Computational Statistics and Data Analysis 169 (2022) 107421

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

www.elsevier.com/locate/csda

Truncated estimation in functional generalized linear regression models

Xi Liu [a,*], Afshin A. Divani [b,1], Alexander Petersen [c]

[a] Department of Statistics and Applied Probability, University of California Santa Barbara, Santa Barbara, CA 93106, United States of America
[b] The University of New Mexico School of Medicine, University of New Mexico, 915 Camino de Salud NE Albuquerque, NM 87106, United States of America
[c] Department of Statistics, Brigham Young University, Provo, UT 84602, United States of America

ARTICLE INFO

Article history:
Received 28 January 2021
Received in revised form 26 October 2021
Accepted 2 January 2022
Available online 7 January 2022

Keywords:
Functional data analysis
Functional generalized linear models
Penalized B-splines
Nested group lasso

ABSTRACT

Functional generalized linear models investigate the effect of functional predictors on a scalar response. An interesting case is when the functional predictor is thought to exert an influence on the conditional mean of the response only through its values up to a certain point in the domain. In the literature, models with this type of restriction on the functional effect have been termed truncated or historical regression models. A penalized likelihood estimator is formulated by combining a structured variable selection method with a localized B-spline expansion of the regression coefficient function. In addition to a smoothing penalty that is typical for functional regression, a nested group lasso penalty is also included which guarantees the sequential entering of B-splines and thus induces the desired truncation on the estimator. An optimization scheme is developed to compute the solution path efficiently when varying the truncation tuning parameter. The convergence rate of the coefficient function estimator and consistency of the truncation point estimator are given under suitable smoothness assumptions. The proposed method is demonstrated through simulations and an application involving the effects of blood pressure values in patients who suffered a spontaneous intracerebral hemorrhage.

1. Introduction

The need to incorporate data features such as curves or longitudinal recordings as predictors for regression has given rise to a large body of research in recent decades on functional regression models (Ramsay and Silverman, 2005; Morris, 2015; Wang et al., 2016; Kokoszka and Reimherr, 2017). In particular, functional generalized linear models investigate the dependency of a scalar response, Y, on a functional predictor, $X \in L^2[0, T]$ for some $T > 0$, where the conditional distribution of Y given X is assumed to belong to a common exponential family. As a direct extension of the classical generalized linear model (McCullagh and Nelder, 1983), this conditional distribution takes the form

$$f(y; X, \alpha_0, \beta_0) = c(y) \exp\left\{\frac{y\theta_0 - b(\theta_0)}{a(\phi)}\right\}, \quad (\alpha_0, \beta_0) \in \mathbb{R} \times L^2[0, T], \tag{1}$$

* Corresponding author.
E-mail address: xliu@pstat.ucsb.edu (X. Liu).
[1] For the ATACH-2 Trial Investigators and the Neurological Emergency Treatment Trials Network.

https://doi.org/10.1016/j.csda.2022.107421
0167-9473/

where a, b, and c are known functions, and θ_0 encodes the connection between X and Y via the identities $E(Y|X) = b'(\theta_0) = g^{-1}\left\{\alpha_0 + \int_0^T \beta_0(t)X(t)\mathrm{d}t\right\}$ for some smooth link function g. The need to go beyond the functional linear model, a special case of which arises when f is a normal distribution and g is the identity function, is dictated by countless examples. For instance, the motivating example for this paper comes from a phase III clinical study for antihypertensive treatment of patients with spontaneous intracerebral hemorrhage (ICH). A current problem in this field is to develop an efficient protocol for the management of systolic blood pressure following hospital admission in order to help reduce mortality and improve functional outcomes. The blood pressure recordings constitute functional regressors, while the outcome is typically measured on a binary or ordinal scale, rendering linear models inadequate.

As for any regression model with functional predictor X, the primary difficulty compared to the classical generalized linear model is the infinite-dimensional nature of X. As a result, regularization in functional regression analysis is necessary regardless of sample size. A typical approach is to represent both X and the functional parameter β via basis expansions. Truncating the expansion at finitely many terms induces an initial regularization that can be followed by a regularized multivariate method on the coefficients of the expansion. James (2002) and Müller and Stadtmüller (2005) employed the eigenbasis arising from the Karhunen-Loève expansion and represented both X and β in that basis. Cardot and Sarda (2005) used B-splines with the usual L^2 penalty on the derivative of the coefficient function β. Within a reproducing kernel Hilbert space framework, Shang et al. (2015) derived the asymptotic distribution of their estimator in order to produce confidence and prediction intervals. Besides these functional extensions of the classical generalized linear model, extensive methodology has been developed for functional generalized additive models (Febrero-Bande and González-Manteiga, 2013; McLean et al., 2014; Greven and Scheipl, 2017).

In this paper, we propose a method for structured estimation of β in the functional generalized linear model under the assumption that X has an effect on the conditional mean of Y only through its values on a subdomain $[0, \delta_0]$ for $T > \delta_0$. That is, variations in X that only occur in the latter end of the interval $(\delta_0, T]$ will not affect the conditional mean of Y. In terms of the quantity $\int_0^T \beta_0(t)X(t)\mathrm{d}t$ involved in the systematic component of the model, this structural assumption effectively states that $\beta(t) \equiv 0$ on $(\delta_0, T]$. Regarding the issue with blood pressure management protocols for treatment of ICH, recent randomized phase III trials have produced conflicting results on the optimum blood pressure target and how fast the blood pressure should be lowered to the target level; see Moullaali et al. (2019); Divani et al. (2019, 2020), and references therein. For this and other similar problems, it is thus crucial to understand not only how the functional covariate influences the regression (the shape of the coefficient function β_0), but also over what time period it has an effect (the support of β_0).

From one perspective, one may view this problem as the functional regression equivalent of simultaneous estimation and model selection in classical or high-dimensional regression models. Indeed, the more general problem of identifying the so-called null region over which $\beta_0(t) = 0$ has been addressed in past work for the functional linear model. James et al. (2009) divided the time period into a fine grid of points and utilized regularized variable selection methods to enforce sparsity on the derivatives of the estimate at the grid points. Zhou et al. (2013) proposed a two-stage shrinkage method, in which the null region is roughly identified in stage one and then adaptively regularized using a group SCAD penalty in stage two. Lin et al. (2017) further designed a functional SCAD penalty to identify the null region and produce a smooth estimate of β_0 outside this region in one single step.

More recently, there has been work on a more structured problem where the null region coincides with the latter portion $(\delta_0, T]$, although still within the framework of the functional linear model. Hall and Hooker (2016) termed it a truncated functional linear model and estimated β_0 and δ_0 simultaneously using a penalized least square method with a penalty on δ^2. The resulting estimate of β is not continuous at $\hat{\delta}$, whereas it is often more reasonable to assume the effect of X on Y changes continuously instead of having a sharp drop at δ_0.

In this paper, we propose a method for simultaneous estimation of δ_0 and β_0 in the functional generalized linear model. We choose to use the localized B-splines to expand the functional parameter as the inherent ordering of B-splines is well-suited to structure a truncated curve. When the tail of the coefficient vector of the spline expansion is identically zero, it induces a corresponding truncation in the coefficient function β. The recent work of Guan et al. (2020) also utilized B-splines in their estimation of the truncated functional linear model. In addition to handling the larger class of generalized responses, our model differs from the method of Guan et al. (2020) in the form of the penalty. Guan et al. (2020) used the nested group bridge penalty resulting in a non-convex optimization problem. Instead, we employ a nested group lasso penalty (Yuan and Lin, 2006), which guarantees the sequential entering of B-splines and yields a convex optimization problem.

The paper is organized as follows. In Section 2, we first introduce the nested group lasso penalty and construct the penalized log-likelihood function. Then we present a fast algorithm designed to optimize the convex loss function in Section 3. The convergence rate of the estimator and the consistency of the truncation point estimator are established in Section 4. In Section 5, the performance of the method is evaluated by simulation studies. A real data example of investigating the effect of blood pressure curve on the outcome of ICH is demonstrated in Section 6.

2. Methodology

Suppose we observe independent pairs $(Y_i, X_i(t))$, $i = 1, \ldots, n$, where the conditional distribution of each Y_i given X_i takes the form in (1) with parameter θ_{0i}. The following developments assume that X_i is fully observed on $[0, T]$ while,

in reality, X_i is usually recorded only at a finite number of points along this continuum. In practice, one can construct an estimate of X_i using smoothing techniques, which would take the place of X_i in the equations below. Recall that a, b, and c are known functions. We assume that the dispersion parameter ϕ is known, as in the common Binomial and Poisson cases. Furthermore, for simplicity, we assume that g is the canonical link, although the method can be applied to any suitable link function. Hence, we have $\theta_{0i} = \alpha_0 + \int_0^T X_i(t)\beta_0(t)\mathrm{d}t$, so that $\alpha_0 \in \mathbb{R}$ and $\beta_0 \in L^2[0,T]$ are the parameters to be estimated. Let $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$ denote the usual $L^2[0,T]$ inner product and norm, respectively.

