The Canadian Journal of Statistics
Vol. 50, No. 4, 2022, Pages 1254–1269
La revue canadienne de statistique

Sparse estimation of historical functional linear models with a nested group bridge approach

Xiaolei XUN[1], Tianyu GUAN[2], and Jiguo CAO[3*]

[1]*Global Statistics and Data Science, BeiGene, Inc., Shanghai, China*
[2]*Department of Mathematics & Statistics, Brock University, St. Catharines, Ontario, Canada*
[3]*Department of Statistics & Actuarial Science, Simon Fraser University, Burnaby, British Columbia, Canada*

Key words and phrases: Finite element; function-on-function regression; functional data analysis; historical linear model; nested group bridge penalty.

MSC 2020: Primary 62G05; secondary 62G08.

Abstract: The conventional historical functional linear model relates the current value of the functional response at time t to all past values of the functional covariate up to time t. Motivated by situations where it is more reasonable to assume that only recent, instead of all, past values of the functional covariate have an impact on the functional response, in this work we investigate the historical functional linear model with an unknown forward time lag into the history. In addition to estimating the bivariate regression coefficient function, we also aim to identify the historical time lag from the data, which is important in many applications. To this end, we propose an estimation procedure that uses the finite element method to conform naturally to the trapezoidal domain of the bivariate coefficient function. We use a nested group bridge penalty to facilitate simultaneous estimation of the bivariate coefficient function and the historical lag, and show that our proposed estimators are consistent. We demonstrate this method of estimation in a real data example investigating the effect of muscle activation recorded via the noninvasive electromyography (EMG) method on lip acceleration during speech production. In addition, we examine the finite sample performance of our proposed method in comparison with the conventional approach to estimation via simulation studies.

Résumé: Le modèle linéaire fonctionnel historique classique relie la valeur actuelle de la réponse fonctionnelle au temps t à toutes les valeurs antérieures de la covariable fonctionnelle jusqu'au temps t. Motivés par des situations où il est plus raisonnable de supposer que seules les valeurs passées récentes de la covariable fonctionnelle ont un impact sur la réponse fonctionnelle, les auteurs de ce travail considèrent un modèle linéaire fonctionnel historique avec décalage temporel inconnu dans l'histoire. En plus d'estimer la fonction du coefficient de régression bivarié, ils cherchent à identifier le décalage historique des données, ce qui est important dans de nombreuses applications. À cette fin, et pour se conformer naturellement au domaine trapézoïdal de la fonction de coefficient bivarié, ils proposent une procédure d'estimation basée sur la méthode des éléments finis. Ils utilisent une pénalité de pont de groupe imbriqué pour faciliter l'estimation simultanée de la fonction du coefficient bivarié et du décalage historique. Ils démontrent en outre la convergence des estimateurs proposés et illustrent leur méthode sur des données concernant l'effet de l'activation musculaire sur l'accélération de la lèvre pendant un discours. Ces données ont été enregistrées par électromyographie non invasive (EMG). Des études de simulations à taille finie leur ont également permis de conclure que la performance des estimateurs proposés est supérieure à celle des méthodes classiques.

Additional Supporting Information may be found in the online version of this article at the publisher's website.

* *Corresponding author: jiguo_cao@sfu.ca*

1. INTRODUCTION

We are interested in a function-on-function linear model that is able to properly describe the relationship of a feed-forward nature between the history of a functional covariate and the current state of a functional response. Given a functional covariate $x_i(t)$ and a functional response $y_i(t)$, both observed over a time interval $[0, T]$ for $i = 1, \ldots, N$, the historical functional linear model (Malfait & Ramsay, 2003) is expressed as

$$y_i(t) = \alpha(t) + \int_{\max(0,t-\delta)}^{t} \beta(s,t)x_i(s)ds + \varepsilon_i(t), \quad t \in [0,T], \tag{1}$$

where $\alpha(t)$ is the intercept function, $\beta(s,t)$ is the bivariate regression coefficient function, $\delta \in [0,T]$ is a constant, and $\varepsilon_i(t)$ is the residual. The bivariate coefficient function $\beta(s,t)$ represents the effect of the functional covariate at time s on the functional response at time t. For a given δ, the model identified in Equation (1) specifies that the response $y(t)$ at current time t is possibly affected by past values of the covariate $x(s)$ during the time interval $[\max(0, t-\delta), t]$, neither beyond a historical window of length δ nor the current time t. Therefore, δ represents a forward time lag into the history during which $y(t)$ is related to $x(t)$. In practical problems, the lag δ is typically unknown and of interest. For example, the status of a patient with chronic disease may well depend on treatment received during the past few days but not longer. The objective of our study is to estimate the historical lag δ and the coefficient function $\beta(s,t)$ from the data.

The model identified in Equation (1) is rather different from alternative models for function-on-function regression and is more reasonable in certain situations. The classical function-on-function linear model $y(t) = \alpha(t) + \int_0^T \beta(s,t)x(s)ds + \varepsilon(t), t \in [0,T]$ uses all past values of the covariate $x(s)$ to explain current response $y(t)$ (Ramsay & Dalzell, 1991). The historical model identified in Equation (1) leads to a much more parsimonious representation when a historical effect exists. Additionally, in the previous classical model, future values of the covariate $x(s)$ with $s > t$ are used to explain the current response $y(t)$. This is reasonable if the underlying process is periodic, but illogical otherwise.

Notice that the degenerate case of the model identified in Equation (1) with $\delta = 0$ reduces to the functional concurrent model $y(t) = \alpha(t) + \beta(t)x(t) + \varepsilon(t)$, for $t \in [0,T]$, under which the predicted response $y(t)$ only depends on the concurrently observed covariate $x(t)$. A restriction on $\beta(s,t)$ is then introduced and a univariate coefficient function $\beta(t)$ suffices to fully describe their relationship (Ramsay & Silverman, 2005). Such a model is applicable only if there is no historical dependence and otherwise becomes too constrained.

Concerning the classic function-on-function regression model, one can apply basis expansion to $x_i(t)$ and $y_i(t)$, and obtain a weighted least squares type estimator, possibly with a bivariate roughness penalty (Besse & Cardot, 1996; Ramsay & Silverman, 2005). Alternatively, one can compute the functional principal component scores for $x_i(t)$ and $y_i(t)$ and base the estimation on the functional principal component scores (Yao, Müller & Wang, 2005; Chen & Wang, 2011; Ivanescu et al., 2015). Cai, Xue & Cao (2020) proposed a variable selection procedure for the function-on-function regression model using the group smoothly clipped absolute deviation regulation method. Cai, Xue & Cao (2021) presented a robust method for the function-on-function regression model using M-estimation and penalized spline regression.

The functional concurrent model, as a type of varying coefficient model, has also been studied in various papers in the literature (Hastie & Tibshirani, 1993; Wu, Chiang & Hoover, 1998; Fan & Zhang, 2000; Wu & Liang, 2004; Zhou, Huang & Carroll, 2008). Hall & Hooker (2016) proposed a truncated linear model when the response is a scalar variable. If repeated measurements are available on multiple subjects, Liu, Wang & Cao (2017) suggested using a functional linear mixed-effects regression model. Jiang et al. (2020) introduced a semiparametric functional single

index model to study the relationship between a response and multiple functional covariates. For a comprehensive introduction to functional regression models, we refer to monographs by Ferraty & Vieu (2006), Hsing & Eubank (2015), Kokoszka & Reimherr (2017), Ramsay & Silverman (2005), and the review papers by Morris (2015) and Wang, Chiou & Müller (2016) and references therein.

