AUTHOR QUERY FORM

ELSEVIER	**Journal:** Computational Statistics and Data Analysis **Article Number:** 5162	**Please e-mail or fax your responses and any corrections to:** **E-mail: corrections.esch@elsevier.river-valley.com** **Fax: +44 1392 285879**

Dear Author,

Please check your proof carefully and mark all corrections at the appropriate place in the proof (e.g., by using on-screen annotation in the PDF file) or compile them in a separate list. Note: if you opt to annotate the file with software other than Adobe Reader then please also highlight the appropriate place in the PDF file. To ensure fast publication of your paper please return your corrections within 48 hours.

For correction or revision of any artwork, please consult http://www.elsevier.com/artworkinstructions.

Location in article	Query / Remark click on the Q link to go Please insert your reply or correction at the corresponding line in the proof
Q1	Please confirm that given names and surnames have been identified correctly.
Q2	The country name has been inserted for the first affiliation. Please check, and correct if necessary.
Q3	Gervini (2010) is/are cited in the text but not provided in the reference list. Please provide it/them in the reference list or delete these citations from the text.
Q4	The citation Maronna (2011) has been changed to match the author name/date in the reference list. Please check here and in subsequent occurrences, and correct if necessary.
Q5	The citation Yohai and Zamar (1998) has been changed to match the author name/date in the reference list. Please check here and in subsequent occurrences, and correct if necessary.
Q6	'.i.' is removed from here. Please check, and correct if necessary.

Thank you for your assistance.

ARTICLE IN PRESS

Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Robust functional linear regression based on splines

Ricardo A. Maronna [a,*], Victor J. Yohai [b]

[a] Department of Mathematics, School of Exact Sciences, Universidad Nacional de La Plata, Argentina

[b] Department of Mathematics, School of Exact and Natural Sciences, Universidad de Buenos Aires and CONICET, Argentina

ARTICLE INFO

Article history:
Received 22 February 2011
Received in revised form 16 November 2011
Accepted 16 November 2011
Available online xxxx

Keywords:
MM estimate
Natural splines
Robust ridge estimator

ABSTRACT

Many existing methods for functional regression are based on the minimization of an L_2 norm of the residuals and are therefore sensitive to atypical observations, which may affect the predictive power and/or the smoothness of the resulting estimate. A robust version of a spline-based estimate is presented, which has the form of an MM estimate, where the L_2 loss is replaced by a bounded loss function. The estimate can be computed by a fast iterative algorithm. The proposed approach is compared, with favorable results, to the one based on L_2 and to both classical and robust Partial Least Squares through an example with high-dimensional real data and a simulation study.[1]

1. Introduction

We consider the analysis of data described by a linear functional regression model. That is, our data are independent identically distributed (i.i.d.) pairs (X_i, y_i), $i = 1, \ldots, n$, where $y_i \in R$ and $X_i(.)$ are random functions defined on an interval I, such that

$$y_i = \alpha_0 + \int_I \alpha(t) X_i(t) dt + e_i, \quad i = 1, \ldots, n, \tag{1}$$

where the number α_0 and the function $\alpha(t)$ are unknown, and $\{e_i\}$ are i.i.d. random errors independent of $\{X_i\}$. In practice one actually observes at given points $t_1 < \cdots < t_p$ in I the values $x_{ij} = X_i(t_j)$. Henceforth we shall denote $\mathbf{X} = [x_{ij}] \in R^{n \times p}$ and $\mathbf{y} = [y_i] \in R^n$;

These data sets are often high-dimensional, in many cases with $p \gg n$. The functional framework allows to profit from qualitative assumptions like smoothness of underlying curves. This type of regression model was first considered in Ramsay and Dalzell (1991). Ramsay and Silverman (2002, 2005) and Ferraty and Vieu (2006), present several case studies demonstrating the advantages of these models. Among recent applications, Goldsmith et al. (2010) present an application to diffusion tensor imaging (DTI) tractography, and Delaigle et al. (2009) deal with a meteorological application. Cardot et al. (2005, 2006) present the theory and applications of quantile regression for functional data.

One of the most important approaches for the estimation of α_0 and α is regularization through a penalized least squares approach after expanding in some basis such as splines: see Ramsay and Dalzell (1991), Eilers and Marx (1996), Marx and Eilers (1999), Cardot et al. (2003). Crambes et al. (2009) proposed a smoothing splines approach prolonging previous work from Cardot et al. (2007). They show that the rates of convergence of their estimators are optimal in the sense

* Correspondence to: Departamento de Matemáticas, Facultad de Ciencias Exactas, C.C. 172, La Plata 1900, Argentina.
E-mail address: rmaronna@retina.ar (R.A. Maronna).

[1] Matlab code for the proposed procedure is provided as supplemental material.

0167-9473/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2011.11.014

Please cite this article in press as: Maronna , R.A., Yohai, V.J., Robust functional linear regression based on splines. Computational Statistics and Data Analysis (2011), doi:10.1016/j.csda.2011.11.014

that they are minimax over large classes of distributions of X_i and of functions α. Their approach boils down to an easy to implement procedure. Recently Wang et al. (2012) proposed a spline-based nonparametric transformation model for functional regression.

Most approaches to functional regression are based on minimizing some L_2 norm, and are therefore sensitive to outliers, which calls for the development of robust methods. There are numerous articles on robust methods for functional data. In particular, Crambes et al. (2008) propose a robust estimator for nonparametric models, and Gervini (2010) deals with robust regression between two stochastic processes; But we are not aware of any robust approach for model (1). The purpose of this article is to propose a robust version of the estimator proposed by Crambes et al. (2009), based on the approach of MM estimation (Yohai, 1987).

Section 2 describes the proposed estimator, the advantages of which are demonstrated in Sections 3 and 4 through their performances with real and simulated data sets, respectively. The computing times of the different estimators are compared in Section 5. Finally Section 6 contains the conclusions of the study.

2. The proposed estimator

We first describe the estimator proposed by Crambes et al. (2009). Let $\widetilde{\mathbf{X}} = [\widetilde{x}_{ij}]$ and $\widetilde{\mathbf{y}} = [\widetilde{y}_i]$ be $\mathbf{X}$ and $\mathbf{y}$ centered by their averages. The estimator is the function $\widehat{\alpha}(t)$ in the Sobolev space $W^{m,2}(I)$ (see e.g. Adams and Fournier, 2003) such that

$$\frac{1}{n}\sum_{i=1}^{n}\left(\widetilde{y}_i - \frac{1}{p}\sum_{j=1}^{p}\widehat{\alpha}\left(t_j\right)\widetilde{x}_{ij}\right)^2 + \lambda\left(\frac{1}{p}\sum_{j=1}^{p}\pi_{\widehat{\alpha}}\left(t_j\right)^2 + \int_I \widehat{\alpha}^{(m)}(t)^2\,dt\right) = \min \tag{2}$$

where in general $\alpha^{(m)}(t)$ denotes the m-th derivative of $\alpha(t)$, $\lambda > 0$ is a penalty parameter and

$$\pi_{\widehat{\alpha}}(t) = \sum_{l=1}^{m}\gamma_{b,l}t^{l-1} = \arg\min_{\pi}\sum_{j=1}^{p}\left(\widehat{\alpha}\left(t_j\right) - \pi\left(t_j\right)\right)^2,$$

where π ranges over the polynomials in t of degree $m-1$. The $\pi_{\widehat{\alpha}}$ terms ensure the existence of a unique solution. The terms with $\int_I \widehat{\alpha}^{(m)}(t)^2dt$ penalize the solutions' roughness.

