Robust functional linear regression based on
splines

Ricardo A. Maronna\textsuperscript{1} and Víctor J. Yohai\textsuperscript{2}

\textsuperscript{1}Department of Mathematics, School of Exact Sciences.
University of La Plata, Argentina (rmaronna@retina.ar)

\textsuperscript{2}Department of Mathematics, School of Exact and Natural Sciences
University of Buenos Aires, and CONICET
(vyohai@uolsinectis.com.ar)

Abstract

Existing methods for functional regression are based on 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. We propose a robust version of the spline–based estimate proposed by Crambes, Kneip and Sarda (Ann. Statist. 2009), 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.

Key words and phrases: MM estimate; Natural splines; Robust ridge estimator.
1 Introduction

We consider the analysis of data described by a functional regression model. That is, our data are $(X, y)$ where $X = [x_{ij}] \in \mathbb{R}^{n \times p}$ and $y = [y_i] \in \mathbb{R}^n$; it is assumed that for random functions $X_i(t)$ ($i = 1, \ldots, n$) and given values $t_1 < \ldots < t_p$ in an interval $I$ the data fulfill $x_{ij} = X_i(t_j)$ and

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

where the number $\alpha_0$ and the function $\alpha(t)$ are unknown, and $\{e_i\}$ are independent identically distributed (i.i.d.) random errors independent of $\{X_i\}$. These data sets are generally high-dimensional, with $p >> n$. This 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.

One of the most important approaches for the estimation of $\alpha_0$ and $\alpha$ 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, Ferraty and Sarda (2003). Recently 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 that they are minimax over large classes of distributions of $X_i$ and of functions $\alpha$. Their approach boils down to an easy to implement procedure.

All approaches to functional regression are based on minimizing some $L_2$ norm, and are therefore sensitive to outliers. Crambes et al. (2008) propose
a robust estimator for nonparametric models. We are not aware of any robust approaches for linear functional regression. 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.

2 The proposed estimator

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

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

where \( \lambda > 0 \) is a penalty parameter and

\[
\pi_{\hat{\alpha}}(t) = \sum_{l=1}^{m} \gamma_{b_l} t^{l-1} = \arg \min_{\pi} \sum_{j=1}^{p} (\hat{\alpha}(t_j) - \pi(t_j))^2,
\]

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

The problem (2) has an explicit solution. Let \( b(t) = (b_1(t), \ldots, b_p(t))^t \) 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 \( B \) the \( p \times p \) matrix with elements \( b_i(t_j) \). Put \( U = \int_I b^{(m)}(t) b^{(m)}(t)^t \, dt \)—where \( b^{(m)} \) is the \( m \)-th derivative— and let \( P_m \) be
the $p \times p$ projection matrix projecting $\mathbb{R}^p$ onto the $m$–dimensional linear space of all (discretized) polynomials of degree $m - 1$; i.e., $P_m = GG^+$, where $G$ has elements $g_{jk} = t_j^k$, $j = 1, ..., p$, $k = 0, ..., m - 1$, and $G^+$ stands for its pseudo–inverse. Let $A^*_m = B^+UB^+$ and let

$$A_m = P_m + pA^*_m,$$  \hspace{1cm} (3)

Then Crambes et al (2009) show that

$$\hat{b}_{NS}^m(t_1, ..., t_p)$$

and that the solution for the vector $\hat{a} = (\hat{a}(t_1), ..., \hat{a}(t_p))'$ is

$$\hat{a} = \arg \min_{a \in \mathbb{R}^p} \left\{ \frac{1}{n} \sum_{i=1}^{n} (\tilde{y}_i - p^{-1}\tilde{X}'_i a)^2 + p^{-1} \lambda a' A_m a \right\},$$  \hspace{1cm} (4)

where $\tilde{X}_i$ is the $i$–th row of $\tilde{X}$.

