Available at
www.ElsevierMathematics.com
POWERED BY SCIENCE DIRECT®
Journal of Multivariate Analysis 92 (2005) 24–41

Journal of Multivariate Analysis

http://www.elsevier.com/locate/jmva

Estimation in generalized linear models for functional data via penalized likelihood

Hervé Cardot[a,*] and Pacal Sarda[b]

[a] Unité Biométrie et Intelligence Artificielle, INRA Toulouse, BP 27, 31326 Castanet-Tolosan Cedex, France
[b] Laboratoire de Statistique et Probabilités, Université Paul Sabatier, 31062 Toulouse Cedex, France

Received 8 January 2002

Abstract

We analyze in a regression setting the link between a scalar response and a functional predictor by means of a Functional Generalized Linear Model. We first give a theoretical framework and then discuss identifiability of the model. The functional coefficient of the model is estimated via penalized likelihood with spline approximation. The L^2 rate of convergence of this estimator is given under smoothness assumption on the functional coefficient. Heuristic arguments show how these rates may be improved for some particular frameworks.

AMS 2000 subject classifications: 62G07; 62G12; 62G20; 65J20

Keywords: Covariance operator; Hilbert space valued random variables; Ill-posed problem; Regularization; Splines

1. Introduction

In many areas of research one has to deal with functional data i.e. with data which are curves. It is especially the case in chemometrics, meteorology or speech analysis. For instance there is the regression setting where the predictor is a random function and the response a scalar. In the past one has mainly developed a "discrete" approach in this context: the discretization points of the curve predictor are considered as the coordinates of a multiple predictor vector. Then procedures that

*Corresponding author. Fax: +33-5-61-2853-35.
E-mail addresses: cardot@toulouse.inra.fr (H. Cardot), sarda@cict.fr (P. Sarda).

0047-259X/$ - see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2003.08.008

take into account the large number of predictors as well as the high correlations between them have been proposed: see for instance [15] for these tools in chemometrics. On the other hand, there has been existing for a long time a "functional" approach for which models aim at taking into acount the functional nature of the data: see for instance the work from [10,12] on Data Analysis in the context of Hilbert spaces theory. Until recently, this approach has been certainly less used than the discrete one in practical studies. The monographs from [26,27] which investigate not only the above regression setting but also a variety of other statistical problems with functional data is an important step for the popularization of these methods. Moreover, an increasing amount of recent papers investigate (functional) models for functional data.

Coming back to the regression problem with a scalar response and a functional predictor, the most natural functional model is the continuous version of the multiple linear model i.e. the functional linear model (see [6,7,18]). However, this model may be too restrictive in several applications for instance when the response is categorical. In the same spirit as in the multivariate setting one can think of a *functional generalized linear model* which is the functional version of the generalized linear model introduced by Nelder and Wedderburn [23]. Such models have been implicitly introduced in the literature. Cardot et al. [5] have proposed a principal components regression to estimate the functional coefficients in a multilogit model in order to recover the land use from a temporal sequence of remote sensing data. Marx and Eilers [21] used a penalized spline procedure in a binomial model for phoneme recognition.

In Section 2, we give a theoretical framework for the generalized functional linear model which involves an exponential family of distributions. We discuss the problems of identifiability of the model and the need of introducing a regularization penalty. An estimation procedure based on B-splines quite similar to the one proposed by Marx and Eilers [21] is introduced in Section 3. Indeed, both procedures are based on penalized likelihood, the difference coming from the penalty which is expressed here as the norm of the derivative of given order of the function. Then we look at asymptotic properties of the estimator which are seldom examined in the literature. Our main result concerns the L^2 rate of convergence for our maximum penalized likelihood estimator. The main strength of this result is that we do not assume any particular structure for the eigenvalues of the covariance operator. In Section 4, a discussion shows, with heuristic arguments, how these rates may depend on the covariance structure of the data and how we could get better rates for some particular situations. Section 5 is devoted to the proofs.

2. The functional generalized linear model

We adopt in the following the same notations as in the paper by Stone [30] by considering an exponential family of the following form:

$$\exp\{b_1(\eta)y + b_2(\eta)\}v(dy), \qquad (1)$$

where v is a nonzero measure on R which is not concentrated at a single point and where the function b_1 is twice continuously differentiable and b_1' is strictly positive on R. Then, the function b_1 is strictly increasing and b_2 is twice continuously differentiable on R. The mean μ of the distribution is

$$\mu = b_3(\eta) = -\frac{b_2'(\eta)}{b_1'(\eta)},$$

where b_3 is continuously differentiable and b_3' is strictly positive on R. The function b_3^{-1} is called the *link function* and one has $\eta = b_3^{-1}(\mu)$.

It is also assumed as in Stone's paper that there is an interval S in R such that v is concentrated on S and

$$\text{(H.1)}\quad b_1''(\eta)y + b_2''(\eta) < 0, \quad \forall \eta \in R, \ \forall y \in S.$$

The reader is referred to [30] for examples of exponential families, such as the Bernouilli or the gamma distribution, satisfying condition (H.1).

Let X and Y be two random variables defined on the same probability space with X valued in the separable Hilbert space $H = L^2_{[0,1]}$ and Y valued in R. Let $\langle \phi, \psi \rangle$ denote the usual inner product of functions ϕ and ψ in H, defined by $\langle \phi, \psi \rangle = \int_0^1 \phi(t)\psi(t)\, dt$ and let $||\phi||$ denote the norm associated with this inner product. We assume that the following *functional generalized linear model* holds, that is to say we assume the existence of a function $\alpha \in H$ such that

$$E(Y|X = x) = b_3(\langle \alpha, x \rangle), \quad x \in H. \tag{2}$$

The conditional distribution of Y given $X = x$ is supposed to belong to the exponential family (1) or at least to satisfy Conditions 2–4 of Stone [30].

Without loss of generality we assume that the functional random variable X is centered i.e. $EX(t) = 0$, for t a.e. We also suppose that X is of second order i.e. $E||X||^2 < \infty$. Thus, the covariance operator Γ of the H-valued random variable X is defined as

$$\Gamma x(t) = \int_0^1 E[X(t)X(s)]x(s)\, ds, \quad x \in H, \ t \in [0,1].$$

The operator Γ is an integral operator whose kernel is the covariance function of X and it is nuclear, self-adjoint and nonnegative [10,11]. Moreover, the operator Γ is assumed to satisfy the condition

(H.2) The eigenvalues of Γ are nonzero.

Condition (H.2) ensures the identifiability of the model (see below) and for instance is assumed in other settings such as the one described in [3]. It is fulfilled when $X(t)$ is a standard brownian motion (see [6]). Let us notice that it can be relaxed when there exist some null eigenvalues by changing the Hilbert space of reference, taking H as the closure of the range of Γ (see [8]). Let us also remark that if we had supposed that there were only a finite number of non null eigenvalues then we would be in a classical parametric framework since it would mean that we would have only a finite number of covariates.

To get identifiability, let us denote by $\lambda_j,\ j=1,2,\ldots$ the eigenvalues of Γ and by $v_j,\ j=1,2,\ldots$ a complete orthonormal sequence of eigenfunctions and let α_1 and α_2 be two functions in H such that

$$b_3(\langle \alpha_1, X\rangle) = b_3(\langle \alpha_2, X\rangle).$$

Since b_3 is strictly increasing one has

$$\langle \alpha_1 - \alpha_2, X\rangle = 0,$$

and then

$$\begin{aligned} E\langle \alpha_1-\alpha_2, X\rangle^2 &= \langle \Gamma(\alpha_1-\alpha_2), \alpha_1-\alpha_2\rangle \\ &= \sum_{j=1}^{+\infty} \lambda_j \langle \alpha_1-\alpha_2, v_j\rangle^2 \\ &= 0. \end{aligned}$$

Now, since $\lambda_j \neq 0,\ \forall j$, one has

$$\langle \alpha_1-\alpha_2, v_j\rangle^2 = 0, \quad \forall j,$$

and then $\alpha_1=\alpha_2$ almost everywhere in H.