The problem (2) has an explicit solution. Let $\mathbf{b}(t) = \left(b_1(t), \ldots, b_p(t)\right)'$ be a functional basis of the space $NS^m(t_1, \ldots, t_p)$ of natural splines of order $2m$ on I with knots $t_1, \ldots, t_p$. Call $\mathbf{B}$ the $p \times p$ matrix with elements $b_i\left(t_j\right)$. Put $\mathbf{U} = \int_I \mathbf{b}^{(m)}(t)\mathbf{b}^{(m)}(t)'dt$ and let $\mathbf{P}_m$ be the $p \times p$ projection matrix projecting R^p onto the m-dimensional linear space of all (discretized) polynomials of degree $m-1$; i.e., $\mathbf{P}_m = \mathbf{G}\mathbf{G}^+$, where $\mathbf{G}$ has elements $g_{jk} = t_j^k$, $j = 1, \ldots, p$, $k = 0, \ldots, m-1$, and $\mathbf{G}^+$ stands for its pseudo-inverse. Let $\mathbf{A}_m^* = \mathbf{B}^{+\prime}\mathbf{U}\mathbf{B}^+$ and let

$$\mathbf{A}_m = \mathbf{P}_m + p\mathbf{A}_m^*, \tag{3}$$

Then Crambes et al. (2009) show that $\widehat{\alpha} \in NS^m(t_1, \ldots, t_p)$ and that the solution for the vector $\widehat{\alpha} = \left(\widehat{\alpha}\left(t_1\right), \ldots, \widehat{\alpha}\left(t_p\right)\right)'$ is

$$\widehat{\alpha} = \arg\min_{\mathbf{a}\in R^p}\left\{\frac{1}{n}\sum_{i=1}^{n}\left(\widetilde{y}_i - p^{-1}\widetilde{\mathbf{x}}_i'\mathbf{a}\right)^2 + p^{-1}\lambda\mathbf{a}'\mathbf{A}_m\mathbf{a}\right\}, \tag{4}$$

where $\widetilde{\mathbf{x}}_i'$ is the i-th row of $\widetilde{\mathbf{X}}$.

As a robustification of the former approach, we propose to find a function $\widehat{\alpha}$ and a number $\widehat{\alpha}_0$ such that

$$\widehat{\sigma}_{\text{ini}}^2\sum_{i=1}^{n}\rho\left(\frac{y_i - \widehat{\alpha}_0 - p^{-1}\sum_{j=1}^{p}x_{ij}\widehat{\alpha}\left(t_j\right)}{\widehat{\sigma}_{\text{ini}}}\right) + \lambda\left(\frac{1}{p}\sum_{j=1}^{p}\pi_{\widehat{\alpha}}\left(t_j\right)^2 + \int_I \widehat{\alpha}^{(m)}(t)^2dt\right) = \min, \tag{5}$$

where ρ is a bounded "ρ-function" in the sense of (Maronna et al., 2006), i.e., $\rho(t)$ is a nondecreasing function of $|t|$ with $\sup_t \rho(t) = 1$; and $\widehat{\sigma}_{\text{ini}}$ is a residual M-estimator of scale from an initial estimator (to be described later). The factor $\widehat{\sigma}_{\text{ini}}^2$ before the summation is employed to make the estimator coincide with the classic one when $\rho(t) = t^2$.

It is not difficult to show that again $\widehat{\alpha} \in NS^m(t_1, \ldots, t_p)$, since $\{\pi_{\widehat{\alpha}}\left(t_j\right), j = 1, \ldots, p\}$ depends only on the values of $\widehat{\alpha}$ at $t_1, \ldots, t_p$, and it is well-known that given these values, the function $\widehat{\alpha}$ minimizing the integral in (5) belongs to $NS^m(t_1, \ldots, t_p)$. Let $\widehat{\alpha}_1 = \left(\widehat{\alpha}\left(t_1\right), \ldots, \widehat{\alpha}\left(t_p\right)\right)'$ and $\mathbf{x}_i = \left(x_{i1}, \ldots, x_{ip}\right)'$. Then it follows that (5) implies

$$\widehat{\sigma}_{\text{ini}}^2\sum_{i=1}^{n}\rho\left(\frac{y_i - \widehat{\alpha}_0 - p^{-1}\mathbf{x}_i'\widehat{\alpha}_1}{\widehat{\sigma}_{\text{ini}}}\right) + p^{-1}\lambda\widehat{\alpha}_1'\mathbf{A}_m\widehat{\alpha}_1 = \min. \tag{6}$$

Please cite this article in press as: Maronna , R.A., Yohai, V.J., Robust functional linear regression based on splines. Computational Statistics and Data Analysis (2011), doi:10.1016/j.csda.2011.11.014

This situation can be reduced to an already studied one. Let $\mathbf{C}$ be a square root of $\mathbf{A}_m$, i.e. a $p \times p$ matrix such that $\mathbf{A}_m = \mathbf{C}'\mathbf{C}$, and put $\mathbf{Z} = \mathbf{X}\mathbf{C}^{-1}$. Let $\widehat{\beta} = \left(\widehat{\beta}_0, \widehat{\beta}_1\right)$ with $\widehat{\beta}_1 = p^{-1}\mathbf{C}\widehat{\alpha}_1$, and $\lambda' = p\lambda$. Then $\widehat{\beta}$ satisfies

$$\widehat{\sigma}_{\text{ini}}^2 \sum_{i=1}^{n} \rho\left(\frac{y_i - \widehat{\beta}_0 - \mathbf{z}_i'\widehat{\beta}_1}{\widehat{\sigma}_{\text{ini}}}\right) + \lambda' \left\|\widehat{\beta}_1\right\|^2 = \min, \tag{7}$$

that is, $\widehat{\beta}$ is an MM ridge estimator, proposed by Maronna (2010). Q4

We now define the initial estimator $\widehat{\beta}_{\text{ini}} = \left(\widehat{\beta}_{\text{ini},0}, \widehat{\beta}_{\text{ini},1}\right)$ needed for (7). For $\mathbf{r} = (r_1, \ldots, r_n)'$ let $S(\mathbf{r})$ be a scale M estimator defined as solution of

$$\frac{1}{n}\sum_{i=1}^{n} \rho_0\left(\frac{r_i}{S(\mathbf{r})}\right) = \delta \tag{8}$$

where ρ_0 is a bounded ρ-function and $\delta \in (0, 1)$ controls the estimator's breakdown point. Then $\widehat{\beta}_{\text{ini}}$ is defined as a penalized regression S estimator, defined as follows. For $\beta = (\beta_0, \beta_1)$ put $r_i(\beta) = y_i - \beta_0 - \mathbf{z}_i'\beta_1$. Then

$$\widehat{\beta}_{\text{ini}} = \arg\min_{\beta} \left(nS(\mathbf{r}(\beta)) + \lambda' \|\beta_1\|^2\right), \tag{9}$$

and $\widehat{\sigma}_{\text{ini}} = cS\left(\mathbf{r}\left(\widehat{\beta}_{\text{ini}}\right)\right)$ where c is a constant that controls the scale's consistency. An S estimate similar to (9) is employed by Tharmaratnam et al. (2010) for nonparametric regression.