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

$$\hat{\sigma}^2 = \sum_{i=1}^{n} \rho \left( \frac{y_i - \hat{\alpha}_0 - p^{-1} \sum_{j=1}^{p} x_{ij} \hat{\alpha}(t_j)}{\hat{\sigma}_0} \right) + \lambda \left( \frac{1}{p} \sum_{j=1}^{p} \pi_{\hat{\alpha}}(t_j)^2 + \int_{t} \hat{\alpha}^{(m)}(t)^2 dt \right) = \min,$$  \hspace{1cm} (5)

where $\rho$ is a bounded “$\rho$–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 $\hat{\sigma}_0$ is a residual M–estimator of scale from an initial estimator. The factor $\hat{\sigma}^2$ 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 $\hat{\alpha} \in NS^m(t_1, ..., t_p)$, since $\{\pi_{\hat{\alpha}}(t_j), j = 1, ..., p\}$ depends only on the values of $\hat{\alpha}$ at $t_1, ..., t_p$, and it is well–known that given these values, the function $\hat{\alpha}$ minimizing the integral in (5) belongs to $NS^m(t_1, ..., t_p)$. Let $\hat{\alpha}_1 = (\hat{\alpha}(t_1), ..., \hat{\alpha}(t_p))'$ and $x_i = (x_{i1}, ..., x_{ip})'$. Then it follows that (5) implies
\[
\hat{\sigma}_{\text{ini}}^2 \sum_{i=1}^{n} \rho \left( \frac{y_i - \hat{\alpha}_0 - p^{-1}x_i^t \hat{\alpha}_1}{\hat{\sigma}_{\text{ini}}} \right) + p^{-1} \lambda' \hat{\alpha}_1 A_m \hat{\alpha}_1 = \min.
\] (6)

This situation can be reduced to an already studied one. Let \( C \) be a square root of \( A_m \), i.e. a \( p \times p \) matrix such that \( A_m = C'C \), and put \( Z = X C^{-1} \). Let \( \hat{\beta} = (\beta_0, \beta_1) \) with \( \beta_1 = p^{-1}C\hat{\alpha}_1 \), and \( \lambda' = p\lambda \). Then \( \hat{\beta} \) satisfies

\[
\hat{\sigma}_{\text{ini}}^2 \sum_{i=1}^{n} \rho \left( \frac{y_i - \hat{\beta}_0 - z_i^t \hat{\beta}_1}{\hat{\sigma}_{\text{ini}}} \right) + \lambda' \left\| \hat{\beta}_1 \right\|^2 = \min,
\] (7)

that is, \( \hat{\beta} \) is an MM ridge estimator, proposed by Maronna (2010). For \( r = (r_1, ..., r_n)' \) let \( S(r) \) be a scale M estimator defined as solution of

\[
\frac{1}{n} \sum_{i=1}^{n} \rho_0 \left( \frac{r_i}{S} \right) = \delta
\]

where \( \rho_0 \) is a bounded \( \rho \)-function and \( \delta \in (0, 1) \) controls the estimator’s robustness. Then the initial estimator \( \hat{\beta}_{\text{ini}} = (\hat{\beta}_{\text{ini},0}, \hat{\beta}_{\text{ini},1}) \) for (7) is a penalized regression S estimator, defined as follows. For \( \beta = (\beta_0, \beta_1) \) put \( r_i(\beta) = y_i - \beta_0 - z_i^t \beta_1 \). Then

\[
\hat{\beta}_{\text{ini}} = \arg \min_{\beta} \left( n S(r(\beta)) + \lambda' \left\| \beta_1 \right\|^2 \right),
\]

and \( \hat{\sigma}_{\text{ini}} = cS(\beta(\hat{\beta}_{\text{ini}})) \) where \( c \) is a constant that controls the scale’s consistency. The MM estimator is computed by an iterative algorithm starting from \( \hat{\beta}_{\text{ini}} \), 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 \( \rho \) and \( \rho_0 \) in the bisquare family, and we take \( m = 2 \) (cubic splines).

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

The minimization is performed over a set of $N_\lambda$ penalty values: $\lambda' \in \Lambda = \{\lambda_1, ..., \lambda_{N_\lambda}\}$. 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 $\lambda$ is obtained as the average of the squared prediction errors. For our MM estimator, approximate prediction errors can be obtained through a Taylor expansion, and an overall error estimator is obtained as a robust scale estimator of the squared prediction errors; see (Maronna 2010) for details.

