbol{B})\boldsymbol{W}\|_2^2$ is the squared error loss with each observation weighted by a $T \times T$ matrix $\boldsymbol{W}$, and $\lambda g(\boldsymbol{\gamma};\alpha)$ is a penalty term to encourage sparsity on $\boldsymbol{\gamma}$ and $\beta(t)$ with tuning parameters $\lambda \geq 0$ and $\alpha > 0$. Each coefficient function $\beta_j(t)$ for $j = 1, \ldots, p$ is estimated by $\widehat{\beta}_j(t) = \sum_{k=1}^{K} \widehat{\gamma}_{jk}\phi_k(t)$.

We use a *group bridge* penalty for $g(\boldsymbol{\gamma};\alpha)$ to achieve sparsity in both $\boldsymbol{\gamma}$ and $\beta_j(t)$. Let $g(\boldsymbol{\gamma};\alpha) = \sum_{j=1}^{p}\sum_{m=1}^{T} g_{j,m}(\boldsymbol{\gamma};\alpha)$ with $g_{j,m}(\boldsymbol{\gamma};\alpha) = \{\sum_{k:\boldsymbol{B}_k^{(m)}\neq 0} |\gamma_{jk}|\}^{\alpha}$, which decomposes $g(\boldsymbol{\gamma};\alpha)$ into pT groups. Within each group, the L_1 penalty on a subset of $\boldsymbol{\gamma}$ leads to sparse estimates $\widehat{\boldsymbol{\gamma}}_j$ (Tibshirani, 1996). At the group level, if $g_{j,m}(\boldsymbol{\gamma};\alpha) = 0$, then $|\beta_j(t_m)| = |\boldsymbol{\gamma}_j^T\boldsymbol{B}^{(m)}| = 0$. Hence, group-level sparsity on $g_{j,m}(\boldsymbol{\gamma};\alpha)$ leads to sparse $\beta_j(t_m)$ at t_m. Knight and Fu (2000) show bridge estimators with $\alpha \in (0,1]$ combine variable selection and parameter estimation for sufficiently large λ. To achieve sparsity in both $\boldsymbol{\gamma}$ and $\beta_j(t)$, we propose to use $\alpha \in (0,1)$ as it has the appealing property to select variables at the individual and group levels simultaneously (Huang et al., 2009). Note that $g(\boldsymbol{\gamma};\alpha)$ becomes the L_1 penalty when $\alpha = 1$, which does not explicitly point to group-level sparsity that is critical to ensure that $\beta_j(t)$ is exactly zero at some t. Nevertheless, the developed algorithm in the following section is applicable for both $\alpha \in (0,1)$ and $\alpha = 1$, and we further compare these two variants in simulations. To use a compact notation for $g_{jm}(\boldsymbol{\gamma};\alpha)$, we indicate with $\mathbf{1}\{\boldsymbol{B}^{(m)}\} = [\mathbf{1}\{B_k^{(m)}\}]$ the indicator function evaluated on $\boldsymbol{B}^{(m)}$, a column vector whose kth element $\mathbf{1}\{B_k^{(m)}\} = 1$ if $B_k^{(m)} \neq 0$, and zero otherwise. Then

$$g_{j,m}(\boldsymbol{\gamma};\alpha) = \left[\sum_{k=1}^{K} \left|\gamma_{jk}\right| \cdot \mathbf{1}\{B_k^{(m)}\}\right]^{\alpha} = \|\mathbf{1}\{\boldsymbol{B}^{(m)}\} \odot \boldsymbol{\gamma}_j\|_1^{\alpha}, \tag{3}$$

where $\|\cdot\|_1$ is the L_1 norm, and $\mathbf{1}\{\boldsymbol{B}^{(m)}\} \odot \boldsymbol{\gamma}_j$ is the element-wise multiplication between $\mathbf{1}\{\boldsymbol{B}^{(m)}\}$ and $\boldsymbol{\gamma}_j$.

One needs to specify $(\lambda, \alpha, \boldsymbol{W})$ in the objective function $L(\boldsymbol{\gamma};\lambda,\alpha,\boldsymbol{W})$. In the following, we omit $(\lambda,\alpha,\boldsymbol{W})$ in $L(\boldsymbol{\gamma};\lambda,\alpha,\boldsymbol{W})$, $f(\boldsymbol{\gamma};\boldsymbol{W})$, and $g(\boldsymbol{\gamma};\alpha)$ to ease exposition, and instead use $L(\boldsymbol{\gamma})$, $f(\boldsymbol{\gamma})$, and $g(\boldsymbol{\gamma})$, respectively, when it does not cause confusion. Data-driven methods to select $\boldsymbol{W}$ and tune parameters such as (α, λ) and K are discussed in Section 2.5.

2.3 | Optimization: Nested ADMM algorithm

We first recast the minimization of $L(\boldsymbol{\gamma})$ into an iterative Lasso problem as in Huang et al. (2009). In particular, we embed $g(\boldsymbol{\gamma})$ in a carefully chosen higher dimensional space then link the solution back to $g(\boldsymbol{\gamma})$ through a particular path. Denote the expanded surface as

$$S(\boldsymbol{\gamma},\boldsymbol{\zeta}) = \sum_{j,m} s_{j,m}(\boldsymbol{\gamma},\boldsymbol{\zeta}), \quad \boldsymbol{\zeta} = \{\zeta_{j,m}\}, \tag{4}$$

$$s_{j,m}(\boldsymbol{\gamma},\boldsymbol{\zeta}) = \frac{\alpha^{\alpha}}{(1-\alpha)^{\alpha-1}}\zeta_{j,m}^{1-\frac{1}{\alpha}}\|\mathbf{1}\{\boldsymbol{B}^{(m)}\}\odot\boldsymbol{\gamma}_j\|_1 + \alpha^{\alpha}(1-\alpha)^{1-\alpha}\zeta_{j,m}. \tag{5}$$

The original nonconvex problem in Equation (2) can be solved by finding the minimizer for

$$\underset{\boldsymbol{\gamma},\boldsymbol{\zeta}}{\operatorname{argmin}}\, L_S(\boldsymbol{\gamma},\boldsymbol{\zeta}) \quad \text{subject to} \quad \boldsymbol{\zeta} = \boldsymbol{\zeta}(\boldsymbol{\gamma}) \text{ and } \boldsymbol{\zeta} > 0, \tag{6}$$

$$\text{where } L_S(\boldsymbol{\gamma},\boldsymbol{\zeta}) = f(\boldsymbol{\gamma}) + S(\boldsymbol{\gamma},\boldsymbol{\zeta}), \text{ and } \zeta_{j,m}(\boldsymbol{\gamma}) = \left(\frac{1-\alpha}{\alpha}\right)^{\alpha}\|\mathbf{1}\{\boldsymbol{B}^{(m)}\}\odot\boldsymbol{\gamma}_j\|_1^{\alpha}. \tag{7}$$

We carry out the optimization by iteratively updating $\boldsymbol{\zeta}$ with fixed $\boldsymbol{\gamma} = \boldsymbol{\gamma}^{\text{old}}$ through the definition $\boldsymbol{\zeta} = \boldsymbol{\zeta}(\boldsymbol{\gamma})$, and updating $\boldsymbol{\gamma}$ by solving a Lasso problem with fixed $\boldsymbol{\zeta} = \boldsymbol{\zeta}^{\text{new}}$.

We propose to solve the iterative Lasso problem using nested alternating direction method of multipliers, or ADMM (Boyd et al., 2011). Algorithm 1 details the nested ADMM algorithm, in which we use common notations for matrices. For any matrix $\boldsymbol{A}$, $\boldsymbol{A}^{+}$ is the positive part and $\boldsymbol{A}^{-}$ is the negative part. The operator $\operatorname{diag}(\cdot)$ extracts the diagonal elements of a square matrix into a vector, and expands a vector to a diagonal square. The operator $\oslash$ between two matrices defines element-wise division.