There also exist intermediate models between the classical functional linear model and the functional concurrent model, which try to model the effect of past values of the predictor on the current response. A class of time-varying functional regression models was discussed by Müller & Zhang (2005), Şentürk & Müller (2008, 2010), etc. Brockhaus et al. (2017) and Greven & Scheipl (2017) discussed a general framework for functional regression, where many aforementioned models are included as special cases, and proposed gradient-boosting-based estimation procedures. Kim, Şentürk & Li (2011) proposed an estimation procedure that was geared towards sparse longitudinal data with irregular observation times and a small number of measurements per subject. Assuming the dependence of $y(t)$ on $x(s)$ for $s \in [t-\delta, t]$ does not change over time, Asencio, Hooker & Gao (2014) introduced the functional convolution model, which is a functional extension of the distributed lag models in time series, and proposed a penalized ordinary least squares estimator for the regression coefficient given the historical time lag δ.

The historical functional linear model specified in Equation (1) is identifiable but not estimable, with effectively an infinite number of covariates, therefore regularization or a roughness penalty on $\beta(s,t)$ is necessary in the estimation process (Ramsay & Silverman, 2005). Malfait & Ramsay (2003) regularized the fit by approximating $\beta(s,t)$ using an expansion with a finite number of basis functions. Such an approach introduces the well-known limitation that a small number of basis functions would decrease the goodness of fit while a large number of basis functions would lead to unstable estimation. Harezlak et al. (2007) improved the fit of the model identified in Equation (1) by imposing a discrete roughness penalty that forces neighbouring coefficients to be similar, which is an extension of the P-spline by Eilers & Marx (1996). Both articles focus on estimation of the coefficient function $\beta(s,t)$ for a given historical lag δ, while the lag δ is decided in a separate manner. Malfait & Ramsay (2003) considered treating the value of δ as a model selection problem. Briefly put, they divided the whole time course $[0,T]$ into M sub-intervals of equal length, fitted the model with $\delta = k/M$ for $k = 0, \ldots, M$, and chose the best model from the $M+1$ candidate models according to a certain criterion.

In this work, we focus on the problem of estimating the historical lag and propose a tailored procedure for the simultaneous estimation of the coefficient function $\beta(s,t)$ and the lag δ in the model identified in Equation (1). Consider the following "full" model,

$$y_i(t) = \alpha(t) + \int_0^t x_i(s)\beta(s,t)ds + \epsilon_i(t), \quad \text{for } t \in [0,T]. \tag{2}$$

Figure 1 illustrates a scenario when the bivariate coefficient function $\beta(s,t)$ becomes zero over the triangular region defined by the vertices $(0,\delta), (0,T)$, and $(T-\delta, T)$. This scenario corresponds to the situation where $x(s)$ does not affect $y(t)$ for $s < t-\delta$. In other words, the full model identified in Equation (2) is equivalent to the historical functional linear model specified in Equation (1) when the support domain of the coefficient function $\beta(s,t)$ is a trapezoidal area S defined by vertices $(0,0), (T,T), (T-\delta, T)$, and $(0,\delta)$, given that the lag $\delta > 0$.

We propose a nested group bridge shrinkage method to estimate the historical lag δ and the coefficient function $\beta(s,t)$ with sparsity, which tackles the problem from a completely different perspective. Our method involves two key features. First of all, we utilize the triangular basis functions from the finite element method theory, which respects the nonrectangular

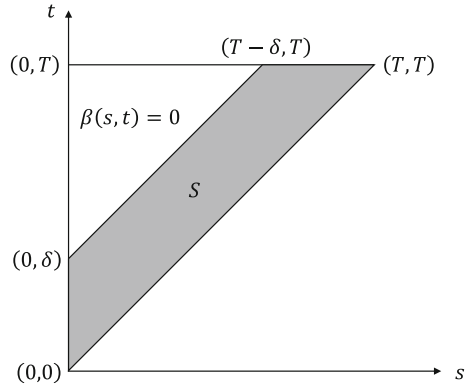

FIGURE 1: The support domain, S, of the coefficient function $\beta(s,t)$ as indicated by the grey area.

support domain of the coefficient function $\beta(s,t)$. Such a basis was also used by Malfait & Ramsay (2003) and Harezlak et al. (2007), as well as in spatial data analysis, where data are distributed over irregularly shaped spatial domains with features like complex boundaries, strong concavities, and interior holes (Ramsay, 2002; Sangalli, Ramsay & Ramsay, 2013). Then under the model specified in Equation (2), we organize the basis coefficients in such a way that the nested group bridge penalty is able to shrink specifically the coefficient function $\beta(s,t)$ over the upper triangular region with vertices $(0,T)$, $(0,\delta)$, and $(T-\delta,T)$ towards zero. The group bridge shrinkage was originally proposed by Huang et al. (2009) for variable selection. Such a penalty has been utilized by Wang & Kai (2015) for locally sparse estimation in nonparametric regression for a scalar-on-function historical functional linear model with a B-spline basis function expansion. The major advantage of our proposed approach is that we can estimate δ automatically without predetermining the set of candidate values. Our simulation studies show that our estimator of δ exhibits better finite sample performance than competing conventional methods of estimation. The nested group bridge shrinkage method is also used by Guan, Lin & Cao (2020) to estimate the scalar-on-function historical functional linear models. Note that this article concerns estimation for the function-on-function historical functional linear models. These two models are fundamentally different in many aspects, as indicated in much of the published literature concerning these two models. For instance, the function-on-function historical functional linear model involves estimating a bivariate functional regression coefficient $\beta(s,t)$, which is much more difficult than estimating the univariate functional coefficient in the scalar-on-function historical functional linear model as shown in Guan, Lin & Cao (2020). We propose to approximate $\beta(s,t)$ by the two-dimensional triangular basis functions, while Guan, Lin & Cao (2020) estimated the univariate functional coefficient with the one-dimensional B-spline basis functions.

The rest of the article is organized as follows. In Section 2 we introduce our proposed estimation procedure and the associated computational details. The asymptotic properties of the derived estimators are identified in Section 3. In Section 4 we describe the use of our method of estimation to analyze speech production data. Section 5 outlines our findings based on simulation studies of the finite sample performance of our method. Section 6 contains some conclusions and summary remarks. All technical details may be found in the accompanying Supplementary Material.

2. METHOD

To ease the notation, the intercept function $\alpha(t)$ can be dropped from the model identified in Equation (2) without loss of generality. Let $y_i^*(t) = y_i(t) - \overline{y}(t)$ and $x_i^*(t) = x_i(t) - \overline{x}(t)$ denote the pointwise-centred response curves and predictor curves, respectively. A centred model that

omits the intercept function is

$$y_i^*(t) = \int_0^t x_i^*(s)\beta(s,t)ds + \varepsilon_i^*(t), \quad t \in [0,T]. \tag{3}$$

Upon obtaining an estimate $\widehat{\beta}(s,t)$, the intercept function is then estimated as

$$\widehat{\alpha}(t) = \overline{y}(t) - \int_0^t \overline{x}(s)\widehat{\beta}(s,t)ds.$$

In what follows, we drop the asterisk from the model specified in Equation (3) and focus our discussion on the estimation of the coefficient function $\beta(s,t)$ in the model

$$y_i(t) = \int_0^t x_i(s)\beta(s,t)ds + \varepsilon_i(t), \quad t \in [0,T]. \tag{4}$$

2.1. Approximation with the Finite Element Method

We propose to approximate the coefficient function $\beta(s,t)$ with the triangular basis, which originates from the finite element method and is widely used in the numerical solution of boundary-value problems involving partial differential equations. Noticing that the support of the coefficient function $\beta(s,t)$ is nonrectangular, approximation with a commonly used bivariate spline generated via a tensor product would result in a jagged shape along the boundary $t = s$. Therefore, finite elements that can approximate arbitrary regions serve as a natural alternative.

