DEPARTAMENTO DE MATEMÁTICA
DOCUMENTO DE TRABAJO

"Linear functional regression:
The case of fixed design and functional response"

Antonio Cuevas, Manuel Febrero y Ricardo Fraiman

D.T.: N° 23 Junio 2001

Vito Dumas 284, (B1644BID) Victoria, Pcia. de Buenos Aires, Argentina
Tel: 4725-7053 Fax: 4725-7010
Email: matematica@udesa.edu.ar

May, 2001

Linear functional regression: the case of fixed design and functional response

By ANTONIO CUEVAS

Departamento de Matemáticas, Universidad Autónoma, 28049-Madrid, Spain

MANUEL FEBRERO

Departamento de Estadística e Investigación Operativa, Universidad de Santiago de Compostela, 15706-Santiago de Compostela, Spain

and RICARDO FRAIMAN

Departamento de Matemática, Universidad de San Andrés, Buenos Aires, Argentina

Acknowledgements: We are very grateful to Kazaros Kazarian for some useful conversations. The useful and constructive comments of three referees are gratefully acknowledged. This work has been partially supported by DGESIC Spanish Grants PB97-0021 (A. Cuevas) and PB98-0182-C02-02, PGIDT99MA 20701 (M. Febrero).

Correspondence address: Antonio Cuevas, Departamento de Matemáticas, Facultad de Ciencias, Universidad Autónoma de Madrid, 28049-Madrid, Spain

e-mail: antonio.cuevas@uam.es

Linear functional regression: the case of fixed design and functional response

Summary

The problem of simple linear regression, with functional explanatory variables and functional response, is considered under a fixed design model. An estimator for the underlying linear operator is proposed. Its consistency is proved under some conditions which ensure that the design is informative enough. The classical calibration (or "inverse regression") problem is considered and a consistent estimator is analyzed. A simulation study is also given. The proposed method is computationally feasible and is not hard to implement in practice.

Some key words: Calibration; Consistent estimation; Functional data; Functional regression; Linear operators; Yurinskii's inequality.

1. Introduction

1.1 Linear functional regression: some background

The technological developments in information processing have made possible the real time monitoring of many processes in different fields including stock markets, audience ratings, meteorology, chemometrics or medicine. As a consequence, there is an increasing availability of functional data and several interesting applications and case studies have been published. In practice, the use of functional data is often preferable to that of large finite-dimensional vectors obtained by discrete approximations of the functions. The reason is that, whereas the smoothness properties can be easily handled in a functional framework, they are likely to cause collinearity problems which invalidate many standard multivariate (finite-dimensional) techniques.

The book by Ramsay and Silverman (1997) has contributed to the popularization of statistical methods with functional data by providing a broad survey of the subject, especially focused on applications and practical aspects.

For additional case-studies or theoretical developments see, e.g., Boente and Fraiman (2000), Bosq (1991), Brumback and Rice (1998), Dauxois, Pousse and Romain (1982),

Fan and Lin (1998), Ferraty and Vieu (1999), Hastie and Mallows (1993), Kneip (1994), Kneip and Gasser (1992), Locantore et al. (1999), Ramsay and Dalzell (1991), Rice and Silverman (1991) and references therein.

This paper is concerned with the problem of regression with functional data. While regression, in different setups (classical, nonparametric, with dependent data...), is a major issue, the corresponding theory for functional data seems to be still scarcely developed in several basic aspects. As we will indicate below, the present paper deals with some of these "elementary" theoretical issues in linear functional regression.

The work by Li (1984) is an early precedent in the study of consistency properties in infinite-dimensional regression. This author considers the following model

$$y = \sum_{i=1}^{\infty} x_i \theta_i + \epsilon, \tag{1}$$

where both the explanatory (or independent) variables $\mathbf{x} = (x_1, x_2, \ldots)^t$ and the regression parameter $\theta = (\theta_1, \theta_2, \ldots)^t$ are assumed to belong to the space ℓ^2 of square-sumable sequences. The aim is to estimate bounded real functionals $T(\theta)$.

Cardot, Ferraty and Sarda (1999) provide consistency results for the case of linear regression (LR) with functional explanatory variables (FE) and scalar response (SR) in the case of random design (RD). More specifically, the explanatory variables are random observations of a continuous-time process with trajectories in $L^2[0, 1]$. This assumption allows the authors to place the problem in the rich mathematical framework of linear functionals on Hilbert spaces. Observe that the spaces L^2 and ℓ^2 are isometrical so that in this respect the model (1) is not very different from that considered by Cardot Ferraty and Sarda (1999). It should be noted however that, unlike these authors, Li (1984) focuses on the estimation of linear real functionals $T(\theta)$ (instead of that the whole parameter θ) assuming a fixed design (instead of a random one).

The model LR/FE/SR/RD is also considered in Ferraty and Vieu (1999), although the mathematical structure for the explanatory variables is that of semi-normed vector spaces instead of the Hilbert space L^2.

The case of functional response (FR) is analyzed (and a real data example is discussed in detail) by Ramsay and Silverman (1997, chap. 11). Their approach is oriented to the practical and computational issues rather than to the asymptotic properties.

1.2 The purpose of this paper

We analyze the case of functional response and fixed design. Thus, in the above notation, our model is LR/FE/FR/FD. This is the functional analogue of the classical elementary simple linear regression model. As far as we know, the basic problem of providing a general method to construct consistent estimates in this model has not been previously considered in the literature. Section 3 below is devoted to this issue.

The fixed design approach is meaningful in many practical situations where there is a controlled input $x(t)$ (a signal message, the concentration of a catalyst in a chemical reaction,...) which produces a random output $y(t)$ affected by distortion and/or random noise. In particular, chemometrics offers a huge potential for applicability of functional regression tools under fixed design. As Frank and Friedman (1993) point out in their survey paper on chemometrics regression tools (we have capitalized the last sentence): *(...) There is generally a high degree of collinearity among the variables, WHICH ARE OFTEN (but not always) DIGITALIZATIONS OF ANALOG SIGNALS.*

Thus in these problems there is a hidden underlying functional structure which accounts for the high collinearity in the data. Given that collinearity in chemometrics typically arises in problems involving multiple fixed design regression models, one might conclude that the development of functional fixed design regression tools would provide alternative models for chemometrics problems, based on a functional (rather than a finite-dimensional) point of view.

As we will see, the conditions on the design are crucial here. They are discussed in Section 2 below. The point is that in a functional (infinite-dimensional) framework there are many different ways in which a design can be "degenerate" in the sense of not providing enough information to estimate the linear operator underlying in the regression model. Finally, Section 4 deals with the classical problem of calibration in the setup of a simple functional regression model. Although the functional framework entails some difficulties for the statement and solution of this problem, we will show that a consistent estimator can be obtained under some conditions.

