Chapter 14
Parameter estimation of the functional linear model with scalar response with responses missing at random

Manuel Febrero-Bande, Pedro Galeano and Wenceslao González-Manteiga

Abstract This contribution considers estimation of the parameters of the functional linear model with scalar response when some of the responses are missing at random. We consider two different estimation methods of the functional slope of the model and analyze their characteristics. Simulations and the analysis of a real data example provides some insight into the behavior of both estimation procedures.

14.1 Introduction

The functional linear model with scalar response is one of the most widely studied model in the literature on functional data analysis. The model establishes a linear relationship between a real response variable and a functional predictor variable. There exist several estimators of the functional slope of the model being the method based on functional principal components the most popular approach. The idea behind this method is that of expanding the functional predictor as well as the functional slope of the model in terms of the eigenfunctions linked to the largest eigenvalues of the functional predictor covariance operator, that allows the response to be written as a finite linear combination of the functional principal components scores. The associated coefficients are then estimated by least squares.

Manuel Febrero-Bande
Department of Statistics, Mathematical Analysis and Optimization, Universidade de Santiago de Compostela, e-mail: manuel.febrero@usc.es

Pedro Galeano (✉)
Department of Statistics and UC3M-BS Institute of Financial Big Data, Universidad Carlos III de Madrid e-mail: pedro.galeano@uc3m.es

Wenceslao González-Manteiga
Department of Statistics, Mathematical Analysis and Optimization, Universidade de Santiago de Compostela, e-mail: wenceslao.gonzalez@usc.es

G. Aneiros et al. (eds.), *Functional Statistics and Related Fields*,
Contributions to Statistics, DOI 10.1007/978-3-319-55846-2_14

Several papers have analyzed the properties of the functional principal components estimation method including [1, 2, 3, 4, 7, 9, 5], among others. See also [6], for a recent overview on the topic.

This contribution considers the case in which some of the responses are missing at random. This case has been little studied in the literature. [10] investigated the asymptotic properties of a kernel type estimator of the regression operator when there are responses missing at random, while [5] considered an imputation method of the missing responses. Here, we propose two estimators of the functional slope of the model. The first one is simplified estimator that only considers the complete pairs of observations. The second one is an imputed estimator that takes into account both the complete pairs and pairs completed with imputed responses.

The rest of this contribution is structured as follows. Section 14.2 presents the functional linear model with scalar response and the estimation method based on the functional principal components approach. Section 14.3 considers the problem of estimating the parameters of the model when there are responses that are missing at random and presents the two estimators that we propose of the functional slope of the model. Properties of the estimators, simulations and the analysis of real data are presented somewhere else.

14.2 The functional linear model with scalar response

Let $L^2(T)$, the separable Hilbertian space of squared integrable functions defined on the closed interval $T = [a,b] \subset \mathbb{R}$. Let χ be a functional random variable valued in $L^2(T)$ and let $\chi(t)$ be the value of χ at any point $t \in T$. We assume, for simplicity, that the functional random variable χ has zero mean function and a covariance operator Γ such that:

$$\Gamma(\eta) = E[(\chi \otimes \chi)(\eta)] = E[\langle \chi, \eta \rangle \chi]$$

for any $\eta \in L^2(T)$, where,

$$\langle \chi, \eta \rangle = \int_T \chi(t)\,\eta(t)\,dt$$

is the usual inner product in $L^2(T)$. We also assume that $E\left[\|\chi\|^2\right] < \infty$, where $\|\cdot\|$ denotes the usual norm in $L^2(T)$. Consequently, Γ has a sequence of non-negative eigenvalues, denoted by $a_1 > a_2 > \cdots > 0$, such that $\sum_{k=1}^{\infty} a_k < \infty$, associated with a sequence of orthonormal eigenfunctions, denoted by $\psi_1, \psi_2, \ldots$, such that $\Gamma(\psi_k) = a_k \psi_k$, for $k = 1, 2, \ldots$

The functional linear model with scalar response relates a real random variable y, defined on the same probability space that χ, with mean 0 and variance σ_y^2, with χ as follows:

$$y = \langle \chi, \beta \rangle + e = \int_T \chi(t)\,\beta(t)\,dt + e, \tag{14.1}$$

where $\beta \in L^2(T)$ is the functional slope of the model, and e is a real random variable with mean 0, finite variance σ_e^2, and uncorrelated with χ. In other words, we assume that the mean and variance of y conditional on χ are given by $E_\chi[y] = \langle \chi, \beta \rangle$ and $Var_\chi[y] = \sigma_e^2$, respectively.