We use the ridge regression estimate $\widehat{\boldsymbol{\gamma}}_{\text{ridge}}$ that minimizes (2) with $g(\boldsymbol{\gamma};\alpha) = \sum_j \|\boldsymbol{\gamma}_j\|_2^2$ and tuning parameter λ_{ridge} as a closed-form "warm-start." Compared to LARS (Efron et al., 2004) adopted by Huang et al. (2009), warm started ADMM can significantly improve the performance: If initialized near the solution, ADMM converges faster to modest accuracy within a few steps (Boyd et al., 2011; Majzoobi et al., 2018). To fully take advantage of this property, we use few ADMM steps within each iteration without checking or waiting till full convergence, leading to "early stopping." The solutions at each iteration are used as warm-starts for the next iteration, further speeding up the convergence. Although the partial derivative $\partial g(\boldsymbol{\gamma})/\partial\gamma_{jk}$ near zero diverges, zero does not tend to be an absorbing state with an increased augmented Lagrangian parameter ρ in Algorithm 1. In addition, the adopted dense initialization via ridge regression and early stopping in ADMM help prevent the coefficients from entering zeros at early stage, leaving enough iterations for the coefficients to prioritize fitting before becoming

ALGORITHM 1 Nested ADMM solver for the case $\alpha \in (0,1)$

```
procedure $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{B}, \lambda, \alpha, \boldsymbol{W}, S_1, S_2$ ▷ Inputs
    Initialize $\boldsymbol{\gamma} \leftarrow \boldsymbol{\gamma}^{(0)} \leftarrow \hat{\boldsymbol{\gamma}}_{\mathrm{ridge}}, \boldsymbol{E} \leftarrow \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\gamma}\boldsymbol{B}$;
    SVD decomposition: $\boldsymbol{X} = \boldsymbol{U}_{\boldsymbol{X}}\boldsymbol{D}_{\boldsymbol{X}}\boldsymbol{V}_{\boldsymbol{X}}^T, \boldsymbol{B}\boldsymbol{W} = \boldsymbol{U}_{\boldsymbol{B}}\boldsymbol{D}_{\boldsymbol{B}}\boldsymbol{V}_{\boldsymbol{B}}^T$;
    Assign $\boldsymbol{V} = \boldsymbol{V}_{\boldsymbol{X}}^T, \boldsymbol{U} = \boldsymbol{U}_{\boldsymbol{B}}$,
    $\boldsymbol{Z} = \mathrm{diag}(\boldsymbol{D}_{\boldsymbol{X}}^2)\mathrm{diag}(\boldsymbol{D}_{\boldsymbol{B}}^2)^T + \rho\mathbf{1}_p\mathbf{1}_K^T$ ;
    for $\upsilon$ in $1, 2, \ldots, S_1$ do ▷ Macro-loop: iterative Lasso
        Calculate $\boldsymbol{D}^{(\upsilon)} = \{D_{jk}^{(\upsilon)}\}$, where
    $D_{jk}^{(\upsilon)} = \alpha^\alpha(1-\alpha)^{1-\alpha}\zeta_{j,m}^{1-\frac{1}{\alpha}}(\boldsymbol{\gamma}^{(\upsilon-1)}) \cdot \mathbf{1}\{B_k^{(m)}\}$;
        Initialize $\boldsymbol{\eta}^{(0)} = \boldsymbol{\gamma}^{(\upsilon-1)}, \boldsymbol{\xi}^{(0)} = \boldsymbol{\gamma}^{(\upsilon-1)}$,
    $\boldsymbol{\psi}^{(0)} = \boldsymbol{\xi}^{(0)} - \boldsymbol{\eta}^{(0)} = \mathbf{0}$;
        for $s$ in $1, 2, \ldots, S_2$ do ▷ ADMM loop, fixed steps
            $\boldsymbol{\phi} \leftarrow \boldsymbol{V}\boldsymbol{X}^T\boldsymbol{Y}\boldsymbol{W}(\boldsymbol{B}\boldsymbol{W})^T\boldsymbol{U} - \boldsymbol{\psi}^{(s-1)} + \rho\boldsymbol{V}\boldsymbol{\eta}^{(s-1)}\boldsymbol{U}$
            $\boldsymbol{\xi}^{(s)} \leftarrow \boldsymbol{V}^T(\boldsymbol{\phi} \oslash \boldsymbol{Z})\boldsymbol{U}^T$;
            $\boldsymbol{\eta}^{(s)} \leftarrow \{\boldsymbol{\xi}^{(s)} + \frac{1}{\rho}\boldsymbol{V}^T\boldsymbol{\psi}^{(s-1)}\boldsymbol{U}^T - \frac{\lambda}{\rho}\boldsymbol{D}^{(\upsilon)}\}^+ - \{\boldsymbol{\xi}^{(s)} +$
    $\frac{1}{\rho}\boldsymbol{V}^T\boldsymbol{\psi}^{(s-1)}\boldsymbol{U}^T + \frac{\lambda}{\rho}\boldsymbol{D}^{(\upsilon)}\}^-$;
            $\boldsymbol{\psi}^{(s)} \leftarrow \boldsymbol{\psi}^{(s-1)} + \rho\boldsymbol{V}\{\boldsymbol{\xi}^{(s)} - \boldsymbol{\eta}^{(s)}\}\boldsymbol{U}$;
        end for
        Assign $\boldsymbol{\gamma} \leftarrow \boldsymbol{\gamma}^{(\upsilon)} \leftarrow \boldsymbol{\eta}^{(S_2)}$;
        Recall $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\boldsymbol{\gamma}^{(\upsilon)}\boldsymbol{B}$, calculate $\widehat{\boldsymbol{E}} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$;
        if $\|\widehat{\boldsymbol{E}}\boldsymbol{W}\|_2$ is stable then Break the loop ▷ Exit
        end if
    end for
    Return $\boldsymbol{\gamma}$;
end procedure
```

sparse. In numerical experiments not reported here, we have found that the proposed algorithm outputs similar estimates with radically different initial values, indicating robustness. Nevertheless, we recommend to use the ridge regression estimate for faster convergence.

We use the following default settings in our numerical experiments, unless otherwise stated. We use fivefold cross validation to select λ_{ridge} in ridge regression. The number of iterations S_1 is 20, and the number of ADMM steps per iteration is $S_2 = 50$. We increase ρ exponentially by letting $\rho = e^{4s/S_2 - 1}$ where s is current ADMM step.

Note that by setting $S_2 = 1$ and forcing $D_{jk}^{(\upsilon)} = 1$ at each iteration, Algorithm 1 also solves the case when $\alpha = 1$.

2.4 | Variance estimation

Algorithm 1 solves an augmented Lagrangian problem L^{aug}:

$$\begin{aligned} L^{\mathrm{aug}}&(\boldsymbol{\xi}, \boldsymbol{\zeta}, \boldsymbol{\psi}, \boldsymbol{\eta}; \boldsymbol{W}, \lambda, \alpha, \rho) \\ &= L_S(\boldsymbol{\xi}, \boldsymbol{\zeta}) + \left\{\mathrm{vec}(\boldsymbol{\psi})^T\mathrm{vec}\left(\boldsymbol{V}\boldsymbol{\xi}\boldsymbol{U} - \boldsymbol{V}\boldsymbol{\eta}\boldsymbol{U}\right)\right\} \\ &\quad + \frac{\rho}{2}\|\boldsymbol{V}(\boldsymbol{\xi} - \boldsymbol{\eta})\boldsymbol{U}\|_2^2, \end{aligned} \tag{8}$$

where vec($\boldsymbol{A}$) vectorizes $\boldsymbol{A}$ by stacking columns of $\boldsymbol{A}$. The dual feasibility in the ADMM optimality condition yields

$$\{\boldsymbol{V}\boldsymbol{\xi}^{(s)}\boldsymbol{U}\} \odot \boldsymbol{Z} = \boldsymbol{V}\boldsymbol{X}^T\boldsymbol{Y}\boldsymbol{W}\boldsymbol{W}^T\boldsymbol{B}^T\boldsymbol{U} - \boldsymbol{\psi}^{(s-1)} + \rho\boldsymbol{V}\boldsymbol{\eta}^{(s-1)}\boldsymbol{U}, \tag{9}$$

$$\mathbf{0} = \left[-\frac{\lambda}{\rho}D(\boldsymbol{\zeta})\mathrm{sign}\{\boldsymbol{\xi}^{(s)}\} + \frac{1}{\rho}\boldsymbol{V}^T\boldsymbol{\psi}^{(s-1)}\boldsymbol{U}^T - \boldsymbol{\eta}^{(s)}\right] \odot \mathrm{sign}\{\boldsymbol{\xi}^{(s)}\}. \tag{10}$$

When Algorithm 1 converges, $\rho(\boldsymbol{\eta}^{(s)} - \boldsymbol{\eta}^{(s-1)}) \to \mathbf{0}$ and $\boldsymbol{\xi}^{(s)} \to \widehat{\boldsymbol{\gamma}}$. Hence, for (j,k) such that $\widehat{\gamma}_{jk} \neq 0$, the preceding optimality conditions in (9) and (10) yield

$$\begin{aligned} &\boldsymbol{X}^{(j)^T}\boldsymbol{X}\widehat{\boldsymbol{\gamma}}\boldsymbol{B}\boldsymbol{W}\boldsymbol{W}^T\boldsymbol{B}_k + \lambda D_{jk}(\widehat{\boldsymbol{\zeta}})\widehat{\gamma}_{jk} \\ &\quad = \boldsymbol{X}^{(j)^T}\boldsymbol{X}\boldsymbol{\gamma}\boldsymbol{B}\boldsymbol{W}\boldsymbol{W}^T\boldsymbol{B}_k + \boldsymbol{X}^{(j)^T}\boldsymbol{E}\boldsymbol{W}\boldsymbol{W}^T\boldsymbol{B}_k. \end{aligned} \tag{11}$$