In this setting, Cardot and Sarda (2005) proposed

$$l_S(\alpha,\beta) = -\frac{1}{n}\sum_{i=1}^{n}[Y_i\theta_i - b(\theta_i)] + \frac{1}{2}\lambda_s\|\beta^{(m)}\|^2, \quad \theta_i = \alpha + \int_0^T X_i(t)\beta(t)\mathrm{d}t, \tag{2}$$

as a target criterion for estimation (although their work omitted the intercept term), where the first term is proportional to the negative log-likelihood, and the superscript m denotes the m-th derivative. Cardot et al. (2003) minimized (2) over the space $S_{q,k}$ of order q B-splines with interior knots $0 < t_1 < \cdots < t_k < T$, which we will take to be equally spaced for simplicity. Let $\Phi(t) = (\phi_1(t),\ldots,\phi_{q+k}(t))^\intercal$ denote the B-spline basis, ordered in the usual way so that ϕ_j is supported on $[t_{j-q}, t_j]$ for $j = 1,\ldots,q+k$, where we define $t_{j-q} = 0$ for $j < q$ and $t_j = T$ for $j > k$ (De Boor, 2001). In our numerical experiments, $m = 2$ and $q = 4$ will be used, corresponding to the space of cubic splines with a penalty on the second derivative, but the methodology can be applied so long as $m \le q-1$. To each $\beta \in S_{q,k}$ there corresponds a unique $\boldsymbol{\eta} = (\eta_1,\ldots,\eta_{q+k})^\intercal$ such that $\beta = \Phi^\intercal\boldsymbol{\eta}$, where we suppress the functional argument t. Hence, with a slight abuse of notation, write $l_S(\alpha,\boldsymbol{\eta}) = l_S(\alpha,\beta)$ and let $(\bar\alpha,\bar{\boldsymbol{\eta}})$ be the minimizer of l_S, yielding the penalized spline estimator $\bar\beta = \Phi^\intercal\bar{\boldsymbol{\eta}}$. The smoothness penalty is natural since it tends to produce interpretable estimates, but does not result in a truncated estimate of β.

Suppose that $\beta_0(t) = 0$ for $t > \delta_0$, so that the systematic component becomes $\theta_{0i} = \alpha_0 + \int_0^{\delta_0} X_i(t)\beta_0(t)\mathrm{d}t$. Hence, we require an estimator $\hat\delta$ and a companion estimator $\hat\beta$ that satisfies $\hat\beta(t) \equiv 0$ for $t > \hat\delta$. Due to the sequential ordering of the supports of ϕ_j, note that $\beta(t) = 0$ for $t > t_j$ if and only if $\eta_l = 0$ for $l > j$. Thus, truncation of β is equivalent to having a zero vector $(\eta_{j+1},\ldots,\eta_{q+k}) = \mathbf{0}$, $j = 0,\ldots,k$. Letting M be the $(q+k)\times(q+k)$ matrix with entries $M_{jl} = \langle\phi_j^{(m)},\phi_l^{(m)}\rangle$, this leads to the objective function

$$\begin{aligned} Q_V(\alpha,\boldsymbol{\eta}) &= l_S(\alpha,\boldsymbol{\eta}) + \lambda_t P_T(\boldsymbol{\eta}) \\ &= -\frac{1}{n}\sum_{i=1}^{n}[Y_i\theta_i - b(\theta_i)] + \frac{1}{2}\lambda_s\boldsymbol{\eta}^\intercal M\boldsymbol{\eta} + \lambda_t\sum_{j=0}^{k}\hat w_j\left(\sum_{l=j+1}^{q+k}\eta_l^2\right)^{1/2}, \end{aligned} \tag{3}$$

where $P_T(\boldsymbol{\eta}) = \sum_{j=0}^{k}\hat w_j\left(\sum_{l=j+1}^{q+k}\eta_l^2\right)^{1/2}$ is the nested group lasso penalty, $\hat w_j$ are positive weights, and λ_s, λ_t are positive tuning parameters that control the amount of regularization that is applied in order to produce a smooth and truncated estimate. In the spirit of Zou (2006), in our analyses we define adaptive weights based on the preliminary consistent estimator $\bar{\boldsymbol{\eta}}$ via $\hat w_j^{-1} = \sum_{l=j+1}\bar\eta_l^2$. Finally, with

$$(\hat\alpha,\hat{\boldsymbol{\eta}}) = \operatorname*{argmin}_{(\alpha,\boldsymbol{\eta})\in\mathbb{R}\times\mathbb{R}^{q+k}} Q_V(\alpha,\boldsymbol{\eta}), \tag{4}$$

we set $\hat\beta = \Phi^\intercal\hat{\boldsymbol{\eta}}$. As will be demonstrated below, (4) will always lead to a truncated estimate $\hat\beta$, so we may define $\hat\delta$ as the smallest knot value after which $\hat\beta$ is identically zero.

3. Computational details

In this section, we describe how the convex nested group lasso penalty leads to a truncated estimate of β_0, without the need for a second thresholding step to zero out small estimates as required by some simultaneous estimation and selection techniques. Section 3.1 demonstrates the properties of hierarchical inclusion of components in $\boldsymbol{\eta}$, after which a computational algorithm is given for approximating the solution path as $\lambda_t \downarrow 0$. When λ_t decreases, the new estimate is computed in two steps. In the first, block norms are used to determine which, if any, basis coefficients should be added to the active set, as described in Section 3.2. Once the active set is updated, Section 3.3 describes an efficient algorithm that overcomes difficulties associated with the nonsmoothness of the truncation penalty. Tuning parameter selection is discussed in 3.4.

3.1. Nested group lasso

Hierarchical structured penalties such as P_T in (3) have been well studied in multivariate regression analysis. In particular, the composite absolute penalty family proposed in Zhao et al. (2009) provided a framework to incorporate grouping or hierarchical information within the regression procedure. For any subset $\mathcal{G} \subset \{1,\ldots,q+k\}$, let $\boldsymbol{\eta}_{\mathcal{G}}$ denote the subvector

specified by the index set $\mathcal{G}$. For $j = 0, \ldots, k$, define the index groups $\mathcal{G}_j = \{j+1, \ldots, q+k\}$, so that the truncation penalty P_T can be viewed as a composition of ℓ^1 and ℓ^2 penalties. Specifically,

$$P_T(\boldsymbol{\eta}) = \sum_{j=0}^{k} \hat{w}_j \left(\sum_{l=j+1}^{q+k} \eta_l^2 \right)^{1/2} = \sum_{j=0}^{k} \hat{w}_j \left| \|\boldsymbol{\eta}_{\mathcal{G}_j}\|_2 \right|.$$

Similar to the rationale of group lasso penalty Yuan and Lin (2006), the ℓ^1 penalty gives rise to sparsity and ℓ^2 penalty keeps the grouping information. Meanwhile, the smaller groups are always nested in the larger group. The following proposition describes two important properties of the truncation penalty.

Proposition 1 *(Theorem 1 and Theorem 2 in Zhao et al. (2009)).*

1. *The truncation penalty $P_T(\boldsymbol{\eta})$ is convex in $\boldsymbol{\eta}$.*
2. *For $i < j$, $\frac{\partial}{\partial \eta_i} P_T(\boldsymbol{\eta}) = 0$ whenever $\eta_j \neq 0$ and $\eta_i = 0$.*

Property 1 allows for the use of convex optimization techniques since the negative log-likelihood function, the smoothness penalty $\boldsymbol{\eta}^\intercal M \boldsymbol{\eta}$, and the truncation penalty P_T are all convex. This also guarantees the uniqueness and existence of the estimator with high probability; see also Theorem 1 in Section 4 below. Property 2 states that, once the jth basis function, ϕ_j, is added to the model, the infinitesimal movements of the coefficient of ϕ_i are not penalized and hence η_i will almost surely deviate from zero. In other words, while the truncation tuning parameter λ_t is decreasing, η_i enters the model no later than η_j, which is exactly the desired hierarchical structure.

3.2. Updating the active set

Benefiting from the convexity and sequential entering property in Proposition 1, we are able to directly detect changes in the active set and compute an approximate solution path of $\hat{\boldsymbol{\eta}}$ as λ_t decreases. At step l, the active set is the set of indices j for which $\hat{\eta}_j \neq 0$. Suppose that, at the current value λ_t^l, the active groups are known to be $(\mathcal{G}_0, \ldots, \mathcal{G}_{k_l - 1})$, i.e. the active set is $\mathcal{A}_l = (1, \ldots, k_l)$. We then optimize the penalized objective function (3) under the restriction that $\eta_j = 0$ for any $j \notin \mathcal{A}_l$ using convex optimization techniques described in Section 3.3 below. The penalty parameter is decreased as $\lambda_t^{l+1} = \Delta \lambda_t^l$ and the active set is assessed to see if more basis coefficients should be added. In our simulations, we take $\Delta = 0.9$ so that the step size shrinks as λ_l^t approaches 0. As Δ gets closer to 1, the solution path approximation becomes more accurate, but requires more time to compute.