Let $\phi_1(s,t), \ldots, \phi_K(s,t)$ denote the K known triangular basis functions. The coefficient function $\beta(s,t)$ is approximated by the expansion,

$$\beta(s,t) \approx \sum_{k=1}^{K} b_k \phi_k(s,t).$$

Plugging the above approximation into the model identified in Equation (3), we obtain

$$\begin{aligned} y_i(t) &= \sum_{k=1}^{K} b_k \int_0^t x_i(s)\phi_k(s,t)ds + \varepsilon_i'(t) \\ &= \sum_{k=1}^{K} b_k \psi_{ik}(t) + \varepsilon_i'(t), \end{aligned}$$

where $\psi_{ik}(t) = \int_0^t x_i(s)\phi_k(s,t)ds$ is known, and $\varepsilon_i'(t)$ includes both the residual $\varepsilon_i(t)$ and the approximation error.

Divide the interval $[0,T]$ on each axis into M subintervals with equidistant nodes $0 = t_0 < t_1 < \ldots < t_M = T$. Further split each square into two triangles by the diagonal parallel to the line $t = s$. This divides the triangular region over which the coefficient function $\beta(s,t)$ is possibly nonzero into M^2 congruent triangles (i.e., the triangular elements or finite elements) with $K = (M+1)(M+2)/2$ nodes. Each node corresponds to a basis function $\phi_k(s,t)$ and a coefficient b_k, both of which are indexed from bottom to top and row-wise from left to right. For example, the left panel in Figure 2 shows the triangulation and indexation of the nodes when $M = 5$. There are 25 triangular elements and 21 nodes corresponding to b_k and $\phi_k(s,t)$ for $k = 1, \ldots, 21$.

 DOI: 10.1002/cjs.11747

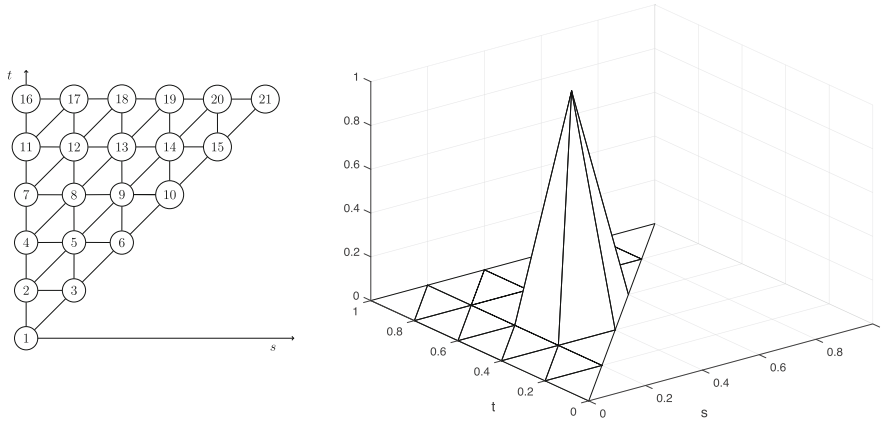

FIGURE 2: Illustration for the triangular basis system when $M = 5$. There are 25 finite elements, 21 nodes, and 21 corresponding basis functions. Left: triangulation and indexation of the nodes. Right: the basis function $\phi_9(s,t)$ corresponding to the ninth node at $(s,t) = (0.4, 0.6)$ shown on the left. It is of a tent shape peaked at the 9th node.

Each basis function $\phi_k(s,t)$ has a compact support, defined over the hexagon centred at the kth node. Basis functions of degree 1 are tent-shaped, piecewise linear, and continuous, with value 1 at node k and value 0 at the boundary of the hexagon. For instance, the right panel in Figure 2 shows a tent-shaped basis function corresponding to the ninth node at $(s,t) = (0.4, 0.6)$. Though the $\phi_k(s,t)$s are not an orthogonal basis, their compact support property provides a certain computational advantage. We refer to Larson & Bengzon (2013) for a comprehensive introduction to the finite element basis system.

2.2. The Nested Group Bridge Approach

The coefficient function $\beta(s,t)$ belonging to the model identified in Equation (4) is formally defined over the triangular region in Figure 1 corresponding to the lag $\delta = T$, while we are trying to identify whether the upper left triangular region is zero or not. Therefore, a proper penalty should be able to shrink the upper left triangular region towards zero specifically, while respecting the nested group structure, to be introduced shortly, among the basis coefficients b_k.

We start by defining a sequence of nested triangular regions. Let $\Delta = T/M$ and D_j denote a triangle with vertices $(0,T), (0,(j-1)\Delta)$, and $((M+1-j)\Delta, T)$, for $j = 1, \ldots, M$, and let $D_{M+1} = \{(0,T)\}$ contain a single point. Notice that $D_1 \supset \ldots \supset D_{M+1}$. Corresponding to these regions, we define a sequence of decreasing groups $A_1 \supset \ldots \supset A_{M+1}$, where A_j consists of the indices of the nodes contained in triangle D_j according to the node indexation described in Section 2.1. Taking $M = 3$ as an example, there are 10 nodes and the index sets are $A_1 = \{7,4,8,2,5,9,1,3,6,10\}$, $A_2 = \{7,4,8,2,5,9\}$, $A_3 = \{7,4,8\}$, and $A_4 = \{7\}$. Furthermore, we denote by $\boldsymbol{b}_{A_j} = \{b_k : k \in A_j\}$ the vector of basis coefficients whose indices belong to set A_j and follow the conventional notation that $||\cdot||_1$ and $||\cdot||_2$ represent the L_1 and L_2 norms, respectively.

We adapt the group bridge approach proposed by Huang et al. (2009) and propose to estimate the vector of basis coefficients $\boldsymbol{b} = (b_1, \ldots, b_K)$ for $\beta(s,t)$ belonging to the model identified in Equation (4) by minimizing the penalized least squares criterion

$$\frac{1}{N}\int_0^T \sum_{i=1}^N \left\{ y_i(t) - \sum_{k=1}^K b_k \psi_{ik}(t) \right\}^2 dt + \lambda \sum_{j=1}^{M+1} c_j ||\boldsymbol{b}_{A_j}||_1^\gamma + \boldsymbol{b}^T \boldsymbol{R} \boldsymbol{b}, \tag{5}$$

with a fixed $\gamma \in (0, 1)$, a nonnegative tuning parameter λ, known weights c_j to offset the effect of different dimensions of A_j, and a known smoothness penalty matrix $\boldsymbol{R}$ to be introduced shortly. Following Huang et al. (2009), a simple choice for the weights is $c_j \propto |A_j|^{1-\gamma}$.

The first term in the criterion specified in Equation (5) is the ordinary least squares taking into account the whole time course $[0, T]$. The second term is the so-called *nested group bridge* penalty, which was introduced by Huang et al. (2009) for simultaneous variable selection at both the group and within-group levels. Consider any two sets A_j and A_k, with $j < k$. Due to the explicit nesting, the vector of coefficients $\boldsymbol{b}_{A_k}$ is always a subvector of $\boldsymbol{b}_{A_j}$, hence the coefficients corresponding to the nodes in region D_k appear in more groups than the ones corresponding only to D_j. This suggests that the nested group bridge penalty shrinks more heavily the coefficients corresponding to regions in a closer proximity to D_{M+1}, as desired. The last term imposes a discrete roughness penalty on the basis coefficients $\boldsymbol{b}$, extending the idea of Eilers & Marx (1996) and Harezlak et al. (2007). Section 1 in the Supplementary Material describes the discrete roughness penalty in greater detail.

