

Chapter 1

Some advances on semi-parametric functional data modelling

Metodi Semi-parametrici per Modelli di Regressione Funzionale

Aldo Goia

Abstract The aim of this work is to present some recent advances in single index modelling when the covariate is a functional variable and the response a scalar. We pay special attention to the situation of possible structural changes which produce unsmooth relevant direction. An estimation procedure combining spline functions and the well-known Nadaraya-Watson approach is illustrated. From an example of interest in the spectrometry, it emerges that the method provides a nice exploratory tool both for analyzing structural changes in the spectrum and for visualizing the most informative directions, still keeping good predictive power.

Key words: additive models; breaking-point; projection pursuit regression; single-index model.

Introduction

In the multivariate regression setting, semi-parametric techniques had been introduced in order to balance the trade-off between very few flexibility of pure parametric modelling and dimensional effects of non-parametric approaches. For a general presentation of semi-parametric ideas in multivariate situations one can refer for instance to [11] and references therein.

In the functional context (see e.g. the books [8] and [14]) one deals with variables belonging to infinite-dimensional spaces (mainly curves but also images, arrays and other complex data). The nature of the problem emphasizes the usual drawbacks described before: from one hand the lacking flexibility of the linear model is much more problematic, and on the other hand the negative impact of the dimensionality on non-parametric approaches appears dramatically. Therefore the semi-parametric

Aldo Goia
University of Eastern Piedmont, e-mail: aldo.goia@eco.unipmn.it

ideas have been extended to the functional framework since a few years, in order to construct models combining both flexibility and dimensionality reduction. In the functional regression context, recent advances in such direction involve for instance the functional single index model (see e.g. [1], [6] or [7]), the functional projection pursuit regression (see [3] or [4]) and functional additive modelling (see e.g., [12] or [13]).

The aim of this work is to focus on one of the simplest semi-parametric functional model, namely the single index modelling, in situations where the relevant functional index is not necessarily smooth due to structural changes in the regression operator. This model will be presented and discussed in Section 1.1. The methodology for constructing estimates of the various components of the model will be developed in Section 1.2. A real data analysis will be carried out along Section 1.3.

1.1 Single functional index model with breaking-points

Consider the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and $\mathcal{L}_{\mathcal{J}}^2$, the Hilbert space of square integrable real functions on $\mathcal{J} = [a, b]$ endowed with the inner product $\langle g, h \rangle = \int_{\mathcal{J}} g(t) h(t) dt$ and the induced norm $\|g\|^2 = \langle g, g \rangle$. Define on Ω the $\mathcal{L}_{\mathcal{J}}^2$ valued random function (r.f.) χ and the real random variable (r.v.) Y . In the Single Index Model approach one postulates that χ acts on Y only through its projection on some single index function $\theta = \theta(t), t \in \mathcal{J}$, such that, for identifiability reasons, $\|\theta\|^2 = 1$ and $\langle \theta, e_1 \rangle = 1$ where e_1 is the first element of a basis of $\mathcal{L}_{\mathcal{J}}^2$. In a general form, the model can be written as

$$Y = \alpha + g(\langle \chi, \theta \rangle) + \varepsilon \quad (1.1)$$

where α is a real constant, ε is a centered real random error uncorrelated with χ . Such a model and estimation techniques for the link function g and the functional direction θ have been studied intensively in the literature: see for instance, [2], [6] or [7].

Suppose now that specific parts of the curve χ operate in a different semi-parametric way for explaining the response Y . This allows us to consider a partition of $\mathcal{J}$ in s sub-intervals $\mathcal{J}_j, j = 1, \dots, s$ (with $\mathcal{J}_1 = [a, \lambda_1], \mathcal{J}_j = (\lambda_{j-1}, \lambda_j], j = 2, \dots, s-1$, and $\mathcal{J}_s = (\lambda_{s-1}, b]$), and to define the restriction θ_j of θ to $\mathcal{J}_j$, which are continuous squared integrable functions on $\mathcal{J}_j$. Moreover, to generalize the above identifiability conditions, we assume that the directions θ_j satisfy

$$\int_{\mathcal{J}_j} \theta_j^2(t) dt = 1 \quad \text{and} \quad \int_{\mathcal{J}_j} \theta_j(t) e_1(t) dt = \int_{\mathcal{J}_h} \theta_h(t) f_1(t) dt = 1 \quad j \neq h$$

where e_1 and f_1 are the first elements of some orthonormal bases of $L^2(\mathcal{J}_j)$ and $L^2(\mathcal{J}_h)$ respectively.