Denote $\boldsymbol{P}(j,k) = \mathrm{vec}\{\boldsymbol{X}^T\boldsymbol{X}^{(j)}\boldsymbol{B}_k^T\boldsymbol{W}\boldsymbol{W}^T\boldsymbol{B}^T + \lambda D_{jk}\boldsymbol{J}(j,k)\}$, where $\boldsymbol{J}(j,k)$ is sparse matrix with only its (j,k) element being 1, and let $\boldsymbol{Q}(j,k) = \boldsymbol{X}^{(j)}\boldsymbol{B}_k^T\boldsymbol{W}\boldsymbol{W}^T$. Then (11) can be written as $\mathrm{vec}\{\boldsymbol{P}(j,k)\}^T\mathrm{vec}(\widehat{\boldsymbol{\gamma}}) = \mathrm{vec}\{\boldsymbol{Q}(j,k)\}^T\mathrm{vec}(\boldsymbol{X}\boldsymbol{\gamma}\boldsymbol{B} + \boldsymbol{Z}\boldsymbol{\theta} + \boldsymbol{E})$, for (j,k) such that $\widehat{\gamma}_{jk} \neq 0$. Letting $\mathrm{vec}_\lambda(\widehat{\boldsymbol{\gamma}})$ be the subvector of $\mathrm{vec}(\widehat{\boldsymbol{\gamma}})$ on its support, that is, all elements in $\mathrm{vec}_\lambda(\widehat{\boldsymbol{\gamma}})$ are nonzeros, and likewise $\boldsymbol{P}_\lambda(j,k)$ be $\boldsymbol{P}(j,k)$ on its support, then we have $\mathrm{vec}\{\boldsymbol{P}(j,k)\}^T\mathrm{vec}(\widehat{\boldsymbol{\gamma}}) = \mathrm{vec}_\lambda\{\boldsymbol{P}(j,k)\}^T\mathrm{vec}_\lambda(\widehat{\boldsymbol{\gamma}})$. Bind $\mathrm{vec}_\lambda\{\boldsymbol{P}(j,k)\}$ by column for all (j,k) such that $\widehat{\gamma}_{jk} \neq 0$, and denote it as $\boldsymbol{P}_\lambda$. It is easy to show that $\boldsymbol{P}_\lambda$ is a square matrix and further $\boldsymbol{P}_\lambda\mathrm{vec}_\lambda(\widehat{\boldsymbol{\gamma}}) = \boldsymbol{Q}\mathrm{vec}(\boldsymbol{X}\boldsymbol{\gamma}\boldsymbol{B} + \boldsymbol{Z}\boldsymbol{\theta} + \boldsymbol{E})$, where $\boldsymbol{Q}$ is also a square matrix column-binded by $\mathrm{vec}\{\boldsymbol{Q}(j,k)\}$. Consequently, the covariance of $\widehat{\boldsymbol{\gamma}}$ is given by

$$\mathrm{Cov}\{\mathrm{vec}_\lambda(\widehat{\boldsymbol{\gamma}})\} = \boldsymbol{P}_\lambda^{-1}\boldsymbol{Q}\mathrm{Cov}\{\mathrm{vec}(\boldsymbol{Z}\boldsymbol{\theta} + \boldsymbol{E})\}\boldsymbol{Q}^T\left(\boldsymbol{P}_\lambda^T\right)^{-1}. \tag{12}$$

2.5 | Choice of weights and parameter tuning

We propose to use $\boldsymbol{W} \propto \boldsymbol{\Sigma}^{-1/2}$ to account for heterogeneity in the functional response, where $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_\theta + \sigma^2\boldsymbol{I}$ is the covariance matrix of $y_i^T - \boldsymbol{X}_i^T\beta$. When $\boldsymbol{\Sigma}_\theta$ and σ^2 are unknown, we substitute using their estimates $\widehat{\boldsymbol{\Sigma}}_\theta$ and $\widehat{\sigma}^2$ to derive $\widehat{\boldsymbol{\Sigma}}$ and subsequently $\boldsymbol{W}$. We propose to employ local linear regression (Fan, 1993; Zhu et al., 2014) to estimate $\boldsymbol{\Sigma}$, assuming the existence of the second-order derivative of $\theta_i(t)$ within each phase $\mathcal{T}^{\mathrm{pa}}$. Denote $\Delta_{m,b}(t) = (t_m - t)/b$ and $\theta'_{i,b}(t) = b \cdot d\theta_i/dt(t)$, and choose a kernel function $K_b(t)$ with the bandwidth parameter b selected by minimizing the generalized cross-validation score (Zhu et al., 2014). For each $t \in \mathcal{T}^{\mathrm{pa}}$, we estimate $\theta_i(t)$ using a weighted least squares procedure (Fan, 1993):

$$\left\{\widehat{\theta}_i(t), \widehat{\theta'}_i(t)\right\}^T = \underset{\theta_i(t),\theta'_i(t)}{\mathrm{argmin}} \sum_{m:t_m \in \mathcal{T}_0^{\mathrm{pa}}} \left[y_i(t_m) - \sum_{j=1}^{p} x_{ij}\widehat{\beta}_j^{ols}(t_m)\right.$$

$$-\{\theta_i(t)+\theta_i'(t)\Delta_{m,b}(t)\}\Big]^2 K_b(t_m-t), \tag{13}$$

where $\widehat{\beta}_j^{ols}(t_m)$ is ordinary lease square estimator of the jth coefficient at time t_m. We then calculate $\widehat{\mathbf{\Sigma}}_\theta$ from the sample covariance of $\{\widehat{\theta}_i(t_m)\}_{i,m}$, and $\widehat{\sigma}^2$ from the residuals $y_i(t_m)-\sum_{j=1}^p x_{ij}\widehat{\beta}_j^{ols}(t_m)-\widehat{\theta}_i(t_m)$. As a passing comment, one may alternatively use $\boldsymbol{W}=\boldsymbol{I}$; this may lead to efficiency loss unless the functional dependence is homogeneous or approximately so, a special case of Model (1) without random effects. The developed methods and theory are applicable for this simplified model with such a choice of $\boldsymbol{W}$.

We next move on to the tuning of α and λ via an adjusted extended Bayesian information criterion (EBIC). We choose an equally spaced sequence of knots for B-splines. For given K, the EBIC proposed by Chen and Chen (2008) is $EBIC_\nu=\|(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})\boldsymbol{W}\|_2^2/(n\sigma^2)+T\log(\sigma^2)+df\cdot\log n/n+\nu\cdot df\cdot\log(pK)/n$, where $\widehat{\boldsymbol{\beta}}$ is the fitted coefficients of a given model, df the number of nonzero elements in $\widehat{\boldsymbol{\gamma}}$, $\nu\in[0,1]$ a constant, and σ^2 an unknown parameter analogue to the error variance in a standard regression model. We use $\nu=\max\{1-\frac{\log(n)}{2\log(pK)},\frac{1}{2}\}$, following Chen and Chen (2008). Huang et al. (2010) and Wang and Kai (2015) substituted $n\sigma^2$ with residual sum of squares based on $\widehat{\boldsymbol{\beta}}$. Since a bridge estimator is not unbiased due to its shrinkage to zero, we instead propose to estimate σ^2 by the weighted residual sum of squares based on the generalized least-square estimator $\widehat{\boldsymbol{\beta}}_{GLS}$, which is unbiased although lacks sparsity. Letting $\widehat{\boldsymbol{\beta}}_{GLS}=(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}\boldsymbol{W}\boldsymbol{W}^T\boldsymbol{B}^T(\boldsymbol{B}\boldsymbol{W}\boldsymbol{W}^T\boldsymbol{B}^T)^{-1}$ and $\widehat{\sigma}^2_{GLS}=\|(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{GLS})\boldsymbol{W}\|_2^2/(nT)$, the proposed adjusted EBIC is

$$\begin{aligned}EBIC_\nu=T\frac{\|(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})\boldsymbol{W}\|_2^2}{\|(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{GLS})\boldsymbol{W}\|_2^2}\\ +df\cdot\frac{\log n}{n}+\nu\cdot df\cdot\frac{\log pK}{n}.\end{aligned} \tag{14}$$

Simulation studies in Section 4 indicate that our proposed adjusted EBIC tends to select parameters that are close to the oracle values. For the tuning parameter K in the B-spline approximation, one may specify a relatively large value that increases with the temporal resolution. Alternatively, the adjusted EBIC can be used to jointly select K and (α,λ) in a data-driven manner.