2.3. Computation

The key difficulty in optimizing the objective function identified in Equation (5) is due to its lack of convexity. We can work on an equivalent constrained optimization problem, which is convex and easier to solve (Huang et al., 2009). The objective function identified in Equation (5) involves an integration of the time-dependent ordinary least squares criterion over time. Dividing the time course $[0, T]$ into Q equally spaced subintervals at time points t_q, $q = 0, 1, \ldots, Q$, the integrated least squares criterion can be approximated by the finite sum

$$\frac{T}{Q}\sum_{q=1}^{Q}\sum_{i=1}^{N}\left\{ y_i(t_q) - \sum_{k=1}^{K} b_k \psi_{ik}(t_q) \right\}^2 .$$

With a sufficiently large Q, the above approximation with Riemann sums over a regular partition of the integral domain provides a satisfactory level of precision with acceptable computational load. Define the Q-by-1 vector $\boldsymbol{y}_i = (y_i(t_1), \ldots, y_i(t_Q)^T$, and the Q-by-K matrix $\boldsymbol{\Psi}_i$ whose (q, k)th element is $\psi_{ik}(t_q)$. By stacking the vectors $\boldsymbol{y}_i$ into a long vector $\boldsymbol{y}$ of length NQ and the matrices $\boldsymbol{\Psi}_i$ into a tall matrix $\boldsymbol{\Psi}$ of size NQ by K, the approximated integrated least squares criterion can be represented as the matrix expression

$$\frac{T}{Q}(\boldsymbol{y} - \boldsymbol{\Psi b})^T (\boldsymbol{y} - \boldsymbol{\Psi b}) .$$

The objective function identified in Equation (5) is then rewritten as the familiar penalized least squares

$$\frac{1}{N}(\boldsymbol{y} - \boldsymbol{\Psi b})^T (\boldsymbol{y} - \boldsymbol{\Psi b}) + \lambda \sum_{j=1}^{M+1} c_j \left|\left|\boldsymbol{b}_{A_j}\right|\right|_1^{\gamma} + \boldsymbol{b}^T \boldsymbol{R b}, \tag{6}$$

where λ differs from earlier notation by a constant factor Q/T.

Define a vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{M+1})$. Then $\widehat{\boldsymbol{b}}_n$ minimizes the criterion specified in Equation (6) if and only if $\left(\widehat{\boldsymbol{b}}_n, \widehat{\boldsymbol{\theta}}\right)$ minimizes

$$\frac{1}{N}(\boldsymbol{y} - \boldsymbol{\Psi b})^T (\boldsymbol{y} - \boldsymbol{\Psi b}) + \sum_{j=1}^{M+1} \theta_j^{1-1/\gamma} c_j^{1/\gamma} \left|\left|\boldsymbol{b}_{A_j}\right|\right|_1 + \tau \sum_{j=1}^{M+1} \theta_j + \boldsymbol{b}^T \boldsymbol{R b}, \tag{7}$$

$$s.t. \quad \theta_j \geq 0, j = 1, \ldots, M + 1,$$

for $0 < \gamma < 1$ and $\tau = [\lambda\gamma^{\gamma}(1-\gamma)^{1-\gamma}]^{1/(1-\gamma)}$ (Huang et al., 2009). Notice that $\left|\left|\boldsymbol{b}_{A_j}\right|\right|_1 = \sum_{k\in A_j}|b_k|$ by definition of the L_1 norm, and that if b_k appears in A_j then it also appears in all A_s with $s < j$. Thus, if we exchange the order of summation in the second term of the expression in Equation (7), and rearrange and collect all the individual terms involving b_k, the objective function becomes

$$\frac{1}{N}(\boldsymbol{y}-\boldsymbol{\Psi b})^T(\boldsymbol{y}-\boldsymbol{\Psi b}) + \sum_{k=1}^{K} g_k|b_k| + \tau\sum_{j=1}^{M+1}\theta_j + \boldsymbol{b}^T\boldsymbol{R}\boldsymbol{b},$$

$$s.t. \quad \theta_j \geq 0, j = 1, \ldots, M+1,$$

where $g_k = \sum_{j=1}^{\ell(k)}\theta_j^{1-1/\gamma}c_j^{1/\gamma}$ with $\ell(k) = \max\{j : b_k \in A_j\}$. Intuitively, $\ell(k)$ is the number of times that coefficient b_k appears in the nested group bridge penalty.

Define $\boldsymbol{G}$ as a $K \times K$ diagonal matrix with the kth diagonal element $(Ng_k)^{-1}$, $\boldsymbol{b}^* = \boldsymbol{G}^{-1}\boldsymbol{b}$, $\boldsymbol{\Psi}^* = \left[\boldsymbol{\Psi}^T, \sqrt{\omega_H N}\boldsymbol{D}_H^T, \sqrt{\omega_V N}\boldsymbol{D}_V^T, \sqrt{\omega_P N}\boldsymbol{D}_P^T\right]^T\boldsymbol{G}$, and $\boldsymbol{y}^* = (\boldsymbol{y}^T, \boldsymbol{0}^T)^T$ of proper length. Finally, the objective function identified in Equation (7) is expressed as

$$\frac{1}{N}\left\{\left(\boldsymbol{y}^* - \boldsymbol{\Psi}^*\boldsymbol{b}^*\right)^T\left(\boldsymbol{y}^* - \boldsymbol{\Psi}^*\boldsymbol{b}^*\right) + \sum_{k=1}^{K}\left|b_k^*\right|\right\} + \tau\sum_{j=1}^{M+1}\theta_j,$$

$$s.t. \quad \theta_j \geq 0, j = 1, \ldots, M+1,$$

where b_k^* is the kth element of vector $\boldsymbol{b}^*$. The following iterative algorithm is used to compute $\widehat{\boldsymbol{b}}_n$.

- *Step* 1: Obtain an initial estimate $\boldsymbol{b}^{(0)}$ with ordinary least squares.
- *Step* 2: At each iteration s, update $\boldsymbol{\theta}$ based on $\boldsymbol{b}^{(s-1)}$ as

$$\theta_j^{(s)} = c_j\left(\frac{1-\gamma}{\tau\gamma}\right)^{\gamma}\left|\left|\boldsymbol{b}_{A_j}^{(s-1)}\right|\right|_1^{\gamma}, \quad j = 1, \ldots, M+1,$$

and correspondingly

$$g_k^{(s)} = \sum_{j=1}^{\ell(k)}\left(\theta_j^{(s)}\right)^{1-1/\gamma}c_j^{1/\gamma}, \quad k = 1, \ldots, K,$$

$$\boldsymbol{G}^{(s)} = N^{-1}\mathrm{diag}\left(1/g_1^{(s)}, \ldots, 1/g_K^{(s)}\right),$$

$$\boldsymbol{\Psi}^{*(s)} = \left[\boldsymbol{\Psi}^T, \sqrt{\omega_H N}\boldsymbol{D}_H^T, \sqrt{\omega_V N}\boldsymbol{D}_V^T, \sqrt{\omega_P N}\boldsymbol{D}_P^T\right]^T\boldsymbol{G}^{(s)}.$$

- *Step* 3: At each iteration s, update $\boldsymbol{b}$ based on $\boldsymbol{\theta}^{(s)}$ by recognizing this is a LASSO problem

$$\boldsymbol{b}^{(s)} = \boldsymbol{G}^{(s)}\mathrm{argmin}_{\boldsymbol{b}^*}\left\{\left(\boldsymbol{y} - \boldsymbol{\Psi}^{*(s)}\boldsymbol{b}^*\right)^T\left(\boldsymbol{y} - \boldsymbol{\Psi}^{*(s)}\boldsymbol{b}^*\right) + \sum_{k=1}^{K}\left|b_k^*\right|\right\}.$$

- Repeat Step 2 and Step 3 until the algorithm converges.

Given $\widehat{\boldsymbol{b}}_N$, the estimators for $\beta(s,t)$ and δ are defined as

$$\widehat{\beta}_N(s,t) = \sum_{k=1}^{K} \widehat{b}_{N,k}\phi_k(s,t), \quad \widehat{\delta}_N = \frac{T}{M}\min_j\{1 \le j \le M : \widehat{\beta}_N(s,t) = 0 \text{ on } D_j\}.$$

After obtaining $\widehat{\delta}_N$, we then refine the estimation for $\beta(s,t)$ by minimizing the criterion identified in Equation (6) with $\lambda = 0$, that is, excluding the nested grouped bridge penalty.