The expected log-likelihood is defined as

$$\Lambda(a) = E(b_1(\langle a, X\rangle) b_3(\langle \alpha, X\rangle) + b_2(\langle a, X\rangle)), \quad a \in H,$$

Hypothesis (H.1) gives directly

$$b_1''(\eta) b_3(\eta_0) + b_2''(\eta) < 0, \quad \forall \eta,\ \eta_0 \in R, \tag{3}$$

which implies that the function $\Psi(\eta) = b_1(\eta) b_3(\eta_0) + b_2(\eta)$ is strictly concave and has a unique maximum at η_0. Then, when model (2) holds, the function α is a maximum of Λ which is essentially uniquely determined under (H.2).

3. Estimation of the functional coefficient

In this section we introduce an estimator of α based on a B-splines expansion maximizing the penalized log-likelihood. First of all, let us describe the space of spline functions defined on $[0,1]$ with equispaced knots. Suppose that q and k are integers and let S_{qk} be the space of *spline* functions defined on $[0,1]$, with degree q and $k-1$ equispaced interior knots. The set S_{qk} is the set of functions s satisfying:

- s is a polynomial of degree q on each interval $[(t-1)/k, t/k]$, $t=1,\ldots,k$;
- s is $q-1$ times continuously differentiable on $[0,1]$.

The set S_{qk} is known to be a linear space with dimension $q+k$ and one can derive a basis by means of normalized B-splines $\{B_{k,j}, j=1,\ldots,k+q\}$ (see [1]). In the following we denote as $\mathbf{B}_k$ the vector of all the B-splines and as $\mathbf{B}_k^{(m)}$ the vector of derivatives of order m of all the B-splines for some integer m $(m<q)$.

Our penalized B-splines estimator of α is thus defined as

$$\begin{aligned}\hat{\alpha}_{\mathrm{PS}} &= \sum_{j=1}^{q+k} \hat{\theta}_j B_{k,j} \\ &= \mathbf{B}_k' \hat{\boldsymbol{\theta}},\end{aligned}$$

where $\hat{\boldsymbol{\theta}}$ is a solution of the following maximization problem:

$$\max_{\boldsymbol{\theta}\in R^{q+k}} \frac{1}{n}\sum_{i=1}^{n}(b_1(\langle \mathbf{B}_k'\boldsymbol{\theta}, X_i\rangle)Y_i + b_2(\langle \mathbf{B}_k'\boldsymbol{\theta}, X_i\rangle)) - \frac{1}{2}\rho\|\mathbf{B}_k^{(m)'}\boldsymbol{\theta}\|^2, \tag{4}$$

with smoothing parameter $\rho>0$. The estimator $\hat{\alpha}_{\mathrm{PS}}$ is of the same type as the one introduced by Marx and Eilers [20], with however a different roughness penalty. Indeed, our penalty, borrowed from [24], allows to obtain a given level of smoothness in the smooth representation following ideas from [26, Chapter 4]. This penalty can also be modified in order to give local measures of roughness [4]. Computation of the estimator is achieved by means of a slight modification of the scoring algorithm for generalized linear models [22]. Marx and Eilers [21] also give formulas for computing a generalized cross validation criterion that allows to choose reasonable values for ρ.

We study now the performance of estimator $\hat{\alpha}_{\mathrm{PS}}$ in terms of the asymptotic behavior of the L^2 norm in H with respect to the distribution of X defined as

$$\|\phi\|_2^2 = \langle \Gamma\phi, \phi\rangle, \quad \phi\in H.$$

Note that since for each ϕ in H, there exists a unique element Φ in the space H' of continuous linear operator from H to R such that $\Phi(X) = \langle \phi, X\rangle$, the corresponding norm in H' is

$$\|\Phi\|_2^2 = E\Phi^2(X), \quad \Phi\in H'.$$

To derive L^2 convergence rates for $\hat{\alpha}_{\mathrm{PS}}$ we assume moreover the following conditions:

$$\text{(H.3)} \quad \|X\|\leqslant C_1 < +\infty, \quad \text{a.s.}$$

The function α is supposed to have p' derivatives for some integer p' with $\alpha^{(p')}$ satisfying

$$\text{(H.4)} \quad |\alpha^{(p')}(y_1) - \alpha^{(p')}(y_2)|\leqslant C_2|y_1 - y_2|^{\nu}, \quad C_2>0,\ \nu\in[0,1].$$

In the following, we note $p = p' + \nu$ and assume that the degree q of the splines is such that $q\geqslant p$.

Theorem 3.1. *Let $\rho\sim n^{(\delta-1)/2}$, for some $0<\delta<1$ and suppose that $\rho^{-1}k^{-2p} + \rho k^{2(m-p)} = O(1)$ and $\rho^2k^{2m} = o(1)$. Under hypothesis* (H.1)–(H.4), *we have*

(i) *A unique solution to the maximization problem* (4) *exists except on an event whose probability tends to zero as $n\to\infty$.*

(ii) $$||\hat{\alpha}_{\mathrm{PS}} - \alpha||_2^2 = O_P\left(\frac{k}{n}\right) + O(k^{-2p}) + O(\rho k^{2(m-p)}) + O(\rho).$$

Corollary 3.1. *Under the assumptions of Theorem* 3.1 *and for* $k \sim n^{1/(2p+1)}$ *and* $\rho \sim n^{(\delta-1)/2}$ *we get for* $m \leqslant p$ *the* L^2 *rate of convergence*

$$||\hat{\alpha}_{\mathrm{PS}} - \alpha||_2^2 = O_P(n^{-2p/(2p+1)}) + O(\rho). \tag{5}$$

Remark 3.1. Let us note that we can also get similar convergence results if we consider an estimator built with the penalty proposed by Marx and Eilers [21]. Indeed, the eigenvalues of the matrix associated to their penalty have the same asymptotic shape as ours (see [4, Lemma 5.2]) and thus conditions on the existence of an estimator and asymptotic rates of convergence are the same for well-chosen values of the regularization parameter ρ.

4. Discussion

4.1. Are these rates optimal?

It is natural to ask if the rates of convergence obtained in Corollary 3.1 are optimal and actually it is still an open problem. Indeed, similar results as the ones stated by Stone [28] in a (multivariate) nonparametric regression setting are not available for the statistical models with functional covariates and for the functional generalized linear model in particular. Following Stone's paper, answering to the general question of optimality for estimators in a model is a hard task and it depends on several things: assumptions on the model, the estimators in hand and the type of loss criterion among others. Our aim in this section is just to give some general insights on this problem.

First of all, let us note that very few asymptotic results have been proved in the context of linear models for functional variables, except [2,3] in the setting of (Hilbertian) autoregressive linear processes and [6,7] for the functional linear regression model. It appears in these papers that the (upper bounds for the) rates of convergence for estimators based on Functional Principal Components are quite poor comparatively to the ones obtained for spline estimators (even with stronger assumptions for the first case). This should be linked with results obtained by means of simulation studies where spline estimators seems to be superior [7].