3 | ASYMPTOTIC PROPERTIES

In this section, we study asymptotic properties of the proposed estimators $\widehat{\beta}_j(t)$ constructed using $\widehat{\boldsymbol{\gamma}}$ in (2). For a function $f(t):[0,1]\to\mathbb{R}$, let $\|f\|_r$ be the L_r norm and $\|f\|_\infty=\sup_{t\in[0,1]}|f(t)|$. Denote by $C^r[0,1]$ the Hölder space on [0,1] with order r, a set of functions $f(t)$ such that for some $L_f>0$, $|f^{(r_0)}(x)-f^{(r_0)}(y)|\le L_f\|x-y\|^{r-r_0}$ for all $x,y\in[0,1]$, where r_0 is the largest integer strictly smaller than r. Let $S(f)=\{t\in\mathcal{T}:f(t)=0\}$ map $f(t)$ to its zero set. Denote by $\delta_{\min}(\boldsymbol{A})$ and $\delta_{\max}(\boldsymbol{A})$ the minimum and maximum eigenvalues for any given matrix $\boldsymbol{A}$. For two sequences $a_{n,T}$ and $b_{n,T}$, $a_{n,T}\lesssim b_{n,T}$ means $a_{n,T}\le Cb_{n,T}$ for some universal constant $C>0$. We write $a_{n,T}\asymp b_{n,T}$ if $a_{n,T}\lesssim b_{n,T}$ and $a_{n,T}\gtrsim b_{n,T}$. Asymptotics in this section are interpreted when n and T go to infinity. Proofs are deferred to the Supporting Information.

We assume the following regularity conditions. Let $C_1,\dots,C_6$ be some positive constants that do not depend on n or T.

Assumption 1. The underlying $\beta_j(t)\in C^r[0,1]$, for $j=1,2,\dots,p$ and $r\ge 2$.

Assumption 2. The integral $\int_{t\in S(\beta_j)^c}|\frac{1}{\beta_j(t)}|^{2(1-\alpha)}dt$ exists and is finite, for all j.

Assumption 3. $\{\epsilon_i(t)\}_{i=1}^n$ and $\{\theta_i(t)\}_{i=1}^n$ are independent across i and sub-Gaussian.

Assumption 4. The design matrix satisfies that $C_1\le\delta_{\min}(\boldsymbol{X}^T\boldsymbol{X}/n)$ and $\delta_{\max}(\boldsymbol{X}^T\boldsymbol{X}/n)\le C_2$, and the weight matrix is chosen such that $C_3\le\delta_{\min}(\boldsymbol{W}\boldsymbol{W}^T)\le\delta_{\max}(\boldsymbol{W}\boldsymbol{W}^T)\le C_4$, for all sufficiently large n and T.

Assumption 5. $\max_m(t_{m+1}-t_m)=O(T^{-1})$.

Assumption 6. $C_5K^{-1}\le t_{m_{k+q}}-t_{m_k}\le C_6K^{-1}$, where $K<T$.

Assumptions 1, 3, 5, and 6, as well as the design matrix condition in Assumption 4, are common in high-dimensional regression; see, for example, Fan and Zhang (2000), Cai and Yuan (2011), and Wang and Kai (2015). Assumptions 5 and 6 are concerned with the spacing of time points and B-spline knots, respectively, and trivially hold when they are equally spaced.

Assumption 2 is in the same vein of Conditions (B') and (C') in Fan and Peng (2004) to ensure that the group bridge penalty does not dominate the least square error on its support, and implies that $\beta_j(t)$ deviates from zero fast enough so that its support and zero set can be well distinguished. Unlike Fan and Peng (2004) and Huang et al. (2009), Assumption 2 decouples the penalty and the regression function, leading to a simpler and more interpretable formulation to regularize regression coefficients. We achieve such simplicity by relying on a carefully modified B-spline approximation that will be detailed in Lemma 1. This assumption also suggests a lower bound of α for $\beta_j(t)$ leaving zeros at polynomial speed. If $\beta_j(t_0) = 0$ for some t_0 and $\beta_j(t)$ satisfies $|\beta_j(t)| \gtrsim |t - t_0|^b$ as t approaches t_0 for some positive constant b, then choosing $\alpha > 1 - 1/(2b)$ is required to comply with this assumption.

Assumption 4 holds for the proposed data-driven $\boldsymbol{W}$ that satisfies $\boldsymbol{W}\boldsymbol{W}^T = \widehat{\boldsymbol{\Sigma}}^{-1}$ if the eigenvalues of $\widehat{\boldsymbol{\Sigma}}$ are bounded from above and below, a condition that is often satisfied following the rich literature of covariance estimation in functional data. For example, letting C_θ be a constant such that $\delta_{\max}(\boldsymbol{\Sigma}_\theta) \leq C_\theta$, Theorem 2 of Zhu et al. (2014) proves the consistency of $\widehat{\boldsymbol{\Sigma}}^{-1}$, and particularly indicates $|\delta_{\max}(\widehat{\boldsymbol{\Sigma}}_\theta) - \delta_{\max}(\boldsymbol{\Sigma}_\theta)| \lesssim C_\theta/2$ and $|\widehat{\sigma}^2 - \sigma^2| \lesssim \sigma^2/2$, yielding $|\delta_{\max}(\widehat{\boldsymbol{\Sigma}}) - \delta_{\max}(\boldsymbol{\Sigma})| \lesssim (C_\theta + \sigma^2)/2$ and thus $\delta_{\max}(\widehat{\boldsymbol{\Sigma}}) \lesssim 3(C_\theta + \sigma^2)/2$. Moreover, the semipositiveness of $\widehat{\boldsymbol{\Sigma}}_\theta$ ensures that $\delta_{\min}(\widehat{\boldsymbol{\Sigma}}) \geq \widehat{\sigma}^2 \gtrsim \sigma^2/2$.

It is well known that if $\beta_j(t) \in C^r[0,1]$ and there are no overlapping spline knots for t_{m_k} for $q \leq k \leq K$, then there exists a B-spline approximation such that L_∞ approximation error is upper bounded by K^{-r} up to a constant; for example, see Schumaker (2007). However, these accurate B-spline approximations do not necessarily capture the sparsity of the true function. In Lemma 1, we propose a sparse modification of such B-splines so that the new approximation preserves the sparsity structure with the same approximation accuracy.

Lemma 1. *Under Assumptions 1 and 6, let* $\widetilde{\beta}_j^*(t) = \sum_{k=1}^K \widetilde{\gamma}_{jk}^* \phi_k(t)$ *be the B-spline approximation in Schumaker (2007) such that* $\|\widetilde{\beta}_j^*(t) - \beta_j(t)\|_\infty \leq C_{\beta_j}^* K^{-r}$ *for a constant* $C_{\beta_j}^*$. *Then there exists a sparse modification* $\widetilde{\beta}_j(t) := \sum_{k=1}^K \widetilde{\gamma}_{jk} \phi_k(t)$ *and a constant* C_{β_j} *such that* $S(\beta_j) \subseteq S(\widetilde{\beta}_j)$, *and* $\|\beta_j(t) - \widetilde{\beta}_j(t)\|_\infty \leq C_{\beta_j} K^{-r}$. *Also for any* $t \notin S(\widetilde{\beta}_j)$, $\sum_{k:\phi_k(t)>0} |\widetilde{\gamma}_{jk}| \geq \frac{C_{\beta_j}^*}{qC_{\beta_j} + C_{\beta_j}^*} |\beta_j(t)|$.

We call the modified B-spline coefficients $\widetilde{\boldsymbol{\gamma}} = \{\widetilde{\gamma}_{jk}\}_{p\times K}$ pseudo true values of $\boldsymbol{\gamma}$. The triangle inequality gives $\|\beta_j(t) - \widehat{\beta}_j(t)\|_\infty \leq \|\beta_j(t) - \widetilde{\beta}_j(t)\|_\infty + \|\widetilde{\beta}_j(t) - \widehat{\beta}_j(t)\|_\infty \leq C_{\beta_j} K^{-r} + \|\widehat{\boldsymbol{\gamma}}_j - \widetilde{\boldsymbol{\gamma}}_j\|_\infty \sum_{k=1}^K \|\phi_k(t)\|_1$, where $\sum_{k=1}^K \|\phi_k(t)\|_1 = 1$. Therefore, convergence rates of $\widehat{\beta}_j(t)$ boils down to convergence rates of $\widehat{\boldsymbol{\gamma}}$ relative to the pseudo true $\widetilde{\boldsymbol{\gamma}}$ and the approximation error $C_{\beta_j} K^{-r}$ of using B-splines. The following Theorem 1 and Theorem 2 establish convergence rates of $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\beta}_j$, respectively. To ease exposition, we relate K and T to n by writing $K = n^{\frac{\kappa}{2r}}$ and $T = n^{\frac{\tau}{2r}}$ where $0 < \kappa \leq \tau$. Here $n^{\frac{\kappa}{2r}}$ and $n^{\frac{\tau}{2r}}$ represent the asymptotic rates of K and T, and any multiplicative constants do not change our results. To state the two theorems in their most general forms, we do not yet assume a specific order of either κ or τ.

Theorem 1 (Convergence rate of $\widehat{\boldsymbol{\gamma}}$). *Let* $c_\kappa = \min(1, \kappa)$. *Suppose Assumptions 1-6 hold and assume* $\frac{\log(\lambda)}{\log(n)} \leq 1 - \frac{c_\kappa}{2} + \frac{\tau}{4r} - \frac{\kappa}{4r}$. *Then* $\|\widehat{\boldsymbol{\gamma}} - \widetilde{\boldsymbol{\gamma}}\|_\infty \leq \|\widehat{\boldsymbol{\gamma}} - \widetilde{\boldsymbol{\gamma}}\|_2 = O_p(n^{\frac{\kappa}{4r} - \frac{c_\kappa}{2}})$ *as* $n \to \infty$.