In our implementation of the algorithm, our numerical studies suggested that the value of γ has little impact on the results. In this article, we set $\gamma = 0.5$ as suggested in Huang et al. (2009). Furthermore, there are two types of tuning parameters, that is, the shrinkage parameter λ (or equivalently τ) and the smoothness parameters as one type, and the number of grids on the time interval $[0,T]$ as the other type. Since the precision of $\widehat{\delta}_N$ obviously depends on M, we desire a relatively large M in order to achieve a reasonably good estimate of the historical lag δ, and at the same time to capture enough local features of the coefficient function $\beta(s,t)$. This is also consistent with the common strategy when applying the penalized least squares approach, where a relatively large number of nodes is preferred and potential overfitting caused by such a choice would be offset via the sparsity and roughness penalty. It is possible to consider a calibration experiment for M as described in Malfait & Ramsay (2003). Briefly put, one can set up some known coefficient functions, create datasets using the observed covariate curves, test the method, and choose an M that works well with the simulated data. The shrinkage parameter and smoothness parameters can be chosen via the Bayesian information criterion (BIC), for example. The effective degrees of freedom for given λ and $\boldsymbol{\kappa} = (\omega_H, \omega_V, \omega_P)$ can be approximated by

$$df(\lambda,\boldsymbol{\kappa}) = \text{trace}\left(\boldsymbol{\Psi}_s\left(\boldsymbol{\Psi}_s^T\boldsymbol{\Psi}_s + N\boldsymbol{R}_s\right)^{-1}\boldsymbol{\Psi}_s^T\right),$$

where $\boldsymbol{\Psi}_s$ consists of the columns of $\boldsymbol{\Psi}$ corresponding to the nonzero coefficients in $\widehat{\boldsymbol{b}}_N$, and $\boldsymbol{R}_s$ is obtained in the same way by properly selecting the columns of $\boldsymbol{D}_H$, $\boldsymbol{D}_V$, and $\boldsymbol{D}_P$. The BIC is then approximated by

$$BIC(\lambda,\boldsymbol{\kappa}) = N\log(||\boldsymbol{y} - \boldsymbol{\Psi}\widehat{\boldsymbol{b}}_N(\lambda,\boldsymbol{\kappa})||_2^2/N) + \log(N)df(\lambda,\boldsymbol{\kappa}),$$

and the tuning parameters are selected as the minimizer of $BIC(\lambda,\boldsymbol{\kappa})$.

2.4. A Confidence Interval for the Historical Lag δ

A 95% confidence interval for the historical lag δ can be constructed via the bootstrap method as follows. The residual can be calculated as $\hat{r}_i(t) = y_i^*(t) - \int_0^t x_i^*(s)\widehat{\beta}_N(s,t)ds, i = 1,\ldots,N$. For the ℓth bootstrap sample, $\ell = 1,\ldots,L$, the new response data are generated via $y_i^{(\ell)}(t) = \int_0^t x_i^*(s)\widehat{\beta}_N(s,t)ds + r_i^{(\ell)}(t)$, where $r_i^{(\ell)}(t)$ is randomly sampled from $\{\hat{r}_1(t),\ldots,\hat{r}_N(t)\}$. The historical lag δ is estimated from the generated data $\left(x_i^*(t), y_i^{(\ell)}(t)\right), i = 1,\ldots,N$, by our iterative algorithm, where we denote the bootstrap estimate as $\widehat{\delta}_N^{(\ell)}$. The 95% confidence interval for the historical lag δ is constructed as $\left[\delta_{2.5\%}, \delta_{97.5\%}\right]$, where $\delta_{2.5\%}$ and $\delta_{97.5\%}$ are the 2.5 and 97.5 percentiles of the bootstrap estimates $\widehat{\delta}_N^{(\ell)}$, $\ell = 1,\ldots,L$.

3. ASYMPTOTIC PROPERTIES

Let δ_0 and $\beta_0(s,t)$ be the true values of δ and $\beta(s,t)$, respectively. Without loss of generality, we assume $T = 1$. If $\delta_0 = 0$, set $J_1 = 0$, and if $\delta_0 = 1$, let $J_1 = M$. Otherwise, let J_1 be an integer such that $\delta_0 \in \left[t_{J_1-1}, t_{J_1}\right)$. Assume $\beta_0(s,t) \in L_2(\Omega)$, where $\Omega = \{(s,t) : 0 \le s \le t \le 1\}$. According

to Theorem 4.9 of Larson & Bengzon (2013), there exists some $\beta_{\text{FEM}}(s,t) = \sum_{k=1}^{K} b_{\text{FEM}k}\phi_k(s,t)$ such that $\|\beta_{\text{FEM}} - \beta_0\|_2 \le C_0 M^{-2}$ for some positive constant C_0. Define $b_{0k} = b_{\text{FEM}k} I_{(k \notin A_{J_1+1})}$, $k = 1, \ldots, K$. Let $\|\cdot\|_2$ denote the L_2 norm in the functional spaces for different domains. That is, for $f \in L_2([0,1])$ and $g, h \in L_2([0,1]\times[0,1])$, $\|f\|_2^2 = \int_0^1 f^2(t)\mathrm{d}t$, $\|g\|_2^2 = \int_0^1\int_0^1 g^2(s,t)\mathrm{d}s\mathrm{d}t$, and $\langle g, h\rangle = \int_0^1\int_0^1 g(s,t)h(s,t)\mathrm{d}s\mathrm{d}t$. In addition, given t_q, $q = 1, \ldots, Q$, let $\|g\|_{2,q}^2 = \int_0^{t_q} g^2(s,t_q)\mathrm{d}s$ and $\langle g, h\rangle_q = \int_0^{t_q} g(s,t_q)h(s,t_q)\mathrm{d}s$. Define Γ and Γ_q as the covariance operators of the random process X, that is, $(\Gamma z)(v) = \int_0^1 E(x(v), x(u))z(u)\mathrm{d}u$ and $(\Gamma_q z)(v) = \int_0^{t_q} E(x(u), x(v))z(u)\mathrm{d}u$. The functions Γ_N and Γ_{Nq} are defined as the empirical versions of Γ and Γ_q, respectively. That is,

$$(\Gamma_N z)(v) = \frac{1}{N}\sum_{i=1}^{N}\int_0^1 x_i(v)x_i(u)z(u)\mathrm{d}u.$$

$$(\Gamma_{N,q} z)(v) = \frac{1}{N}\sum_{i=1}^{N}\int_0^{t_q} x_i(v)x_i(u)z(u)\mathrm{d}u.$$

Let ϕ_{kq} denote the function $\phi_k(s,t_q)$ and let $\boldsymbol{H}_q$ be the $K\times K$ matrix with element $h_q(i,j) = \langle \Gamma_{N,q}\phi_{iq}, \phi_{jq}\rangle_q$. In order to establish the desired asymptotic properties, we assume that the following conditions hold:

C.1 $E\|X\|_2^4 < \infty$ and $E\|\epsilon\|_2^2 < \infty$.

C.2 $M = o(N^{1/2})$, $M = \omega\left(N^{\frac{1}{4}}\right)$, $\omega_H = o(N^{-1/2}M^{-1/2})$, $\omega_V = o(N^{-1/2}M^{-1/2})$, and $\omega_P = o(N^{-1/2}M^{-1/2})$.

C.3 There are constants $C_{\max} > C_{\min} > 0$ such that

$$C_{\min}M^{-1} \le \rho_{\min}\left(\frac{1}{Q}\sum_{q=1}^{Q}\boldsymbol{H}_q\right) \le \rho_{\max}\left(\frac{1}{Q}\sum_{q=1}^{Q}\boldsymbol{H}_q\right) \le C_{\max}M^{-1}$$

with probability tending to one as $N\to\infty$, where $\rho_{\min}$ and $\rho_{\max}$ denote the smallest and largest eigenvalues of a matrix, respectively.