2. Statement of the problem. Basic assumptions

2.1 The linear model

We consider a fixed design model. Hence, our data are $(x_{in}(s), Y_{in}(t))$, $x_{in} \in L^2[a,b]$,

$Y_{in} \in L^2[c,d]$ and the design is given by the triangular array

$$\{x_{in}, 1 \leq i \leq n\}_n,$$

We will sometimes omit the subscript n in x_{in}, Y_{in}.

The assumed model is

$$Y_{in}(t) = (Tx_{in})(t) + e_{in}(t), \tag{2}$$

$T : L^2[a,b] \longrightarrow L^2[c,d]$ being a linear operator given by

$$(Tx)(t) = \int_a^b x(s)\beta(s,t)ds, \tag{3}$$

where the "kernel function" $\beta(s,t)$ satisfies

$$\int_a^b \int_c^d \beta(s,t)^2 ds dt < \infty, \tag{4}$$

and the "error functions" e_{in} are independent processes satisfying

(E1) For $i = 1, \ldots, n$ we have $E(e_{in}(t)) = 0$, $E\|e_{in}\|^2 := b^2 < \infty$ and there exists $A > 0$ such that

$$E\|e_{in}\|^p \leq \frac{p!}{2} b^2 A^{p-2}, \text{ for all } p \geq 2. \tag{5}$$

Observe that the "error trajectories" e_{in} need not be identically distributed; just the values $E(e_{in}(t))$ and $E\|e_{in}\|^2$ are required to be constant. Note also that assumption (5) is fulfilled when e_{in} are the trajectories of a Gaussian process.

We wish to estimate the linear operator T, that is, we have to define appropriate estimators $T_n(x_{1n}, \ldots, x_{nn}; Y_{1n}, \ldots, Y_{nn})$ which must converge to T in the norm of the operator's space,

$$\|T\| = \sup_{\|x\|=1} \frac{\|Tx\|}{\|x\|}.$$

Assumption (4) ensures (see, e.g., Conway (1990), p. 41) that T is a compact operator with

$$\|T\| \leq \int_a^b \int_c^d \beta(s,t)^2 dt ds.$$

2.2 The design

The problem of design choice is much more crucial here than in the ordinary finite--dimensional regression. The point is that in the functional setup we must estimate the projection of the underlying linear operator T on an infinite number of orthogonal directions. Hence, the design must achieve a trade-off between two partially opposite targets: on the one hand, as the sample size n grows, the design should "fill the space" in the sense that an increasing number of orthogonal directions should be represented in $\{x_{1n}, \ldots, x_{nn}\}$ ("completeness" condition). On the other hand, the design functions x_{in} should have, often enough as n grows, a "significant" projection on every particular direction in the space $L^2[a, b]$ so that the estimator T_n can "learn" about the behavior of T in such direction ("learning condition"). These ideas are more formally expressed in the following conditions.

We will assume that, for every n, the set of indexes $\{1, \ldots, n\}$ can be partitioned into $m = m_n$ subsets $J_{1n}, \ldots, J_{mn}$, with $\mathrm{card}(J_{in}) = k_{in}$, and $k_n = \min_i k_{in} \to \infty$, $m = m_n \to \infty$, such that for $i = 1, \ldots, m$ the averages

$$\bar{x}_i := \bar{x}_{in} = \frac{1}{k_{in}} \sum_{j \in J_{in}} x_{jn},$$

are linearly independent and the closed linear spans, $\mathcal{H}_n = \mathcal{L}\{\bar{x}_{1n}, \ldots, \bar{x}_{mn}\}$, fulfil

(D1) $\mathcal{H}_n \subset \mathcal{H}_{n+1}$, $\mathcal{L}(\cup \mathcal{H}_n) = L^2[a, b]$.

(D2) The smallest eigenvalue of the matrix with elements $\langle \bar{x}_{in}, \bar{x}_{jn} \rangle$, $1 \leq i, j \leq m$, is bounded below by a constant $a_0 > 0$ for all n.

It is easily seen that this condition is equivalent to the following one

(D2a) There exists an orthonormal basis of $\mathcal{H}_n$, denoted by $\{\varphi_{1n}, \ldots, \varphi_{mn}\}$ such that the matrix $B_n = (\beta_{jk}^{(n)})$ associated with the change of basis

$$\bar{\varphi}_{jn} = \sum_k \beta_{jk}^{(n)} \bar{x}_{kn},$$

satisfies

$$\|B_n\| \leq a_0, \tag{6}$$

where a_0 is a constant not depending on n.

Assumption (D1) can be interpreted as the formal expression of the "completeness condition" roughly defined at the beginning of this section. In particular, assumption

$\mathcal{L}(\cup\mathcal{H}_n) = L^2[a,b]$ is clearly necessary in order to consistently estimate the linear operator T. If this condition does not hold we only can estimate the restriction of T to a proper subspace of $L^2[a,b]$.

On the other hand, (D2) is the "learning condition" above mentioned. In intuitive terms, the assumption that the spaces $\mathcal{H}_n$ are generated by the averages $\bar{x}_i$ entails a redundancy in the design: every orthogonal direction in the space $L^2[a,b]$ must be "repeated" infinitely often if we want to estimate the restriction of T in such direction. On the other hand, (6) establishes that the sets $\{x_{1n},\ldots,x_{mn}\}$ are not very far from orthogonal systems of functions with norms bounded from below. In particular, if $\{x_{1n},\ldots,x_{mn}\}$ is orthonormal, then (6) holds with $a_0 = 1$ and $B_n = I_n$, the identity matrix. We can move away from this situation in two different ways: either the norms of the averages $\bar{x}_{in}$ can be very small or some of the $\bar{x}_{in}$ can be nearly collinear. In both cases condition (D2) fails. In geometrical terms, (D2) is satisfied when the design functions are concentrated on narrow cones whose edges are close to those of an orthogonal system. Then this property will be also inherited by the averages $\bar{x}_{in}$.

If $\{\bar{x}_1,\ldots,\bar{x}_m\}$ are orthogonal (D2) reduces to

(D2b) There exists a constant a_0, not depending on n such that

$$\min_{1\le i\le m}\|\bar{x}_{in}\| \ge a_0, \text{ for all } n \ge 1.$$

Note that, since the averages in $\bar{x}_{in}$ ensure the reduction of variance in the estimation of the responses in every direction, condition (D2b) essentially coincides with the classical consistency condition that the design points do not concentrate at zero, imposed in the simple linear regression model (of course in this case there is just one direction to be considered).

Observe also that the norm $\|\bar{x}_{in}\|$ of the average $\bar{x}_{in} = (1/k_i)\sum_{j\in J_i} x_{jn}$ is larger when we choose the x_{jn}, for $j \in J_i$, to be nearly collinear. This can be interpreted again as a "learning condition"; some directions in the space $L^2[a,b]$ are "repeated" infinitely often in order to estimate the restriction of T in such direction.