The MM estimator is computed by an iterative algorithm starting from $\widehat{\beta}_{\text{ini}}$, such that the criterion (7) decreases at each iteration. Details on the computation of the MM and S estimators are given in Maronna (2010). We choose the functions ρ and ρ_0 in the bisquare family, i.e., they are of the form $\rho(r) = \rho_{\text{bis}}(r/c)$, where c is some positive constant and

$$\rho_{\text{bis}}(r) = 1 - \left(1 - r^2\right)^3 \mathrm{I}(|r| \le 1),$$

where I () stands for the indicator function; the choice of c is described in Maronna (2010). We take $m = 2$ (cubic splines).

The matrix $\mathbf{C}$ is theoretically positive definite, but we have found that for $p > n$ it may be numerically ill-conditioned. To overcome this difficulty, let $\gamma_{\min}$ and $\gamma_{\max}$ be the smallest and largest eigenvalues of $\mathbf{C}$, and δ a small tolerance (we take $\delta = 10^{-6}$). Then if $\gamma_{\min} < \delta\gamma_{\max}$ we replace $\mathbf{C}$ by $\mathbf{C} + \delta\gamma_{\max}\mathbf{I}_p$.

The degree of "shrinking" is measured by the "equivalent degrees of freedom" (edf); see e.g. (Friedman et al., 2009). Call $\mathbf{y}$ and $\widehat{\mathbf{y}}$ the vectors of observed and fitted values, respectively. Then we can write $\widehat{\mathbf{y}} = \mathbf{H}\mathbf{y}$, where the "hat matrix" $\mathbf{H}$ depends on the data. The edf are defined as the trace of $\mathbf{H}$; see (Maronna, 2010) for details.

The minimization is performed over a set of N_λ penalty values: $\lambda' \in \Lambda = \{\lambda_1, \ldots, \lambda_{N_\lambda}\}$. Crambes et al. (2009) employ generalized cross-validation (GCV), taking advantage of the fact that for linear estimators based on the L_2 norm, the leave-one-out prediction errors can be computed explicitly. An overall GCV estimator of the prediction error for each λ is obtained as the average of the squared prediction errors. For our MM estimator, approximate prediction errors can be obtained through a Taylor expansion, given in Eq. (20) of (Maronna, 2010). An overall error estimator is obtained as a robust scale estimator of the squared prediction errors; the complete method employed to select λ is described in Sections 2.4 and 2.5 of (Maronna, 2010).

For both the L_2 and the MM estimator we have observed in both real and simulated data that the estimated α is rather rough. At the same time it was observed the curve of prediction error as a function of λ to be very flat around its minimum. This fact leads to the thought that λ – and hence the smoothness of α – may be increased above its "optimum" value without serious damage to prediction. To this end we use an approach similar to the "one standard deviation rule" (Friedman et al., 2009, pp. 21 and 218), implemented as follows.

For each trial value of $\lambda \in \Lambda$ call $\mathbf{u}(\lambda) = (u_1(\lambda), \ldots, u_n(\lambda))'$ the vector of cross-validation prediction errors. For the L_2 estimator put

$$M(\lambda) = \text{ave}_i\{u_i(\lambda)^2\}, \qquad s(\lambda) = \frac{1}{\sqrt{n}}\text{sd}_i\{u_i(\lambda)^2\}, \tag{10}$$

where ave and sd stand respectively for the average and the sample standard deviation, so that $s(\lambda)$ is the estimated sd of $M(\lambda)$. Let $\lambda_{\text{opt}} = \arg\min_{\lambda\in\Lambda} M(\lambda)$ and $s_0 = s\left(\lambda_{\text{opt}}\right)$. For a given C_{CV} we take $\widehat{\lambda} = \max\{\lambda \in \Lambda : M(\lambda) \le M\left(\lambda_{\text{opt}}\right) + C_{\text{CV}}s_0\}$. We tried $C_{\text{CV}} = 0$, 1 and 2.

For the robust estimator we use a robust version of (10), replacing the average of squares by a robust and efficient scale, namely

$$M(\lambda) = S_i\{u_i(\lambda)\}^2, \qquad s(\lambda) = \frac{1}{\sqrt{n}}S\{u_i(\lambda)^2 - M(\lambda)\},$$

where S is a τ-scale with constant 5 (Yohai and Zamar, 1988), which may be considered as a robust mean squared error. Q5

Please cite this article in press as: Maronna , R.A., Yohai, V.J., Robust functional linear regression based on splines. Computational Statistics and Data Analysis (2011), doi:10.1016/j.csda.2011.11.014

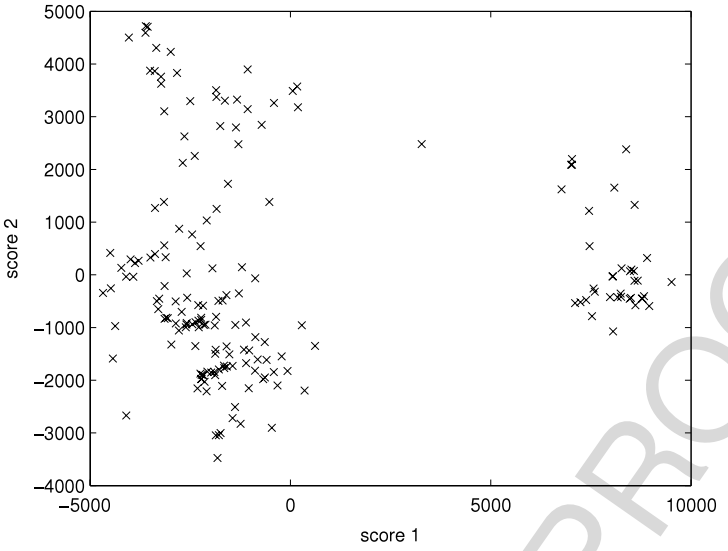

Fig. 1. Vessel data: first two principal components of predictors.