As a matter of fact, the results obtained in Section 3 are stated in a quite large setting with no assumptions on the decay on the eigenvalues of the covariance operator (it is not the case for estimators based on Functional Principal Components), no assumptions on the regularity of the sample paths of $X(t), t \in [0,1]$ and no assumptions on the distribution of X (for instance no stationarity assumption or bounded density), except that X admits a finite second moment. Additional assumptions on the model can lead to improved asymptotic results. We will come back to this point in Sections 4.3 and 4.4 below.

For the moment let us note that another way to improve our results could be to consider the number of covariates in the model introduced in Section 2 as infinite. These "covariates" are however highly correlated. There exist theoretical works on the rates of convergence for generalized linear models when the number of covariates tends to infinity (see e.g. [25]) and one could think of using these ideas in our setting. Nevertheless the main point in the above works is to suppose that the covariance matrix is bounded below. That is not the case for functional data since the covariance operator is compact. Note that it is also for this reason that one has to add a penalization term in the likelihood to get consistent estimators (different but related arguments for introducing a penalty may be found in [19]).

Now, let us compare our results to the ones obtained by Ferraty and Vieu [14] which deal with a fully functional nonparametric model. These authors suppose that a fractal type assumption on X holds. Roughly speaking, that means that X belongs locally to a functional space with finite dimension. Under this condition they obtain rates of convergence which can be related to the ones of Stone [28]. It is important to see that this condition deals not only with the process X but also with the estimation procedure *via* a semi-norm used to evaluate the proximity between curves. Then, an important question in practical situations is to find a data-driven approach to adapt the estimator (to find a semi-norm) to the process in hand. Note for instance that this fractal type assumption is not fulfilled for the Brownian motion when the proximity between curves is measured by means of the usual $L^2_{[0,1]}$ norm.

As a conclusion of this section we will say that the points raised above deserve further investigations in several directions: adapting (or introducing new) estimation procedures as well as finding conditions under which better rates hold (see Sections 4.3 and 4.4 below for some partial answers to the latter point). In this sense, one can even ask if actually it is possible to find estimates/conditions for which the usual parametric rate is achieved. Indeed, both the covariates and the functional coefficient α belong to the same functional space and thus are vectors of the same space. As a consequence, one could imagine that we are in a parametric framework, in an infinite dimension case, and thus parametric rates may occur. This is true for instance when estimating the eigenfunctions of a covariance operator (see [11]).

4.2. *Do the rates depend on the eigenvalues of the covariance operator?*

One can reasonably think that the rates of convergence should depend on the eigenvalues of the covariance operator (or the second derivative operator of the likelihood). Actually, we will see that this dependency is "hidden" in our results.

Indeed, on the one hand, our loss criterion

$$\langle \Gamma f, f \rangle = E\langle f, X \rangle^2, \tag{6}$$

can be seen as the standard squared L^2 norm in the usual nonparametric setting. But this criterion can also be written as follows:

$$\langle \Gamma f, f \rangle = \sum_{j=1}^{+\infty} \lambda_j \langle f, v_j \rangle^2, \tag{7}$$

where the sequence of eigenvalues λ_j satisfies $\sum_j \lambda_j < +\infty$. Since these eigenvalues are the variance of the projection of X onto the eigenfunctions v_j, this latter relation means that our criterion gives more importance to the directions in which X has a larger variance, that is to say in which X "often goes" and gives a "neglictible" importance to the directions where it is most rarely. Maybe it is why conditions on the decay of the eigenvalues are dropped in our theorem.

From another point of view, the shape of the eigenvalues have some incidence on the existence of our estimator.

4.3. Some heuristic arguments for better rates

The limitation of the speed of convergence in Corollary 3.1 comes from the condition $\rho \sim n^{-(1-\delta)/2}$ (to get existence of the estimator) together with the bias term: it follows from this condition that the term $O(\rho)$ cannot be eliminated. Thus, there are two ways in getting better rates of convergence. On the one hand, we may find better bound for the bias (see Section 4.4) and on the other hand, we may improve the bound on the smallest eigenvalue of the information matrix in order to weaken the condition $\rho \sim n^{-(1-\delta)/2}$.

From Eqs. (29), (30), (32) and (33) below we find a lower bound for the smallest eigenvalue of order ρ/k. In fact, we can show in some cases that it is larger than ρ/k and thus weaker conditions on ρ can lead to the existence of a solution except on a space whose probability tends to zero as n tends to infinity. For this, we will consider some particular covariance structures for which convergence rates may be improved.

For instance, let us consider the covariance structure $E(X(s)X(t)) = \exp(-|s-t|)$, i.e. the covariance function of the Ornstein–Uhlenbeck process defined in $[0,1]$. One can show (see [16] or [3]) that the eigenelements of the covariance operator satisfy

$$v_j(t) = \sin(v_j(t-0.5) + (j+1)\pi/2), \quad t \in [0,1],$$

and

$$\lambda_j = 2(1+v_j^2)^{-1},$$

where the v_j's are the solutions of the transcendental equation $\tan v_j = -2v_j(1-v_j^2)^{-1}$ sorted in ascending order. An explicit solution cannot be found but it is easy to check that asymptotically, $v_j \sim j\pi$ and $\lambda_j \sim j^{-2}$.

Consider now the more general case in which the eigenfunctions are not necessarily known explicitly but satisfy

$$D^m v_j = \pm v_j^m v_j, \tag{8}$$

where D^m is the derivative operator of order m with m even. Other examples of eigenfunctions satisfying (8) can be found in [16]. Let us notice that these eigenfunctions v_j, $j = 1, 2, \ldots,$ also diagonalize the penalization operator. Thus, the eigenvalues of the penalized covariance operator are

$$\lambda_j^\rho \sim \lambda_j + \rho v_j^{2m}.$$

Suppose moreover that $v_j \sim j$ asymptotically.

In the case of an arithmetic decay for the eigenvalues, $\lambda_j = cj^{-\gamma},\ p \geqslant \gamma > 1$, it is easy to check that $\min_j \lambda_j^\rho \sim \rho^{\gamma/(2m+\gamma)}$. Now, taking $\rho \sim n^{-(1-\delta)(1/2+m/\gamma)+\eta}$ for $0 < \eta \leqslant (1-\delta)(1/2+m/\gamma) - 2p/(2p+1)$ and adequate values for δ and m implies the existence of the estimator on a space whose probability tends to zero as n tends to infinity. Moreover in this case ρ is negligible with respect to $n^{-2p/(2p+1)}$. This tells us that choosing adequate basis of functions to build our estimator (instead of splines) may leads to Stone's "optimal" rate of convergence.

4.4. The particular case of the linear model

For the linear model, we have an explicit expression for our estimator and it can be shown that better bound for the bias occurs if we suppose moreover that the "projection" of α onto the space S_{qk} belongs to the range of the covariance operator Γ. It allows us to get Stone's "optimal" rates of convergence.