First, we compute

$$\begin{aligned} dl_{S,j}^l &:= \left. \frac{\partial l_S}{\partial \eta_j} \right|_{\boldsymbol{\eta} = \hat{\boldsymbol{\eta}}^l}, \quad 1 \le j \le q+k, \\ dP_{T,j}^l &:= \begin{cases} \hat{\eta}_j \sum_{i=0}^{j-1} \hat{w}_i \|\hat{\boldsymbol{\eta}}_{\mathcal{G}_i}\|_2^{-1}, & 1 \le j \le k_l, \\ \sum_{i=k_l}^{j-1} \hat{w}_i, & k_l < j \le q+k. \end{cases} \end{aligned} \tag{5}$$

Since l_S is differentiable, $dl_{S,j}^l$ corresponds to the j-th partial derivative of l_S evaluated at $\hat{\boldsymbol{\eta}}^l$. However, the partial derivatives of P_T at $\hat{\boldsymbol{\eta}}^l$ only exist for active indices, so that $dP_{T,j}^l$ for $j > k_l$ are defined as subdifferentials, i.e. as the limit when $\eta_j \downarrow 0$. However, taking the alternative limit as $\eta_j \uparrow 0$ does not affect the procedure, which depends only on the norms

$$\begin{aligned} Bl_{S,j}^l &= \|(dl_{S,1}^l, \ldots, dl_{S,j}^l)\|_2, \\ BP_{T,j}^l &= \|(dP_{T,1}^l, \ldots, dP_{T,j}^l)\|_2, \; j = 1, \ldots, k. \end{aligned} \tag{6}$$

Similar to most of the stepwise optimization algorithms which update the unknown parameter from the origin, when decreasing the tuning parameter of the convex penalty, the convex feasible space gets larger. Therefore, when λ_t gets updated, there is a chance that more η's can be active. Specifically, the block norms in (6) represent the magnitude of the change in the smoothed likelihood and truncation penalty terms, respectively, in the objective function. Hence, for a currently inactive index $j > k_l$, the condition $Bl_{S,j}^l \ge \lambda_t^{l+1} BP_{T,j}^l$ indicates that, by adding η_j to the active set, the overall objective value can be decreased. By Proposition 1, if η_j is active, all $\eta_i, i < j$, are also active.

Algorithm 1 describes the steps to update the active set, and can also be used to determine the starting point of the solution path as a special case. Define $\lambda_t^0 = \infty$, so that $\hat{\boldsymbol{\eta}}^0 = 0$. According to the above arguments, the tuning parameter value

$$\lambda_t^1 = \max_{j=1,\ldots,k} \frac{Bl_{S,j}^0}{BP_{T,j}^0} \tag{7}$$

marks the first point at which Algorithm 1 yields a nonempty active set.

Algorithm 1: Updating the Active Set.

input : Current Estimate $\hat{\boldsymbol{\eta}}^l$
output: New active set $\mathcal{A}_{l+1}$

for $i \leftarrow 1$ **to** $q+k$ **do**
  Compute $dl^l_{S,i}$ and $dP^l_{T,i}$ as in (5)
end
for $j \leftarrow k_l$ **to** k **do**
  Compute $Bl^l_{S,j}$ and $BP^l_{T,j}$ as in (6)
end
$k_{l+1} \leftarrow \max\left\{j : k_l \le j \le k \text{ and } Bl^l_{S,j} \ge \lambda_t^{l+1} BP^l_{T,j}\right\}$
$\mathcal{A}_{l+1} \leftarrow \{1, \ldots, k_{l+1}\}$

3.3. Convex optimization with fixed dimension

Given a value λ_t, Algorithm 1 can be used to determine the active set $\mathcal{A} = \{1, \ldots, k_0\}$. With the active set known, it remains to minimize (3) with $\boldsymbol{\eta}$ ranging over the set $\{\boldsymbol{\eta} \in \mathbb{R}^{q+k} : \eta_{k_0+1} = \cdots = \eta_{q+k} = 0\}$. Thus, the dimension of the minimization problem is $k_0 + 1$, including the intercept. However, the nonsmoothness of (3) induced by the truncation penalty prevents the use of classical coordinate descent, Newton-Raphson, or similar first- and second-order methods, since these are required to have the continuous first order derivative and will have irregular behavior near the points of singularity. To overcome this, we employ the dual formulation of minimizing P_T over the constrained active set, and introduce an auxiliary quadratic function to construct a smooth approximation of this dual problem (Chen et al., 2012). Then we are safe to apply the accelerated gradient descent onto the smooth convex function. Details are presented in Section S.1 in the Supplementary Material.

3.4. Selection of tuning parameters

Utilizing the conditions in Algorithm 1 and the optimization informed by knowledge of the active set, the solution path can be computed efficiently. Thanks to the continuous solution path, the former step solution can be used as a warm start in next step optimization algorithm. We thus propose an approximate BIC quantity to determine the optimal tuning parameter pair (λ_s, λ_t). Letting $\mathcal{D}(\hat{\alpha}, \hat{\boldsymbol{\eta}})$ be the model deviance, and $\mathrm{df}(\hat{\alpha}, \hat{\boldsymbol{\eta}})$ approximate degrees of freedom, set

$$\mathrm{BIC}(\lambda_s, \lambda_t) = \mathcal{D}(\hat{\alpha}, \hat{\boldsymbol{\eta}}) + \log(n) \times \mathrm{df}(\hat{\alpha}, \hat{\boldsymbol{\eta}}).$$

In the linear model, the degrees of freedom is equal to the trace of the hat matrix. In the generalized case, we consider an approximate hat matrix from the method of scoring iterative equations, as an extension of the degrees of freedom definition used by Guan et al. (2020). Let U be the matrix with entries $U_{i1} = 1$ and $U_{i,j+1} = \langle X_i, \phi_j \rangle$, $i = 1, \ldots, n$, $j = 1, \ldots, q+k$, set $\hat{V} = \mathrm{diag}\{b''(\hat{\theta}_i)\}_{i=1}^n$, and let M_t be the diagonal matrix of dimension $1+q+k$, with the first diagonal element being 0 and all others set to $P_T(\hat{\boldsymbol{\eta}})/\|\hat{\boldsymbol{\eta}}\|^2$. Then the degrees of freedom approximation is

$$\mathrm{df}(\hat{\alpha}, \hat{\boldsymbol{\eta}}) = \mathrm{trace}[\widehat{H}(\lambda_s, \lambda_t)], \quad \widehat{H}(\lambda_s, \lambda_t) = U\left(U^\intercal \hat{V} U + \lambda_s \begin{pmatrix} 0 & 0 \\ 0 & M \end{pmatrix} + \lambda_t M_t\right)^{-1} U^\intercal \hat{V}.$$

4. Asymptotic properties

In this section, we provide the convergence rates of the estimators in (4). Following Cardot and Sarda (2005), we measure the estimation error of the functional parameter in terms of the norm $\|\cdot\|_X$ induced by the inner product

$$\langle \beta_1, \beta_2 \rangle_X = \mathrm{E}\left(\langle \beta_1, X \rangle \langle \beta_2, X \rangle\right) = \int_0^T \int_0^T C(s,t)\beta_1(s)\beta_2(t)\mathrm{d}s\mathrm{d}t,$$

where $C(s,t) = \mathrm{E}(X(s)X(t))$ is the kernel function of the operator $\Gamma(f)(t) = \int_0^T C(s,t)f(s)\mathrm{d}s$, for $f \in L^2[0,T]$. Let (X, Y) be a generic pair distributed identically to the (X_i, Y_i), and define

$$L_S(\alpha, \beta) = \mathrm{E}[l_S(\alpha, \beta)] = -\mathrm{E}[Y\theta - b(\theta)] + \frac{1}{2}\lambda_s \|\beta^{(m)}\|^2, \tag{8}$$

where $\theta = \alpha + \langle \beta, X \rangle$. The stochastic behavior of the estimates will be investigated relative to two intermediary approximations of (α_0, β_0). The first is defined via

$$(\tilde{\alpha}, \tilde{\beta}) = \underset{(\alpha,\beta) \in \mathbb{R} \times S_{q,k}}{\mathrm{argmin}} \; L_S(\alpha, \beta), \tag{9}$$

from which one can extract oracle weights

$$\tilde{w}_j = \left(\sum_{l=j+1}^{q+k} \tilde{\eta}_l^2 \right)^{-1}, \quad \tilde{\beta} = \Phi^\intercal \tilde{\boldsymbol{\eta}}. \tag{10}$$

The second is a truncated spline approximation of β_0. Denoted $\tilde{\beta}_0 = \Phi^\intercal \tilde{\boldsymbol{\eta}}_0$, this approximation is defined in (S.19) and its properties are detailed in Lemma 1 in Section S.3 of the Supplementary Material. In particular, suppose the true truncation point δ_0 lies in (t_{j_0}, t_{j_0+1}). If

$$j_{0k} = \begin{cases} j_0 + q/2, & q \text{ even}, \\ j_0 + (q+1)/2, & q \text{ odd}, \end{cases} \tag{11}$$

then $\tilde{\eta}_{0l} = 0$ for all $l \geq j_{0k}$, while $\tilde{\eta}_{0,j_{0k}-1} \neq 0$. Hence, the smoothed approximation $\tilde{\beta}_0$ will satisfy $\tilde{\beta}_0(t) = 0$ for all $t > t_{j_{0k}}$. Setting $\tilde{\delta}_{0k} = t_{j_{0k}}$, this approximation of the true truncation point δ_0 satisfies $|\delta_0 - \tilde{\delta}_{0k}| = O(k^{-1})$.