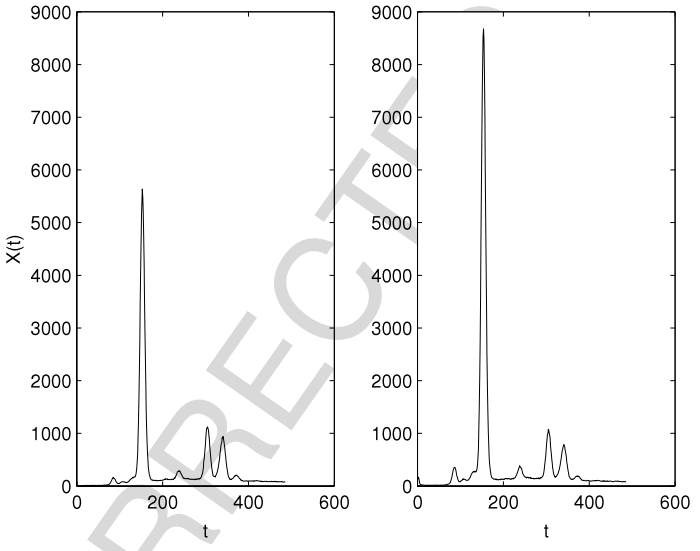

Fig. 2. Vessel data: "Typical" X curves from left-hand and right-hand clusters.

3. An example with real data

In this section the performances of L_2 and MM are compared through a high-dimensional real dataset corresponding to electron-probe X ray microanalysis of archeological glass vessels (Janssens et al., 1998), where each of $n = 180$ vessels is represented by a spectrum on 1920 frequencies. For each vessel the contents of 13 chemical compounds are registered. Since for frequencies below 15 and above 500 the values of x_{ij} are almost null, we keep frequencies 15–500, so that we have $p = 486$.

In order to gain some insight on the data, Fig. 1 shows the representation of the X data on the first two principal components, which account for 85% of the variability. The plot shows two clusters, of which the one on the right contains 37 observations. Fig. 2 shows a "typical" $X(t)$ curve from each cluster. It is seen that the shapes of both curves are similar, but they differ on the sizes of the main peaks.

The estimators considered were the ones proposed by Crambes et al. (2009) and our proposal with nominal efficiency 0.85 (henceforth "L_2" and "MM" respectively), with C_{CV} equal to 0, 1 and 2. Henceforth L_2 and MM with $C_{CV} = C$ will be abbreviated by $L_2(C)$ and MM(C), respectively.

We also included for comparison two versions of Partial Least Squares (PLS): the classical one, and the "partial robust M-regression" proposed by Serneels et al. (2005). In both versions the number of components was chosen through 5-fold CV. The classical and robust versions will be henceforth denoted by C-PLS and R-PLS respectively. There are other proposals for

Please cite this article in press as: Maronna , R.A., Yohai, V.J., Robust functional linear regression based on splines. Computational Statistics and Data Analysis (2011), doi:10.1016/j.csda.2011.11.014

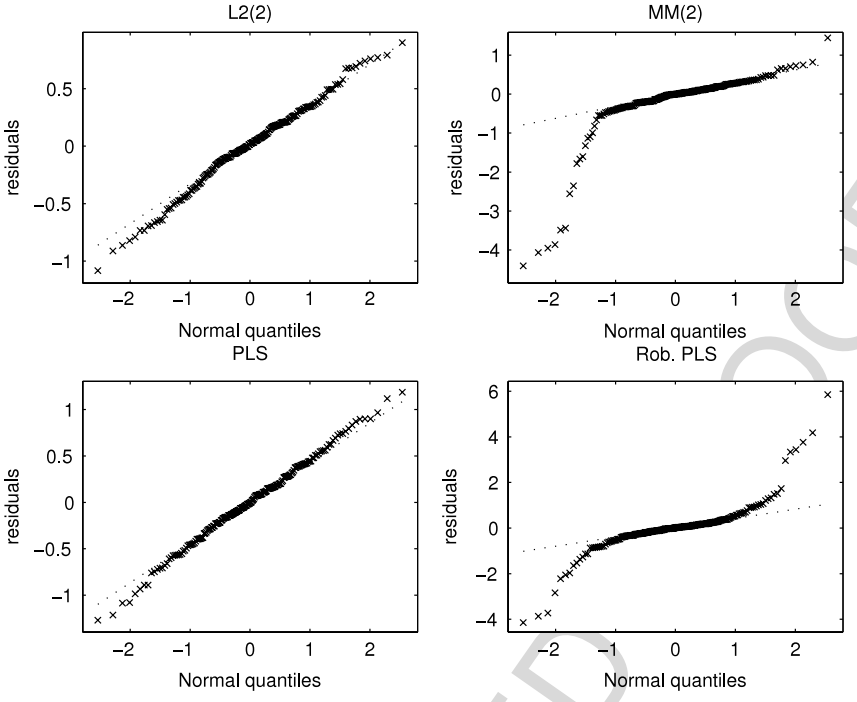

Fig. 3. Vessel data: residual QQ plots for compound 4.

Table 1
Vessel data: prediction CV RMSEs of estimators for compound 5.

L_2				MM			C-PLS	R-PLS
C_{CV} =	0	1	2	0	1	2		
RMSE	0.68	0.66	0.65	0.85	0.87	0.91	0.71	1.44
RMSE (0.9)	0.50	0.49	0.48	0.37	0.38	0.40	0.51	0.69

robust PLS, such as Hubert and Vanden Branden (2003), but we preferred the partial robust M-regression since exploratory simulations indicated that it had a better performance.

Except for compounds 2, 3 and 6, the residual QQ plots from MM showed several clear outliers, while those from L_2 showed none or almost none. Fig. 3 shows the residual QQ plots for compound 4 (SiO_2). For both L_2 and MM, the results for $C_{CV} = 0$, 1 and 2 were similar. We give those for $C_{CV} = 2$ for brevity. It is seen that $L_2(2)$ and C-PLS point out no outliers; both MM(2) and R-PLS point out about 10–12 outliers, but the respective configurations are quite different. All outliers correspond to X data in the right-hand cluster of Fig. 1.

The equivalent degrees of freedom for $L_2(2)$ and MM(2) are respectively 62 and 26, which suggests that the latter gives smoother parameters. R-PLS chooses 14 components.

The left-hand panel of Fig. 4 compares the ordered absolute residuals from MM(2) with those from $L_2(2)$; the dotted line corresponds to the identity. The center panel gives an enlarged view of the lower left-hand corner of the former panel, corresponding to the smallest 152 (out of 180)values; it is seen that the values from MM(2) are smaller than those from $L_2(2)$. The right-hand panel compares MM and R-PLS; here we see that *all* ordered absolute residuals from MM are smaller than those from R-PLS, which indicates a much better fit.

We now compare the degrees of smoothness of the different parameter sequences. Fig. 5 displays, from top to bottom, the functions $\widehat{\alpha}(t)$ corresponding to $L_2(2)$, MM(2) and R-PLS. Since the purpose of the plot is just to compare degrees of smoothness, the curves have been shifted to avoid overlapping. It is seen that MM(2) gives a smoother curve than $L_2(2)$ – as could be expected from the respective edf's – and that R-PLS gives the roughest curve, which can also be expected since its definition does not involve any smoothness restrictions. We may thus say that MM(2) gives a better fit than $L_2(2)$ for 90% of the data, while yielding much smoother estimated parameters.