A simple design satisfying (D1) and (D2) can be defined as follows: let $\{v_1, v_2, \ldots\}$ be a basis of $L^2[a,b]$. Define $x_{1n} = v_1, x_{2n} = v_1, \ldots, x_{kn} = v_1, x_{(k+1)n} = v_2, x_{(k+2)n} = v_2, \ldots, x_{2k,n} = v_2, \ldots$, where $n = km$ is a multiple of m and $k = k_n \to \infty$, $m = m_n \to \infty$. Therefore, we can take $k_i = k$ and $\bar{x}_{in} = v_i$ where the norms $\|v_i\|$ are bounded away from zero. In what follows, this design will be called "design of repeated observations".

3. A simple estimator: strong consistency

With the notation of the above section, define

$$\bar{Y}_{in} := \bar{Y}_i = \frac{1}{k_{in}} \sum_{j \in J_{in}} Y_{jn},\ \bar{e}_{in} := \bar{e}_i = \frac{1}{k_{in}} \sum_{j \in J_{in}} e_{jn}.$$

Observe that, under the assumed model (2), $\bar{Y}_i(t) = (T\bar{x}_i)(t) + \bar{e}_i(t)$.

We propose an estimator $\hat{T}_n$ of T defined as the linear operator which interpolates the pairs $(\bar{x}_i, \bar{Y}_i)$, $i = 1, \ldots, m_n$ and such that the restriction of $\hat{T}_n$ on $\mathcal{H}_n^{\perp}$ (the closed subspace orthogonal to $\mathcal{H}_n$) is 0.

An explicit construction of $\hat{T}_n$ can be done as follows. Since $\bar{x}_1, \ldots, \bar{x}_m$ are linearly independent, we can choose a conjugate system of functions $x_1^*, \ldots, x_m^*$ such that

$$\langle x_r^*, \bar{x}_i \rangle = \delta_{ir}. \tag{7}$$

Then $\hat{T}_n$ is the linear operator associated with the kernel $\hat{\beta}_n(s,t) = \sum_{i=1}^m x_i^*(s)\bar{Y}_i(t)$, that is,

$$(\hat{T}_n x)(t) = \int_a^b \hat{\beta}_n(s,t)x(s)ds. \tag{8}$$

Remark 1.- Note that when the $\bar{x}_i$ are orthogonal, the computation of the conjugate functions x_i^* becomes trivial since $x_i^* = \bar{x}_i/\|\bar{x}_i\|$. The general case can be also easily handled: in practice, we can evaluate the function $\hat{\beta}(s,t)$ on a equally-spaced square $m \times m$ grid,

$$\left\{ \left(a + \frac{(b-a)j}{m}, c + \frac{(d-c)l}{m} \right) :\ j, l = 1, \ldots, m \right\}$$

just by solving a discrete version of the linear system (7) where the inner product $\langle x_r^*, \bar{x}_i \rangle$ is replaced by the scalar product on $\Re^m$ of the vectors obtained evaluating the functions x_r^* and $\bar{x}_i$ in the grid points $c + \frac{(d-c)l}{m}$ and $a + \frac{(b-a)j}{m}$, respectively.

The strong consistency of the estimator (8) is next established.

Theorem 1.- *Let us consider the model defined by (2), (3) and (4). Assume (E1) together with*

(i) The design conditions (D1) and (D2) hold.

(ii)

$$\frac{\min_{1 \le i \le m} k_i}{m_n \log m_n} \to \infty,\ \text{as } n \to \infty$$

.

Then, $\|\hat{T}_n - T\| \to 0$ *completely as* $n \to \infty$, that is,

$$\sum_{n=1}^{\infty} P\left\{\|\hat{T}_n - T\| > \eta\right\} < \infty, \text{ for all } \eta > 0.$$

Proof: Denote $m := m_n$, $\bar{x}_{in} := \bar{x}_i$. Let us consider an (unobservable) auxiliary operator $T_n : L^2[a,b] \to L^2[c,d]$, such that,

$$T_n(\bar{x}_i) = T(\bar{x}_i) := \tilde{Y}_i, \text{for all } i = 1, \ldots, m.$$

This defines T_n on the subspace $\mathcal{H}_n = \mathcal{L}\{\bar{x}_{1n}, \ldots, \bar{x}_{mn}\}$. For all $z \in \mathcal{H}_n^{\perp}$ we put $T_n(z) = 0$. We have

$$\|\hat{T}_n - T\| \leq \|\hat{T}_n - T_n\| + \|T_n - T\|. \tag{9}$$

Since T is a compact operator between Hilbert spaces we have, from assumption (D1)

$$\|T_n - T\| \to 0, \ n \to \infty. \tag{10}$$

This is a consequence of the well-known property of approximation of a compact T by finite-range operators (see, e.g., Theorem 4.4 in Conway (1990), p. 41).

In order to prove the convergence of the stochastic term in (9), let us observe that if we denote

$$u_1 = \sum_{i=1}^{m} \lambda_i \bar{x}_i \in \mathcal{H}_n$$

$$\|\hat{T}_n - T_n\| = \sup_{\|u_1\| \leq 1, u_1 \in \mathcal{H}_n} \|(\hat{T}_n - T_n)(u_1)\| \leq \sup_{\|u_1\| \leq 1, u_1 \in \mathcal{H}_n} \sum_{i=1}^{m} |\lambda_i| \|\bar{Y}_i - \tilde{Y}_i\| \leq$$

$$\max_{1 \leq i \leq m} \|\bar{Y}_i - \tilde{Y}_i\| \sup_{\|u_1\| \leq 1, u_1 \in \mathcal{H}_n} \sum_{i=1}^{m} |\lambda_i| \leq \sqrt{m} \max_{1 \leq i \leq m} \|\bar{Y}_i - \tilde{Y}_i\| \sup_{\|u_1\| \leq 1, u_1 \in \mathcal{H}_n} \left(\sum_{i=1}^{m} \lambda_i^2\right)^{1/2} \leq$$

$$\sqrt{m} \max_{1 \leq i \leq m} \|\bar{Y}_i - \tilde{Y}_i\| a_0. \tag{11}$$

The third inequality in this expression is based on the fact that, from Cauchy-Schwartz inequality, $\sum_i |\lambda_i| \leq \sqrt{m} \left(\sum_i \lambda_i^2\right)^{1/2}$.

To check the last inequality observe that, if $\{\varphi_i\}$ is the "canonical" basis mentioned in (D2a) and $u_1 = \sum_{i=1}^{m} \alpha_i \varphi_i$, we necessarily have $\sum_{i=1}^{m} \alpha_i^2 \leq 1$ (as $\|u_1\| \leq 1$) and $\sum_{i=1}^{m} \lambda_i^2 \leq \|B\alpha\|^2$ (since $\lambda = B\alpha$).

Since the second term in the right side of (9) is non-stochastic, the conclusion of the theorem will follow, from (10), from the complete convergence $\|\hat{T}_n - T_n\| \to 0$. We now establish this convergence.