Let ρ_k be the smallest eigenvalue of the $(q+k) \times (q+k)$ matrix $\mathbf{C}$ with entries $\mathbf{C}_{jl} = \langle \Gamma(\phi_j), \phi_l \rangle$, so that $\rho_k = o(1)$ as k diverges, and set

$$a_{n1} = \left\{ \lambda_s + \lambda_s k^{2(m-p)} + k^{-2p} \right\}^{1/2}, \quad a_{n2} = \rho_k^{-1/2} k^{-1/2} n^{-1/2}. \tag{12}$$

We require the following assumptions for estimation consistency.

(A1) The function b is three times continuously differentiable, with $b'' > 0$ and $\mathrm{E}\left\{ b''(\theta_0) X \right\} = 0$.
(A2) $\|X\|$ is bounded almost surely, and the eigenvalues of Γ are nonzero.
(A3) β_0 has p' derivatives for some integer p' satisfying $|\beta_0^{(p')}(t_1) - \beta_0^{(p')}(t_2)| \leq C|t_1 - t_2|^v$, $C > 0$, $v \in [0, 1]$, and the degree $q - 1$ of the splines satisfies $q - 1 \geq p = p' + v$. Furthermore, $\|\beta_0^{(m)}\| < \infty$.
(A4) $k = k(n) \to \infty$, $\rho_k = \omega(k^{-(2p+1)})$, and $\lambda_s = o(k\rho_k) = o(1)$ as $n \to \infty$.
(A5) $\tilde{w}_{j_{0k}-1} \rho_k^{-1} a_{n2} = o(1)$, and $\lambda_t = O\left(\rho_k^{1/2} \upsilon_k^{-1} [a_{n1} + a_{n2}] \right)$, where $\upsilon_k = \sum_{j=1}^{j_{0k}-1} \tilde{w}_j$.

Assumptions (A1)–(A3) are regularity and smoothness conditions, and are also made in Cardot and Sarda (2005). We note that the last requirement in (A1) can be satisfied by properly centering the functional variable X, similar to the zero mean assumption typically made for linear models. Next, (A4) requires that minimum eigenvalue ρ_k decay at a polynomial rate in k, and can be viewed as a compatibility assumption between the operator Γ and the B-spline basis. To avoid confusion, we emphasize that these eigenvalues are different from the eigenvalues of penalty matrices typically encountered in penalized spline estimation for nonparametric regression models involving scalar predictors (Xiao, 2019). In this latter scenario, the eigenvalues are not dependent on the distribution of the predictors, so that explicit lower bounds may be calculated. In the scenario considered in this paper with a functional predictor X, they are directly related to the second moment operator of X, and do not admit an explicit lower bound. As a comparison, in the special case of a functional linear model, Guan et al. (2020) made the stronger assumption that $\rho_k \sim k^{-1}$, whereas we assume only a polynomial lower bound related to the smoothness of β_0. The decay of the smoothing parameter λ_s must then be of a smaller order than $k\rho_k$. Assumption (A5) ensures that a unique minimizer $(\hat{\alpha}, \hat{\boldsymbol{\eta}})$ of the estimation criterion exists with high probability, and determines the final rate of convergence. The proofs of all theoretical results are in Section S.3 of the Supplementary Material.

Theorem 1. *Suppose that (A1)–(A4) hold. Then*

$$|\tilde{\alpha} - \alpha_0| = O(a_{n1}), \quad \|\tilde{\beta} - \beta_0\|_X = O(a_{n1}).$$

Furthermore, if (A5) holds, the minimizers in (4) exist and are unique with probability tending to one, and

$$|\hat{\alpha} - \alpha_0| = O_p(a_{n1} + a_{n2}), \quad \|\hat{\beta} - \beta_0\|_X = O_p(a_{n1} + a_{n2}).$$

We compare this result with other penalized spline-based regression methods involving functional predictors. As a comparison to the rate obtained by Guan et al. (2020) in the case of a functional linear model with a nested group bridge approach, suppose that $\rho_k \sim k^{-1}$ and $\lambda_s = O(k^{-2p})$. Then the bias is $a_{n1} = k^{-p}$, matching the result of Guan et al. (2020). However, the stochastic term a_{n2} obtained in Theorem 1 is the parametric rate $n^{-1/2}$, compared to the rate of $kn^{-1/2}$ obtained in Guan et al. (2020). The ability to attain the parametric rate under this particular scenario is due to our use of the norm $\|\cdot\|_X$, which effectively measures prediction error, as was also discovered in Cai and Hall (2006) in the functional linear regression model. One must keep in mind, however, that this comparison does not tell the whole story, since the two methods give distinct requirements regarding the divergence of the number of knots. Additionally, the proof in Guan et al. (2020) is quite different from our proof of Theorem 1, which relies heavily on the convexity of (3), and somewhat follows

the proofs in Cardot et al. (2003). Although our bias term matches that of Cardot et al. (2003), the stochastic rates differ mainly due to our inclusion of a truncation penalty. Indeed, the stochastic rate in Cardot et al. (2003) is independent of ρ_k, so that our rate is more efficient only in the regime $\rho_k = \omega(k^{-2})$.

Lastly, we prove consistent estimation of the truncation point δ_0. Specifically, we define $\hat{\delta}$ as the unique knot $t_{\hat{j}}$ for which the estimator $\hat{\beta} = \Phi^\intercal \hat{\boldsymbol{\eta}}$ satisfies $\hat{\beta}(t) \neq 0$ for some $t \in (t_{\hat{j}-1}, t_{\hat{j}})$, while $\hat{\beta}(t) = 0$ for all $t \geq t_{\hat{j}}$. Although similar in spirit to model selection consistency in the sparse estimation of high-dimensional generalized linear models, this type of result requires a different nuance. In particular, we do not show that all nonzero elements of $\tilde{\boldsymbol{\eta}}_0$ in $\tilde{\beta}_0 = \Phi^\intercal \tilde{\boldsymbol{\eta}}_0$ have nonzero estimates with high probability. Such a result may be difficult to obtain given that the individual elements $\tilde{\eta}_{0j}$ shrink to zero as the corresponding knots approach $\tilde{\delta}_{0k}$, a condition typically prohibited in the high-dimensional case. Rather, we leverage the functional nature of the problem to obtain the weaker but sufficiently strong result that $\hat{\delta}$ converges in probability.

This is proved in two steps. First, using Theorem 1, one can immediately conclude that $\hat{\delta}$ must lie above $\delta_0 - \epsilon$ with probability approaching one for any $\epsilon > 0$. This is the usual consistency argument, and requires that λ_t decay to zero quickly enough, as specified in (A5). On the other hand, in order to prove the so-called "sparsistency" result, which ensures that $\hat{\delta}$ does not lie too far above δ_0, a lower bound on the decay of λ_t is needed, as follows.

(A6) $\rho_k^{-1} k^{-1/2} \upsilon_k^{-1}(a_{n1} + a_{n2}) = o(1)$, and $\lambda_t \rho_k^{3/2} k^{1/2}(a_{n1} + a_{n2})^{-3} \to \infty$.

The first part of this assumption makes it possible to choose a sequence λ_t satisfying both (A5) and the second part of (A6). In terms of the functional parameter estimate, the following result guarantees that $\hat{\beta}$ will be zero in any neighborhood of δ_0 with probability tending to one.

Theorem 2. *Let $\epsilon \in (0, \delta_0)$ be arbitrary. Under assumptions (A1)–(A5), $P(\hat{\delta} < \delta_0 - \epsilon) \to 0$. If, in addition, (A6) holds, then $P\left(\hat{\delta} > \tilde{\delta}_{0k}\right) \to 0$, so that*

$$\hat{\delta} - \delta_0 = o_p(1).$$

5. Simulation studies

Data were simulated for functional linear and functional logistic regression models. The systematic components, $\theta_0 = \alpha_0 + \int_0^1 \beta_0(t) X(t) \mathrm{d}t$, were generated from the following two cases.

- Case I: $\alpha_0 = 1$ and $\beta_0(t)$ is a linear combination of B-splines with order $q = 4$ and 14 equally spaced interior knots. The first eight coefficients of B-splines are set as (-0.45, -0.29, -1.79, -3.08, -2.11, 0.64, 5.72, 3.02), and the last ten are all 0. Thus, $\beta_0(t) = 0$ if $t > 0.54$.
- Case II: $\alpha_0 = 1$ and $\beta_0(t) = \sin(2\pi t) I_{\{t<=0.5\}}$. Therefore, $\beta_0(t)$ truncates at 0.5.

The random predictor curves were generated as $X(t) = \sum_{j=1}^{30} \xi_j \phi_j(t)$, with $\{\phi_j\}_{j=1}^{30}$ being the cubic B-spline functions. The coefficients ξ_j were generated sequentially, with $\xi_1 \sim N(0, \sigma_x^2)$ and for $j \geq 2$, $\xi_j \sim \sqrt{1/j} N(\xi_{j-1}, \sigma_x^2)$. For the linear regression simulations, $\sigma_x = 3$ and $\sigma_x = 5.5$ in Cases I and II, respectively, and the response was generated as $Y|X \sim \mathcal{N}(\theta_0, 1)$. The signal-to-noise ratio is approximately 0.55 in Cases I and 0.7 in Case II, where signal-to-noise ratio is defined as $\mathrm{Var}(\langle X, \beta_0 \rangle)/\mathrm{Var}(Y)$. The proposed method is compared with Guan et al. (2020), which utilized a nested group bridge (NGR) penalty to achieve truncation point estimation and Cardot and Sarda (2005), which simply is $\lambda_t = 0$ in our method. In logistic regression models, we set $\sigma_x = 4$ and $\sigma_x = 16$ in Cases I and II, respectively, and the response Y was generated as a Bernoulli random variable with probability of success $(1 + \exp\{-\theta_0\})^{-1}$.