Theorem 2 (Convergence rate of $\widehat{\beta}_j(t)$). *Under the same assumptions of Theorem 1, as* $n \to \infty$,

$$\|\widehat{\beta}_j(t) - \beta_j(t)\|_\infty = O_p\left(n^{\frac{\kappa}{4r} - \frac{c_\kappa}{2}}\right), \tag{15}$$

$$\|\widehat{\beta}_j(t) - \beta_j(t)\|_2 = O_p\left(n^{-\frac{c_\kappa}{2}}\right). \tag{16}$$

We remark that there is a phase transition phenomenon at $\tau = 1$ for the L_∞ rate in (15). When $\tau < 1$, because $\kappa < \tau$, we have $c_\kappa = \kappa$, and the optimal L_∞ rate is attained at the largest κ, that is, $\kappa = \tau$, which gives the rate $O_p(n^{(1-2r)\frac{\tau}{4r}})$. In this case, the rate improves as τ increases. When $\tau > 1$, the optimal L_∞ rate is $O_p(n^{(1-2r)\frac{1}{4r}})$, which is achieved at $\kappa = 1$, and increasing τ does not improve the rate. The same phase transition also applies to the L_2 rate in (16), which coincides with the observation made in Cai and Yuan (2011). In addition, our rate calculation implies that with $\kappa = \min(\tau, 1)$, the L_2 rate in (16) becomes $O_p(n^{-\min(\tau,1)/2})$, that is, $O_p(n^{-1/2} + T^{-r})$, which is minimax optimal (Cai and Yuan, 2011).

The rate under the L_∞ norm in (15) indicates that $\widehat{\beta}_j(t)$ converges to $\beta(t)$ at each t for $\kappa < 2r$. This is particularly useful in detecting sparse regions as pointwise convergence suggests low false positive rates (FPRs) and low false negative rates in finding the support of $\beta_j(t)$. In particular, we consider δ-sparsity denoted by $S_\delta(h) = \{t \in \mathcal{T} : |h(t)| \leq \delta\}$ for $\delta > 0$. Then Equation (15) suggests that for arbitrary $\delta > 0$, as $n \to \infty$,

$$P\{S(\widehat{\beta}_j) \subseteq S_\delta(\beta_j)\} \to 1, \quad P\{S(\beta_j) \subseteq S_\delta(\widehat{\beta}_j)\} \to 1. \tag{17}$$

In addition, the following theorem establishes one side of strict sparsity for the proposed method, that is, the support of $\widehat{\beta}_j(t)$ is a subset of $\beta_j(t)$ with probability approaching to 1. This leads to low FPRs under strict sparsity. It is an interesting future direction to study under what conditions the other side of strict sparsity also holds.

Theorem 3. *Under the same assumptions as in Theorem 1 and the additional condition that* $\frac{\log(\lambda)}{\log(n)} > 1 + \frac{c_\kappa(\alpha-2)}{2} + \frac{\tau}{2r} - \frac{\kappa\alpha}{4r}$, *there holds* $P\{S(\beta_j) \subseteq S(\widehat{\beta}_j)\} \to 1$ *as* $n \to \infty$.

While we have focused on $\alpha < 1$, a close inspection into the proofs of Theorems 1 and 2 suggests that they also hold for $\alpha = 1$ (note that Assumption 2 is not needed in this case). As discussed in Section 2.2, choosing $\alpha < 1$ induces exactly zero estimates of functional coefficients. Theorem 3 reassuringly shows that the recovered exactly zero regions tend to contain the zero regions of the true regression functions at least asymptotically. The proof of Theorem 3 crucially relies on the choice of $\alpha < 1$; see the Supporting Information for details.

Our theoretical developments for the estimators defined in Section 2.2 are disengaged from Algorithm 1 that was used to find such estimators. There is a literature on coupling the derivation of asymptotics for statistical estimators with the employed optimization algorithm; interested readers may refer to Agarwal et al. (2012) and Fan et al. (2014) for examples.

4 SIMULATIONS

We conduct simulations to compare finite sample performances of the proposed approach with competing methods in terms of function estimation and sparse region detection. We generate data according to Model (1) using three coefficients with different sparsity levels: $\beta_1(t) = 0$ for global sparsity, $\beta_2(t) = \sin(\pi t)$ for no sparsity (a dense coefficient), and $\beta_3(t) = \sin(\frac{5\pi}{2}t - \frac{\pi}{2})\mathbf{1}_{[0.2,0.4)}(t) + \mathbf{1}_{[0.4,0.6)}(t) + \sin(\frac{5\pi}{2}t - \pi)\mathbf{1}_{[0.6,0.8)}(t)$ for local sparsity. The most interesting coefficient is $\beta_3(t)$, but including $\beta_1(t)$ and $\beta_2(t)$ in the model allows us to study the performance of the proposed method in the presence of other coefficient functions with various sparsity levels. The design matrix $\boldsymbol{X}$ is generated from the standard normal distribution $N(0, 1)$. To simulate different phases of experiments, the random effects $\theta_i(t)$ are generated from an $AR(1) \times \sigma(t)$ process, where $\sigma(t) = 0.1 + 0.2 \times \mathbf{1}_{[0.4,0.8)}(t) + 0.5 \times \mathbf{1}_{[0.8,1)}(t)$ is a nondecreasing step function and $AR(1)$ is an order one autoregressive process with correlation .9. The three coefficient functions and $\sigma(t)$ are visualized in Figure S1 in the Supporting Information. The errors $\epsilon_i(t_m) \sim N(0, 1)$ are independent across observations and time points. We use $T = 100$ as the time resolution to generate equally spaced t_m. We consider two sample sizes: $n = 100$ (small sample size) and $n = 1000$ (large sample size). We run 100 simulations for each sample size.

In addition to the proposed weighted function group bridge approach, we include its two variants: homogeneous weight ($\boldsymbol{W} = \boldsymbol{I}$) and $\alpha = 1$. Other competing methods include Group Lasso (gLasso) proposed by Barber et al. (2017) and Group MCP (gMCP) by Chen et al. (2016). We also implement two-step function-on-scalar (2-Step FoS) regression (Fan and Zhang, 2000), which first obtains regression coefficients at each time point then smooths these estimates. Although 2-Step FoS is not designed for functions with sparsity, we include it to compare estimation accuracy. We use B-splines with $K = 30$ for all methods when smoothing is needed. Tuning parameters α and λ in the proposed method are selected by minimizing the adjusted EBIC in (14) using grid search, unless otherwise stated. We derive joint confident bands for $\beta(t)$s based on the variance estimation in Section 2.4. In particular, we perturb sparse estimates $\hat{\gamma}_{jk}$ with small random numbers to expand Equation (12) into all $\hat{\gamma}_{jk}$s, which gives the covariance of $\hat{\boldsymbol{\gamma}}$ and subsequently a joint confident band for each $\beta(t)$.

Figure 2 visualizes the estimates and their 95% joint confidence intervals using one randomly selected replication. The proposed method appears to have tighter and more adaptive confidence intervals than gMCP and gLasso, partly due to its accounting for heterogeneous errors. Two-step FoS also enjoys tight joint confidence bands, but it cannot recover sparsity. Although all other methods contain sparsity constraints in their design, gLasso fails to detect the support of $\beta_3(t)$. Functional group bridge methods and gMCP succeed in recovering the globally sparse signal $\beta_1(t)$. The proposed method is the only method to recover the locally sparse coefficient $\beta_3(t)$ at $t \in (0.8, 1]$, when the error variance is large, indicating its adaptivity to various sparse and noise levels.

We assess the proposed adjusted EBIC method for parameter tuning and observe that the selected parameters give the best or close to the best accuracy in root mean squared error (RMSE); see Figure S2 and the discussion therein in the Supporting Information for more details.

We next focus on $\beta_3(t)$ and compare each method in terms of both accuracy and support detection. For each method, we calculate the RMSE $\|\hat{\beta}_3(t) - \beta_3(t)\|_2$ to measure the overall accuracy, and $\|\hat{\beta}_3(t) - \beta_3(t)\|_\infty$ to measure the maximum difference, reported in Table 1. We can see that our approach as well as its variant with homogeneous weight outperforms other methods in estimating the locally sparse coefficient $\beta_3(t)$ for both sample sizes. Table 1 also presents the coverage of the confident bands produced by each method. All methods except $\alpha = 1$ attain the nominal coverage without significant deviation at $n = 1000$. Both gMCP and gLasso lead to the largest coverage at the expense of wider confidence bands; this can be clearly observed in Figure 2. In contrast, the proposed method gives much tighter confidence bands while maintaining a coverage that is close to the nominal level.