Given $\eta > 0$ define $\eta_1 = \eta/a_0$. Then, from (11),

$$P\{\|\hat{T}_n - T_n\| > \eta\} \leq P\{\sqrt{m} \max_{1\leq i\leq m} \|\bar{Y}_i - \tilde{Y}_i\| > \eta_1\} =$$
$$1 - \prod_{i=1}^{m} P\left\{\|\bar{Y}_i - \tilde{Y}_i\| \leq \frac{\eta_1}{\sqrt{m}}\right\} = 1 - \prod_{i=1}^{m} P\left\{\|\frac{1}{k_i}\sum_{j\in J_i} e_j\| \leq \frac{\eta_1}{\sqrt{m}}\right\}. \qquad (12)$$

By using the Bernstein-type inequality for random variables in Hilbert spaces due to Yurinskii (1976), we get that, for m larger than some m_0 which does not depend on i,

$$P\left\{\|\sum_{j\in J_i} e_j\| > \frac{k_i\eta_1}{\sqrt{m}}\right\} \leq 2\exp\left(\frac{-c_1 k_i}{2m(1+\frac{1.62c_1A}{\sqrt{m}b})}\right) \leq 2\exp\left(\frac{-c_2 k_i}{m}\right), \qquad (13)$$

where c_1 and c_2 are appropriate constants and A and b are defined in (E1).

Then, from (12) and (13),

$$\prod_{i=1}^{m} P\left\{\|\bar{Y}_i - \tilde{Y}_i\| \leq \frac{\eta_1}{\sqrt{m}}\right\} \geq \left(1 - 2\exp\left(\frac{-c_2 k_n}{m}\right)\right)^m. \qquad (14)$$

This bound has the same order as

$$\exp\left(-m\exp\left(-\frac{c_2 k_n}{m}\right)\right),$$

where $k_n = \min_{1\leq i\leq m} k_i$. Therefore, to establish the complete convergence, it suffices to prove

$$\sum_{n=1}^{\infty}\left(1 - \exp\left(-m\exp\left(-\frac{c_2 k_n}{m}\right)\right)\right) < \infty. \qquad (15)$$

As a consequence of the inequality $1 - t \leq \exp(-t)$, a sufficient condition for (15) is

$$\sum_{n=1}^{\infty} m\exp\left(-\frac{c_2 k_n}{m}\right) < \infty,$$

and this follows from assumption (ii) in the theorem.

Remark 2.- A sufficient condition for the validity of assumption (ii) in the theorem is $\min_{1\leq i\leq n} k_i = Cn^{\alpha}$, where C and α are constants with $\alpha > 1/2$.

Remark 3.- The above proof can be easily adapted to obtain convergence rates for the "stochastic term" $\|\hat{T}_n - T_n\|$. If assumption (ii) is replaced by

$$\frac{\min_{1\leq i\leq m} k_i}{\beta_n m_n \log m_n} \to \infty,$$

as $n \to \infty$, then one can show that $\beta_n \|\hat{T}_n - T_n\| \longrightarrow 0$, as $n \to \infty$. Let us observe, however, that this does not provide a convergence rate for $\|\hat{T}_n - T\|$ since the convergence to zero of the "bias term" $\|T_n - T\|$ may be arbitrarily slow, even for the best choice of the subspaces $\mathcal{H}_n$, unless additional conditions be imposed on the target operator T.

Remark 4.- As Li (1984) points out "...in the case that Θ [the parameter space] is infinite-dimensional, *it is typical that the least squares estimate no longer works"*. It can be easily seen, however, that when the design of "repeated observations" is used, our estimator coincides with that of least squares obtained by minimizing the sum of residuals. So, in this particular case, least squares provides consistency. However, for this particular design we are able to decrease variance by averaging in the same controlled directions. Under other very simple designs (for example, when $x_i = e_i$, the elements of an orthogonal basis), the least squares estimator will indeed lead to overfitting and inconsistency.

4. Calibration

4.1 Statement of the problem. A proposal for the functional setup

Calibration, also called "inverse regression" is a classical problem which appears often in the regression setup under fixed design. A review of the literature on this subject can be found in Osborne (1991). In our functional regression setup the problem can be stated as follows.

At a first stage, n pairs of functional data (x_i, Y_i), $i = 1, \ldots, n$, are observed. It is assumed that these data satisfy the functional linear model defined by the above expressions (2) and (3) with $[a, b] = [c, d]$. A consistent estimator $\hat{T}_n$ of the underlying linear operator T is obtained from these data.

In a second step N replications, $Y_{n+1}, \ldots, Y_{n+N}$ of the response variable Y, corresponding to the same value x_0 of the explanatory variable, are observed. That is,

$$Y_{n+j} = Tx_0 + e_j, \ j = 1, \ldots, N.$$

The final aim is estimating x_0. For instance, the functions $x = x(t)$ could correspond to the exact measurement of some process (as given by an exact, expensive method) and the response functions $Y = Y(t)$ could be the respective approximate measurements obtained with a less expensive procedure. This type of situation arises often in the experimental sciences and, in particular, is typical in chemometrics; see, e.g., the book by Martens and Naes (1989) where chemical applications appear as a major motivation for calibration methods. A number of references dealing with interesting case studies of the calibration methodology can be found in Gruet (1996).

In order that the calibration problem be well-defined, the operator T should be injective. A simple condition for this is (see, e.g., Riesz and Nagy (1955), p. 234)

(C1) The operator T is compact and symmetric and the sequence of normalized eigenfunctions φ_i of T corresponding to the eigenvalues λ_i different from zero form an orthonormal basis in $L^2[a, b]$.

Note that if T is compact it can not be invertible as a linear operator, that is, there is no linear and continuous inverse transformation T^{-1}. However, we only need $Tx \neq 0$ for $x \neq 0$ and this is guaranteed by (C1).

A natural proposal would be estimating x_0 by $\hat{T}_n^{-1}(\bar{Y}_N)$, where $\bar{Y}_N = \sum_{j=1}^{N} Y_{n+j}/N$. However, this direct approach presents an obvious shortcoming: the inverse transformation $\hat{T}_n^{-1}$ need not to exist. This is the case, in particular, of the estimator $\hat{T}_n$ defined in Section 4, which is of finite rank and hence non-injective. A simple way to overcome this problem is to perturb the operator $\hat{T}_n$ in the form $\hat{T}_n + \lambda I$, where I is the identity operator and λ is a real number tending to zero as $n, N \to \infty$. Then, provided that the inverse transformation $(\hat{T}_n + \lambda I)^{-1}$ does exist, we could use

$$(\hat{T}_n + \lambda I)^{-1}(\bar{Y}_N), \tag{16}$$

as an estimator of x_0.

These ideas are developed in the following subsection.

4.2 A consistency result

We next establish the almost sure consistency of the estimator defined in (16). In particular, we will show that the operator $\hat{T}_n + \lambda I$ is indeed invertible (in the sense of having a linear and continuous inverse) for large enough n.