C.4 $\lambda = O(N^{-1/2}M^{-1/2}\eta^{-1})$, where $\eta = \left(\sum_{j=1}^{J_1} c_j^2 \left|\left|\boldsymbol{b}_{0A_j}\right|\right|_1^{2\gamma-2}|A_j|\right)^{1/2}$ with $c_j \propto |A_j|^{1-\gamma}$.

C.5 $\dfrac{\lambda}{M^{1-\gamma}N^{\gamma/2-1}} \to \infty$.

Condition C.1 assures the consistency of the covariance function of X. In Condition C.2, we set the growth rate for the smoothing tuning parameters ω_H, ω_V, and ω_P. Conditions C.4 and C.5 impose certain constraints on the decay rate of the truncation parameter λ. We state the main results below; the corresponding proofs may be found in the Supplementary Material.

Theorem 1 (Convergence rate)**.** *Suppose that conditions C.1–C.5 hold. Then,* $\|\hat{\beta}_N - \beta_0\|_2 = O_P(MN^{-1/2} + M^{-2})$.

Theorem 1 identifies the convergence rate of the estimator $\hat{\beta}_N$. The convergence rate consists of two competing parts, the variance term $MN^{-1/2}$ and the bias term M^{-2}. This implies that when M increases, the approximation to $\beta(s,t)$ by the finite element method improves at the cost of increased variance.

We observe that our convergence rate and the ones provided in Guan, Lin & Cao (2020), Lin et al. (2017), and Zhou, Wang & Wang (2013) are slower than the convergence rate of

the penalized B-spline estimator in Cardot, Ferraty & Sarda (2003). This is partially due to the fact that the penalized B-spline estimator considers only a roughness penalty, whereas the other methods also use shrinkage penalties in the estimation procedure, such as the nested group bridge penalty used in our proposed method. These shrinkage penalties increase the estimation variability and have main effects on the convergence rate.

The next result shows that the estimators of the null region of $\beta(s,t)$ and δ enjoy estimation consistency.

Theorem 2 (Estimation consistency)**.** *Suppose that conditions C.1–C.5 hold.*

(i) *For any* $\zeta \in (0, 1-\delta_0)$, $\hat{\beta}_N(s,t) = 0$ *for* $s + \delta_0 + \zeta \le t \le 1$ *with probability tending to 1.*

(ii) $\hat{\delta}_N$ *converges to* δ_0 *in probability.*

4. ANALYSIS OF THE SPEECH PRODUCTION DATA

In this analysis, we apply our proposed nested group bridge approach to data from a speech production experiment. It is known that there are over 100 muscles that must be controlled centrally during speech production, such as the muscles of the thoracic and abdominal walls, the neck and face, the oral cavity, etc. The timing of activation of different muscle groups is a central issue in anatomical and physiological research of speech. Due to the fact that muscle contractions generate electrochemical changes, the noninvasive electromyography (EMG) method is used to collect data on muscle activation. With electrodes attached to the skin over the muscle, potentials that arise as a result of muscle stimulation can be picked up. In this experiment, a subject said the syllable "bob" 32 times. By Newton's second law, the accelerations of the centre of the lower lip reflect the force applied to tissue by muscle contraction, which was recorded as the response curves. Ramsay & Silverman (2002) provided a comprehensive introduction to the background of this physiology study.

A random subset of 10 smoothed observations for EMG recordings (top) and corresponding lower lip accelerations (bottom) is shown in Figure 3, where the time range is $[0, 0.64]$. Considering that the current lip acceleration is very likely a result of recent muscle movement, this motivates us to fit the historical functional linear model identified in Equation (1) with an unknown lag. The EMG recordings are the functional covariates and the lip accelerations are the functional response. The goal of this analysis was to explore the association between EMG recording and lip acceleration, and identify a historical lag if there is any. To apply the nested group bridge approach, we divided the time range [0,0.64] into 20 subintervals with equidistant nodes, that is, $M = 20$. This led to $K = 231$ nodes and triangular basis functions. By minimizing the criterion specified in Equation (5), as described in Section 2, we obtained estimated values for both the regression coefficient and the historical lag.

The estimated historical lag is $\widehat{\delta} = 0.352$ seconds, indicating there is a historical effect of muscle activation on the lip movement. A 95% confidence interval for the historical lag δ is $(0.348, 0.406)$, constructed via the bootstrap method by resampling the residuals and then re-estimating the model. Figure 4 shows the corresponding estimate of the bivariate coefficient function $\beta(s,t)$. The estimate of $\beta(s,t)$ has extraordinarily large values along the diagonal direction of two regions, $(s,t) \in (0, 0.06) \times (0, 0.06)$ and $(s,t) \in (0.33, 0.44) \times (0.33, 0.44)$. This is as expected and corresponds to the two productions of the /b/ syllable when the lip is closed and the muscle activation is most influential. The width of the diagonal band with large values varies approximately from 50 to 60 milliseconds, which corresponds to the delay for a neural signal to be transduced into muscle contraction. And peaks at larger lags suggest possible covariation of EMG and lip movement that requires solid physiological knowledge for interpretation.

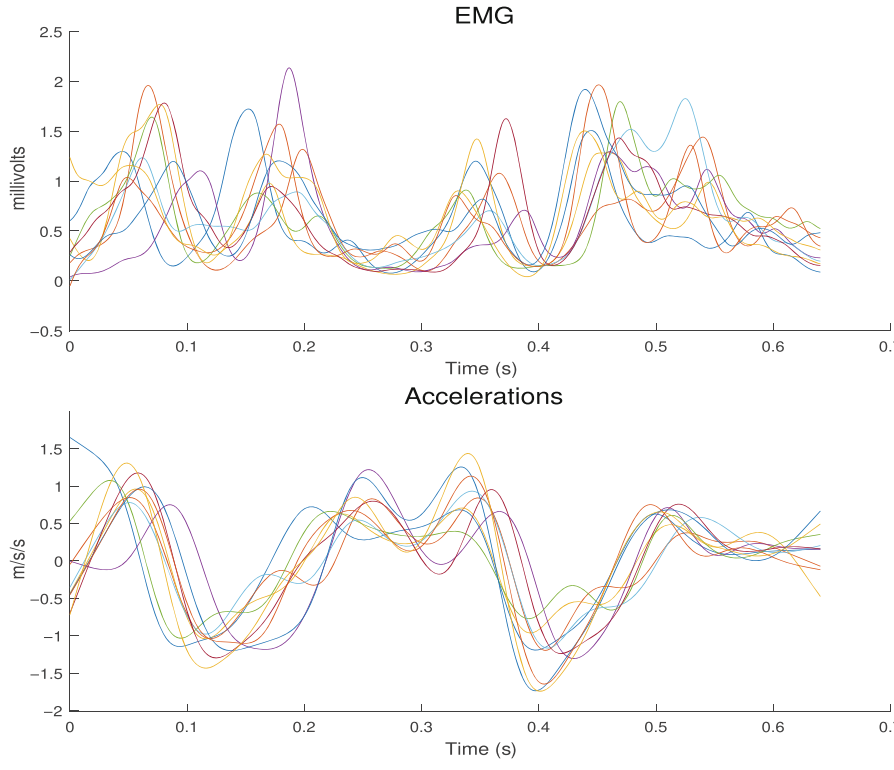

FIGURE 3: A random subset of 10 pairs of EMG and lip acceleration curves. The observation time range is $[0, 0.64]$. Top: the EMG activities associated with the depressor labii inferioris muscle. Bottom: the acceleration of the centre of the lower lip.