For support detection, we calculate the FPR or recall by $|S(\beta_3) \cap S_\delta(\hat{\beta}_3)^c|/|S(\beta_3)|$, true positive rate (TPR) $|S(\beta_3)^c \cap S_\delta(\hat{\beta}_3)^c|/|S(\beta_3)^c|$, and precision $|S(\beta_3)^c \cap S_\delta(\hat{\beta}_3)^c|/|S_\delta(\hat{\beta}_3)^c|$ for some $\delta \geq 0$, where $|\cdot|$

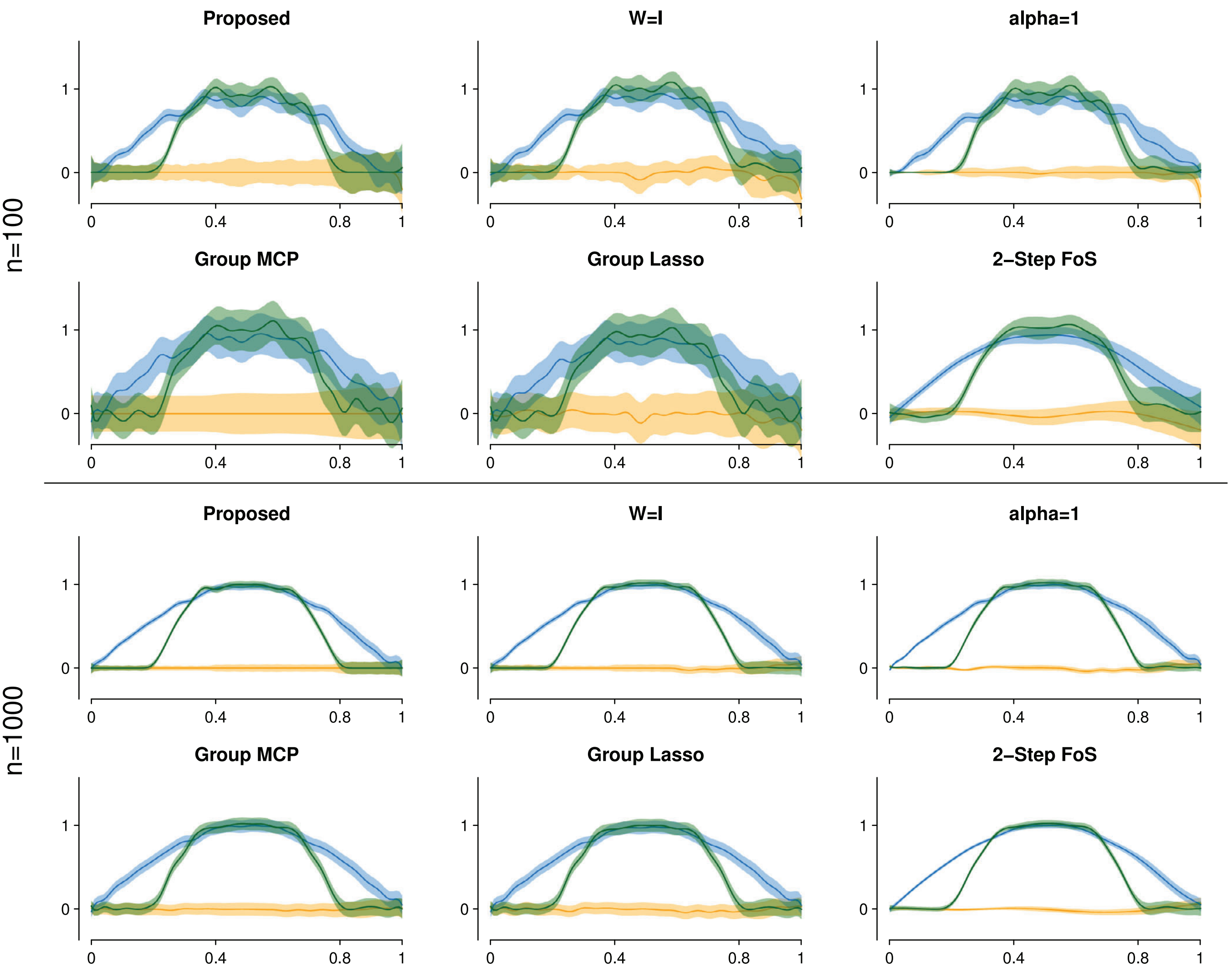

FIGURE 2 Fitted coefficients with 95% joint confidence intervals at $n = 100$ (top two rows) and $n = 1000$ (bottom two rows). This figure appears in color in the electronic version of this paper, and any mention of color refers to that version

TABLE 1 Performance comparison of various methods. For each sample size $n \in \{100, 1000\}$, RMSE calculates $\|\beta_3(t) - \widehat{\beta}_3(t)\|_2$, L_∞ measures $\|\beta_3(t) - \widehat{\beta}_3(t)\|_\infty$, Coverage is the pointwise coverage of 95% joint confidence intervals, and F_1 score assesses support recovery. All results are averaged over 100 simulations. Standard errors are reported in parentheses

n	Metrics	Proposed	$W = I$	$\alpha = 1$	gMCP	gLasso	2-Step FoS
100	RMSE (×0.01)	8.6 (0.3)	8.1 (0.3)	8.8 (0.3)	9.5 (0.2)	10.2 (0.3)	8.8 (0.2)
	L_∞ (×0.01)	21.2 (0.9)	23.0 (0.9)	22.4 (0.7)	26.2 (0.9)	25.6 (0.7)	23.1 (0.8)
	Coverage	94.8 (1.0)	93.8 (0.9)	87.3 (1.0)	99.8 (0.1)	99.5 (0.2)	93.6 (0.6)
	F_1 score	0.85 (0.0)	0.82 (0.0)	0.79 (0.0)	0.75 (0.0)	0.75 (0.0)	0.75 (0.0)
1000	RMSE (×0.01)	2.4 (0.1)	2.5 (0.1)	3.2 (0.1)	3.3 (0.1)	3.6 (0.1)	3.0 (0.1)
	L_∞ (×0.01)	7.6 (0.3)	8.1 (0.3)	8.5 (0.2)	9.8 (0.3)	9.7 (0.3)	8.5 (0.3)
	Coverage	94.6 (0.4)	96.4 (0.4)	86.3 (0.8)	99.6 (0.1)	99.2 (0.2)	94.6 (0.5)
	F_1 score	0.94 (0.0)	0.90 (0.0)	0.85 (0.0)	0.75 (0.0)	0.75 (0.0)	0.75 (0.0)

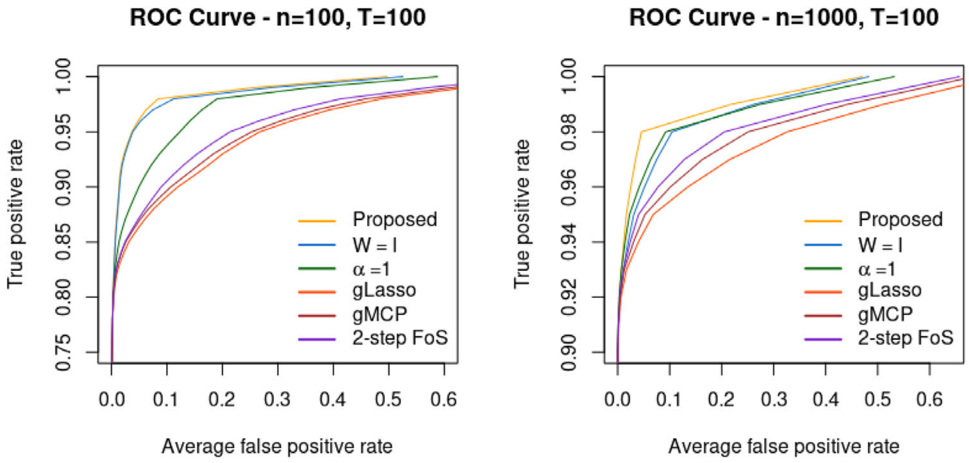

FIGURE 3 ROC curve for each method with $n = 100$ (left) and $n = 1000$ (right), averaged over 100 simulations. This figure appears in color in the electronic version of this paper, and any mention of color refers to that version

counts the number of time points in an interval. Table 1 reports the F_1 score $2/(\text{recall}^{-1} + \text{precision}^{-1})$ under strict sparsity ($\delta = 0$), and Figure 3 shows the receiver operating characteristic (ROC) curve by varying δ, both averaged over 100 simulations. The proposed method gives the highest F_1 score for both sample sizes, corroborating sparsity recovery of functional group bridge. Additional results (Table S1 in the Supporting Information) show that the proposed method leads to the lowest FPR, while the other three competing methods do not produce strict sparsity, thus yielding the same F_1 score. In Figure 3, group bridge–based methods, particularly the weighted version, tend to dominate other methods under δ-sparsity. Overall, the substantial performance gain of the proposed method over $\alpha = 1$ and other methods may be partly due to the functional group bridge penalty for local sparsity and data-dependent weighting for heterogeneous volatility.