As mentioned in the introduction, the functional principal components estimation method is the most popular approach to estimate the functional slope β of the model in (14.1). This is because the functional principal components allows the functional linear model to be more easily written. The functional principal components scores, given by $s_k = \langle \chi, \psi_k \rangle$, for $k = 1, 2 \ldots$, are uncorrelated univariate random variables with mean 0 and variance a_k that allows the Karhunen-Loève expansion of the functional random variable χ to be written as follows:

$$\chi = \sum_{k=1}^{\infty} s_k \psi_k. \tag{14.2}$$

Similarly, the functional slope β can be also written in terms of the eigenfunctions $\psi_1, \psi_2, \ldots$ as:

$$\beta = \sum_{k=1}^{\infty} b_k \psi_k, \tag{14.3}$$

where $b_k = \langle \beta, \psi_k \rangle$, for $k = 1, 2 \ldots$ are constant coefficients. Now, (14.2) and (14.3) allows the functional linear model with scalar response to be written as:

$$y = \sum_{k=1}^{\infty} b_k s_k + e,$$

which shows that the coefficients b_k can be written as:

$$b_k = \frac{Cov[y, s_k]}{a_k}, \tag{14.4}$$

for $k = 1, 2 \ldots$ where $Cov[y, s_k] = E[y s_k]$ is the covariance between the real response y and the k-th functional principal component score s_k.

Assume now that we are given a random sample of independent pairs, given by $\{(\chi_i, y_i),\ i = 1, \ldots, n\}$, drawn from the random pair (χ, y). Then, the functional slope β in the model (14.1) can be estimated with the functional principal component estimation method as follows. Let $\chi_C = \{\chi_1, \ldots, \chi_n\}$ and $y_C = \{y_1, \ldots, y_n\}$ be the complete sequences of predictors and responses, respectively. Then, the sample covariance operator of the complete sample χ_C, that converts any function $\eta \in L^2(T)$ into another function in $L^2(T)$ given by:

$$\widehat{\Gamma}_{\chi_C}(\eta) = \frac{1}{n} \sum_{i=1}^{n} \langle \chi_i, \eta \rangle \chi_i,$$

is an estimate of the covariance operator of χ, Γ. The sample covariance operator $\widehat{\Gamma}_{\chi_C}$ also has a sequence of non-negative eigenvalues, denoted by $\widehat{a}_{1,C} \geq \widehat{a}_{2,C} \geq \cdots$,

such that $\widehat{a}_{k,C} = 0$, for $k > n$, and a set of orthonormal eigenfunctions, denoted by $\widehat{\psi}_{1,C}, \widehat{\psi}_{2,C}, \ldots$, such that $\widehat{\Gamma}_{\chi_C}\left(\widehat{\psi}_{k,C}\right) = \widehat{a}_{k,C}\widehat{\psi}_{k,C}$, for $k = 1, 2, \ldots$ Additionally, the k-th sample functional principal component score of χ_i, $i = 1, \ldots, n$, based on the complete sample χ_C, is given by $\widehat{s}_{i,k,C} = \left\langle \chi_i, \widehat{\psi}_{k,C} \right\rangle$, for $k = 1, 2, \ldots$ The set of sample functional scores $\widehat{s}_{1,k,C}, \ldots, \widehat{s}_{n,k,C}$ has sample mean 0 and sample variance $\widehat{a}_{k,C}$. Now, the functional principal components estimate of the functional slope β is given by:

$$\widehat{\beta}_{k_C,C} = \sum_{k=1}^{k_C} \widehat{b}_{k,C}\widehat{\psi}_{k,C}, \tag{14.5}$$

where $\widehat{b}_{k,C}$ is an estimate of the coefficient b_k in (14.4) given by:

$$\widehat{b}_{k,C} = \begin{cases} \frac{1}{n\widehat{a}_{k,C}} \sum_{i=1}^{n} y_i \widehat{s}_{i,k,C} & \text{for } k = 1, \ldots, k_C \\ 0 & \text{for } k = k_C + 1, \ldots \end{cases}$$

and k_C is a certain threshold such that $\widehat{a}_{k_C,C} > 0$. Consequently, given a new value χ, say χ_{n+1}, the prediction of the corresponding response under the model (14.1), denoted by y_{n+1}, is given by:

$$\widehat{y}_{n+1,k_C,C} = \left\langle \chi_{n+1}, \widehat{\beta}_{k_C,C} \right\rangle.$$

See [9], [7] and [6], for finite sample properties of the slope estimate (14.5).