THEOREM 2.- *Let us consider the model defined by (2) and (3), with* $[a, b] = [c, d]$. *Assume that (C1) holds and that* $\langle Tx, x\rangle \geq 0$ *for all* x. *Let* $\hat{T}_n$ *be a strongly consistent*

estimator of T (that is, $\|\hat{T}_n - T\| \to 0$, a.s.). Assume further that

(C2) $\lambda = \lambda(n, N)$ is a sequence of parameters such that, as $n, N \to \infty$,

$$\lambda \to 0, \quad \frac{\log N}{\lambda\sqrt{N}} \to 0 \quad \text{and} \quad \frac{\|\hat{T}_n - T\|}{\lambda^2} \longrightarrow 0, \ a.s.$$

(C3) The "error trajectories" $e_{n+1}, \ldots, e_{n+N}$ are independent realizations of a process satisfying hypothesis (E1).

Then

$$\|(\hat{T}_n + \lambda I)^{-1}\bar{Y}_N - x_0\| \longrightarrow 0, \ a.s. \tag{17}$$

Proof: We first prove that, for n large enough, the operator $\hat{T}_n + \lambda I$ is invertible. We use the following result (see Nashed (1976), Proposition 3.3):

Assume that A, $B \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ are linear operators between two Banach spaces $\mathcal{X}, \mathcal{Y}$. If A is invertible and $\Delta := \|A^{-1}\|\|A - B\| < 1$, then B is also invertible and

$$\|B^{-1}\| \leq \frac{\|A^{-1}\|}{1-\Delta}, \ \|B^{-1} - A^{-1}\| \leq \frac{\Delta\|A^{-1}\|}{1-\Delta}. \tag{18}$$

We use this result with $A = T + \lambda I$ and $B = \hat{T}_n + \lambda I$. The operator $T + \lambda I$ is indeed invertible and its inverse is defined by

$$(T + \lambda I)^{-1}(\varphi_i) = \frac{1}{\lambda + \lambda_i}\varphi_i.$$

In particular, $\|(T + \lambda I)^{-1}\| \leq \frac{1}{\lambda}$, since the definiteness assumption $\langle Tx, x\rangle \geq 0$ ensures that $\lambda_i > 0$ (see, e.g., Riesz and Nagy (1955), p. 237). Thus

$$\Delta := \|(T + \lambda I)^{-1}\|\|\hat{T}_n - T\| \leq \frac{1}{\lambda}\|\hat{T}_n - T\|$$

and, from (C2), the right-hand side is smaller than (say) 1/2 eventually almost surely. Hence, from (18), the operator $\hat{T}_n + \lambda I$ is invertible and

$$\|(\hat{T}_n + \lambda I)^{-1} - (T + \lambda I)^{-1}\| \leq \frac{\Delta}{\lambda(1-\Delta)} \leq \frac{1}{\lambda^2}\frac{\|\hat{T}_n - T\|}{1-\Delta} \to 0 \,\text{a.s.}, (\text{by (C2)}) \tag{19}$$

Now observe that

$$(\hat{T}_n + \lambda I)^{-1}\bar{Y}_N = (\hat{T}_n + \lambda I)^{-1}(Tx_0 + \bar{e}_N) = (\hat{T}_n + \lambda I)^{-1}(Tx_0) + (\hat{T}_n + \lambda I)^{-1}\bar{e}_N. \tag{20}$$

Then, in view of (19), the conclusion (17) of the theorem will follow from both

$$\|(T+\lambda I)^{-1}(Tx_0)-x_0\| \to 0, \text{ as } n, N \to \infty \tag{21}$$

and

$$\|(T+\lambda I)^{-1}\bar{e}_N\| \to 0, \quad \text{a.s., as } n, N \to \infty. \tag{22}$$

If $x_0 = \sum_i \alpha_i \varphi_i$ then, from (C1), $Tx_0 = \sum_i \alpha_i \lambda_i \varphi_i$ and

$$(T+\lambda I)^{-1}(Tx_0)-x_0 = \sum_i \frac{\alpha_i \lambda_i}{\lambda_i+\lambda}\varphi_i - \sum_i \alpha_i \varphi_i = \sum_i \alpha_i \left(\frac{\lambda_i}{\lambda_i+\lambda}-1\right)\varphi_i$$

Thus,

$$\|(T+\lambda I)^{-1}(Tx_0)-x_0\|^2 = \sum_i \alpha_i^2 \left(\frac{\lambda_i}{\lambda_i+\lambda}-1\right)^2$$

and, since $\sum_i \alpha_i^2 < \infty$, the dominated convergence theorem (recall $\lambda_i > 0$) entails that the right-hand side tends to zero as $\lambda \to 0$. Hence (21) holds.

Finally, (22) follows from Yurinskii's (1976) inequality, since $\frac{\log N}{\lambda\sqrt{N}} \to 0$ as $n, N \to \infty$. So the proof is complete.

Remark 5.- The use of the parameter λ here is reminiscent of some methods, called "inverse iteration" and "successive over-relaxation" in numerical analysis (see, e.g., Golub and Van Loan (1983), pages 238 and 357). However, a closer analogy can be established with the ridge regression methodology, first proposed by Hoerl and Kennard (1970) (see also Frank and Friedman 1993), where a "ridge parameter" (often denoted by λ) is introduced playing a very similar role to that of λ in our approach. In both cases the basic idea is to avoid a problem of non-invertibility. The effective choice of the ridge parameter λ is known to be a non trivial problem. Golub, Heath and Wahba (1979) consider λ as a "meta parameter" of the procedure and apply a cross-validation methodology to select it.

Although a complete study of the best choice for the parameter λ in our calibration procedure is beyond the scope of this paper, Theorem 2 provides some guide in this respect. Remember that the role of λ is to make invertible $T_n + \lambda I$. According to the proof of Theorem 2, this is achieved when $\|\hat{T}_n - T\| < \lambda$. An evaluation of $\|\hat{T}_n - T\|$ can be obtained by a leave-one-out device. However, assumption (C2) in Theorem 2

suggests that a more appropriate choice of λ could be

$$\lambda > \max\left\{\frac{\log N}{\sqrt{N}}, \|\hat{T}_n - T\|, \sqrt{\|\hat{T}_n - T\|}\right\}.$$

5. A simulation study

We have carried out a simulation study aimed at two basic targets. First, we want to show that the estimator can be effectively implemented (along the lines suggested in Remark 1). Second, we wish to check its efficiency for moderate sample sizes.

The details of the study are as follows.

- We use the functional linear model (2), where $[a, b] = [c, d] = [0, 2\pi]$ and the target operator (3), $T : L^2[0, 2\pi] \to L^2[0, 2\pi]$ is defined by the kernel $\beta(s, t) = \log(1 + s)\log(1 + t)$.
- We use a design of repeated observations, defined in subsection 2.2. The design functions are

$$\begin{aligned} x_1(s) &= I_{[0,2\pi]}(s), \\ x_{2l}(s) &= 1 + 10\sin(ls), \\ x_{2l+1}(s) &= 1 + 10\cos(ls), \end{aligned}$$

  for $l = 1, \ldots, L$. We have two different values for L: 7 and 12. In the notation of subsection 2.2, this corresponds to take $m = 2L + 1 = 15, 25$ functions in the design.