The predictive behavior of the estimators was assessed through 5-fold CV; the criteria used were the root mean squared error (RMSE) and RMSE with upper 10% trimming (RMSE(0.9)), considered to be safer in the presence of outliers. Table 1 shows the results.

It is seen that increasing C_{CV} does not damage the behavior of the estimators. According to RMSE, L_2 is superior to MM, while the values of RMSE(0.9) for L_2 are about 20%–35% higher than those for MM. The difference is attributable to the outliers shown in the former figures. C-PLS behaves similarly to L_2. It is surprising to see that the behavior of R-PLS is the worst.

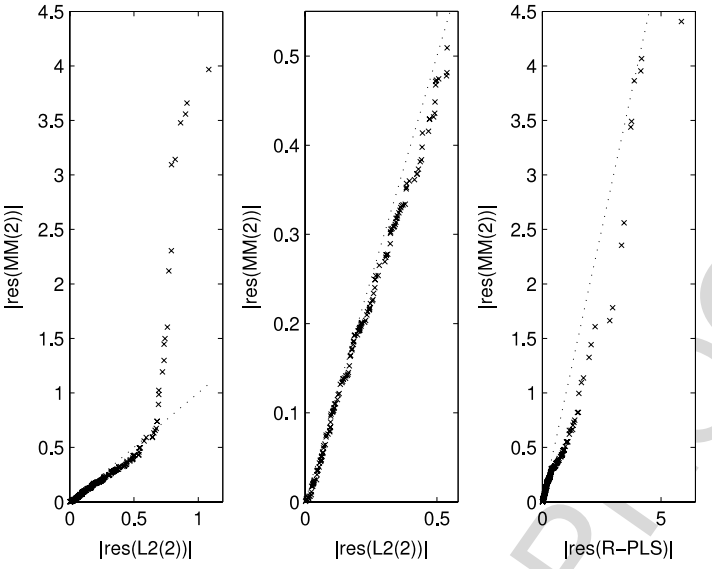

Fig. 4. Vessel data, ordered absolute residuals. Left panel: MM (2) vs. $L_2(2)$, together with the identity line (dotted). Center panel: lowest 152 values in left panel. Right panel: MM (2) vs. R-PLS.

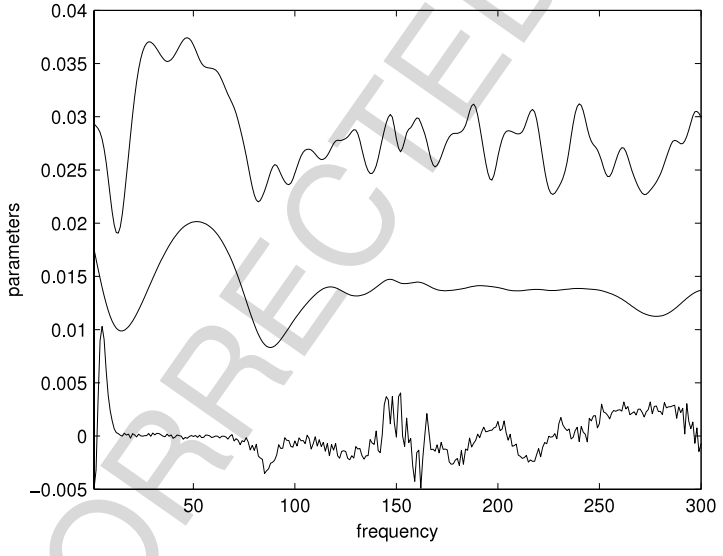

Fig. 5. From top to bottom: parameters (shifted) of $L_2(2)$, MM (2) and R-PLS as functions of the frequency.

4. Simulation

The estimators were assessed through a reduced simulation study. The estimators considered were the same as in the former section. We considered two scenarios.

4.1. Scenario 1

For each situation, $N_{\text{rep}} = 500$ samples of size n and dimension p were generated according to (1) to emulate the data in the former example. The following characteristics were observed in the vessel **X**:

- all spectra (rows) have two or three main peaks
- the correlation between columns j and k decays approximately exponentially with $|j - k|$
- the dispersions of the columns are approximately proportional to their means
- the columns' distributions look approximately lognormal.

Please cite this article in press as: Maronna , R.A., Yohai, V.J., Robust functional linear regression based on splines. Computational Statistics and Data Analysis (2011), doi:10.1016/j.csda.2011.11.014

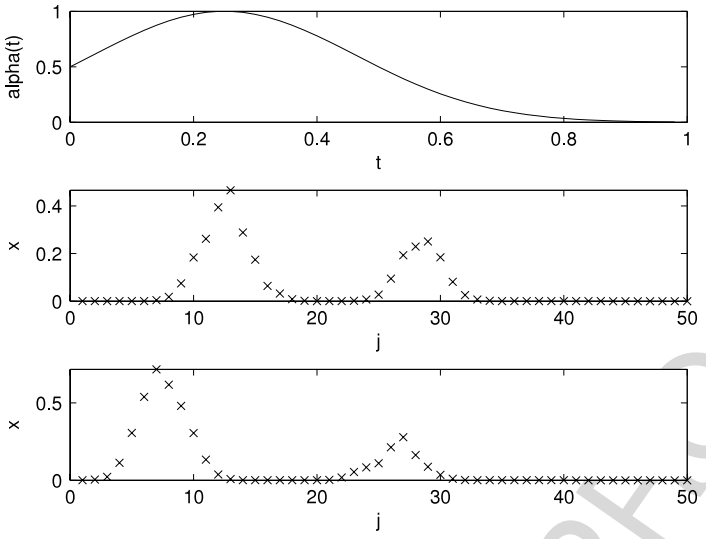

Fig. 6. Simulation: "True" function $\alpha(t)$ and two typical rows of $\mathbf{X}_0$.

To implement these features, let $t_j = j/p$ for $j = 1, \ldots, p$. Define for $t \in [0, 1]$

$$h(t, u_1, u_2) = a \exp\left[-\left(\frac{t-u_1}{b}\right)^2\right] + (1-a)\left[-\left(\frac{t-u_2}{b}\right)^2\right],$$

where a and b are fixed constants. Call $\mathbf{X}_0$ the matrix with elements

$$x_{0,ij} = h\left(t_j, u_{1i}, u_{2i}\right) \exp\left(c v_{ij}\right),$$

where u_{1i} and u_{2i} are independent uniform variables in (0, 0.4) and (0.5, 0.9) respectively, and for each i, $\left(v_{i1}, \ldots, v_{ip}\right)$ have a p-variate normal distribution with zero means and covariances $\sigma_{jk} = \rho^{|j-k|}$. We have chosen $a = 0.8$, $b = 0.05$, $c = 0.5$ and $\rho = 0.5, 0.7$ and 0.9. Let $\alpha(t) = \exp\left[-\left((t-0.25)/0.3\right)^2\right]$ and call α the vector with elements $\alpha\left(t_j\right)$. Fig. 6 plots α and two typical rows of $\mathbf{X}_0$.