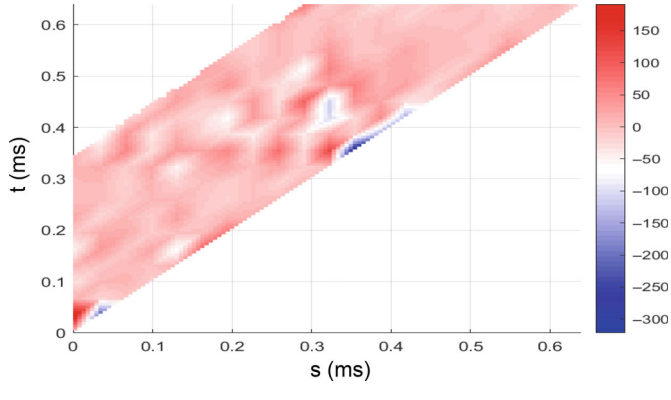

FIGURE 4: The estimated regression coefficient function $\widehat{\beta}(s,t)$ with a historical lag of 0.352 estimated from the data.

Note that there seems to be a mixture of amplitude variation and phase variation in the speech production data. One reviewer suggested incorporating time warping in to our analysis of this dataset. However, we chose not to include time warping for the following reasons. If we incorporated time warping first, it would be hard to interpret the historical effect of $x_i(t)$ on $y_i(t)$. In addition, the predictor $x_i(t)$ and response $y_i(t)$ were paired measurements. If we aligned $x_i(t)$

and $y_i(t)$ separately, we would destroy the relationship between $x_i(t)$ and $y_i(t)$. It is hard to align $x_i(t)$ and $y_i(t)$ simultaneously for each subject. Malfait & Ramsay (2003) analyzed this dataset using the same historical functional linear model, and they also did not include time warping in their analysis.

5. SIMULATION STUDIES

We carried out simulation studies in order to evaluate our proposed estimation procedure with a nested group bridge penalty (NGB) compared with the penalized approach by Harezlak et al. (2007) based on a linear model approximation (PLMA). Set $T = 1$. Let S denote the trapezoidal support with vertices $(0, 0)$, $(1, 1)$, $(1 - \delta, 1)$, and $(0, \delta)$ as shown in Figure 1, and $I(\cdot)$ be the indicator function. We considered the following three scenarios, and the true coefficient function $\beta(s, t)$ is shown in Figure 5.

Scenario 1. *$\beta(s,t) = 10I\{(s,t) \in S_\varepsilon\} + 10\big(\delta/\varepsilon + s/\varepsilon - t/\varepsilon\big)I\{(s,t) \in S \setminus S_\varepsilon\}$, where S_ε is the trapezoid with vertices $(0, 0)$, $(1, 1)$, $(1 - \delta + \varepsilon, 1)$, and $(0, \delta - \varepsilon)$. The coefficient function $\beta(s, t)$ is constant over the region S_ε, linear over $S \setminus S_\varepsilon$, and vanishes at the line $t = s + \delta$.*

Scenario 2. *$\beta(s,t) = 10\big(1 + s/\delta - t/\delta\big) \times I\{(s,t) \in S\}$. The coefficient function $\beta(s, t)$ is linear over its support S, with maximum along the line $t = s$ and vanishes at the line $t = s + \delta$.*

Scenario 3. *We randomly created some "holes" in the coefficient function $\beta(s, t)$ of Scenario 2, such that $\beta(s, t)$ can vanish inside S_ε.*

In Scenario 1, we took a small value of $\varepsilon = 0.05$, which leads to a sharp drop in the coefficient function $\beta(s, t)$ towards zero when going outside from the region S_ε towards the line $t = s + \delta$. We considered this scenario to be the easiest situation in which to determine the historical lag δ due to the sharp change, and expected both methods of estimation to exhibit comparable, satisfactory performance. Scenario 2 was slightly more difficult, since the boundary $t = s + \delta$ becomes more blurred than that in Scenario 1 when the data are contaminated with errors. Scenario 3 added yet another difficulty with its irregularly shaped coefficient function $\beta(s, t)$ and also a more general assumption that the dependence of the response on the historical predictor varied with time. For all scenarios, the observed covariate curves from the example that we analyzed in Section 4 were rescaled to the interval [0, 1] with the true coefficient function $\beta(s, t)$ to generate the mean response curves, possibly to mimic real situations. A random subset of 10 covariate curves is shown in the top panel of Figure 3. We refer readers to Section 4 for details about the example. The errors were additive and generated pointwise from a $N(0, 0.5^2)$ distribution.

For our proposed NGB approach, we set $\gamma = 0.5$ and used the ordinary least squares estimate by Malfait & Ramsay (2003) as an initial value for the algorithm. We chose $c_j = |A_j|^{1-\gamma}/||\boldsymbol{b}_{A_j}^{(0)}||_2^\gamma$ following the suggestion for the adaptive LASSO (Zou, 2006). The tuning parameters were selected based on the BIC criterion over a grid of candidate values. The BIC for our proposed NGB approach is specified near the end of Section 2.3. The BIC for the PLMA approach was similar, with the degrees of freedom defined as $\text{trace}(\boldsymbol{\Psi}(\boldsymbol{\Psi}^T\boldsymbol{\Psi} + N\boldsymbol{R})^{-1}\boldsymbol{\Psi}^T)$. For each scenario, we simulated 100 independent replications. The results that we observed are summarized in Table 1. Concerning the estimated RMSE of $\widehat{\delta}$, our proposed NGB estimator outperformed the PLMA estimator in all three scenarios, exhibiting a much smaller bias and comparable SD. While the PLMA estimator performed well in Scenario 1 with a small bias and a reasonably small SD, the performance deteriorated badly in more difficult scenarios and the estimator also tended to underestimate δ. To assess the accuracy of the estimated bivariate coefficient function,

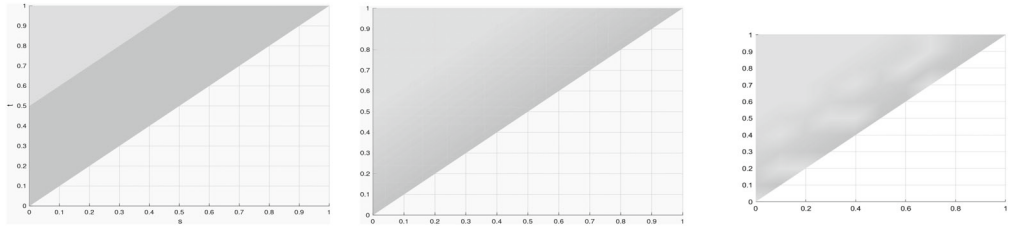

FIGURE 5: The true coefficient function $\beta(s,t)$ used in the simulation scenarios. From left to right are the actual coefficient functions for Scenario 1, 2, and 3, respectively, with increasing difficulty in the estimation.

TABLE 1: Summary of the 100 simulation replications for each scenario, including the root mean squared error (RMSE), percent bias (%bias), and standard deviation (SD) of $\widehat{\delta}$, and the square root of the mean integrated squared errors (RISE) of $\widehat{\beta}(s,t)$ along with the corresponding SD in parentheses. The two methods implemented are the proposed nested group bridge approach (NGB) and the penalized linear model approximation approach (PLMA).

	RMSE of $\widehat{\delta}$		%bias of $\widehat{\delta}$		SD of $\widehat{\delta}$		RISE of $\widehat{\beta}(s,t)$	
	NGB	PLMA	NGB	PLMA	NGB	PLMA	NGB	PLMA
Scenario 1	0.0173	0.0206	1.2	1.7	0.0162	0.0188	1.29 (0.25)	0.83 (0.35)
Scenario 2	0.0480	0.1587	−5.4	−31.5	0.0396	0.0192	69.86 (7.04)	69.33 (7.54)
Scenario 3	0.0548	0.2426	−9.2	−48.3	0.0297	0.0235	8.74 (1.33)	9.75 (1.34)

we calculated the square root of the mean integrated squared errors (RISEs) over a meshgrid, and these results are also listed in Table 1. The NGB and PLMA approaches did not exhibit any discernible difference in terms of the estimated RISE for $\widehat{\beta}(s,t)$.