- The parameter k (the number of appearances of every design function) has been taken equal to m. Hence, the considered sample sizes are $n = 15^2, 25^2 = 125, 625$.
- As for the errors e_{in} in (2) we have considered two cases:
  (a) Independent normal errors $N(0, \sigma)$.
  (b) A Brownian motion $B(t) = \sigma W(t)$, where W is the standard Wiener process on $[0, 2\pi]$, evaluated also at 15, 25 equally spaced points, respectively.

  In both cases we have taken 10 equally spaced different values for the parameter σ, from 0.1 to 1.

Observe that the case (a) correspond to a particularly "wild" situation in which the error process $e(t)$ is an approximation to a white noise. Of course, as the white noise is not realizable in continuous time, it is not a suitable model for the errors. However, for finite sample sizes it is interesting to see the behavior of the estimator under approximate versions of this process.

- One hundred independent replications have been performed for each of the $2 \times 10 \times 2$=40 different cases.
- The effective computation of the estimated kernel $\hat{\beta}(s,t)$ relies on the ideas described in Remark 1.

A typical realization (for just one run), with Brownian error and $\sigma = 1$, is illustrated in Figures 1 to 5: Figure 1 is a plot of the design x's functions (with $m = 25$), Figure 2 are the response variables, Figure 3 is the true $\beta(s,t)$ kernel surface, Figure 4 is the estimated kernel function $\hat{\beta}(s,t)$; finally, in Figure 5 we plot the differences between $\hat{\beta}(s,t)$ and $\beta(s,t)$.

INSERT FIGURES 1-5 AROUND HERE

The simulation results are reported in Table 1 (independent normal residuals; see case (a) above) and Table 2 (Brownian residuals; case (b)) where the rows $n = 125$ and $n = 625$, correspond to the averages (over the 100 runs) of the respective L^2-distance between the estimated kernel $\hat{\beta}(s,t)$ and the true one $\beta(s,t)$ for the values of the parameter σ in the error variables e_{in}.

σ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
$n = 125$	0.0317	0.0633	0.0966	0.1275	0.1599	0.1918	0.2263	0.2579	0.2949	0.3172
$n = 625$	0.0284	0.0557	0.0841	0.1121	0.1397	0.1709	0.1977	0.2231	0.2516	0.2751

Table 1: L^2-distances between $\hat{\beta}_n(t,s)$ and $\beta_n(t,s)$ for the case of independent normal residuals

σ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
$n = 125$	0.0139	0.0287	0.043	0.0549	0.0686	0.0853	0.1017	0.1114	0.1257	0.1389
$n = 625$	0.0071	0.0143	0.0210	0.0277	0.0354	0.0425	0.0492	0.0575	0.0635	0.0705

Table 2: L^2-distances between $\hat{\beta}_n(t,s)$ and $\beta_n(t,s)$ for the case of Brownian residuals

As it might be expected, the results are much better in the case of Brownian errors although, even in the worst case of independent errors, the quotient $\sigma/$(L2 error) is nearly constant about 3.6 for $n = 225$ and 3.1 for $n = 625$. In the Brownian case the respective quotients are about 7.1 and 14.1, showing both a higher efficiency and a faster improvement as the sample size increases.

A remarkable fact in the simulations is the almost perfect linear dependence (see Figure 6) between the parameter σ an the efficiency of the estimator, measured by its expected L^2 error. In view of this empirical result, a plausible conjecture is that the standard deviation of the asymptotic distribution would be proportional to σ.

INSERT FIGURE 6 AROUND HERE

References

Boente, G. & Fraiman, R. (2000). Kernel-based functional principal components. *Statistics and Probability Letters.*

Bosq, D. (1991). Modelization, non-parametric estimation and prediction for continuous time processes. In *Nonparametric Functional Estimation and Related Topics*, G.G. Roussas et al. eds., 509-529. Kluwer Academic Publishers.

Brumback, B.A. & Rice, J.A. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves (with discussion). *Journal of the American Statistical Association* **93**, 961-994.

Cardot, H., Ferraty, F. & Sarda, P. (1999). Functional linear model. *Statistics and Probability Letters* **45**, 11-22.

Conway, J.B. (1990). *A Course in Functional Analysis.* Springer-Verlag.

Dauxois, J., Pousse, A. & Romain, Y. (1982). Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference. *Journal of Multivariate Analysis* **12**, 136-154.

Fan, J. & Lin, S.K. (1998). Test of significance when the data are curves. *Journal of the American Statistical Association* **93**, 1007-1021.

Ferraty, F. & Vieu, P. (2000). Functional nonparametric model for scalar response. *Manuscript.*

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

Golub, G.H., Heath, M. & Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **21**, 215-224.

Golub, G.H. & Van Loan, C.F. (2001). *Matrix Computations.* The Johns Hopkins University Press.

Gruet, M.A. (1996). A nonparametric calibration analysis. *Annals of Statistics* **24**, 1474-1492.

Hastie, T. & Mallows, C. (1993). A discussion of "A statistical view of some chemometrics regression tools" by I.E. Frank and J.H. Friedman. *Technometrics* **35**, 140-143.

Hoerl, A.E. & Kennard, R.W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **8**, 27-51.

Kneip, A. (1994). Nonparametric estimation of common regressors for similar curve data. *Annals of Statistics* **22**, 1386-1428.

Kneip, A. & Gasser, T. (1992). Statistical tools to analyze data representing a sample of curves. *Annals of Statistics* **20**, 1266-1305.

Li, K. (1984). Regression models with infinitely many parameters: consistency of bounded linear functionals. *Annals of Statistics* **12**, 601-611.

Locantore, N., Marron, J.S., Simpson, D.G., Tripoli, N., Zhang, J.T. & Cohen, K.L. (1999). Robust principal component analysis for functional data (with discussion). *Test* **8**, 1-74.

Martens, H. & Naes, T. (1989). *Multivariate calibration*. Wiley.

Nashed, M.Z. (1976). Perturbations and approximations for generalized inverses and linear operator equations. In *Generalized Inverses and Applications*, M.Z. Nashed ed., 325-396. Academic Press.

Osborne, C. (1991). Statistical calibration: a review. *International Statistical Review* **59**, 309-336.

Ramsay, J.O. & Dalzell, C.J. (1991). Some tools for functional data analysis (with discussion). *Journal of the Royal Statistical Society B* **52**, 539-572.

Ramsay, J.O. & Silverman, B.W. (1997). *Functional Data Analysis*. Springer-Verlag.

Rice, J.A. & Silverman, B.W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society B* **53**, 233-243.