Let $\mathbf{y}_0 = \mathbf{X}_0\alpha$. In the case of no contamination, the data are $(\mathbf{X}, \mathbf{y})$ with $\mathbf{X} = \mathbf{X}_0$ and $\mathbf{y} = \mathbf{y}_0 + \sigma\mathbf{e}$, where $\mathbf{e}$ has i.i.d. N (0, 1) elements. For the value of σ, we choose a "noise-to-signal" parameter NSR, and then take $\sigma = \text{NSR} \times \text{sd}(\mathbf{y}_0)$. The values of NSR were chosen as 0.02, 0.05 and 0.1, the choice of this range being guided by examination of this and other spectral data sets, such as the diesel data of (Eigenvector Research, Inc., 2007). In the case of a contamination rate ε, let $m = [n\varepsilon]$ where [.] stands for the integer part. Then the first rows of $\mathbf{X}$ are multiplied by k_{lev}, where k_{lev} is a parameter that regulates the outlier leverage and is fixed at 2; and for $i = 1, \ldots, m$ we put $y_i = y_{0i}k_{\text{lev}}k_{\text{slo}}$, where the parameter k_{slo} regulates the outliers' slope. The effect of this contamination is to pull the estimator $\widehat{\alpha}$ towards $k_{\text{slo}}\alpha$. The values of k_{slo} were taken on a grid between 1.1 and 1.8, in order to find the largest MSE of the robust estimators.

To evaluate an estimator $\widehat{\alpha}$, let for a given sample $\text{MPE} = n^{-1}\left\|\mathbf{y}_0 - \mathbf{X}_0\widehat{\alpha}\right\|^2$, which measures the mean squared prediction error of $\widehat{\alpha}$ if applied to the uncontaminated data. The N_{rep} values of MPE can be summarized through their average. However, it was observed that in contamination situations, this average was frequently heavily influenced by a few very large values (as was seen in Table 1). For this reason we also computed the 10% (upper) trimmed average, henceforth referred to as "trimmed MPE".

This scenario involves three values of ρ, three of NSR, two values of n and three of p for each n, two values of ε and about 15 of k_{slo} for $\varepsilon = 1$, i, which total some 800 sampling situations. In order to simplify the exposition, for each combination (n, p, ρ, NSR) we took the MPE for $\varepsilon = 0$ and the maximum (over k_{slo} MPE for $\varepsilon = 0.1$. Q6

We give the values of MPE; the trimmed MPE yielded results more favorable to MM. The three values of ρ yielded gave qualitatively similar results; we exhibit the results for $\rho = 0.7$. Tables 2 and 3 display the simulation results, where the MPEs have been multiplied by 100 to improve legibility.

We postpone the discussion of these results to Section 4.3.

4.2. Scenario 2

The next scenario is defined as follows. Let $\mu(t) = \sin(6\pi t)(t+1)$; given $\rho \in (0, 1)$, let $\mathbf{u}_i \in R^p$, $i = 1, \ldots, n$ be i.i.d. such that the correlation between u_{ij} and u_{ik} is

$$\frac{1}{1 + a(j-k)^2} \quad \text{with } a = \frac{1}{\rho} - 1,$$

so that the lag-one correlation is ρ, and the correlations decay more slowly than the exponential rate of Scenario 1.

Table 2
Simulation scenario 1: maximum MPEs (×100) for $\rho = 0.7$ and $n = 50$.

p	NSR	L_2				MM			PLS	
		C_{CV} =	0	1	2	0	1	2	C-PLS	R-PLS
		ε								
25	0.02	0	0.59	0.51	0.59	0.52	0.57	0.62	0.52	0.60
		0.1	28.25	27.45	27.64	2.03	1.78	1.52	32.17	1.46
	0.05	0	1.43	1.16	1.37	1.47	1.41	1.51	1.20	1.29
		0.1	28.98	27.71	27.96	5.49	4.78	4.11	33.64	3.11
	0.10	0	2.76	2.11	2.52	2.82	2.41	2.76	2.22	2.48
		0.1	30.17	28.41	28.27	8.97	7.22	5.28	34.75	5.35
60	0.02	0	2.47	1.82	2.00	2.62	2.36	2.74	3.40	3.61
		0.1	125.68	134.10	140.54	6.16	5.22	4.32	162.26	9.20
	0.05	0	6.15	4.42	4.68	6.28	5.64	6.19	7.40	8.43
		0.1	128.96	136.53	142.35	15.07	10.28	9.11	169.85	18.36
	0.10	0	12.20	8.57	9.67	11.07	11.19	12.45	13.28	14.34
		0.1	134.38	140.54	146.49	28.99	24.23	19.75	180.07	29.84
100	0.02	0	5.88	4.99	5.65	5.77	7.24	8.21	9.96	9.58
		0.1	305.06	324.76	345.75	14.91	12.39	13.07	416.37	25.89
	0.05	0	14.62	12.14	13.45	13.73	14.56	18.01	21.12	20.43
		0.1	311.47	331.54	352.97	48.30	32.23	29.24	446.32	48.18
	0.10	0	29.00	23.53	26.52	27.70	33.29	41.39	39.81	37.33
		0.1	325.13	342.19	363.59	69.61	58.48	62.22	523.07	75.30

Table 3
Simulation scenario 1: maximum MPEs (×100) for $\rho = 0.7$ and $n = 100$.

p	NSR	L_2				MM			PLS	
		C_{CV} =	0	1	2	0	1	2	C-PLS	R-PLS
		ε								
25	0.02	0	0.31	0.31	0.41	0.33	0.36	0.38	0.29	0.35
		0.1	35.82	31.56	29.38	0.87	0.67	0.61	38.44	0.86
	0.05	0	0.73	0.71	0.95	0.77	0.81	0.93	0.68	0.80
		0.1	36.18	31.66	29.45	1.92	1.53	1.38	39.54	1.71
	0.10	0	1.41	1.32	1.82	1.45	1.56	1.78	1.25	1.41
		0.1	36.73	31.87	29.59	4.04	3.11	2.52	40.45	3.11
50	0.02	0	1.58	1.25	1.46	1.48	1.50	1.60	1.39	1.55
		0.1	106.73	103.88	102.14	4.36	3.13	2.51	124.48	3.41
	0.05	0	3.83	2.74	3.34	3.12	3.27	3.77	2.89	3.10
		0.1	108.96	104.84	102.76	7.67	6.1	5.17	129.92	7.04
	0.10	0	7.39	4.97	6.14	5.72	5.89	6.73	5.13	5.44
		0.1	112.43	106.33	103.57	17.69	11.9	10.11	133.41	11.72
150	0.02	0	10.20	6.50	5.28	6.75	6.54	7.56	12.71	12.56
		0.1	554.79	597.27	639.53	14.24	11.48	11.43	777.44	30.52
	0.05	0	25.42	15.92	14.79	15.44	15.12	17.75	25.68	26.34
		0.1	567.16	608.12	649.67	39.25	33.56	31.22	820.58	57.03
	0.10	0	50.63	30.86	29.49	29.43	30.49	33.83	46.65	47.43
		0.1	588.29	626.78	666.10	75.73	54.59	48.57	935.89	94.08