We measure the estimation accuracy as well as the prediction accuracy according to the following definition.

$$\mathrm{D}_{\mathrm{esti}}(\hat{\beta}) = \|\hat{\beta} - \beta_0\|^2$$
$$\mathrm{D}_{\mathrm{pred}}(\hat{\beta}) = \|\hat{\beta} - \beta_0\|_X^2.$$

The number of cubic B-spline basis used for estimation is 100, where $k = 96$ and $q = 4$. Data were generated for three different sample sizes, $n = 200, 400, 600$, with 200 simulation runs in each case. The averaged estimation and prediction error of the estimates with and without truncation penalty are summarized in Table 1, with estimates from randomly selected simulation runs visualized in Fig. 1.

Comparing with Cardot and Sarda (2005), the proposed method has a smaller estimation error in general, but has a larger prediction error in logistic model Case II. The reason is that β_0 in Case II actually can not satisfy the smoothness assumption, and the selection of two tuning parameters adds extra difficulties and variation. In linear model cases, the proposed method is the most accurate in terms of estimation and prediction error, followed closely by the method of Guan et al. (2020).

Table 1
Estimation error and prediction error of β_0.

		Linear model					
		D^{sp}_{esti}	D^{tr}_{esti}	D^{ngr}_{esti}	D^{sp}_{pred}	D^{tr}_{pred}	D^{ngr}_{pred}
I	$n = 200$	0.584(0.27)	0.283(0.18)	0.364(0.35)	0.060(0.03)	0.033(0.02)	0.040(0.03)
	$n = 400$	0.353(0.13)	0.157(0.08)	0.183(0.16)	0.029(0.01)	0.016(0.01)	0.018(0.01)
	$n = 600$	0.285(0.11)	0.120(0.06)	0.125(0.11)	0.021(0.01)	0.011(0.01)	0.012(0.01)
II	$n = 200$	0.029(0.02)	0.012(0.01)	0.015(0.03)	0.234(0.13)	0.157(0.12)	0.171(0.15)
	$n = 400$	0.019(0.01)	0.007(0.01)	0.008(0.01)	0.131(0.07)	0.080(0.06)	0.088(0.07)
	$n = 600$	0.016(0.01)	0.006(<0.01)	0.006(<0.01)	0.115(0.06)	0.063(0.05)	0.057(0.04)
		Logistic model					
		D^{sp}_{esti}	D^{tr}_{esti}	D^{ngr}_{esti}	D^{sp}_{pred}	D^{tr}_{pred}	D^{ngr}_{pred}
I	$n = 200$	0.550(0.41)	0.357(0.44)	-	1.244(1.54)	1.132(1.73)	-
	$n = 400$	0.329(0.14)	0.186(0.11)	-	0.643(0.49)	0.580(0.49)	-
	$n = 600$	0.263(0.11)	0.139(0.08)	-	0.450(0.30)	0.412(0.33)	-
II	$n = 200$	0.025(0.02)	0.022(0.02)	-	8.315(12.63)	11.244(14.34)	-
	$n = 400$	0.019(0.01)	0.013(0.01)	-	5.337(7.51)	6.560(7.15)	-
	$n = 600$	0.013(0.01)	0.009(0.01)	-	3.517(3.78)	4.574(4.85)	-

Superscripts indicate the method of estimation: (sp) proposed estimator with $\lambda_t = 0$; (tr) proposed estimator with λ_t chosen by BIC; (ngr) estimator of Guan et al. (2020) for the functional linear model.

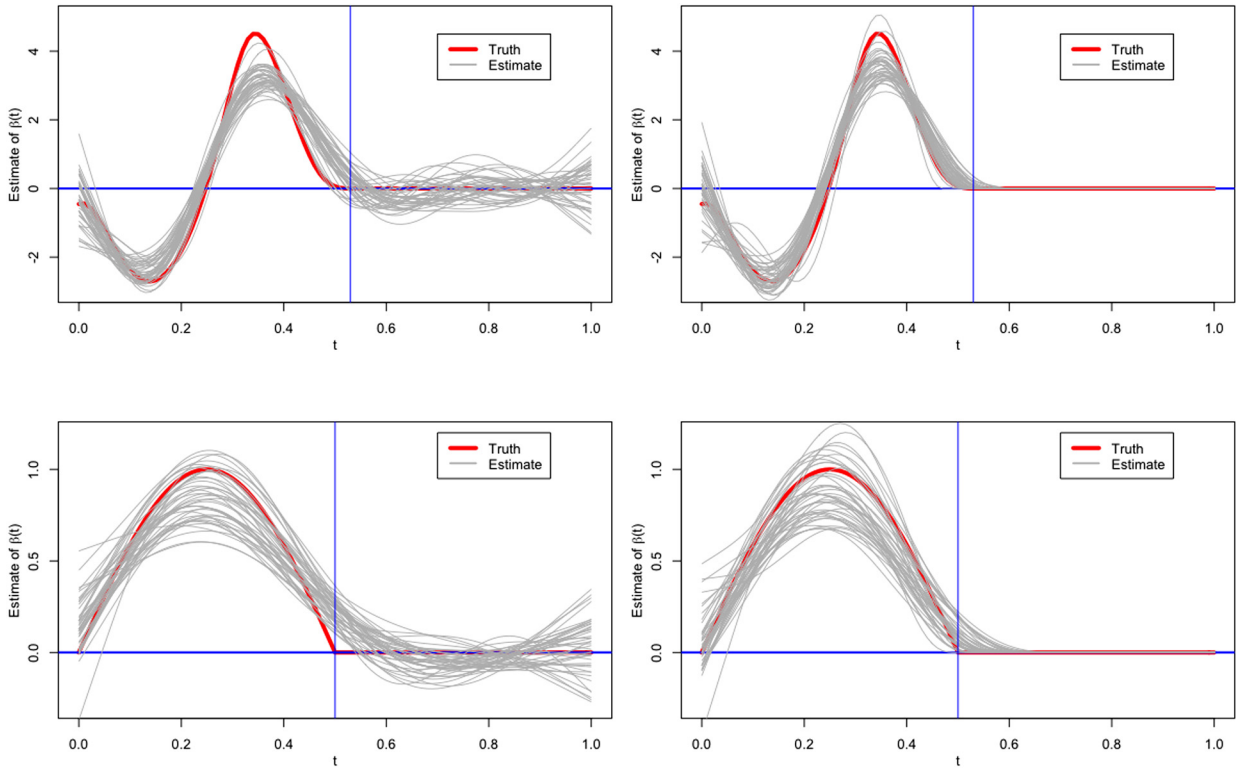

Fig. 1. Comparison between the proposed truncated estimates, with $\lambda_t = 0$ (left column, equivalent to the method of Cardot and Sarda (2005)) and $\lambda_t \neq 0$ chosen according to Section 3.4 (right). The top (bottom) row corresponds to linear (logistic) regression under the setting of Case I (Case II). Within each row, estimates from 40 randomly selected simulation runs are shown.

To assess the quality of the truncation point estimates, histograms of $\hat{\delta}$ for the functional linear model under Case I are given in Fig. 2. The corresponding plots for Case II and also those for functional logistic regression are given in Figures S.2 and S.2 in Section S.2 of the Supplementary Material. We also provide the mean absolute error (MAE), root mean square error (RMSE) and standard deviation (SD) of $\hat{\delta}$ in Table 2. In all but one case, the proposed method has lower errors than the NGR method. For both NGR and the proposed methods, truncation point estimation in Case I is far more accurate than Case II since β_0 in Case II does not satisfy the smoothness assumption on derivatives. Similar to Guan et al. (2020), performance is expected to deteriorate as the functional change at the true truncation point δ_0 becomes sharper, or even discontinuous.