Riesz, F. & Nagy, B.S. (1955). *Functional Analysis*. Frederick Ungar.

Yurinskii, V.V. (1976). Exponential inequalities for sums of random vectors. *Journal of Multivariate Analysis* **6**, 473-499.

FIGURES FOR THE PAPER
"Linear functional regression: the case of fixed design and functional response",
by A. Cuevas, M. Febrero and R. Fraiman

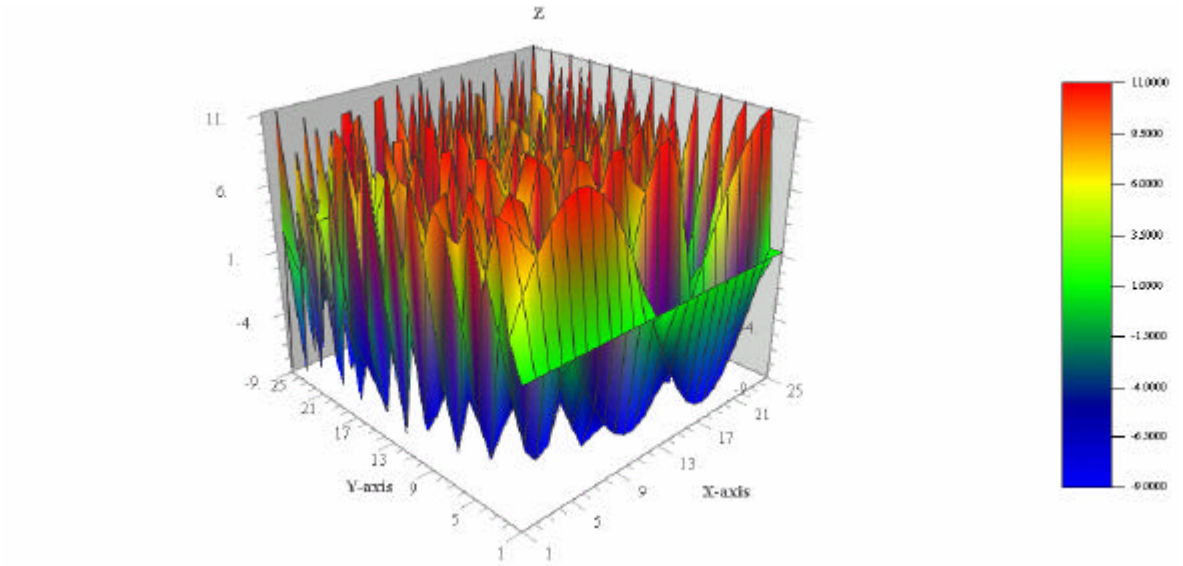