5 | APPLICATION TO iEEG DATA

In this section, we apply the proposed method to a human iEEG) data set that is collected in Beauchamp's Lab to investigate multisensory integration and has been extensively described in Ozker et al. (2018), Karas et al. (2019), and Magnotti et al. (2020). In this experiment, participants either listened to recordings of words (auditory-only condition, A) or viewed videos and listened to recordings of words (AV condition). We are interested in analyzing the contrasting effect of A versus AV in different brain areas.

Each experimental session contains $n = 64$ to 128 trials due to varied subject hospitalization duration. There are eight participating patients with a total of 65 Superior Temporal Gyrus (STG) electrodes. We fit the proposed model on each subject and electrode separately as the highly precise iEEG measurements localize activities of a small population of neurons nearest each electrode and lead to drastically different signals across electrodes. This separate analysis also allows us to study transferability of our findings. Information borrowing through jointly modeling all electrodes may yield further efficiency gain. This can be achieved by accounting for the spatial feature in a hierarchical model that links trials of various electrodes. The main challenges include the need to formulate the spatial dependence of signals and stochastic random effects across electrodes, and develop scalable algorithms for the increased parameter space.

The original analog traces are measured at 2000 Hz. We apply notch filters to remove line noise and its harmonics (60, 120, and 180 Hz, etc.). Then a common average reference is used to remove common shifts introduced by patient activities. High-gamma oscillations usually stay above 75 Hz; hence, we apply wavelet transform to extract 75-150 Hz activities from the raw analogue traces. The transformed data are further down-sampled to 100 Hz for storage purposes. Each session is sliced into trials according to epoch information. All the trials are aligned to auditory onset, that is, the time when audio stimuli started to emerge. Because there might be visual information such as mouth movement before audio onset, we collect 3 s of data for each trial, with 1 s prior and 2 s posterior to audio onset. Since brain activity levels often shift for each trial and frequency, we calibrate the signals of high-gamma activities against their own baselines (the average signals during the baseline period from −1 to −0.3 s) for each trial and frequency. After the baseline period, we collapse the data by frequencies, resulting in a $T = 301$ time-point functional data for each trial and electrode.

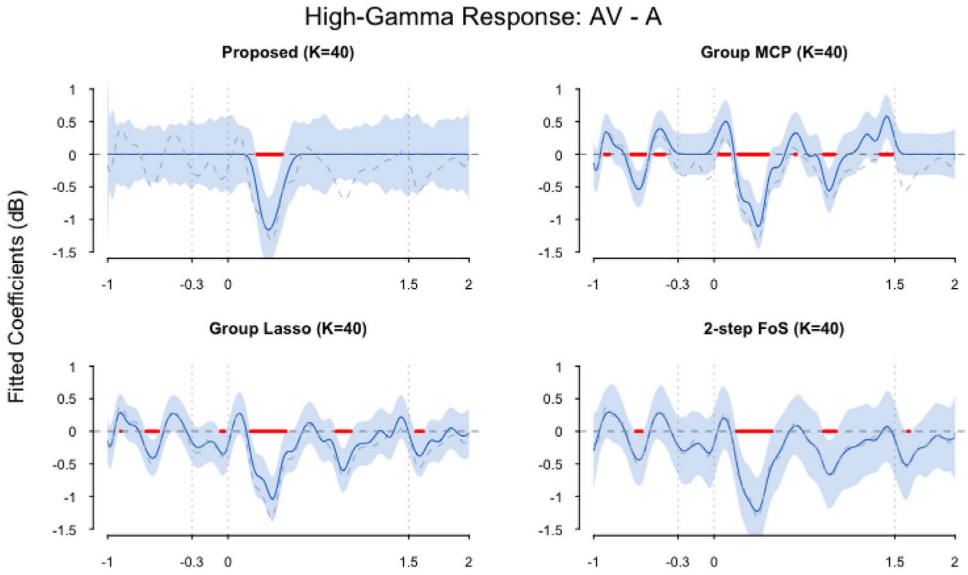

FIGURE 4 Effect of AV versus A estimated by various methods using the data in Figure 1. In each plot, the solid line (blue) is the estimated effect $\widehat{\beta}_2(t)$, shaded area (light-blue) is the joint 95% confidence band, and dashed line (gray) is the estimated effect from ordinary least squares at each time point. This figure appears in color in the electronic version of this paper, and any mention of color refers to that version

The functional response $\boldsymbol{Y}$ is an $n \times 301$ matrix (n ranges from 64 to 128) for each electrode. The design matrix $\boldsymbol{X}$ is $n \times 2$, with the first column being constant one for the intercept and the second column indicating whether visual stimuli are present. The second regression coefficient $\beta_2(t)$ reflects the effect of AV stimuli versus auditory only (A) stimuli, and is of primary interest in this study. The time domain $\mathcal{T}$ is partitioned into four parts: $\mathcal{T}^1 = [-1, -0.3)$ as the baseline period, $\mathcal{T}^2 = [-0.3, 0)$ containing video onset but without audio in, $\mathcal{T}^3 = [0, 1.5)$ when both auditory and visual stimuli are present, and $\mathcal{T}^4 = [1.5, 2]$ as clip offset. Since each trial is calibrated to the baseline, differential brain activities to experiment stimuli are expected to be zero during the baseline period, and nonzero when experiment stimuli present and exhibit effects. Consequently, the estimation of the locally sparse function $\beta_2(t)$ as well as detection of its support is of particular interest.

We implement the proposed method and three other methods considered in the proceeding section. We use $K = 40$ in the B-spline approximation for all methods in view of the increased temporal resolution. For the proposed method, we select λ and α by the adjusted EBIC. Figure 4 plots the fitted coefficients for $\beta_2(t)$, AV versus A effect, using the electrode reported in Figure 1. In the proposed approach, no significant signals are seen in the baseline window, while the other methods deviate from this expectation. The Supporting Information contains a sensitivity analysis suggesting that the proposed method is robust to the choice of K (Figure S4), and additional results for jointly selecting (K, α, λ) using the adjusted EBIC.

According to Karas et al. (2019), we should expect negative AV-A responses in some brain areas after the auditory onset as visual stimuli may suppress activities. Because the proposed functional group bridge method is sparse on nonsignificant responses, it becomes easy not only to observe the suppression (AV < A), but to locate the starting time of that suppression as well as to automate the detection of duration of significant AV versus A effect. Figure 5 visualizes all the 65 STG electrodes using the N27 template brain (Holmes et al., 1998). The lower 5% quantile values within 500 ms after auditory onset indicate that the suppression by additional visual stimuli do exist in the posterior STG area when audio is present. We further test the null hypothesis that there is no suppression (AV ≥ A) for each electrode. The z-score is derived from $\min_{t\in(0,0.5)}\{\widehat{\beta}_2(t)/\widehat{s}_2(t)\}$, where $\widehat{s}_2(t)$ is the estimated standard derivation of $\widehat{\beta}_2(t)$ as described in Section 4 based on (12). The p-values are displayed in the third row of Figure 5. There are 10 electrodes in the posterior STG with p-values smaller than or equal to .01. A closer inspection into the results show that most participants are associated with at least one significant electrode, indicating consistent visual suppression in the posterior STG area for AV stimuli. Our analysis for cross-modal suppression of auditory cortex complements the work by Ozker et al. (2018) and Karas et al. (2019), showing that multisensory interactions are a powerful modulator of activity throughout the speech perception network. Compared with the traditional methods used by Ozker et al. (2018) and Karas et al. (2019), the proposed nonparametric method is more flexible with