For $t_j = j/p$ $(j = 1, \ldots, p)$, put $x_{ij} = \mu\left(t_j\right) + \gamma_1 u_{ij}\sqrt{|\mu\left(t_j\right)|}$, where γ_1 determines the roughness of $X(t)$; we employed $\gamma_1 = 0.9$ and 0.5. Define $\alpha(t) = \sqrt{t}$ and let $\alpha \in R^p$ have elements $\alpha_j = \alpha\left(t_j\right)$. Finally put $\mathbf{y} = \mathbf{X}\alpha + \gamma_2\mathbf{e}$ with $\mathbf{e} \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$, where γ_2 controls the NSR.

Tables 4 and 5 gives the maximum MPEs for $n = 50$ and 100, respectively, with $\gamma_1 = 0.9$ and NSR = 0.05. While in Scenario 1, the different values of ρ yielded qualitatively similar results, here ρ has a more important influence, namely, that situations with large ρ are more favorable to PLS. Therefore we display the results for all three values of ρ.

4.3. Discussion

4.3.1. Scenario 1

$L_2(C)$ with $C = 0$ is always worse than with $C = 1$ or 2. In some cases, $L_2(0)$ even has a higher MSE for $\varepsilon = 0$ than MM(0). In these cases we have repeated the simulation with a larger N_{rep} and a different seed, obtaining the same pattern. We have not found an explanation for this phenomenon.

Among the three versions of MM, MM(1) and MM(2) behave in general better than MM(0), and the overall behavior of MM(2) is slightly better than that of MM(1). For $n < p$ both L_2 and MM are more efficient than both versions of PLS. It is curious that R-PLS seems slightly more efficient than C-PLS, which might be due to the differences in the respective algorithms. For $n < p$, R-PLS is slightly more robust than MM(2), but quite the opposite holds for $n > p$.

Please cite this article in press as: Maronna , R.A., Yohai, V.J., Robust functional linear regression based on splines. Computational Statistics and Data Analysis (2011), doi:10.1016/j.csda.2011.11.014

Table 4
Simulation scenario 2: maximum MPEs for $n = 50$.

p	ρ	L_2				MM			PLS	
		$C_{CV} =$	0	1	2	0	1	2	C-PLS	R-PLS
		ε								
25	0.5	0	0.46	0.31	0.39	0.38	0.40	0.45	0.61	0.62
		0.1	3.92	3.37	3.09	0.63	0.55	0.58	4.12	0.90
	0.7	0	0.78	0.50	0.57	0.64	0.66	0.77	0.78	0.68
		0.1	6.54	5.75	5.23	1.05	0.84	0.98	6.66	1.04
	0.9	0	1.75	0.96	1.13	1.21	1.28	1.37	1.27	0.93
		0.1	13.92	12.37	11.44	2.27	1.80	1.83	13.32	1.57
60	0.5	0	1.57	1.12	1.17	1.21	1.27	1.46	3.03	3.04
		0.1	10.3	9.97	9.44	2.09	1.92	2.04	11.75	4.87
	0.7	0	2.77	1.88	1.85	2.17	2.08	2.34	4.00	3.52
		0.1	17.75	17.14	16.18	3.35	2.95	3.12	19.15	4.97
	0.9	0	7.81	4.60	4.15	4.90	4.68	5.77	6.16	4.58
		0.1	46.81	45.11	42.60	7.93	6.48	6.79	49.48	7.11
100	0.02	0	2.79	2.43	2.66	2.62	2.86	3.33	6.27	7.19
		0.1	16.42	16.12	15.91	4.19	4.30	5.02	19.26	15.28
	0.05	0	5.06	4.16	4.53	4.70	4.97	5.52	10.56	9.73
		0.1	29.67	28.82	28.15	7.37	7.15	8.10	33.02	14.79
	0.10	0	14.65	10.73	11.34	11.68	12.78	14.25	16.41	12.56
		0.1	83.31	80.39	75.93	19.17	16.97	17.48	87.02	16.56

Table 5
Simulation scenario 2: maximum MPEs for $n = 100$.

p	ρ	L_2				MM			PLS	
		$C_{CV} =$	0	1	2	0	1	2	C-PLS	R-PLS
		ε								
25	0.5	0	0.22	0.16	0.26	0.16	0.21	0.30	0.33	0.36
		0.1	3.80	2.85	2.22	0.30	0.24	0.32	3.94	0.54
	0.7	0	0.38	0.27	0.31	0.25	0.29	0.45	0.45	0.42
		0.1	6.24	4.80	3.73	0.49	0.37	0.48	6.12	0.61
	0.9	0	0.85	0.52	0.54	0.51	0.58	1.03	0.61	0.44
		0.1	13.57	10.68	8.62	1.03	0.89	0.98	12.21	0.83
50	0.5	0	0.92	0.49	0.52	0.52	0.53	0.76	1.20	1.25
		0.1	8.47	7.76	7.20	0.93	0.68	0.87	8.86	1.84
	0.7	0	1.64	0.77	0.81	0.82	0.89	0.89	1.63	1.57
		0.1	14.25	12.97	12.18	1.40	1.12	1.13	14.05	2.06
	0.9	0	4.32	1.79	1.53	1.91	1.92	1.87	2.10	1.78
		0.1	36.15	32.59	30.51	3.64	2.54	2.40	32.45	2.48
150	0.5	0	4.04	2.85	2.81	2.71	3.08	3.77	8.97	9.07
		0.1	28.55	28.69	28.18	4.74	4.69	5.31	31.99	15.72
	0.7	0	7.43	5.02	4.81	4.66	5.33	6.52	12.65	11.08
		0.1	51.66	51.68	50.51	7.79	7.50	8.54	55.78	16.05
	0.9	0	21.72	13.94	12.15	11.78	13.87	16.19	17.46	14.05
		0.1	142.21	141.93	138.92	19.61	17.67	19.90	146.90	19.71

For $\varepsilon = 0$, MM(0) and MM(1) have similar best performances. In most cases, MM(2) has an acceptable efficiency compared to them. All MMs generally outperform RPLS. For $\varepsilon = 0.1$, in the great majority of cases, MM(2) is the best MM; otherwise it is close to the best; and MM(0) the worst. MM(1) and MM(2) outperform RPLS except when $p = 25$ and $n = 50$.