14.3 Estimation and prediction with responses missing at random

Assume now the situation in which we are given a random sample of independent triplets $\{(\chi_i, y_i, r_i), i = 1, \ldots, n\}$ drawn from the random triplet (χ, y, r), where r is a Bernoulli variable that acts as an indicator of the missing responses. Thus, for $i = 1, \ldots, n$, $r_i = 1$, if y_i is observed, and $r_i = 0$, if y_i is missing. Specifically, we assume a missing at random (MAR) mechanism, i.e.:

$$\Pr(r = 1|y, \chi) = \Pr(r = 1|\chi) = p(\chi),$$

where $p(\chi)$ is an unknown function operator of χ. As a consequence, the response y and the binary variable r are independent given the predictor χ.

Now, the goal is to estimate the functional slope β in (14.1) using the sample $\{(\chi_i, y_i, r_i), i = 1, \ldots, n\}$. For that, let $r_C = (r_1, \ldots, r_n)'$ be the complete sequence of missing indicators. Two different estimates are introduced next based on the missing indicators r_C.

The first estimate is the simplified functional principal component estimate, that uses only the complete pairs, i.e., those pairs with $r_i = 1$, for $i = 1, \ldots, n$. Then, let

$I_S = \{i : r_i = 1, i = 1, \ldots, n\}$, i.e., the indices of the complete pairs and let $n_S = \#I_S$, i.e., the number of observed complete pairs. Additionally, let $\chi_S = \{\chi_i : i \in I_S\}$ and $y_S = \{y_i : i \in I_S\}$, i.e., the sequences of predictors and responses, respectively, corresponding to the complete pairs. Then, the sample covariance operator of χ_S, that converts any function $\eta \in L^2(T)$ into another function in $L^2(T)$ given by:

$$\widehat{\Gamma}_{\chi_S}(\eta) = \frac{1}{n_S} \sum_{i=1}^{n} r_i \langle \chi_i, \eta \rangle \chi_i = \frac{1}{n_S} \sum_{i \in I_S} \langle \chi_i, \eta \rangle \chi_i,$$

is an estimate of Γ. As in the complete case developed in Section 14.2, $\widehat{\Gamma}_{\chi_S}$ has a sequence of non-negative eigenvalues, denoted by $\widehat{a}_{1,S} \geq \widehat{a}_{2,S} \geq \cdots$, such that $\widehat{a}_{k,S} = 0$, for $k > n_S$, and a set of orthonormal eigenfunctions, denoted by $\widehat{\psi}_{1,S}, \widehat{\psi}_{2,S}, \ldots$, such that $\widehat{\Gamma}_{\chi_S}\left(\widehat{\psi}_{k,S}\right) = \widehat{a}_{k,S}\widehat{\psi}_{k,S}$, for $k = 1, 2, \ldots$ Additionally, the k-th sample functional principal component score for χ_i, $i \in I_S$, based on the simplified sample χ_S, is given by $\widehat{s}_{i,k,S} = \left\langle \chi_i, \widehat{\psi}_{k,S} \right\rangle$, for $k = 1, 2, \ldots$ The set of sample functional components scores $\left\{\widehat{s}_{i,k,S} : i \in I_S\right\}$ have sample mean 0 and sample variance $\widehat{a}_{k,S}$. Now, the simplified functional component estimate of the functional slope β is given by:

$$\widehat{\beta}_{k_S,S} = \sum_{k=1}^{k_S} \widehat{b}_{k,S}\widehat{\psi}_{k,S}, \tag{14.6}$$

where $\widehat{b}_{k,S}$ is an estimate of the coefficient b_k in (14.4) given by:

$$\widehat{b}_{k,S} = \begin{cases} \frac{1}{n_S \widehat{a}_{k,S}} \sum\limits_{i \in I_S} y_i \widehat{s}_{i,k,S} & \text{for } k = 1, \ldots, k_S \\ 0 & \text{for } k = k_S + 1, \ldots \end{cases}$$

and k_S is a certain threshold such that $\widehat{a}_{k_S,S} > 0$. Prediction of the response y_{n+1} corresponding to a new predictor χ_{n+1} under the model (14.1), is given by:

$$\widehat{y}_{n+1,k_S,S} = \left\langle \chi_{n+1}, \widehat{\beta}_{k_S,S} \right\rangle.$$