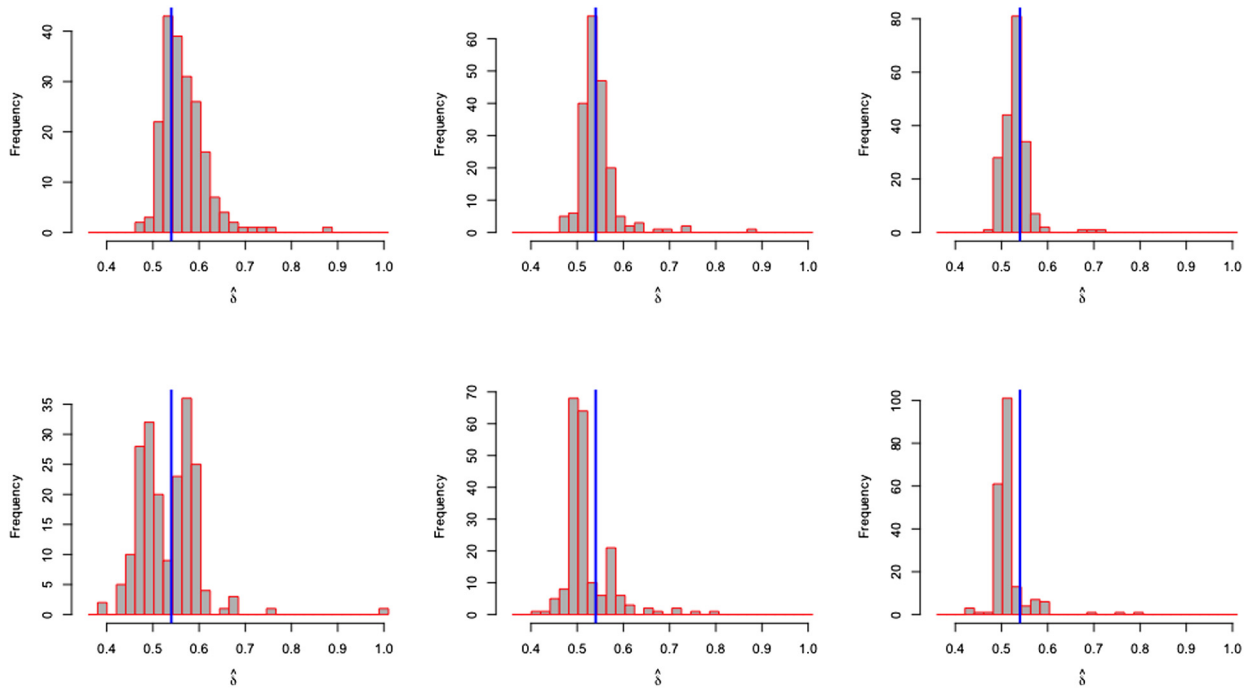

Fig. 2. Histograms of the estimated truncation points in linear model under Case I, using sample sizes 200 (left), 400 (middle), and 600 (right), for the proposed nested group lasso estimator (top row) and the NGR estimator of Guan et al. (2020) (bottom row).

Table 2
Accuracy of the estimate of the truncation point.

			Case I			Case II		
			$n = 200$	$n = 400$	$n = 600$	$n = 200$	$n = 400$	$n = 600$
Linear	TR	MAE	0.038	0.024	0.020	0.107	0.084	0.080
		RMSE	0.056	0.043	0.030	0.133	0.095	0.087
		SD	0.048	0.043	0.029	0.081	0.046	0.034
	NGR	MAE	0.049	0.043	0.035	0.110	0.093	0.084
		RMSE	0.064	0.052	0.044	0.132	0.105	0.093
		SD	0.064	0.050	0.038	0.082	0.055	0.043
Logistic	TR	MAE	0.040	0.024	0.022	0.099	0.097	0.087
		RMSE	0.071	0.035	0.029	0.127	0.121	0.105
		SD	0.070	0.034	0.026	0.080	0.072	0.060

TR is the truncated estimate proposed in this paper and NGR is the method for functional linear model Guan et al. (2020).

In addition, we report the computational time of calculating BIC plus getting the final estimates for a single data set. Because these results depend on the chosen grids of tuning parameters, they provide a practical comparison of the implementations of the two competing methods, as opposed to a purely algorithmic comparison. On average, for the linear model cases, the proposed method takes 13.3 s, 15.3 s and 14.8 s for sample sizes $n = 200, 400, 600$, respectively, while the NGR method had corresponding averages of 4.9 s, 8.1 s, and 12.8 s. For logistic models, the proposed method takes 13.3 s, 19.8 s and 21.1 s for sample size $n = 200, 400, 600$, on average. These computational times were measured on a 2.6 GHz 6-core i7 Mac. The proposed method was coded in python and Guan et al. (2020)'s optimization employed the `glmnet` R package written in Fortran. It is also worth mentioning that, for the NGR method, the optimal tuning parameters were searched on a fixed grid, which was provided by the author. For the proposed method, λ_s was selected from a fixed length sequence of ten values. The λ_t sequence was chosen as outlined in Section 3.2, and was thus more sensitive to differences between data sets, explaining the larger computation time for our method as well as its nonlinear dependence on the sample size.

6. Application to blood pressure curves

It has been reported that the yearly overall incidence of spontaneous intracerebral hemorrhage (ICH) worldwide is 2.5 per 10,000 people, with approximately 40,000 to 67,000 cases per year in the United States (Caceres and Goldstein, 2012). People with ICH often have an acute hypertensive response which may be associated with hematoma expansion that can lead to

increased risk of mortality. (Dandapani et al., 1995). The current guidelines from the American Heart Association/American Stroke Association suggest that, for ICH patients presenting with systolic blood pressure (SBP) between 150 and 220 mm Hg and without contraindication to acute SBP treatment, acute lowering of SBP to 140 mm Hg is safe (Morgenstern et al., 2010). Benefits of early reduction in blood pressure have been shown in several studies (Qureshi et al., 2012; ATLANTIS et al., 2004), though findings are not entirely consistent (Qureshi et al., 2010).

The Antihypertensive Treatment of Acute Cerebral Hemorrhage (ATACH II) is a designed randomized trial aiming to test whether aggressive SBP reduction in acute settings significantly decreases the likelihood of death or disability at three months after ICH. From 2011 to 2015, 1000 patients underwent randomization, which took place within 4.5 hours after ICH onset. The detailed design, method, and rationale were clearly described in Qureshi and Palesch (2011). Roughly speaking, the goal for the standard group was to reduce and maintain SBP between 140 and 180 mm Hg for 24 hours after randomization, while that of the intensive group is to reduce and maintain SBP between 110 and 140 mm Hg for 24 hours. A uniform SBP reduction strategy using calcium channel blocker was implemented to reach the goal in the standard and intensive groups. The final conclusion of ATACH II experiment was that treatment in the intensive group did not result in a lower rate of death or disability than the standard group (Qureshi et al., 2016).

In most of the previous investigations, the information in the SBP curve, which is a functional covariate, was summarized to scalar variables, such as a dummy variable to indicate treatment group (Qureshi et al., 2016), or numerical variables to measure magnitude or variability of SBP (Divani et al., 2019; Meeks et al., 2019). However, the time-varying SBP curves constitute a richer source of information than these scalar summaries. In this analysis, we will use ATACH II data to investigate the time-period in which SBP management can have an effect on the 3-month rate of mortality and disability. Due to the different management strategies in the two groups, we will analyze them separately.

Mortality and disability are measured by the modified Rankin scale (mRS) score on an integer scale ranging from 0 (no symptoms) to 6 (death) at 3 months after randomization. The dichotomized mRS score (mRS = 0–3 vs. 4–6) is the scalar response in the FGLM. The functional predictors are the SBP curves, with recordings made at 15-minute intervals during the first hour and one measurement per hour thereafter during the first 24 hours after randomization. The R function `smooth.spline` was used to estimate the smooth curve $\mathrm{SBP}(t)$ from the discrete observations. Using a simple modification of the algorithms described in Section 3.2, the analysis also adjusts for relevant scalar covariates, including age, Glasgow Coma Scale (GCS) score, observed hematoma expansion (HE, binary) and logarithm of initial hematoma volume (log(Vol)). Thus, the linear component in the logistic regression is

$$\theta = \alpha + \int_0^{\delta} \mathrm{SBP}(t)\beta_0(t)\mathrm{d}t + \gamma_1 \mathrm{Age} + \gamma_2 \mathrm{GCS} + \gamma_3 \mathrm{HE} + \gamma_4 \log(\mathrm{Vol}).$$

After omitting the subjects with missing values, the sample size of standard group is 340 and intensive group is 376.

We compare the estimates of the proposed truncated model ($k = 47, q = 4$ and $m = 2$) with a penalized spline estimator implemented in the `refund` package. The 47 interior knots are equally spaced on $[0, 24]$, so that the possible estimates for the truncation point occur at half an hour or on the hour. Fig. 3 plots the estimates of β_0 for comparison, with a third FPCA-based estimator also included for the sake of interpretation, as described below. The estimated truncation points are $\hat{\delta} = 16.5$ and $\hat{\delta} = 13.5$ for the standard and intensive groups, respectively, so that the SBP after these time periods is estimated to have no significant effect on binarized 3-month mRS with the current management of ICH patients. To assess the uncertainty of the truncated estimate $\hat{\beta}$, nonparametric bootstrap was applied by sampling pairs (X_i, Y_i) with replacement. For each bootstrapped data set, the estimation was carried out using the same set of tuning parameters as the full data set, with the active set of B-splines also fixed to match the support of $\hat{\beta}$. These bootstrapped estimates were then used to obtain a 95% pointwise confidence bands shown in red dotted lines in Fig. 3. For both groups, the truncated estimate is mostly positive, confirming the intuition that higher SBP is associated with higher likelihood of disability or mortality. The penalized spline and the proposed estimators are similar for the standard group, but the truncation regularization yields a more logical estimate that does not take negative values toward the end of the 24 hour period. For the intensive group, the estimates are again qualitatively the same, with the penalized spline estimator being clearly oversmoothed as it is nearly linear. Thus, the proposed truncated estimator has advantages over the penalized spline estimator in both groups. The estimates of standard and intensive group mostly differ in the initial period after randomization, where the estimate in the intensive group is markedly larger during this period, consistent with the fact that patients in the intensive group were treated so as to lower SBP to below 140 mm Hg within the first two hours.