Maximizing the expected log-likelihood is equivalent to minimizing the following criterion

$$\min_{\beta \in S_{qk}} E(\langle \alpha - \beta, X \rangle^2), \tag{9}$$

that is to say $\min_{\beta \in S_{qk}} ||\alpha - \beta||_2^2$. Let us denote by $\tilde{\alpha}$ the minimizer of (9). Consider to simplify the ridge regression approximation (i.e. $m = 0$) $\tilde{\alpha}_{\mathrm{PS}}$ defined as

$$\mathrm{Arg} \min_{\beta \in S_{qk}} E(\langle \alpha - \beta, X \rangle^2) + \frac{1}{2}\rho ||\beta||^2. \tag{10}$$

Since S_{qk} is a finite dimensional function space and the eigenvalues of Γ are supposed to be strictly positive, it is easy to show that $\tilde{\alpha}$ and $\tilde{\alpha}_{\mathrm{PS}}$ are uniquely determined. Moreover, they satisfy respectively the functional normal equations

$$\Gamma\tilde{\alpha}(t) = E(YX(t)), \quad t \in [0,1], \tag{11}$$

and

$$\Gamma\tilde{\alpha}_{\mathrm{PS}}(t) + \rho\tilde{\alpha}_{\mathrm{PS}}(t) = E(YX(t)), \quad t \in [0,1]. \tag{12}$$

Combining equalities (11), (12) and expanding $\tilde{\alpha}$ and $\tilde{\alpha}_{\mathrm{PS}}$ in the basis of the orthonormal eigenfunctions of Γ, we get

$$\langle v_j, \tilde{\alpha}_{\mathrm{PS}} \rangle = \frac{\lambda_j}{\lambda_j + \rho} \langle v_j, \tilde{\alpha} \rangle$$

and

$$||\tilde{\alpha} - \tilde{\alpha}_{\mathrm{PS}}||_2^2 = \sum_{j=1}^{\infty} \lambda_j \frac{\rho^2}{(\lambda_j + \rho)^2} \langle v_j, \tilde{\alpha} \rangle^2. \tag{13}$$

Suppose now that $\tilde{\alpha}$ belongs to the range of Γ, that is to say there exists a function $g_{\tilde{\alpha}} \in H$ such that $\Gamma g_{\tilde{\alpha}} = \tilde{\alpha}$. We have that

$$\sum_{j=1}^{\infty} \frac{\langle v_j, \tilde{\alpha} \rangle^2}{\lambda_j^2} < +\infty \tag{14}$$

and thus with (13), we can bound

$$||\tilde{\alpha} - \tilde{\alpha}_{\mathrm{PS}}||_2^2 = O(\rho^2). \tag{15}$$

With similar arguments as those used in the beginning of the proof of Lemma (5.1), one can find a function $s \in S_{kq}$ such that $\sup_{t \in [0,1]} |s(t) - \alpha(t)| \leqslant C_3 k^{-p}$ and thus $||\alpha - \tilde{\alpha}||_2^2 \leqslant ||\alpha - s||_2^2 = O(k^{-2p})$. Consequently, the squared bias is of order $O(\rho^2) + O(k^{-2p})$ and the rates of convergence of the ridge regression estimator is

$$||\alpha - \hat{\alpha}_{\mathrm{PS}}||_2^2 = O_p(n^{-2p/(2p+1)}),$$

provided that ρ and k are well chosen.

5. Proof of Theorem 3.1

The proof is based on similar arguments as the proofs in [13] and [30] for generalized linear models and generalized additive models, respectively. The difference is that the parameter and the data belong to an infinite dimension space and thus estimation is an ill-posed problem. The novelty consists in adding a penalty term in the log-likelihood in order to get a consistent estimator.

To avoid confusions, matrices and vectors are denoted with bold faces letters and usual norms for these objects are denoted by $||.||$.

For some function $a \in H$, let us define Λ_ρ the expected penalized log-likelihood

$$\Lambda_\rho(a) = E(b_1(\langle a, X \rangle) b_3(\langle \alpha, X \rangle) + b_2(\langle a, X \rangle)) - \frac{1}{2}\rho ||a^{(m)}||^2,$$

and $\Lambda_{n,\rho}(a)$ the empirical penalized log-likelihood

$$\Lambda_{n,\rho}(a) = \frac{1}{n} \sum_{i=1}^{n} (b_1(\langle a, X_i \rangle) Y_i + b_2(\langle a, X_i \rangle)) - \frac{1}{2}\rho ||a^{(m)}||^2.$$

The proof of Theorem 3.1 will be complete after showing the following lemmas.

Lemma 5.1. *If* $\rho^{-1}k^{-2p} + \rho k^{2(m-p)} = O(1)$,

(i) *there is a unique* $\tilde{\alpha}_{\mathrm{PS}} \in S_{kq}$ *such that*

$$\tilde{\alpha}_{\mathrm{PS}} = \arg \max_{\beta \in S_{kq}} \Lambda_\rho(\beta),$$

(ii) $||\alpha - \tilde{\alpha}_{\mathrm{PS}}||_2^2 = O(k^{-2p}) + O(\rho k^{2(m-p)}) + O(\rho)$.

Lemma 5.2. *If conditions in Lemma* 5.1 *are fulfilled and* $\rho \sim n^{(\delta-1)/2}$ *for some* $\delta > 0$,

(i) *a unique solution to the maximization problem* (4) *exists except on an event whose probability tends to zero as* $n \to \infty$.

(ii) *If moreover* $\rho^2 k^{2m} = o(1)$, *then*

$$||\tilde{\alpha}_{\text{PS}} - \hat{\alpha}_{\text{PS}}||_2^2 = O_P\left(\frac{k}{n}\right).$$

5.1. Proof of Lemma 5.1

From Theorem XII.1 in [1] and (H.4), there exists a function $s \in S_{kq}$ such that $\sup_{t\in[0,1]} |s(t) - \alpha(t)| \leqslant C_3 k^{-p}$. Then, one can deduce from (H.3), (H.4) and Lemma 8 of [29] that

$$||s - \alpha||_2^2 + \rho||s^{(m)}||^2 \leqslant C_4(k^{-2p} + \rho k^{2(m-p)} + \rho). \tag{16}$$

Let $\delta_n = k^{-2p} + \rho k^{2(m-p)} + \rho$ and c be a positive constant that will be determined later and consider the space of functions $\beta \in S_{qk}$ such that

$$||\beta - \alpha||_2^2 + \rho||\beta^{(m)}||^2 \leqslant c\delta_n.$$

Define $\beta^{(t)} = t\beta + (1-t)\alpha, t \in [0,1]$. One has

$$\frac{d^2}{dt^2}\Lambda(\beta^{(t)}) = E(\langle \beta - \alpha, X\rangle^2 [b_1''(\langle \beta^{(t)}, X\rangle) b_3(\langle \alpha, X\rangle) + b_2''(\langle \beta^{(t)}, X\rangle)]).$$

On the other side

$$\begin{aligned} |\langle \beta^{(t)}, X\rangle| &\leqslant ||\beta^{(t)}|| ||X|| \\ &\leqslant (t||\beta|| + (1-t)||\alpha||)||X||. \end{aligned}$$

Now, let us expand β as follows:

$$\beta(t) = \tilde{P}(t) + \tilde{R}(t), \quad t \in [0,1], \tag{17}$$

where $\tilde{P}(t) = \sum_{\ell=0}^{m-1} \frac{t^\ell}{\ell!}\beta^{(\ell)}(0)$ and $\tilde{R}(t) = \int_0^t \beta^{(m)}(u)\frac{(t-u)^{m-1}}{(m-1)!}\,du$. Since $\tilde{P}$ belongs to the m-dimensional space of polynomial functions on $[0,1]$ with degree less or equal to $m-1$, one obtains easily with assumption (H.2)

$$\begin{aligned} ||\beta||^2 &\leqslant 2||\tilde{P}||^2 + 2||\tilde{R}||^2 \\ &\leqslant C_5||\tilde{P}||_2^2 + 2||\tilde{R}||^2 \\ &\leqslant 2C_5||\beta||_2^2 + 2C_5||\Gamma|| ||\tilde{R}||^2 + 2||\tilde{R}||^2. \end{aligned}$$