For both the $L_2$ and the MM estimator we have observed in both real and simulated data that the estimated $\alpha$ is rather rough. At the same time it was observed the curve of prediction error as a function of $\lambda$ to be very flat around its minimum. This fact leads to the thought that $\lambda$ –and hence the smoothness of $\alpha$– 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 $\mathbf{u}(\lambda) = (u_1(\lambda), ..., u_n(\lambda))'$ the vector of cross–validation prediction errors. For the $L_2$ estimator put

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

where sd stands for 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 $s = s_0(\lambda_{\text{opt}})$. For a given $C_{\text{CV}}$ we take $\hat{\lambda} = \max\{\lambda \in \Lambda : M(\lambda) \leq M(\lambda_{\text{opt}}) + C_{\text{CV}} s_0\}$. We tried $C_{\text{CV}} = 0$, 1 and 2.
For the robust estimator we use a robust version of (8), replacing the average of squares by a robust and efficient scale, namely

\[ M(\lambda) = S\{u_i(\lambda)\}^2, \quad s(\lambda) = \frac{1}{\sqrt{n}}S\{u_i(\lambda)^2 - M(\lambda)\}, \]

where \( S \) is a \( \tau \)-scale with constant 5 (Yohai and Zamar 1998), which may be considered as a robust mean squared error.

### 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 to 500, so that we have \( p = 486 \). The estimators considered where the one 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).

Working in Matlab on a PC with a 3.0 GH Intel Core Duo processor, the computing time of MM was 153 seconds.

We also included for comparison two versions of Partial Least Squares (PLS): the classical one and the robust estimator proposed by Hubert and Vanden Branden (2003), as implemented in the functions csimpls and rsimpls in the library LIBRA (see http://wis.kuleuven.be/stat/robust/LIBRA.html); in both cases the number of components was chosen through cross-validation. The classical and robust versions will be henceforth denoted by C–PLS and R–PLS.
respectively.

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. Figure 1 shows the residual QQ plots for compound 5 ($P_2O_5$). 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) points out an outlier, C–PLS points out none; MM(2) and R–PLS point out about 10 outliers, but the respective configurations are quite different.

The left–hand panel of Figure 2 compares the ordered absolute residuals from MM(2) with those from $L_2$ (2); the dotted line corresponds to the identity. It is seen that the MM residuals are in general smaller than the $L_2$ ones, except for the outliers. 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.
Figure 2: Vessel data: ordered absolute residuals from MM(2) vs. those from $L_2(2)$ (left) and robust PLS (right), together with the identity line (dotted lines).

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 50% higher than those for MM. The difference is attributable to the outliers shown in the former Figures. C–PLS behaves similarly to $L_2$. The behavior of R–PLS is disappointing.
4 Simulation

The estimators were assessed through a reduced simulation study. The estimators considered were the same as in the former Section.

For each scenario, $N_{\text{rep}}$ 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 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.

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 $X_0$ the matrix with elements

$$x_{0,ij} = h(t_j, u_{1i}, u_{2i}) \exp(c_{ij}),$$

Table 1: Vesel data: Prediction CV RMSEs of estimators for compound 5

<table>
<thead>
<tr>
<th>$C_{\text{CV}} = L_2$</th>
<th>MM</th>
<th>C–PLS</th>
<th>R–PLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE</td>
<td>4.30</td>
<td>3.98</td>
<td>9.51</td>
</tr>
<tr>
<td>RMSE(0.9)</td>
<td>3.03</td>
<td>2.77</td>
<td>1.83</td>
</tr>
</tbody>
</table>
where \( u_{1i} \) and \( u_{2i} \) are independent uniform variables in \((0, 0.4)\) and \((0.5, 0.9)\) respectively, and for each \( i \), \((v_{i1}, \ldots, v_{ip})\) 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 \( \rho = 0.9 \). Let \( \alpha(t) = \exp\left[ -\left( (t - 0.25)/0.3 \right)^2 \right] \) and call \( \alpha \) the vector with elements \( \alpha(t_j) \). Figure 3 plots \( \alpha \) and two typical rows of \( X_0 \).

Figure 3: Simulation: “True” function \( \alpha(t) \) and two typical rows of \( X_0 \).

Let \( y_0 = X_0 \alpha \). In the case of no contamination, the data are \((X, y)\) with \( X = X_0 \) and \( y = y_0 + \sigma e \), where \( e \) has i.i.d. \( N(0, 1) \) elements. For the value of \( \sigma \), we choose a “noise-to-signal” parameter NSR, and then \( \sigma = \text{NSR} \times \text{std}(y_0) \).