The fitted functional coefficient functions require careful interpretation. For instance, in the standard group, the fitted coefficient function is negative during the first two hours, but we can not conclude that the higher initial SBP is associated with lower probability of disability or mortality. To aid the interpretation, functional principal component analysis (FPCA) of the SBP curves was performed on the domains $[0, 16.5]$ and $[0, 13.5]$. After retaining the first four FPC scores, which explain 87.92% and 91.51% of the variability in standard and intensive groups, respectively, these were combined with the other scalar covariates and served as predictors in a standard logistic regression model.

Using the standard group as an example, we denote the estimated eigenfunctions of the standard group as $\{\varphi_j^S\}_{j=1}^4$, and j-th FPC score as FPC_j. These FPCA-based regression results are summarized in Table 3, with the corresponding estimates of β_0 depicted in Fig. 3. It can be seen that the individual p-values for scalar predictors are generally significant, with the

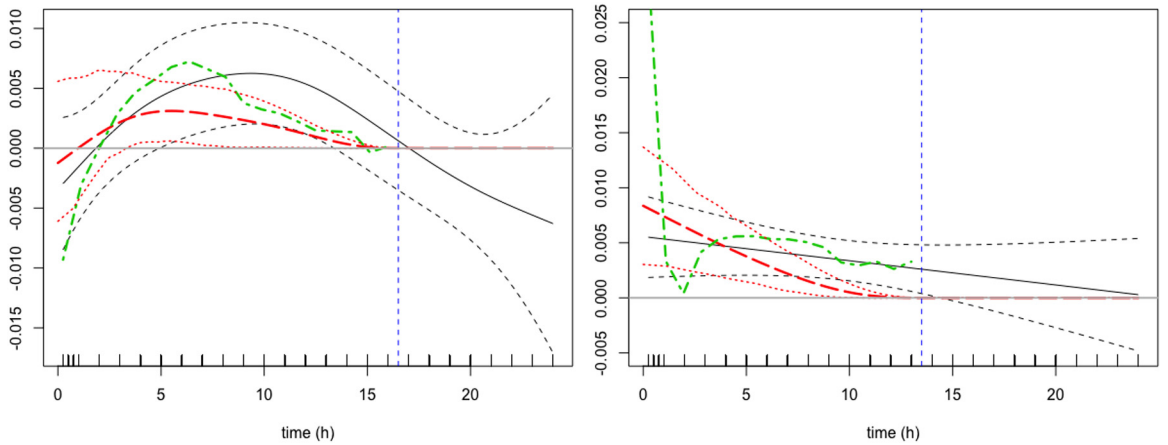

Fig. 3. Estimates of the functional parameter β_0 in the functional logistic regression model for the standard (left) and intensive (right) groups. (Solid Black) Penalized spline estimate produced by the function `pfr` in the `refund` package, with pointwise 95% confidence bands (Black Dash); (Red Dash) Proposed truncated estimate, with 95% pointwise bootstrap confidence bands (Red Dot); (Green Dash-Dot) FPCA-based truncated estimate. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

Table 3
Summary table of logistic regression models using four functional principal component scores. The p-values in bold indicate components of the blood pressure curves with p-values below 0.15.

	Standard Group				Intensive Group			
	Estimate	SE	z value	Pr(> |z|)	Estimate	SE	z value	Pr(> |z|)
GCS	-0.156	0.070	-2.217	0.027	-0.089	0.074	-1.213	0.225
Age	0.052	0.011	4.758	< 0.001	0.081	0.012	6.529	< 0.001
log(Vol)	0.914	0.181	5.060	< 0.001	1.341	0.191	7.02	< 0.001
HE	1.164	0.290	4.016	< 0.001	1.242	0.313	3.966	< 0.001
FPC_1	0.010	0.004	2.363	0.018	0.022	0.006	3.57	< 0.001
FPC_2	-0.006	0.007	-0.819	0.413	-0.008	0.01	-0.763	0.446
FPC_3	-0.012	0.008	1.473	0.141	0.015	0.015	1.009	0.313
FPC_4	-0.002	0.009	-0.232	0.816	-0.009	0.013	-0.669	0.503

exception of the GCS score in the intensive group. Likelihood ratio tests were conducted to compare the full model to a submodel containing only the scalar predictors and no FPC scores, resulting in p-values, 0.018 and 0.003, respectively, for standard and intensive groups, providing evidence for the significant effect of SBP on outcome after controlling other scalar variables. Furthermore, it emerges that the strongest SBP component affecting the outcome is FPC_1, with p-values of 0.01 and 0.022 in the standard and intensive groups, respectively, while the other FPC scores were not significant in either group.

In order to relate the FPC regression results in Table 3 to the truncated estimates $\hat{\beta}^S$ and $\hat{\beta}^I$, these estimates were projected onto the spaces spanned by the first four FPC basis functions within each group. Ranking the projection coefficients by magnitude, the smallest coefficients were set to zero, yielding approximations

$$\begin{aligned} \tilde{\beta}^S &= 0.006\varphi_1^S - 0.005\varphi_3^S, \\ \tilde{\beta}^I &= 0.014\varphi_1^I. \end{aligned} \tag{13}$$

Unsurprisingly, both groups retain a component corresponding to FPC_1. However, only the standard group retains a component along FPC_3. Although this component was not statistically significant in the FPC regression model, the proposed truncated estimator clearly contains a measurable signal in this direction, given that its coefficient of -0.005 is nearly equal in magnitude to that of the first component. We remark that $\|\tilde{\beta}^S\|^2 \approx 0.963\|\hat{\beta}^S\|^2$ and $\|\tilde{\beta}^I\|^2 \approx 0.918\|\hat{\beta}^I\|^2$, so that these projections are very near to the original estimates, and that the signs of the coefficients in these projections match those of the corresponding estimates in Table 3. The salient effects of SBP on the probability of disability or death can then be understood by modes of variation of the relevant eigenfunctions in Fig. 4, i.e. by adding and subtracting multiples of the eigenfunction from the mean SBP curve.

In both the standard and intensive groups, variation in the first eigenfunction direction is of the most significance. From Fig. 4, the first eigenfunction represents the overall magnitude of SBP curves, where larger values of FPC_1 associate with higher SBP level. Since the projection coefficients of the truncated estimates onto the first eigenfunctions in (13) are positive, we conclude that higher overall SBP leads to poor 3-month outcome. In the standard group, the third eigenfunction primarily differentiates between SBP patterns that are below/above the mean between hours 2–9 after randomization. Thus,

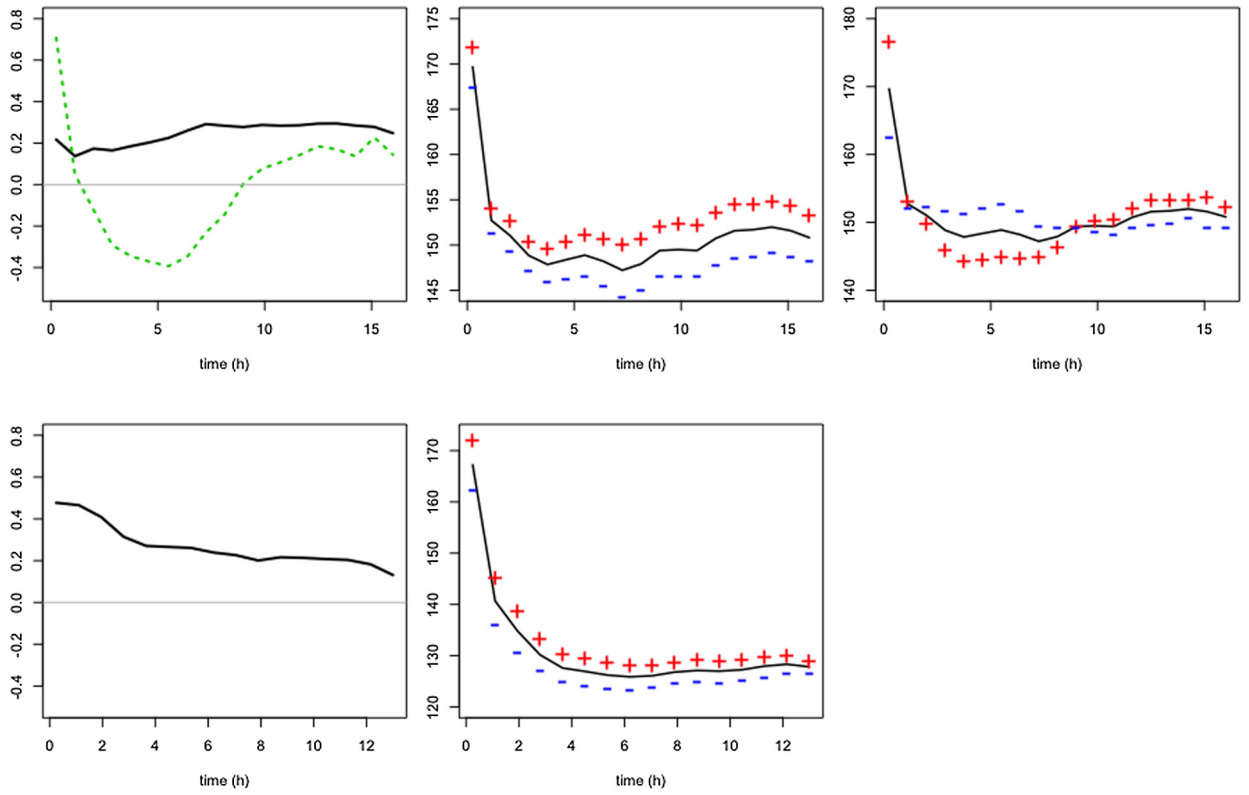

Fig. 4. Eigenfunction and mode of variation (mean $\pm 10 \times$ eigenfunction) plots for standard (top row) and intensive (bottom row) groups. (Top Left) First (Solid Black) and third (Green Dash) eigenfunctions, corresponding to FPC scores with p-value smaller than 0.15 in Table 3; (Top Middle) Mode of variation of the first eigenfunction; (Top Right) Mode of variation of the third eigenfunction; (Bottom Left) First eigenfunction, corresponding to the FPC score with p-value smaller than 0.15 in Table 3; (Bottom Middle) Mode of variation of the first eigenfunction.

the negative projection coefficient onto the third eigenfunction in (13) further emphasizes the importance of maintaining low SBP levels during this specific time period.