Since we have with the Schwarz inequality

$$(\tilde{R}(t))^2 \leqslant C_6 \int_0^t (\beta^{(m)}(u))^2\,du,$$

one gets

$$||\beta||^2 = O(1) + O(\delta_n \rho^{-1}), \tag{18}$$

which gives us, with the condition $\rho^{-1}k^{-2p} + \rho k^{2(m-p)} = O(1)$,

$$||\beta||^2 \leqslant C_7, \tag{19}$$

for some positive constant C_7. It follows from (3) and a continuity argument that there are two positive constants C_8 and C_9 such that

$$-C_8||\beta - \alpha||_2^2 \leqslant \frac{d^2}{dt^2}\Lambda(\beta^{(t)}) \leqslant -C_9||\beta - \alpha||_2^2. \tag{20}$$

Since α maximizes Λ one gets

$$\frac{d}{dt}\Lambda(\beta^{(t)})\bigg|_{t=0} = 0,$$

and then

$$\Lambda(\beta) - \Lambda(\alpha) = \int_0^1 (1-t)\frac{d^2}{dt^2}\Lambda(\beta^{(t)})\,dt,$$

which gives us with (20)

$$\Lambda_\rho(\beta) - \Lambda(\alpha) \leqslant -C_{10}(||\beta - \alpha||_2^2 + \rho||\beta^{(m)}||^2), \tag{21}$$

where $C_{10} = \min(C_9, 1/2)$. On the other hand we have

$$\Lambda_\rho(s) - \Lambda(\alpha) \geqslant -C_{11}(||s - \alpha||_2^2 + \rho||s^{(m)}||^2), \tag{22}$$

with $C_{11} = \max(C_8, 1/2)$. Let us consider a function $a \in S_{kq}$ such that

$$||a - \alpha||_2^2 + \rho||a^{(m)}||^2 = c\delta_n. \tag{23}$$

One has

$$\Lambda_\rho(a) - \Lambda_\rho(s) = \Lambda_\rho(a) - \Lambda(\alpha) + \Lambda(\alpha) - \Lambda_\rho(s),$$

which gives us, with (21) and (22),

$$\Lambda_\rho(a) - \Lambda_\rho(s) \leqslant C_{11}(||s - \alpha||_2^2 + \rho||s^{(m)}||^2) - C_{10}c\delta_n < 0, \tag{24}$$

provided that c is chosen sufficiently large. We then have for a such that $||a - \alpha||_2^2 + \rho||a^{(m)}||^2 = c\delta_n$ the strict inequality

$$\Lambda_\rho(a) < \Lambda_\rho(s).$$

From the strict concavity of Λ_ρ on $\{\beta \in S_{qk}: ||\beta - \alpha||_2^2 + \rho||\beta^{(m)}||^2 \leqslant c\delta_n\}$ it follows that there is a unique $\tilde{\alpha}_{\mathrm{PS}}$ in S_{qk} such that

$$\tilde{\alpha}_{\mathrm{PS}} = \arg\max_{\beta \in S_{kq}} \Lambda_\rho(\beta),$$

and

$$||\tilde{\alpha}_{\mathrm{PS}} - \alpha||_2^2 + \rho||\tilde{\alpha}_{\mathrm{PS}}^{(m)}||^2 = O(\delta_n). \tag{25}$$

Finally we get

$$||\tilde{\alpha}_{\mathrm{PS}} - \alpha||_2^2 = O(\delta_n),$$

which achieves the proof of Lemma 5.1.

5.2. Proof of Lemma 5.1

(i) Let $\tau = \tau_n$ be a sequence of positive reals tending to zero such that τ/ρ is bounded and define the space

$$B_n(\tau) = \{\beta \in S_{kq}\colon\ ||\tilde{\alpha}_{\rm PS} - \beta||_2^2 + \rho||\tilde{\alpha}_{\rm PS}^{(m)} - \beta^{(m)}||^2 \leqslant \tau\}.$$

Let us consider $\beta \in B_n(\tau)$ and write $\beta(t) = \sum_{j=1}^{q+k} \theta_j B_{kj}(t) = \boldsymbol{\theta}'\mathbf{B}_k(t)$. Let us also write $\tilde{\alpha}_{\rm PS}(t) = \sum_{j=1}^{q+k} \tilde{\theta}_j B_{kj}(t) = \tilde{\boldsymbol{\theta}}'\mathbf{B}_k(t)$. The score $s_n(\boldsymbol{\theta})$ is given by

$$\begin{aligned} s_n(\boldsymbol{\theta}) &= \frac{\partial \Lambda_{n,\rho}(\beta)}{\partial \boldsymbol{\theta}} \\ &= \frac{1}{n}\sum_{i=1}^{n} \langle \mathbf{B}_k, X_i \rangle (b_1'(\langle \beta, X_i \rangle) Y_i + b_2'(\langle \beta, X_i \rangle)) - \rho \mathbf{G}_k \boldsymbol{\theta}, \end{aligned} \tag{26}$$

where $\langle \mathbf{B}_k, X_i \rangle$ is the vector with generic element $\langle B_{kj}, X_i \rangle$ and $\mathbf{G}_k$ is the matrix with elements $[\mathbf{G}_k]_{lj} = \langle B_{kj}^{(m)}, B_{kl}^{(m)} \rangle$. The second derivative $\mathbf{H}_n(\boldsymbol{\theta})$ satisfies

$$\begin{aligned} \mathbf{H}_n(\boldsymbol{\theta}) &= \frac{\partial^2 \Lambda_{n,\rho}(\beta)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \\ &= \frac{1}{n}\sum_{i=1}^{n} \langle \mathbf{B}_k, X_i \rangle \langle \mathbf{B}_k, X_i \rangle' (b_1''(\langle \beta, X_i \rangle) Y_i + b_2''(\langle \beta, X_i \rangle)) - \rho \mathbf{G}_k. \end{aligned}$$

Let us define the operator Γ_β mapping H to H as

$$\begin{aligned} \Gamma_\beta x &= E((b_1''(\langle \beta, X \rangle) Y + b_2''(\langle \beta, X \rangle)) \langle X, x \rangle X) \\ &= E((b_1''(\langle \beta, X \rangle) b_3(\langle \alpha, X \rangle) + b_2''(\langle \beta, X \rangle)) X \otimes X(x)). \end{aligned} \tag{27}$$

Now, by (25) we can bound above $||\tilde{\alpha}_{\rm PS}^{(m)}||^2$ by a positive constant and we get that $||\beta^{(m)}||^2 = O(\tau/\rho) + O(1)$ is bounded. Appealing to the same arguments as those used for showing (19), one can show with assumption (H.3) that there exists some $\eta_0 > 0$ such that

$$P(|\langle \beta, X \rangle| \leqslant \eta_0) = 1. \tag{28}$$

Thus condition (H.1) implies by continuity arguments that there exist two strictly positive constants such that the following inequalities hold almost surely

$$-C_{12} \leqslant (b_1''(\langle \beta, X \rangle) b_3(\langle \alpha, X \rangle) + b_2''(\langle \beta, X \rangle)) \leqslant -C_{13},$$

and thus defining the matrices $\mathbf{C} = \langle \Gamma \mathbf{B}_k, \mathbf{B}_k \rangle$, with generic elements $[\mathbf{C}]_{lj} = \langle \Gamma B_{kl}, B_{kj} \rangle$, and $\mathbf{C}_\beta = \langle \Gamma_\beta \mathbf{B}_k, \mathbf{B}_k \rangle$, with generic elements $[\mathbf{C}_\beta]_{lj} = \langle \Gamma_\beta B_{kl}, B_{kj} \rangle$, one gets

$$-C_{12}\mathbf{u}'\mathbf{C}\mathbf{u} \leqslant \mathbf{u}'\mathbf{C}_\beta \mathbf{u} \leqslant -C_{13}\mathbf{u}'\mathbf{C}\mathbf{u} \quad \text{for } \mathbf{u} \in R^{q+k}. \tag{29}$$