4.3.2. Scenario 2

$L_2(0)$ exhibits the same puzzling poor behavior as in Scenario 1. For $\varepsilon = 0$, MM(0) is generally the best MM, with MM(1) not far behind. Comparing each MM(C) with the corresponding $L_2(C)$, MM(1) appears as reasonably efficient, and MM(2) less so. All MMs outperform RPLS for $\rho = 0.5$ and 0.7, and for $\rho = 0.9$ they are only slightly worse except for $p = 25$. For $\varepsilon = 0.1$, MM(1) is generally best among MMs, with MM(2) not far behind. All MMs outperform RPLS, except in some cases when $\rho = 0.9$. Recall that PLS takes advantage of high correlations among the predictors, while the spline approach does not, and this may explain the comparatively better behavior of RPLS for $\rho = 0.9$.

5. Computing times

Table 6 gives the mean computing times in seconds for 10 random samples generated as in Scenario 1, for the three estimators, run in Matlab on a PC with a 3.0 GH Intel Core Duo processor. It can be seen that the MM approach is feasible for

Please cite this article in press as: Maronna , R.A., Yohai, V.J., Robust functional linear regression based on splines. Computational Statistics and Data Analysis (2011), doi:10.1016/j.csda.2011.11.014

Table 6
Computing times in seconds.

p	n	L_2	MM	R-PLS
25	50	0.028	4.44	0.25
	100	0.030	4.75	0.29
	200	0.042	5.51	0.30
50	50	0.046	8.39	1.48
	100	0.045	10.06	1.51
	200	0.060	11.87	1.61
100	50	0.089	8.49	1.69
	100	0.114	29.13	15.25
	200	0.138	36.66	15.30
150	50	0.209	8.40	1.91
	100	0.244	28.86	15.72
	200	0.306	85.46	103.00
200	50	0.358	8.64	1.73
	100	0.485	30.25	16.05
	200	0.530	161.02	268.18

large data sets. R-PLS is much faster than MM for small and moderate data sets, but MM is faster for $n = 200$ and $p = 150$ and 200.

6. Conclusions

In the vessel data example, the MM approach showed a better predictive performance than L_2 and both versions of PLS. In the simulations, MM showed a reasonable efficiency (compared to L_2 and PLS) for normal data, and was in general more robust than R-PLS for contaminated data. In the trade-off between efficiency and robustness, it seems that MM(2) is the estimator of choice.

Acknowledgments

This research was partially supported by grants X–018 from University of Buenos Aires, PID 5505 from CONICET and PICTs 21407 and 00899 from ANPCyT,

Appendix. Supplementary data

Supplementary material related to this article can be found online at doi:10.1016/j.csda.2011.11.014.

References

Adams, R.A., Fournier, J.J., 2003. Sobolev Spaces, second ed. Academic Press.
Cardot, H., Crambes, C., Kneip, A., Sarda, P., 2007. Smoothing splines estimators in functional linear regression with errors-in-variables. Comput. Statist. Data Anal. 51, 4832–4848.
Cardot, H., Crambes, C., Sarda, P., 2005. Quantile regression when the covariates are functions. J. Nonparametr. Stat. 17, 841–856.
Cardot, H., Crambes, C., Sarda, P., 2006. Ozone pollution forecasting using conditional mean and conditional quantiles with functional covariates. In: Härdle, , Vieu, (Eds.), Statistical Methods for Biostatistics and Related Fields. Springer.
Cardot, H., Ferraty, F., Sarda, P., 2003. Spline estimators for the functional linear model. Statist. Sinica 13, 571–591.
Crambes, C., Delsol, L., Laksaci, A., 2008. Robust nonparametric estimation for functional data. In: Dabo-Niang, S., Ferraty, F. (Eds.), Functional and Operatorial Statistics. Springer, Heidelberg, pp. 109–116.
Crambes, C., Kneip, A., Sarda, P., 2009. Smoothing splines estimators for functional linear regression. Ann. Statist. 37, 35–72.
Delaigle, A., Hall, P., Apanasovich, T.V., 2009. Weighted least squares methods for prediction in the functional data linear model. Electron. J. Stat. 3, 865–885.
Eigenvector Research, Inc. 2007. http://software.eigenvector.com/Data/SWRI/index.html.
Eilers, P.H., Marx, B.D., 1996. Flexible smoothing with *B*-splines and penalties. Statist. Sci. 11, 89–102.
Ferraty, F., Vieu, P., 2006. Nonparametric Functional Data Analysis: Methods, Theory, Applications and Implementations. Springer, London.
Friedman, J., Hastie, T., Tibshirani, R., 2009. The Elements of Statistical Learning: Data Mining, Inference and Prediction, second ed. Springer, London.
Goldsmith, J., Federy, J., Crainiceanuz, C.M., Caffo, B., Reichyy, D., 2010. Penalized functional regression, Johns Hopkins University, Dept. of Biostatistics Working Papers, #104.
Hubert, M., Vanden Branden, K., 2003. Robust methods for partial least squares regression. J. Chemom. 17, 537–549.
Janssens, K., Deraedt, I., Freddy, A., Veekman, J., 1998. Composition of 15–17th century archeological glass vessels excavated in Antwerp, Belgium. Mikrochim. Acta 15 (Suppl.), 253–267.
Maronna, R.A., 2010. Robust ridge regression for high-dimensional data. Technometrics 53, 44–53.
Maronna, R.A., Martin, R.D., Yohai, V.J., 2006. Robust Statistics: Theory and Methods. John Wiley and Sons, New York.
Marx, B.D., Eilers, P.H., 1999. Generalized linear regression on sampled signals and curves: a *P*-spline approach. Technometrics 41, 1–13.
Ramsay, J.O., Dalzell, C.J., 1991. Some tools for functional data analysis. J. R. Stat. Soc. Ser. B 53, 539–572.
Ramsay, J.O., Silverman, B.W., 2002. Functional Data Analysis. Springer, New York.
Ramsay, J.O., Silverman, B.W., 2005. Applied Functional Data Analysis, second ed. Springer, New York.
Serneels, S., Croux, C., Filzmoser, P., Van Espen, P.J., 2005. Partial robust *M*-regression. Chemometr. Intell. Lab. Syst. 79, 55–64.
Tharmaratnam, K., Claeskens, G., Croux, C., Salibian-Barrera, M., 2010. *S*-estimation for penalized regression splines. J. Comput. Graph. Statist. 5, 609–625.
Wang, G., Lin, N., Zhang, B., 2012. Functional linear regression after spline transformation. Comput. Statist. Data Anal. 56, 587–601.
Yohai, V.J., 1987. High breakdown-point and high efficiency estimates for regression. Ann. Statist. 15, 642–665.
Yohai, V.J., Zamar, R., 1988. High breakdown point estimates of regression by means of the minimization of an efficient scale. J. Amer. Statist. Assoc. 86, 403–413.

Please cite this article in press as: Maronna , R.A., Yohai, V.J., Robust functional linear regression based on splines. Computational Statistics and Data Analysis (2011), doi:10.1016/j.csda.2011.11.014