6. SUMMARY

In analyzing two functional objects, a response and a possible covariate, it may be of great interest to estimate the extent to which the covariate history affects the current value of the response. For this purpose, the historical functional linear model serves as a natural alternative to the commonly used functional regression model. Also, we ought to estimate any possible historical lag from observed data.

We have proposed an NGB approach, tailored for the simultaneous estimation of the historical lag and the regression coefficient function. The NGB penalty is able to shrink not only a group of coefficients corresponding to the designated area towards zero, but also individual coefficients corresponding to the remaining area towards zero. We adopted the triangular basis from the finite element method in order to conform naturally to the nonrectangular domain of the regression coefficient function. The triangular basis system is computationally efficient in the sense that the compact support of the basis functions leads to a sparse design matrix. In estimating the bivariate coefficient function, our proposed method of estimation behaves like existing, competing alternative approaches. However, it outperforms existing methods in estimating the historical lag δ of practical interest. The proposed NGB estimators of the coefficient function and δ are both consistent estimators.

Under the historical functional linear model that we identified in Equation (1), we assumed that the historical lag is independent of time. It may be of interest to describe more precisely the historical effect of the functional covariate on the functional response via a time-dependent lag

$\delta(t)$. Another point of interest is the alternative choice of the smoothness penalty. The discrete penalty was first suggested by Eilers & Marx (1996) in connection with P-splines, and involves a large matrix and its inversion. It may be of interest to explore other possibilities. Each of these points we have mentioned represents a topic that would benefit from further research.

ACKNOWLEDGEMENTS

The authors would like to thank the Editor, the Associate Editor, and two anonymous referees for many insightful comments that gave rise to improvements in the paper. This research was supported by Discovery grants from the Natural Sciences and Engineering Research Council of Canada (NSERC) to Tianyu Guan and Jiguo Cao.

REFERENCES

Asencio, M., Hooker, G., & Gao, H. O. (2014). Functional convolution models. *Statistical Modelling: An International Journal*, 14(4), 315–335.

Besse, P. C. & Cardot, H. (1996). Approximation spline de la prévision d'un processus fonctionnel autorégressif d'ordre 1. *Canadian Journal of Statistics*, 24(4), 467–487.

Brockhaus, S., Melcher, M., Leisch, F., & Greven, S. (2017). Boosting flexible functional regression models with a high number of functional historical effects. *Statistics and Computing*, 27(4), 913–926.

Cai, X., Xue, L., & Cao, J. (2021). Robust penalized M-estimation for function-on-function linear regression. *Stat*, 10, e390.

Cai, X., Xue, L., & Cao, J. (2022). Variable selection for multiple function-on-function linear regression. *Statistica Sinica*, 32, 1435–1465.

Cardot, H., Ferraty, F., & Sarda, P. (2003). Spline estimators for the functional linear model. *Statistica Sinica*, 13, 571–591.

Chen, H. & Wang, Y. (2011). A penalized spline approach to functional mixed effects model analysis. *Biometrics*, 67(3), 861–870.

Eilers, P. H. & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2), 89–102.

Fan, J. & 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(2), 303–322.

Ferraty, F. & Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice*, Springer, New York.

Greven, S. & Scheipl, F. (2017). A general framework for functional regression modelling. *Statistical Modelling: An International Journal*, 17(1-2), 1–35.

Guan, T., Lin, Z., & Cao, J. (2020). Estimating truncated functional linear models with a nested group bridge approach. *Journal of Computational and Graphical Statistics*, 29(3), 620–628.

Hall, P. & Hooker, G. (2016). Truncated linear models for functional data. *Journal of the Royal Statistical Society, Series B*, 78(3), 637–653.

Harezlak, J., Coull, B. A., Laird, N. M., Magari, S. R., & Christiani, D. C. (2007). Penalized solutions to functional regression problems. *Computational Statistics & Data Analysis*, 51(10), 4911–4925.

Hastie, T. & Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(4), 757–779.

Hsing, T. & Eubank, R. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*, John Wiley & Sons, New York.

Huang, J., Ma, S., Xie, H., & Zhang, C.-H. (2009). A group bridge approach for variable selection. *Biometrika*, 96(2), 339–355.

Ivanescu, A. E., Staicu, A.-M., Scheipl, F., & Greven, S. (2015). Penalized function-on-function regression. *Computational Statistics*, 30(2), 539–568.

Jiang, F., Baek, S., Cao, J., & Ma, Y. (2020). A functional single index model. *Statistica Sinica*, 30, 303–324.

Kim, K., Şentürk, D., & Li, R. (2011). Recent history functional linear models for sparse longitudinal data. *Journal of Statistical Planning and Inference*, 141(4), 1554–1566.

Kokoszka, P. & Reimherr, M. (2017). *Introduction to Functional Data Analysis*, CRC Press, New York.
Larson, M. G. & Bengzon, F. (2013). *The Finite Element Method: Theory, Implementation and Applications*, Springer, New York.
Lin, Z., Cao, J., Wang, L., & Wang, H. (2017). Locally sparse estimator for functional linear regression models. *Journal of Computational and Graphical Statistics*, 26(2), 306–318.
Liu, B., Wang, L., & Cao, J. (2017). Estimating functional linear mixed-effects regression models. *Computational Statistics & Data Analysis*, 106, 153–164.
Malfait, N. & Ramsay, J. O. (2003). The historical functional linear model. *Canadian Journal of Statistics*, 31(2), 115–128.
Morris, J. S. (2015). Functional regression. *Annual Review of Statistics and Its Application*, 2(1), 321–359.
Müller, H.-G. & Zhang, Y. (2005). Time-varying functional regression for predicting remaining lifetime distributions from longitudinal trajectories. *Biometrics*, 61(4), 1064–1075.
Ramsay, J. O. & Dalzell, C. J. (1991). Some tools for functional data analysis. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53(3), 539–561.
Ramsay, J. O. & Silverman, B. W. (2002). *Applied Functional Data Analysis: Methods and Case Studies*, Springer, New York.
Ramsay, J. O. & Silverman, B. W. (2005). *Functional Data Analysis*, 2nd ed., Springer, New York.
Ramsay, T. O. (2002). Spline smoothing over difficult regions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(2), 307–319.
Sangalli, L. M., Ramsay, J. O., & Ramsay, T. O. (2013). Spatial spline regression models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(4), 681–703.
Şentürk, D. & Müller, H.-G. (2008). Generalized varying coefficient models for longitudinal data. *Biometrika*, 95(3), 653–666.
Şentürk, D. & Müller, H.-G. (2010). Functional varying coefficient models for longitudinal data. *Journal of the American Statistical Association*, 105(491), 1256–1264.
Wang, H. & Kai, B. (2015). Functional sparsity: Global versus local. *Statistica Sinica*, 25(4), 1337–1354.
Wang, J.-L., Chiou, J.-M., & Müller, H.-G. (2016). Functional data analysis. *Annual Review of Statistics and Its Application*, 3(1), 257–295.
Wu, C. O., Chiang, C.-T., & Hoover, D. R. (1998). Asymptotic confidence regions for kernel smoothing of a varying-coefficient model with longitudinal data. *Journal of the American Statistical Association*, 93(444), 1388–1402.
Wu, H. & Liang, H. (2004). Backfitting random varying-coefficient models with time-dependent smoothing covariates. *Scandinavian Journal of Statistics*, 31(1), 3–19.
Yao, F., Müller, H.-G., & Wang, J.-L. (2005). Functional linear regression analysis for longitudinal data. *The Annals of Statistics*, 33(6), 2873–2903.
Zhou, J., Wang, N.-Y., & Wang, N. (2013). Functional linear model with zero-value coefficient function at sub-regions. *Statistica Sinica*, 23, 25–50.
Zhou, L., Huang, J. Z., & Carroll, R. J. (2008). Joint modelling of paired sparse functional data using principal components. *Biometrika*, 95(3), 601–619.
Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476), 1418–1429.

Received 12 January 2022
Accepted 5 August 2022