By condition (H.2), the matrix $\mathbf{C}_\rho = \mathbf{C} + \rho \mathbf{G}_k$ is strictly positive and from Lemma 6.2 in [7] its smallest eigenvalue satisfies $\lambda_{\min}(\mathbf{C}_\rho) \geqslant C_{14}\rho k^{-1}$. Consequently, one easily obtains with (29) that

$$\lambda_{\max}(\mathbf{C}_\beta - \rho\mathbf{G}_k) \leqslant -C_{15}\rho k^{-1}, \tag{30}$$

where $\lambda_{\max}(\mathbf{A})$ stands for the largest eigenvalue of the symmetric matrix $\mathbf{A}$. From Theorem 1.19 in [9] one gets

$$\|\mathbf{C}_\beta - \rho\mathbf{G}_k - \mathbf{H}_n(\boldsymbol{\theta})\| \leqslant \sup_{1\leqslant l\leqslant q+k} \sum_{j=1}^{q+k} \|\Gamma_{n,\beta} - \Gamma_\beta\| \, |\langle B_{k,j}, B_{k,l}\rangle|,$$

where $\Gamma_{n,\beta}$ is the empirical version of Γ_β. Now using the corollary from [31, p. 491] one can deduce with the assumption on the conditional distribution of Y and (28) that

$$\|\Gamma_{n,\beta} - \Gamma_\beta\| = o_P(n^{(\delta-1)/2}).$$

For $|l-j| > q+1$, we have $B_{k,j}B_{k,l} \equiv 0$, then

$$\sup_{1\leqslant l\leqslant q+k} \sum_{j=1}^{q+k} |\langle B_{k,j}, B_{k,l}\rangle| = O(k^{-1}),$$

from which we get

$$\|\mathbf{C}_\beta - \rho\mathbf{G}_k - \mathbf{H}_n(\boldsymbol{\theta})\| = o_P\left(\frac{1}{kn^{(1-\delta)/2}}\right). \tag{31}$$

Appealing to Corollary 2.3 of Gohberg and Krein [17] we get that

$$|\lambda_{\max}(\mathbf{C}_\beta - \rho\mathbf{G}_k) - \lambda_{\max}(\mathbf{H}_n(\boldsymbol{\theta}))| = o_P(k^{-1}n^{(\delta-1)/2}). \tag{32}$$

Thus, from (30) we can deduce that

$$\lambda_{\max}(\mathbf{H}_n(\boldsymbol{\theta})) \leqslant -C_{15}\rho k^{-1} + o_P(k^{-1}n^{(\delta-1)/2}) \tag{33}$$

and, taking $\rho \sim n^{(\delta-1)/2}$, the strict concavity of the empirical log-likelihood on $B_n(\tau)$ except on an event whose probability tends to zero with n.

Now, let $\beta \in \partial B_n(\tau)$; we have

$$\Lambda_{n,\rho}(\beta) - \Lambda_{n,\rho}(\tilde{\alpha}_{\mathrm{PS}}) = (\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})' s_n(\tilde{\boldsymbol{\theta}}) + \frac{1}{2}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})' \mathbf{H}_n(\boldsymbol{\theta}_1)(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}),$$

where $\boldsymbol{\theta}_1 = t_1\boldsymbol{\theta} + (1-t_1)\tilde{\boldsymbol{\theta}}$ for some $t_1 \in [0,1]$. Decomposing $\beta(t) - \tilde{\alpha}_{\mathrm{PS}}(t)$ as in (17), one gets that $\|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|^2 = O(\tau k) + O(k\tau/\rho)$. Thus, using now (31), for $\beta = \beta_1$ where $\beta_1 = t_1\beta + (1-t_1)\tilde{\alpha}_{\mathrm{PS}}$, we can write

$$\Lambda_{n,\rho}(\beta) - \Lambda_{n,\rho}(\tilde{\alpha}_{\mathrm{PS}}) \leqslant |(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})' s_n(\tilde{\boldsymbol{\theta}})| + \frac{1}{2}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})'(\mathbf{C}_{\beta_1} - \rho\mathbf{G}_k)(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}) + o_P(\tau).$$

Then using (29) we get the inequality

$$\begin{aligned} &\Lambda_{n,\rho}(\beta) - \Lambda_{n,\rho}(\tilde{\alpha}_{\mathrm{PS}}) \\ &\quad \leqslant |(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})' s_n(\tilde{\boldsymbol{\theta}})| - \frac{1}{2}\min(C_{13}, 1)(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})'(\mathbf{C} + \rho\mathbf{G}_k)(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}) + o_P(\tau). \end{aligned}$$

Using now the fact that β belongs to $\partial B_n(\tau)$, we have that except on an event whose probability tends to zero with n

$$\Lambda_{n,\rho}(\beta) - \Lambda_{n,\rho}(\tilde{\alpha}_{\mathrm{PS}}) \leqslant |(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})' s_n(\tilde{\boldsymbol{\theta}})| - \tau C_{16}, \tag{34}$$

C_{16} being a strictly positive constant. By the Markov inequality we have

$$P(|(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})' s_n(\tilde{\boldsymbol{\theta}})| \leqslant \tau C_{16}) \geqslant 1 - \frac{E((\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})' s_n(\tilde{\boldsymbol{\theta}}))^2}{\tau^2 C_{16}^2}.$$

Noticing that by the definition of $\tilde{\alpha}_{\mathrm{PS}}$, $E[s_n(\tilde{\boldsymbol{\theta}})] = 0$, we obtain

$$\begin{aligned} E((\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})' s_n(\tilde{\boldsymbol{\theta}}))^2 = & E \frac{1}{n^2} \sum_{i=1}^{n} \psi(X_i)^2 (\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})' \langle \mathbf{B}_k, X_i \rangle \langle \mathbf{B}_k, X_i \rangle' (\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}) \\ & - \frac{\rho^2}{n} ((\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})' \mathbf{G}_k \tilde{\boldsymbol{\theta}})^2, \end{aligned}$$

where $\psi(X_i)^2 = (b_1'(\langle \tilde{\alpha}_{\mathrm{PS}}, X_i \rangle) Y_i + b_2'(\langle \tilde{\alpha}_{\mathrm{PS}}, X_i \rangle))^2$. Hypothesis on the conditional distribution of Y_i given X_i allows to write $E(\psi(X_i)^2 | X_i) \leqslant C_{17}$ and since $\beta \in \partial B_n(\tau)$, it is easy to check with the Schwarz inequality that

$$E((\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})' s_n(\tilde{\boldsymbol{\theta}}))^2 \leqslant \frac{C_{18}}{n} (||\beta - \tilde{\alpha}_{\mathrm{PS}}||_2^2 + \rho\tau),$$

where C_{17} and C_{18} are strictly positive constants. Then one has

$$E((\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})' s_n(\tilde{\boldsymbol{\theta}}))^2 \leqslant \tau \frac{C_{18}}{n} (1 + \rho)$$

and

$$P(|(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})' s_n(\tilde{\boldsymbol{\theta}})| \leqslant \tau C_{16}) \geqslant 1 - C_{19} \tau^{-1} \frac{1}{n}. \tag{35}$$

Inequalities (34) and (35) imply that, for every $\eta > 0$, one can find τ such that for n sufficiently large

$$P(\Lambda_{n,\rho}(\beta) < \Lambda_{n,\rho}(\tilde{\alpha}_{\mathrm{PS}}) \quad \text{for } \beta \in \partial B_n(\tau)) \geqslant 1 - \eta,$$

which implies that except on an event whose probability tends to zero with n, $\Lambda_{n,\rho}(\beta) < \Lambda_{n,\rho}(\tilde{\alpha}_{\mathrm{PS}})$ for $\beta \in \partial B_n(\tau)$. It follows with the strict concavity of $\Lambda_{n,\rho}$ on $B_n(\tau)$ that the spline estimator $\hat{\alpha}_{\mathrm{PS}}$ exists and is unique, except on an event whose probability tends to zero with n and, moreover, $\hat{\alpha}_{\mathrm{PS}}$ belongs to $B_n(\tau)$.