The values of NSR were chosen as 0.02, 0.05 and 0.1, the choice of this range being guided by examination of real data. In the case of a contamination rate \( \varepsilon \), let \( m = \lfloor n \varepsilon \rfloor \) where \( \lfloor \cdot \rfloor \) stands for the integer part. Then the first rows of \( X \) are multiplied by \( k_{\text{lev}} \), where \( k_{\text{lev}} \) is a parameter that regulates the outlier leverage and is fixed at 2; and for \( i = 1, \ldots, m \) we put \( y_i = y_{i0} k_{\text{lev}} k_{\text{slo}} \), where the parameter \( k_{\text{slo}} \) regulates the outliers’ slope. The effect of this contamination is
to pull the estimator $\hat{\alpha}$ towards $k_{slo} \alpha$. 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 $\hat{\alpha}$, let for a given sample $\text{MPE} = n^{-1} \| y_0 - X_0 \hat{\alpha} \|^2$, which measures the mean squared prediction error of $\hat{\alpha}$ if applied to the uncontaminated data. The $N_{\text{rep}}$ 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. For this reason we preferred to employ the 10% (upper) trimmed average, henceforth referred to as "trimmed MPE". Table 2 displays the simulation results, where the trimmed MPEs have been multiplied by 100 to improve legibility.

<table>
<thead>
<tr>
<th>NSR</th>
<th>$\varepsilon$</th>
<th>$\mathbb{L}_2$</th>
<th>MM</th>
<th>PLS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$C_{CV}$ = 0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>$p = 25$, $n = 50$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.02</td>
<td>0.57</td>
<td>0.49</td>
<td>0.60</td>
<td>0.68</td>
</tr>
<tr>
<td>0.1</td>
<td>11.18</td>
<td>10.76</td>
<td>10.75</td>
<td>1.96</td>
</tr>
<tr>
<td>0.05</td>
<td>0.35</td>
<td>1.12</td>
<td>1.40</td>
<td>1.58</td>
</tr>
<tr>
<td>0.1</td>
<td>30.85</td>
<td>29.05</td>
<td>28.39</td>
<td>4.63</td>
</tr>
<tr>
<td>0.1</td>
<td>32.12</td>
<td>29.75</td>
<td>28.93</td>
<td>8.22</td>
</tr>
<tr>
<td>$p = 50$, $n = 30$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.02</td>
<td>2.35</td>
<td>2.07</td>
<td>2.43</td>
<td>2.73</td>
</tr>
<tr>
<td>0.1</td>
<td>40.00</td>
<td>41.59</td>
<td>44.67</td>
<td>6.62</td>
</tr>
<tr>
<td>0.05</td>
<td>5.78</td>
<td>4.94</td>
<td>5.78</td>
<td>6.58</td>
</tr>
<tr>
<td>0.1</td>
<td>43.42</td>
<td>44.26</td>
<td>47.12</td>
<td>16.18</td>
</tr>
<tr>
<td>0.1</td>
<td>11.45</td>
<td>9.35</td>
<td>11.23</td>
<td>13.39</td>
</tr>
<tr>
<td>0.1</td>
<td>78.30</td>
<td>78.83</td>
<td>83.44</td>
<td>29.69</td>
</tr>
</tbody>
</table>

Table 2: Simulation: trimmed MPEs ($\times 100$)

In general we observe that $L_2$ with $C_{CV} = 1$ (henceforth “$L_2 (1)$”) is the most efficient estimator for $\varepsilon = 0$, and MM(1) the most efficient one among MM, with MM(2) close by. But MM(2) is clearly the most robust one. The efficiencies of MM(1) with respect to $L_2 (1)$ range between 75% and 81%, which are not far from the nominal 0.85. When $p = 25$, C–PLS is rather efficient, while
R–PLS exhibits low efficiency and low robustness. When $p = 50$ both versions of PLS exhibit very low efficiency and no robustness, since under contamination their behavior is as bad as that of $L_2$. We also tried using R–PLS with several fixed values of the number of components, but its performance did not improve. We performed some exploratory simulations which indicated that R–PLS works reasonably well when $p \ll n$ but fails when $p \gg n$; these features appeared also in the simulations of (Maronna 2010).

In the trade-off between efficiency and robustness, it seems that MM(2) is the estimator of choice.

ACKNOWLEDGEMENTS

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,

REFERENCES


Yohai, V.J. and Zamar, R., 1988, High breakdown point estimates of re-
gression by means of the minimization of an efficient scale, Jr. Amer. Statist. Assoc., 86. 403-413.