Journal of Multivariate Analysis 151 (2016) 54–68

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Best estimation of functional linear models

Giacomo Aletti [a,b], Caterina May [c,*], Chiara Tommasi [b]

[a] ADAMSS Center, Milan, Italy
[b] Università degli Studi di Milano, Milan, Italy
[c] Università del Piemonte Orientale, Novara, Italy

ARTICLE INFO

Article history:
Received 16 July 2015
Available online 29 July 2016

AMS subject classifications:
primary 62H12
secondary 47B06
62M15
62M99

Keywords:
Functional data analysis
Sobolev spaces
Linear models
Repeated measurements
Gauss–Markov theorem
Riesz representation theorem
Best linear unbiased estimator

ABSTRACT

Observations that are realizations of some continuous process are frequently found in science, engineering, economics, and other fields. In this paper, we consider linear models with possible random effects and where the responses are random functions in a suitable Sobolev space. In particular, the processes cannot be observed directly. By using smoothing procedures on the original data, both the response curves and their derivatives can be reconstructed, both as an ensemble and separately. From these reconstructed functions, one representative sample is obtained to estimate the vector of functional parameters. A simulation study shows the benefits of this approach over the common method of using information either on curves or derivatives. The main theoretical result is a strong functional version of the Gauss–Markov theorem. This ensures that the proposed functional estimator is more efficient than the best linear unbiased estimator (BLUE) based only on curves or derivatives.

1. Introduction

Observations which are realizations of some continuous process are ubiquitous in many fields like science, engineering, economics and other fields. For this reason, the interest in statistical modelling of functional data is increasing, with applications in many areas. Reference monographs on functional data analysis are, for instance, the books of Ramsay and Silverman [11] and Horváth and Kokoszka [7], and the book of Ferraty and Vieu [5] for the non-parametric approach. They cover topics like data representation, smoothing and registration; regression models; classification, discrimination and principal component analysis; derivatives and principal differential analysis; and many other.

Regression models with functional variables can cover different situations; for example, functional responses, or functional predictors, or both. In this paper, linear models with functional response and multivariate (or univariate) regressor are examined. We consider the case of repeated measurements, where the theoretical results remain valid in the standard case. The focus of the work is to find the best estimation of the functional coefficients of the regressors.

The use of derivatives is very important in exploratory analysis of functional data; as well as for inference and prediction methodologies. High quality derivative information may be determined, for instance, by reconstructing the functions with spline smoothing procedures. Recent developments in the estimation of derivatives are contained in Sangalli et al. [12] and

* Corresponding author.
E-mail addresses: giacomo.aletti@unimi.it (G. Aletti), caterina.may@uniupo.it (C. May), chiara.tommasi@unimi.it (C. Tommasi).

http://dx.doi.org/10.1016/j.jmva.2016.07.005
0047-259X/

in Pigoli and Sangalli [10]. See also Baraldo et al. [3], who have obtained derivatives in the context of survival analysis, and Hall et al. [6] who have estimated derivatives in a non-parametric model.

Curves and derivatives are reconstructed from a set of observed values. The reason for this is that the response processes cannot be observed directly. In the literature, the usual space for functional data is L^2 and the observed values are used to reconstruct either curve functions or derivatives.

The most common method to reconstruct derivatives is to build the sample of functions by using a smoothing procedure on the data, and then to differentiate these curve functions. However, the sample of functions and the sample of derivatives may be obtained separately. For instance, different smoothing techniques may be used to obtain the functions and the derivatives. Another possibility is when two sets of data are available, which are suitable to estimate functions and derivatives, respectively.

Some examples of curve and derivative data are: studying how the velocity of a car on a particular street is influenced by some covariates, the velocity measured by a police radar; GPS-tracked position estimation. In chemical experiments, data on reaction velocity and concentration may be collected separately. The novelty of the present work is that both information on curves and derivatives (that are not obtained by differentiation of the curves themselves) are used to estimate the functional coefficients.