(ii) Write $\hat{\alpha}_{\mathrm{PS}} = \hat{\boldsymbol{\theta}}' \boldsymbol{B_k}$. By definition of $\hat{\alpha}_{\mathrm{PS}}$, $s_n(\hat{\boldsymbol{\theta}}) = 0$ and then a Taylor expansion of the score gives us

$$s_n(\tilde{\boldsymbol{\theta}}) = -\mathbf{H}_n(\boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}), \tag{36}$$

where $\boldsymbol{\theta}^* = t\hat{\boldsymbol{\theta}} + (1-t)\tilde{\boldsymbol{\theta}}$, for some $t \in [0,1]$. Since $\mathbf{H}_n$ is a strictly negative matrix except on an event whose probability tends to zero with n, one has equivalently

$$\mathbf{H}_n(\boldsymbol{\theta}^*)^{-1/2} s_n(\tilde{\boldsymbol{\theta}}) = -\mathbf{H}_n(\boldsymbol{\theta}^*)^{1/2}(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}). \tag{37}$$

Using inequalities (29) and (31) we obtain, since $\hat{\alpha}_{\mathrm{PS}}$ belongs to $B_n(\tau)$ except on an event whose probability tends to zero with n,

$$||\mathbf{H}_n(\boldsymbol{\theta}^*)^{1/2}(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}})||^2 \geqslant C_{20}(||\tilde{\alpha}_{\mathrm{PS}} - \hat{\alpha}_{\mathrm{PS}}||_2^2 + \rho||(\tilde{\alpha}_{\mathrm{PS}} - \hat{\alpha}_{\mathrm{PS}})^{(m)}||^2) + o_p(\tau) \qquad (38)$$

the constant C_{20} being strictly positive. On the other hand, we have

$$||\mathbf{H}_n(\boldsymbol{\theta}^*)^{-1/2}s_n(\tilde{\boldsymbol{\theta}})||^2 = |\mathrm{tr}(\mathbf{H}_n(\boldsymbol{\theta}^*)^{-1}s_n(\tilde{\boldsymbol{\theta}})s_n(\tilde{\boldsymbol{\theta}})')|. \qquad (39)$$

Before pursuing the calculus let us give some properties of the score vector $s_n(\tilde{\boldsymbol{\theta}})$. By definition, the score vector can be written as

$$s_n(\tilde{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1}^{n} s(X_i, Y_i, \tilde{\boldsymbol{\theta}})$$

where $s(X_i, Y_i, \tilde{\boldsymbol{\theta}}),\ i = 1, \dots, n$, are centered independent random variables. Furthermore, with hypothesis (H.3), the conditions on the distribution of Y and the fact that by (25) we can bound above $||\tilde{\alpha}_{\mathrm{PS}}^{(m)}||^2$ by a positive constant, the $s(X_i, Y_i, \tilde{\boldsymbol{\theta}})$ have finite moments and

$$E||s(X_i, Y_i, \tilde{\boldsymbol{\theta}})||^2 = O(k^{-1}) + O(\rho^2 k^{2m-1}).$$

The part due to the penalty term comes from the fact that by (25) we can bound above $||\tilde{\alpha}_{\mathrm{PS}}^{(m)}||^2 = \tilde{\boldsymbol{\theta}}'\mathbf{G}_k\tilde{\boldsymbol{\theta}}$ by a positive constant and $||\mathbf{G}_k|| = O(k^{2m-1})$ (see [4]). Then, since $\rho^2 k^{2m} = o(1)$, a normalization by nk and a direct calculus give us

$$nk||s_n(\tilde{\boldsymbol{\theta}})s_n(\tilde{\boldsymbol{\theta}})' - E(s_n(\tilde{\boldsymbol{\theta}})s_n(\tilde{\boldsymbol{\theta}})')|| = o_P\left(\frac{1}{n^{(1-\delta)/2}}\right) \qquad (40)$$

for $\delta > 0$, where

$$E(s_n(\tilde{\boldsymbol{\theta}})s_n(\tilde{\boldsymbol{\theta}})) = \frac{1}{n}E(\psi(X)^2\langle \mathbf{B}_k, X\rangle\langle \mathbf{B}_k, X\rangle') - \frac{\rho^2}{n}\mathbf{G}_k\tilde{\boldsymbol{\theta}}\tilde{\boldsymbol{\theta}}'\mathbf{G}_k. \qquad (41)$$

Expanding (39), we get using (40)

$$\begin{aligned} ||\mathbf{H}_n(\boldsymbol{\theta}^*)^{-1/2}s_n(\tilde{\boldsymbol{\theta}})||^2 = {} & |\mathrm{tr}(\mathbf{H}_n(\boldsymbol{\theta}^*)^{-1}\, E(s_n(\tilde{\boldsymbol{\theta}})s_n(\tilde{\boldsymbol{\theta}})'))| \\ & + |\mathrm{tr}(\mathbf{H}_n(\boldsymbol{\theta}^*)^{-1})|o_P\left(\frac{1}{nkn^{(1-\delta)/2}}\right). \end{aligned} \qquad (42)$$

With (30), (31) and $\rho \sim n^{(\delta-1)/2}$, one gets $||\mathbf{H}_n(\boldsymbol{\theta}^*)^{-1}|| = O_P(k/\rho)$, except on an event whose probability tends to zero as n tends to infinity. Thus $|\mathrm{tr}(\mathbf{H}_n(\boldsymbol{\theta}^*)^{-1})| = O_p(k^2/\rho)$ and

$$||\mathbf{H}_n(\boldsymbol{\theta}^*)^{-1/2}s_n(\tilde{\boldsymbol{\theta}})||^2 = |\mathrm{tr}(\mathbf{H}_n(\boldsymbol{\theta}^*)^{-1}\, E(s_n(\tilde{\boldsymbol{\theta}})s_n(\tilde{\boldsymbol{\theta}})'))| + o_P\left(\frac{k}{n}\right). \qquad (43)$$

Now

$$|\mathrm{tr}(\mathbf{H}_n(\boldsymbol{\theta}^*)^{-1}E(s_n(\tilde{\boldsymbol{\theta}})s_n(\tilde{\boldsymbol{\theta}})'))| \leqslant \frac{1}{n}|\mathrm{tr}(\mathbf{H}_n(\boldsymbol{\theta}^*)^{-1}E(\psi(X)^2\langle \mathbf{B}_k, X\rangle\langle \mathbf{B}_k, X\rangle'))| + \frac{\rho^2}{n}|\mathrm{tr}(\mathbf{H}_n(\boldsymbol{\theta}^*)^{-1}\mathbf{G}_k\tilde{\boldsymbol{\theta}}\tilde{\boldsymbol{\theta}}'\mathbf{G}_k)|. \tag{44}$$