7. Discussion

In this paper, we develop a new methodology for producing a truncated estimate of the coefficient function in functional generalized linear models, where the true coefficient β_0 is assumed to be identically zero after some fixed but unknown timepoint $\delta_0 \in [0, T]$. With a minor modification to the proposed truncation penalty, the methodology can also be applied to regression models in which the coefficient is assumed to be zero prior to δ_0, rather than after, so that the functional covariate only has an effect toward the end of the domain. All theoretical results will hold in this scenario under essentially the same assumptions.

In simulations, the proposed estimator was found to be competitive with the recent method of Guan et al. (2020) in the setting of functional linear regression, and exhibited improved accuracy with a larger sample size in all simulation cases. Unlike previous methods, which either directly penalized the value of the truncation point (Hall and Hooker, 2016) or resulted in a nonconvex optimization problem (Guan et al., 2020), the proposed approach utilizes a nested group lasso penalty (Yuan and Lin, 2006) and leads to a convex optimization problem. We have demonstrated how the nonsmooth lasso penalty can be expressed in its dual formulation, and subsequently smoothed so that the objective function can be optimized by accelerated gradient descent algorithm. Furthermore, the explicit conditions for updating the active set after a change of the truncation tuning parameter lead to efficient computation of the entire solution path.

Acknowledgement

This work is supported by NSF grant DMS-2128589. The authors would like to thank Renee L. Martin and Lydia D. Foster from Medical University of South Carolina for their help in the use of the ATACH II database.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csda.2022.107421.

References

Atlantis, T., et al., 2004. Association of outcome with early stroke treatment: pooled analysis of atlantis, ecass, and ninds rt-pa stroke trials. Lancet 363, 768–774.

Caceres, J.A., Goldstein, J.N., 2012. Intracranial hemorrhage. In: Emergency Medicine Clinics of North America, vol. 30, p. 771.

Cai, T., Hall, P., 2006. Prediction in functional linear regression. Ann. Stat. 34, 2159–2179.

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

Cardot, H., Sarda, P., 2005. Estimation in generalized linear models for functional data via penalized likelihood. J. Multivar. Anal. 92, 24–41.

Chen, X., Lin, Q., Kim, S., Carbonell, J.G., Xing, E.P., et al., 2012. Smoothing proximal gradient method for general structured sparse regression. Ann. Appl. Stat. 6, 719–752.

Dandapani, B.K., Suzuki, S., Kelley, R.E., Reyes-Iglesias, Y., Duncan, R.C., 1995. Relation between blood pressure and outcome in intracerebral hemorrhage. Stroke 26, 21–24.

De Boor, C., 2001. A Practical Guide to Splines. Springer-Verlag.

Divani, A.A., Liu, X., Di Napoli, M., Lattanzi, S., Ziai, W., James, M.L., Jafarli, A., Jafari, M., Saver, J.L., Hemphill, J.C., et al., 2019. Blood pressure variability predicts poor in-hospital outcome in spontaneous intracerebral hemorrhage. Stroke 50, 2023–2029.

Divani, A.A., Liu, X., Petersen, A., Lattanzi, S., Anderson, C.S., Ziai, W., Torbey, M.T., Moullaali, T.J., James, M.L., Jafarli, A., et al., 2020. The magnitude of blood pressure reduction predicts poor in-hospital outcome in acute intracerebral hemorrhage. Neurocrit. Care, 1–10.

Febrero-Bande, M., González-Manteiga, W., 2013. Generalized additive models for functional data. Test 22, 278–292.

Greven, S., Scheipl, F., 2017. A general framework for functional regression modelling. Stat. Model. 17, 1–35.

Guan, T., Lin, Z., Cao, J., 2020. Estimating truncated functional linear models with a nested group bridge approach. J. Comput. Graph. Stat., 1–20.

Hall, P., Hooker, G., 2016. Truncated linear models for functional data. J. R. Stat. Soc., Ser. B, Stat. Methodol. 78, 637–653.

James, G.M., 2002. Generalized linear models with functional predictors. J. R. Stat. Soc. B 64, 411–432.

James, G.M., Wang, J., Zhu, J., et al., 2009. Functional linear regression that's interpretable. Ann. Stat. 37, 2083–2108.

Kokoszka, P., Reimherr, M., 2017. Introduction to Functional Data Analysis. CRC Press.

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

McCullagh, P., Nelder, J.A., 1983. Generalized Linear Models. Monographs on Statistics and Applied Probability. Chapman & Hall, London.

McLean, M.W., Hooker, G., Staicu, A.M., Scheipl, F., Ruppert, D., 2014. Functional generalized additive models. J. Comput. Graph. Stat. 23, 249–269.

Meeks, J.R., Bambhroliya, A.B., Meyer, E.G., Slaughter, K.B., Fraher, C.J., Sharrief, A.Z., Bowry, R., Ahmed, W.O., Tyson, J.E., Miller, C.C., et al., 2019. High in-hospital blood pressure variability and severe disability or death in primary intracerebral hemorrhage patients. Int. J. Stroke 14, 987–995.

Morgenstern, L.B., Hemphill III, J.C., Anderson, C., Becker, K., Broderick, J.P., Connolly Jr, E.S., Greenberg, S.M., Huang, J.N., Macdonald, R.L., Messe, S.R., et al., 2010. Guidelines for the management of spontaneous intracerebral hemorrhage: a guideline for healthcare professionals from the American heart association/American stroke association. Stroke 41, 2108–2129.

Morris, J.S., 2015. Functional regression. Annu. Rev. Stat. Appl. 2, 321–359.

Moullaali, T.J., Wang, X., Martin, R.H., Shipes, V.B., Robinson, T.G., Chalmers, J., Suarez, J.I., Qureshi, A.I., Palesch, Y.Y., Anderson, C.S., 2019. Blood pressure control and clinical outcomes in acute intracerebral haemorrhage: a preplanned pooled analysis of individual participant data. Lancet Neurol. 18, 857–864.

Müller, H.G., Stadtmüller, U., 2005. Generalized functional linear models. Ann. Stat. 33, 774–805.

Qureshi, A.I., Palesch, Y., 2011. Antihypertensive treatment of acute cerebral hemorrhage (atach) ii: design, methods, and rationale. Neurocrit. Care 15, 559–576.

Qureshi, A.I., Palesch, Y., Investigators, A.I., et al., 2012. Expansion of recruitment time window in antihypertensive treatment of acute cerebral hemorrhage (atach) ii trial. J. Vasc. Interv. Neurol. 5, 6.

Qureshi, A.I., Palesch, Y.Y., Barsan, W.G., Hanley, D.F., Hsu, C.Y., Martin, R.L., Moy, C.S., Silbergleit, R., Steiner, T., Suarez, J.I., et al., 2016. Intensive blood-pressure lowering in patients with acute cerebral hemorrhage. N. Engl. J. Med. 375, 1033–1043.

Qureshi, A.I., Palesch, Y.Y., Martin, R., Novitzke, J., Cruz-Flores, S., Ezzeddine, M.A., Goldstein, J.N., Hussein, H.M., Suri, M.F.K., Tariq, N., et al., 2010. Effect of systolic blood pressure reduction on hematoma expansion, perihematomal edema, and 3-month outcome among patients with intracerebral hemorrhage: results from the antihypertensive treatment of acute cerebral hemorrhage study. Arch. Neurol. 67, 570–576.

Ramsay, J.O., Silverman, B.W., 2005. Functional Data Analysis, second ed. Springer Series in Statistics. Springer, New York.

Shang, Z., Cheng, G., et al., 2015. Nonparametric inference in generalized functional linear models. Ann. Stat. 43, 1742–1773.

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

Xiao, L., 2019. Asymptotic theory of penalized splines. Electron. J. Stat. 13, 747–794.

Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. J. R. Stat. Soc., Ser. B, Stat. Methodol. 68, 49–67.

Zhao, P., Rocha, G., Yu, B., et al., 2009. The composite absolute penalties family for grouped and hierarchical variable selection. Ann. Stat. 37, 3468–3497.

Zhou, J., Wang, N.Y., Wang, N., 2013. Functional linear model with zero-value coefficient function at sub-regions. Stat. Sin. 23, 25.

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