The heuristic justification for this choice is that the data may provide different information about curve functions and their derivatives; it is therefore always recommended to use all available information. In fact, in this paper we prove that if we take into consideration both information about curves and their derivatives, we obtain the best linear unbiased estimates for the functional coefficients. Therefore, the common method of using information on either curve functions or their associated derivatives provides always a less efficient estimate (see Theorem 3 and Remark 2). For this reason, our theoretical results may have a relevant impact in practice.

Analogous to the Riesz Representation Theorem, we can find a representative function in H^1 which incorporates the information provided by a curve function and a derivative (which belong to L^2). Hence, from the two samples of reconstructed functions and their associated derivatives, only one representative sample is obtained and we use this representative sample to estimate the functional parameters. Once this method is given, the consequential theoretical results may appear as a straightforward extension of the well-known classical ones; their proof, however, requires much more technical effort and is not a straightforward extension.

The OLS estimator (based on both curves and derivatives through their Riesz representatives in H^1) is provided and some practical considerations are drawn. In general, the OLS estimator is not a BLUE because of the possible correlation between curves and derivatives. Therefore, a different representation of the data is provided (which takes into account this correlation). The resulting version of the Gauss–Markov theorem is proven in the proper infinite-dimensional space (H^1), showing that our sample of representative functions carries all the relevant information on the parameters. We propose an unbiased estimator which is linear with respect to the new sample of representatives and which minimizes a suitable covariance matrix (called global variance). This estimator is denoted H^1-functional SBLUE.

A simulation study numerically demonstrates the superiority of the H^1-functional SBLUE with respect to both the OLS estimators which are based only on curves or derivatives. This suggests that both sources of information should be used jointly, when available. A rough way of considering information on both curves and derivatives is to make a convex combination of the two OLS estimators. Simulation results show that the H^1-functional SBLUE is more efficient, as expected.

Let us finally remark that the results in this paper provide a strong theoretical foundation to generalize the theory of optimal design of experiments when functional observations occur (see Aletti et al. [1,2]).

The paper is organized as follows. Section 2 describes the model and proposes the OLS estimator obtained from the Riesz representation of the data. Section 3 explains some considerations which are fundamental from a practical point of view. Section 4 presents the construction of the H^1-functional SBLUE. Finally, Section 5 is devoted to the simulation study. Section 6 is a summary together with some final remarks. Some additional results and the proofs of theorems are deferred to Appendix A.1.

2. Model description and Riesz representation

Let us first consider a regression model where the response y is a random function that depends linearly on a known variable $\mathbf{x}$, which is a vector (or scalar) through a functional coefficient, that needs to be estimated. In particular, we assume that there are n units (subjects or clusters), and $r \geq 1$ observations per unit at a condition $\mathbf{x}_i$ ($i = 1, \dots, n$). Note that $\mathbf{x}_1, \dots \mathbf{x}_n$ are not necessarily different. In the context of repeated measurements, we consider the following random effect model:

$$y_{ij}(t) = \mathbf{f}(\mathbf{x}_i)^\top \boldsymbol{\beta}(t) + \alpha_i(t) + \varepsilon_{ij}(t) \quad i = 1, \dots, n; \; j = 1, \dots, r, \tag{1}$$

where: t belongs to a compact set $\tau \subseteq \mathbb{R}$; $y_{ij}(t)$ denotes the response curve of the jth observation at the ith experiment; $\mathbf{f}(\mathbf{x}_i)$ is a p-dimensional vector of known functions; $\boldsymbol{\beta}(t)$ is an unknown p-dimensional functional vector; $\alpha_i(t)$ is a zero-mean process which denotes the random effect due to the ith experiment and takes into account the correlation among the r repetitions; $\varepsilon_{ij}(t)$ is a zero-mean error process. Let us note that we are interested in precise estimation of the fixed effects $\boldsymbol{\beta}(t)$; herein the random effects are nuisance parameters.