On the one hand, the construction of $\mathbf{H}_n(\boldsymbol{\theta}^*)$ implies that $||\mathbf{H}_n(\boldsymbol{\theta}^*)^{-1}\ \rho\mathbf{G}_\mathbf{k}|| \leqslant 1$ and

$$\frac{\rho^2}{n}|\tilde{\boldsymbol{\theta}}'\mathbf{G}_\mathbf{k}\mathbf{H}_n(\boldsymbol{\theta}^*)^{-1}\mathbf{G}_\mathbf{k}\tilde{\boldsymbol{\theta}}| \leqslant C_{21}\frac{\rho}{n}\tilde{\boldsymbol{\theta}}'\mathbf{G}_\mathbf{k}\tilde{\boldsymbol{\theta}} \leqslant C_{22}\frac{\rho}{n}. \tag{45}$$

On the other hand, since $E(\psi(X)^2|X) \leqslant C_{17}$ we get directly

$$E(\psi(X)^2\langle \mathbf{B}_k, X\rangle\langle \mathbf{B}_k, X\rangle') \leqslant \frac{1}{n}C_{17}\mathbf{C}, \tag{46}$$

where inequalities between matrices are defined as in (29). Let us consider now $\mathbf{C}_{n,\beta^*}$ (resp. $\mathbf{C}_n$) the empirical version of $\mathbf{C}_{\beta^*}$ (resp. $\mathbf{C}$) where $\beta^* = \mathbf{B}_k\boldsymbol{\theta}^*$. From (29) and Yurinskii's Lemma we have

$$\begin{aligned}\mathbf{C} &\leqslant -\frac{1}{C_{13}}(\mathbf{C}_{n,\beta^*} + \mathbf{C}_{\beta^*} - \mathbf{C}_{n,\beta^*}) \\ &\leqslant -\frac{1}{C_{13}}\mathbf{C}_{n,\beta^*} + o_P\left(\frac{1}{kn^{(1-\delta)/2}}\right).\end{aligned} \tag{47}$$

By construction of $\mathbf{H}_n(\boldsymbol{\theta}^*) = \mathbf{C}_{n,\beta^*} - \rho\mathbf{G}_k$ and since it is a negative matrix we have $||\mathbf{H}_n(\boldsymbol{\theta}^*)^{-1}\mathbf{C}_{n,\beta^*}|| \leqslant 1$. Consequently, taking $\rho \sim n^{(\delta-1)/2}$ we have

$$\frac{C_{17}}{n}|\mathrm{tr}(\mathbf{H}_n(\boldsymbol{\theta}^*)^{-1}\mathbf{C})| = O_p\left(\frac{k}{n}\right) + o_P\left(\frac{k}{n}\right) \tag{48}$$

that completes the proof of point (ii) with (37), (38), (43), (44) and (45).

Acknowledgments

We thank Apratim Guha for a careful reading of a previous version of the manuscript that improved significantly some results and the participants in the working group "staph" in Toulouse for stimulating discussions.

References

[1] C. de Boor, A Practical Guide to Splines, Springer, New York, 1978.

[2] D. Bosq, Modelization, non-parametric estimation and prediction for continuous time processes, in: G. Roussas (Ed.), Nonparametric Functional Estimation and Related Topics, NATO, ASI Series, 1991, pp. 509–529.

[3] D. Bosq, Linear processes in function spaces, in: Lecture Notes in Statistics, Vol. 149, Springer, Berlin, 2000.

[4] H. Cardot, Spatially adaptive splines for statistical linear inverse problems, J. Multivariate Anal. 81 (2002) 100–119.

[5] H. Cardot, R. Faivre, M. Goulard, Functional approaches for predicting land use with the temporal evolution of coarse resolution remote sensing data, J. Appl. Statist. 30 (2003a) 1185–1199.

[6] H. Cardot, F. Ferraty, P. Sarda, Functional linear model, Statist. Probab. Lett. 45 (1999) 11–22.

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

[8] H. Cardot, F. Ferraty, A. Mas, P. Sarda, Testing hypotheses in the functional linear model, Scand. J. Statist. 30 (2003b) 241–255.

[9] F. Chatelin, Spectral Approximation of Linear Operators, Academic Press, New York, 1983.

[10] J. Dauxois, A. Pousse, Les analyses factorielles en calcul des probabilités et en statistique: Essai d'étude synthétique, Thèse, Université Paul sabatier, Toulouse, France.

[11] J. Dauxois, A. Pousse, Y. Romain, Asymptotic theory for the principal component analysis of a random vector function: some applications to statistical inference, J. Multivariate Anal. 12 (1982) 136–154.

[12] J.C. Deville, Méthodes statistiques et numériques de l'analyse harmonique, Ann. Insee 15 (1974) 1974.

[13] L. Fahrmeir, H. Kaufmann, Consistency and asymptotic normality of the maximum likelihood estimators in generalized linear models, Ann. Statist. 13 (1985) 342–368.

[14] F. Ferraty, P. Vieu, The functional nonparametric model and application to spectrometric data, Comput. Statist. 17 (2002) 545–564.

[15] I.E. Frank, J.H. Friedman, A statistical view of some chemometrics regression tools, Technometrics 35 (1993) 109–148.

[16] R.G. Ghanem, P.D. Sanos, Stochastic Finite Elements: A Spectral Approach, Springer, New York, 1991.

[17] I.C. Gohberg, M.G. Krein, Opérateurs Linéaires non Auto-adjoints Dans un Espace Hilbertien, Dunod, Paris, 1971.

[18] C. Goutis, Second-derivative functional regression with applications to near infra-red spectroscopy, J. Roy. Statist. Soc. B 60 (1998) 103–114.

[19] S. Leurgans, R. Moyeed, B. Silverman, Canonical correlation analysis when the data are curves, J. Roy. Statist. Soc. B 55 (1993) 725–740.

[20] B.D. Marx, P.H. Eilers, Generalized linear regression on sampled signals with penalized likelihood, in: A. Forcina, G.M. Marchetti, R. Hatzinger, G. Galmacci (Eds.), Statistical Modelling, Proceedings of the 11th International Workshop on Statistical Modelling, Orvietto, 1996.

[21] B.D. Marx, P.H. Eilers, Generalized linear regression on sampled signals and curves: a *P*-Spline approach, Technometrics 41 (1999) 1–13.

[22] P. McCullagh, J.A. Nelder, Generalized Linear Models, 2nd Edition, Chapman & Hall, London, 1989.

[23] J.A. Nelder, R.W.M. Wedderburn, Generalized linear models, J. Roy. Statist. Soc. A 135 (1972) 370–384.

[24] F. O'Sullivan, A statistical perspective on ill-posed inverse problems (with discussions), Statist. Sci. 4 (1986) 502–527.

[25] S. Portnoy, Asymptotic behaviour of likelihood methods for exponential families when the number of parameters tends to infinity, Ann. Statist. 16 (1988) 356–366.

[26] J.O. Ramsay, B.W. Silverman, Functional Data Analysis, Springer, Berlin, 1997.

[27] J.O. Ramsay, B.W. Silverman, Applied Functional Data Analysis: Methods and Case Studies, Springer, Berlin, 2002.

[28] C.J. Stone, Optimal global rates of convergence for nonparametric regression, Ann. Statist. 10 (1982) 1040–1053.

[29] C.J. Stone, Additive regression and other nonparametric models, Ann. Statist. 13 (1985) 689–705.

[30] C.J. Stone, The dimensionality reduction principle for generalized additive models, Ann. Statist. 14 (1986) 590–606.

[31] V.V. Yurinskiĭ, Exponential inequalities for sums of random vectors, J. Multivariate Anal. 6 (1976) 473–499.