The second estimate is the imputed functional principal component estimate, that uses both the complete pairs and the pairs obtained after imputing the missing responses with the estimate (14.6). Then, let $I_I = \{i : r_i = 0, i = 1, \ldots, n\}$, i.e., the indices of the pairs with missing responses and let $n_I = \#I_I$, i.e., the number of pairs with missing responses. Additionally, let $\chi_I = \{\chi_i : i \in I_I\}$ and $y_I = \{y_i : i \in I_I\}$, i.e., the sequences of predictors and responses, respectively, corresponding to the pairs with missing responses. Therefore, imputation of the missing responses using the simplified estimate $\widehat{\beta}_{k_S,S}$ in (14.6) can be done as follows:

$$\widehat{y}_{i,I} = \left\langle \chi_i, \widehat{\beta}_{k_S,S} \right\rangle,$$

for $i \in I_I$. Now, given the set of pairs $\{(\chi_i, y_{i,I}), i = 1, \ldots, n\}$ where:

$$y_{i,I} = r_i y_i + (1 - r_i)\widehat{y}_{i,I},$$

for $i = 1, \ldots, n$, the imputed functional principal component estimate of the functional slope β is given by:

$$\widehat{\beta}_{k_I,I} = \sum_{k=1}^{k_I} \widehat{b}_{k,I}\widehat{\psi}_{k,C}, \tag{14.7}$$

where $\widehat{b}_{k,I}$ is an estimate of the coefficient b_k in (14.4) given by:

$$\widehat{b}_{k,I} = \begin{cases} \frac{1}{n\widehat{a}_{k,C}} \sum_{i=1}^{n} y_{i,I}\widehat{s}_{i,k,C} & \text{for } k = 1, \ldots, k_I \\ 0 & \text{for } k = k_I + 1, \ldots \end{cases}$$

and k_I is a certain threshold such that $\widehat{a}_{k_I,C} > 0$. Two important comments are in order. First, $\widehat{\beta}_{k_I,I}$ depends on the eigenfunctions and eigenvalues of the sample covariance operator $\widehat{\Gamma}_{\chi_C}$ based on the complete set of predictors χ_C. Second, the threshold k_I in (14.7) does not necessarily coincides with the threshold k_S in (14.6). Prediction of the response y_{n+1} corresponding to a new predictor χ_{n+1} under the model (14.1), is given by:

$$\widehat{y}_{n+1,k_I,I} = \left\langle \chi_{n+1}, \widehat{\beta}_{k_I,I} \right\rangle.$$

Acknowledgements The first and third author acknowledges financial support from Ministerio de Economía y Competitividad grant MTM2013-41383-P. The second author acknowledges financial support from Ministerio de Economía y Competitividad grant ECO2015-66593-P.

References

[1] Cai, T.T., Hall, P.: Prediction in functional linear regression. Ann. Stat. **34**, 2159–2179 (2006)
[2] Cardot, H., Ferraty, F., Sarda, P.: Functional linear model. Stat. Probabil. Lett. **45**, 11–22 (1999)
[3] Cardot, H., Ferraty, F., Sarda, P.: Spline estimators for the functional linear model. Stat. Sinica. **13**, 571–591 (2003)
[4] Cardot, H., Mas, A., Sarda, P.: CLT in functional linear regression models. Probabil. Theory and Relat. Fields. **138**, 325–361 (2007)
[5] Crambes, C., Henchiri, Y.: Regression imputation in the functional linear model with missing values in the response. Manuscript.
[6] Febrero-Bande, M., Galeano, P., González-Manteiga, W.: Functional principal component regression and functional partial least-squares regression: an overview and a comparative study. Int. Stat. Rev. (2016) doi: 10.1111/insr.12116
[7] Ferraty, F., González-Manteiga, W., Martínez-Calvo, A., Vieu, P.: Presmoothing in functional linear regression. Stat. Sinica. **22**, 69–94 (2012)

[8] Hall, P., Horowitz, J. L.: Methodology and convergence rates for functional linear regression. Ann. Stat. **35**, 70–91 (2007)
[9] Hall, P., Hosseini-Nasab, M.: On properties of functional principal components analysis. J. Roy. Stat. Soc. B **68**, 109–126 (2006)
[10] Ling, N., Ling, L., Vieu, P.: Nonparametric regression estimation for functional stationary ergodic data with missing at random. J. Stat. Plan. Infer. **162**, 75–87 (2015)