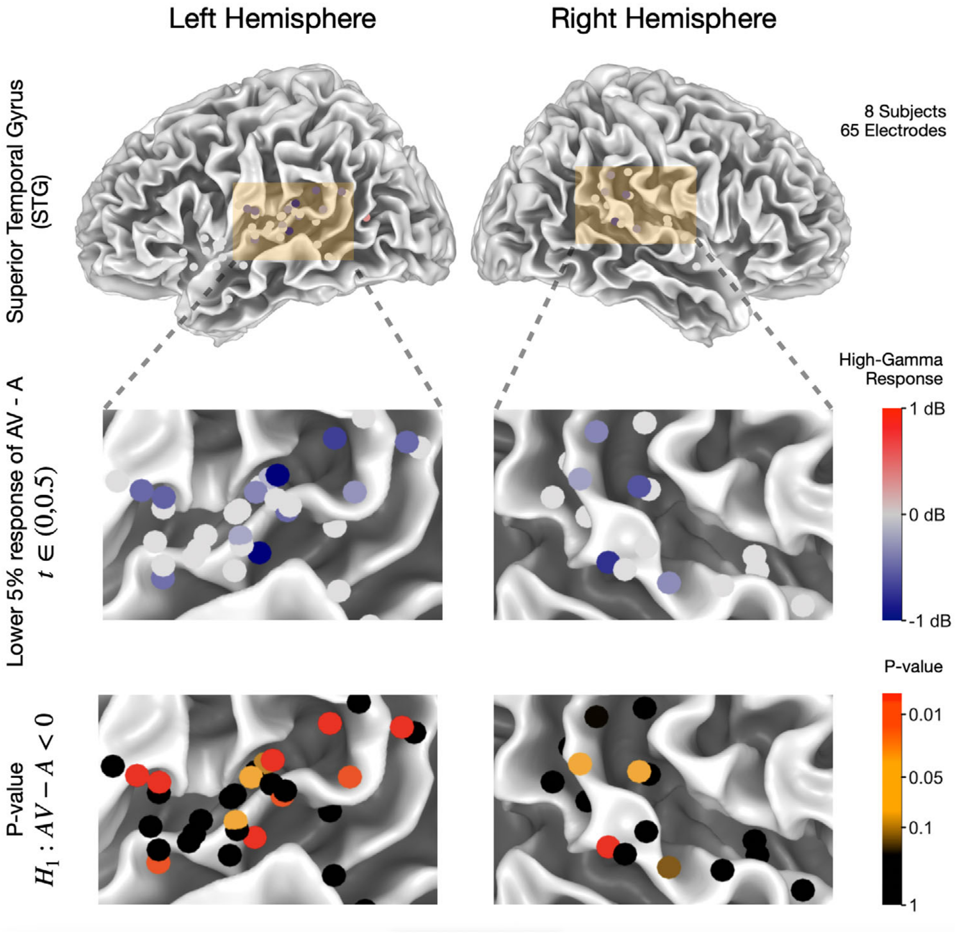

FIGURE 5 First row: visualization of all 65 electrodes mapped onto N27 template brain. Second row: lower 5% quantile value of AV-A from auditory onset to 500 ms after the auditory onset. Color coding: blue for suppression introduced by visual stimuli, gray for little to no differences, and red for that AV is greater than A. Third row: *p*-values for each electrodes with alternative hypothesis of AV less than A response. This figure appears in color in the electronic version of this paper, and any mention of color refers to that version

theoretical support. In addition, the proposed method provides a data-driven approach to find the time window where the brain response to each experimental condition differs, rather than relying on a manually defined window as was done in the initial publications.

ACKNOWLEDGMENTS

The authors are grateful to the editor, associate editor, and three referees for their valuable comments and suggestions. This research was partly supported by the Grant DMS-2015569 from the National Science Foundation and 1R24MH117529 from the BRAIN Initiative of the United States National Institutes of Health.

DATA AVAILABILITY STATEMENT

The intracranial electroencephalography data that support the findings of this paper are openly available at https://doi.org/10.5281/zenodo.6363319 (Wang et al., 2022).

ORCID

Zhengjia Wang https://orcid.org/0000-0001-5629-1116
John Magnotti https://orcid.org/0000-0003-2093-0603
Michael S. Beauchamp https://orcid.org/0000-0002-7599-9934
Meng Li https://orcid.org/0000-0003-2123-2444

REFERENCES

Agarwal, A., Negahban, S. and Wainwright, M.J. (2012) Fast global convergence of gradient methods for high-dimensional statistical recovery. *Annals of Statistics*, 40(5), 2452–2482. https://doi.org/10.1214/12-AOS1032

Barber, R.F., Reimherr, M. and Schill, T. (2017) The function-on-scalar Lasso with applications to longitudinal GWAS. *Electronic Journal of Statistics*, 11, 1351–1389.

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

Cai, T.T. and Yuan, M. (2011) Optimal estimation of the mean function based on discretely sampled functional data: phase transition. *Annals of Statistics*, 39, 2330–2355.

Chen, J. and Chen, Z. (2008) Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95, 759–771.

Chen, Y., Goldsmith, J. and Ogden, R.T. (2016) Variable selection in function-on-scalar regression. *Stat*, 5, 88–101.

De Boor, C. (1978) *A Practical Guide to Splines*, Volume 27. New York: Springer.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004) Least angle regression. *Annals of Statistics*, 32, 407–499.

Fan, J. (1993) Local linear regression smoothers and their minimax efficiencies. *Annals of Statistics*, 21, 196–216.

Fan, J. and Peng, H. (2004) Nonconcave penalized likelihood with a diverging number of parameters. *Annals of Statistics*, 32, 928–961.

Fan, J., Xue, L. and Zou, H. (2014) Strong oracle optimality of folded concave penalized estimation. *Annals of Statistics*, 42, 819.

Fan, J. and Zhang, J.-T. (2000) Two-step estimation of functional linear models with applications to longitudinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62, 303–322.

Frank, I.E. and Friedman, J.H. (1993) A statistical view of some chemometrics regression tools. *Technometrics*, 35, 109–135.

Holmes, C.J., Hoge, R., Collins, L., Woods, R., Toga, A.W. and Evans, A.C. (1998) Enhancement of mr images using registration for signal averaging. *Journal of Computer Assisted Tomography*, 22, 324–333.

Huang, J., Horowitz, J.L. and Wei, F. (2010) Variable selection in nonparametric additive models. *Annals of Statistics*, 38, 2282–2313.

Huang, J., Ma, S., Xie, H. and Zhang, C.-H. (2009) A group bridge approach for variable selection. *Biometrika*, 96, 339–355.

James, G.M., Wang, J. and Zhu, J. (2009) Functional linear regression that's interpretable. *Annals of Statistics*, 37, 2083–2108.

Kaiju, T., Doi, K., Yokota, M., Watanabe, K., Inoue, M. Ando, Ando, H. et al., (2017) High spatiotemporal resolution ECoG recording of somatosensory evoked potentials with flexible micro-electrode arrays. *Frontiers in Neural Circuits*, 11, 20.

Karas, P.J., Magnotti, J.F., Metzger, B.A., Zhu, L.L., Smith, K.B., Yoshor, D. et al. (2019) The visual speech head start improves perception and reduces superior temporal cortex responses to auditory speech. *eLife*, 8, e48116.

Knight, K. and Fu, W. (2000) Asymptotics for Lasso-type estimators. *Annals of Statistics*, 28, 1356–1378.

Lachaux, J.-P., Axmacher, N., Mormann, F., Halgren, E. and Crone, N.E. (2012) High-frequency neural activity and human cognition: past, present and possible future of intracranial EEG research. *Progress in Neurobiology*, 98, 279–301.

Lin, Z., Cao, J., Wang, L. and Wang, H. (2017) Locally sparse estimator for functional linear regression models. *Journal of Computational and Graphical Statistics*, 26, 306–318.

Ma, X. and Kundu, S. (2022) Multi-task learning with high-dimensional noisy images. *arXiv preprint 2103.03370v3*.

Majzoobi, L., Shah-Mansouri, V. and Lahouti, F. (2018) Analysis of distributed ADMM algorithm for consensus optimisation over lossy networks. *IET Signal Processing*, 12, 786–794.

Magnotti, J.F., Wang, Z. and Beauchamp, M.S. (2020) RAVE: Comprehensive open-source software for reproducible analysis and visualization of intracranial EEG data. *NeuroImage*, 223, 117341. https://doi.org/10.1016/j.neuroimage.2020.117341

Ozker, M., Yoshor, D. and Beauchamp, M.S. (2018) Frontal cortex selects representations of the talker's mouth to aid in speech perception. *eLife*, 7, e30387.

Ramsay, J. and Silverman, B.W. (2005) *Functional Data Analysis*, 2nd edition. Springer Series in Statistics. New York: Springer-Verlag.

Schumaker, L. (2007) *Spline Functions: Basic Theory*. Cambridge: Cambridge University Press.

Tibshirani, R. (1996) Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B: Methodological*, 58, 267–288.

Wang, H. and Kai, B. (2015) Functional sparsity: global versus local. *Statistica Sinica*, 25, 1337–1354.

Wang, Z., Magnotti, J., Beauchamp, M. and Li, M. (2022) iEEG data for "Functional Group Bridge for Simultaneous Regression and Support Estimation." *Zenodo*, Version 0.9.0, https://doi.org/10.5281/zenodo.6363319.

Zhang, C.-H. (2010) Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, 38, 894–942.

Zhou, J., Wang, N.-Y. and Wang, N. (2013) Functional linear model with zero-value coefficient function at sub-regions. *Statistica Sinica*, 23, 25–50.

Zhu, H., Fan, J. and Kong, L. (2014) Spatially varying coefficient model for neuroimaging data with jump discontinuities. *Journal of the American Statistical Association*, 109, 1084–1098.

SUPPORTING INFORMATION

Web Appendices on proofs and additional simulations, Figures, and Tables referenced in Sections 3, 4 and 5 are available with this paper at the Biometrics website on Wiley Online Library. The proposed method is implemented in the R package `spfda`, posted online with this paper and also available on CRAN (https://cran.r-project.org/web/packages/spfda/index.html).

How to cite this article: Wang, Z., Magnotti, J., Beauchamp, M.S. and Li, M. (2022) Functional group bridge for simultaneous regression and support estimation. *Biometrics*, 1–13. https://doi.org/10.1111/biom.13684