Figure 1: Design functions

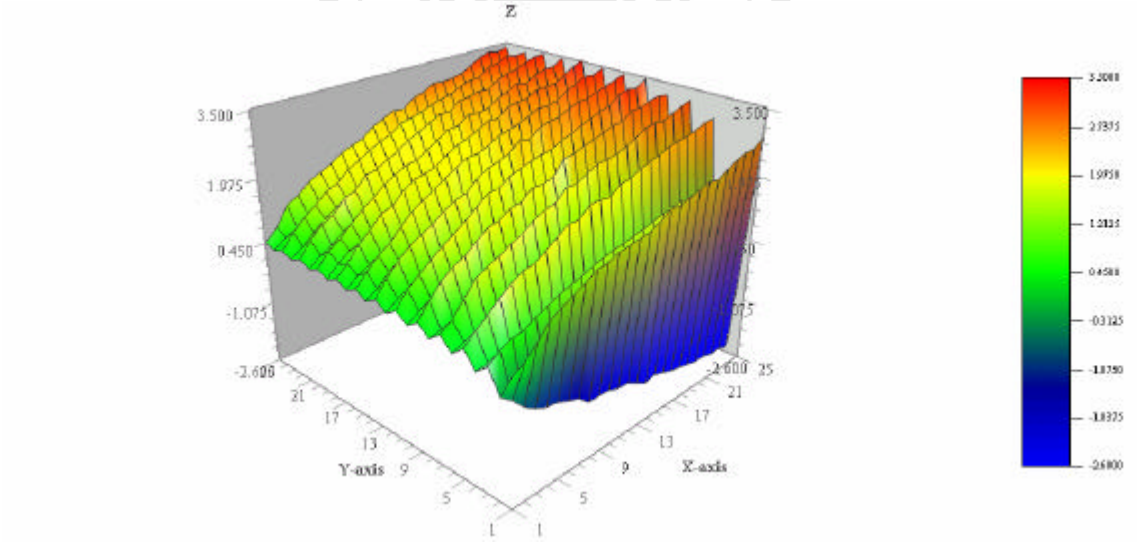

Figure 2: Response functions

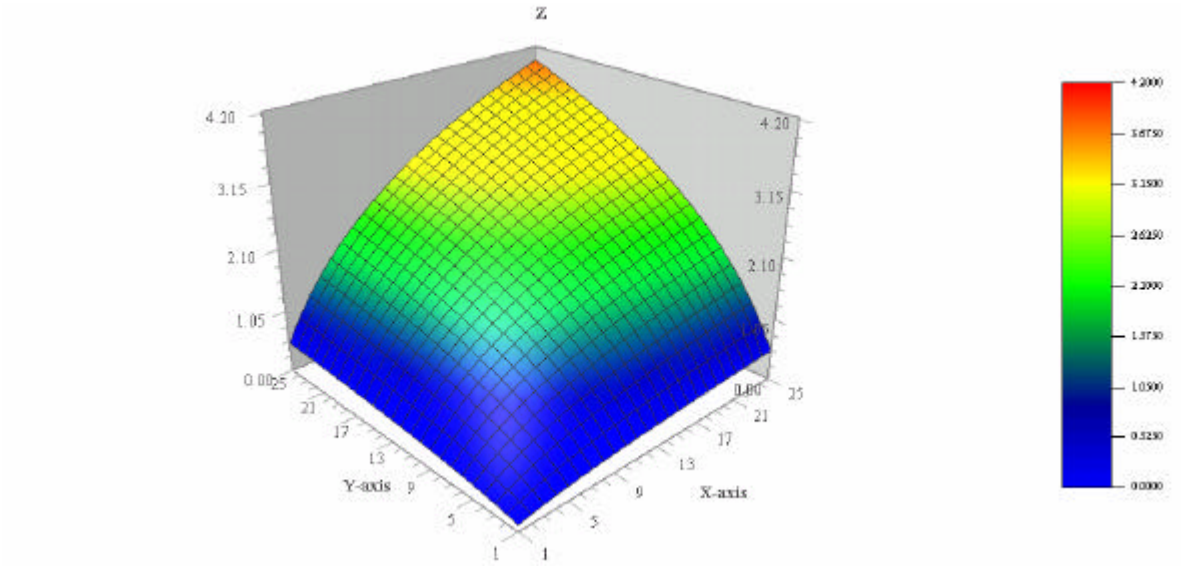

Figure 3: True $b(s,t)$

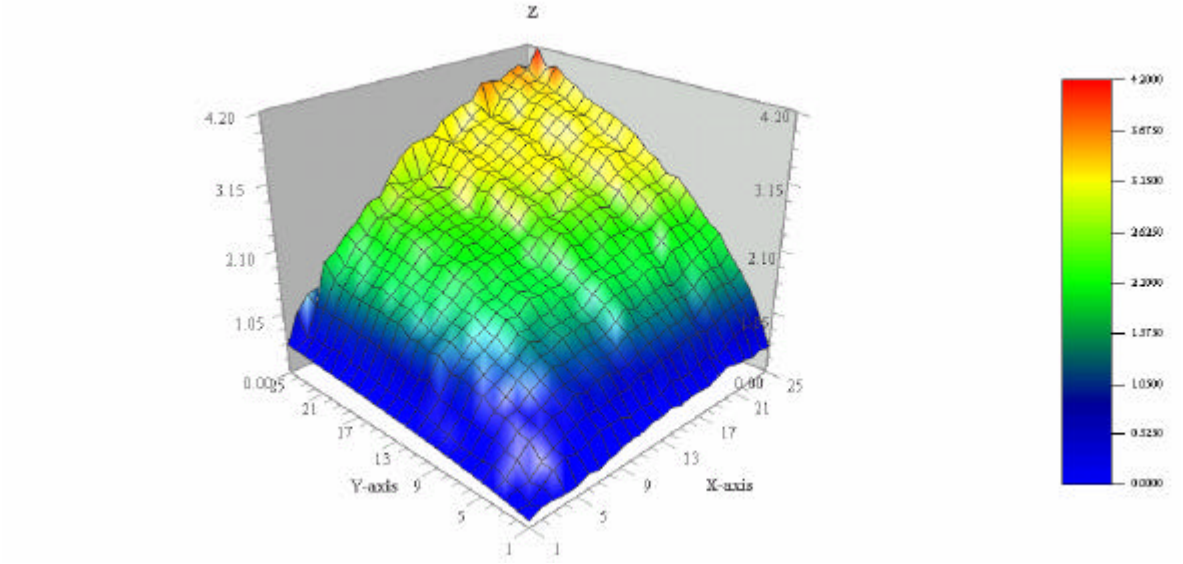

Figure 4: Estimated $b(s,t)$

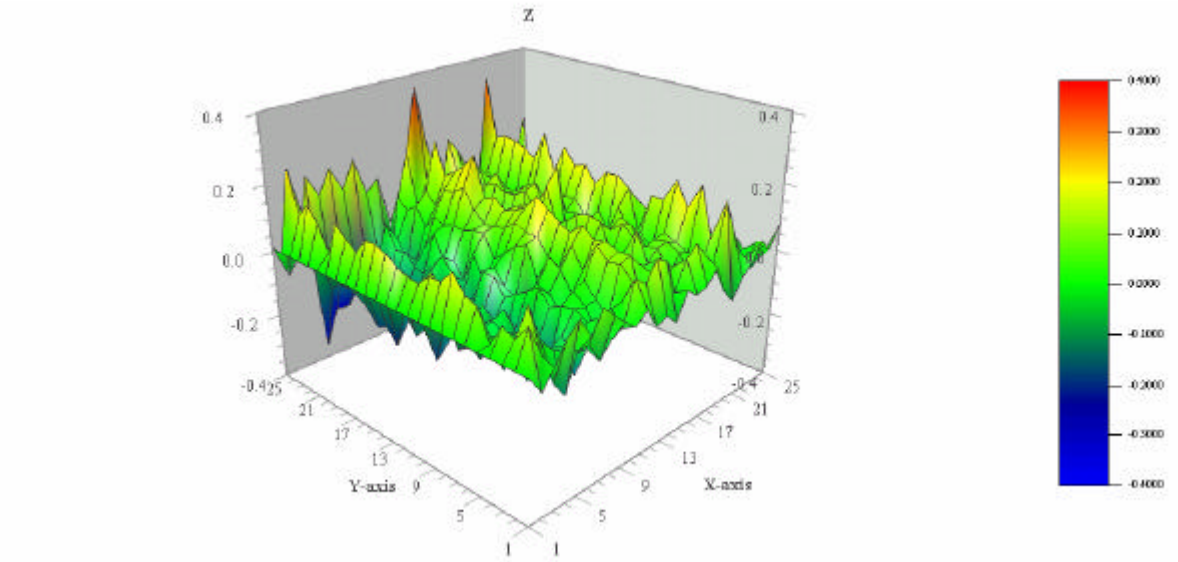

Figure 5: Error surface

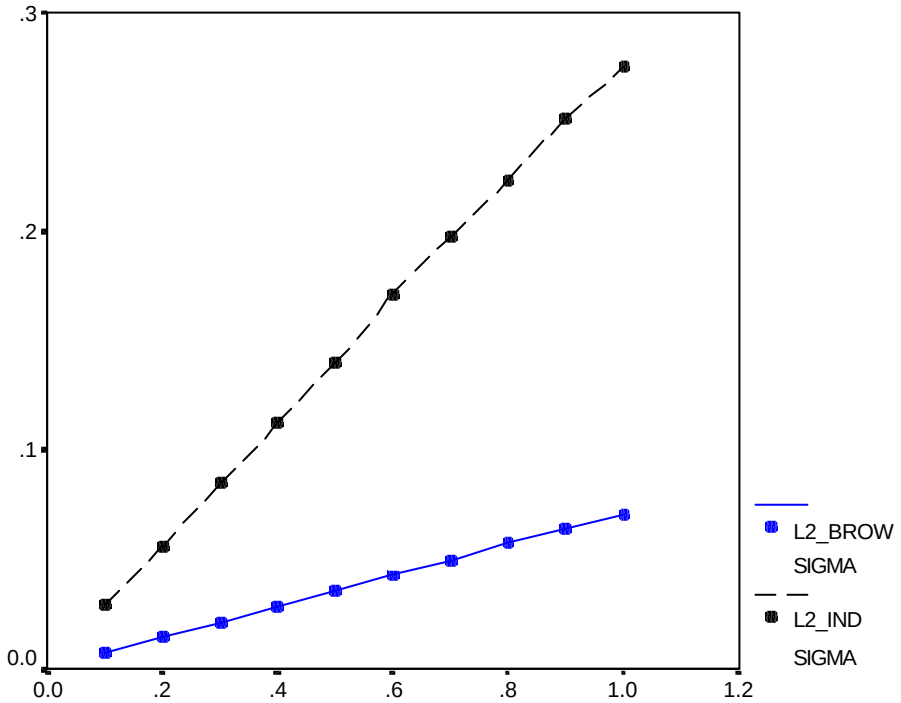

Figure 6: Plots of s versus L2-errors