Assuming that g acts additively on the projections of χ over θ_j 's, the SIM model with breaking-points can be written as:

$$Y = \alpha + \sum_{j=1}^s g_j \left(\int_{\mathcal{I}_j} \chi(t) \theta_j(t) dt \right) + \varepsilon \quad (1.2)$$

where α is a real constant, g_j is an unknown real link function and ε is a centered real r.v. uncorrelated with the regressor. When $s = 2$ one can refer to the recent work [10]. Note that the proposed approach can play an important role in the selection of the most important parts of the functional covariate that explain the variability of the scalar response. This aspect of relevant practical interest will be highlighted through real data analysis in Section 1.3.

1.2 Estimation procedure

Let $\{(\chi_i, Y_i), i = 1, \dots, n\}$ be a sample of i.i.d. replications of (χ, Y) , and let $\{(x_i, y_i), i = 1, \dots, n\}$ the corresponding observed values. In what follows we introduce an estimation procedure for model (1.2): first we consider the case where the break points are known, then we introduce a method to find break points.

Break points are known. Suppose that λ_{js} (and then s) are known: for each $\mathcal{I}_j$ we define a suitable space of spline functions with degree q_j and with $k_j - 1$ interior equispaced knots ($q_j > 2$ and $k_j > 1$, integers) and denote by $\{B_{j,l_j}(t), l_j = 1, \dots, q_j + k_j\}$ the normalized B-splines basis of such spaces. In such basis $\theta_j(t)$ can be represented as $\theta_j(t) \approx \delta_j^T \mathbf{B}_j(t)$, where $\mathbf{B}_j(t)$ is the vector of all the B-splines. In order to remove trivial ambiguity, the vector δ_j of coefficients is such that its first element is positive, and satisfies the normalization condition $\delta_j^T \int_{\mathcal{I}_j} \mathbf{B}_j(t) \mathbf{B}_j(t)^T dt \delta_j = 1$. To estimate the model (1.2) we introduce the following backfitting algorithm:

INITIALIZE – $\hat{\alpha} = n^{-1} \sum_{i=1}^n y_i$, $\hat{g}_j(u) = 0$, $\hat{\delta}_j = \mathbf{0}$, for $j = 1, \dots, s$.

CYCLE – For $j = 1, \dots, s$, find $\hat{\delta}_j$ which minimizes

$$CV_j(\mathbf{d}) = \frac{1}{n} \sum_{i=1}^n \left(y_i - \hat{\alpha} - \sum_{h \neq j} \hat{g}_h \left(\hat{\delta}_h^T \mathbf{b}_{h,i} \right) - \hat{g}_j^{[-i]} \left(\mathbf{d}^T \mathbf{b}_{j,i} \right) \right)^2$$

where $\mathbf{b}_{j,i} = \langle \mathbf{B}_j, x_i \rangle$ and $\hat{g}_j^{[-i]}$ is the leave-one-out Nadarya-Watson kernel estimate of the link function g_j , excluding the i -th observation.

Compute $\hat{g}_j$ from $\hat{g}_j^{[-i]}$ by subtracting its empirical mean.

Cycle the procedure until stabilization of the quadratic criterion $\sum_{i=1}^n (y_i - \hat{y}_i)^2$, where $\hat{y}_i = \hat{\alpha} + \sum_j \hat{g}_j(\hat{\delta}_j^T \mathbf{b}_{j,i})$.

Data-driven break point selection. The previous procedure is defined for fixed values of the parameters λ_{js} : we propose a method to choose them in practice. The idea is to use a cascade algorithm which provides a partitioning tree based on the minimal prediction error. In brief, this algorithm work as follows:

INITIALIZE – State $\mathcal{I}^* = \mathcal{I}$.
 STEP 1 – Computing $\hat{\lambda}_{opt}$.
 Using the backfitting algorithm introduced above, estimate the model (1.2) restricted to $\mathcal{I}^*$ with $s = 2$ for every $\ell \in \Lambda \subset \mathcal{I}^*$ (a grid of possible candidates), and compute the cross-validated prediction error $CV(\ell)$. Choose $\hat{\lambda}_{opt}$ which minimizes $CV(\ell)$.
 STEP 2 – Splitting the interval $\mathcal{I}^*$.
If $CV(\hat{\lambda}_{opt}) < \tau$ (a fixed threshold) split $\mathcal{I}^*$ in two subinterval $\mathcal{I}_1^*$ and $\mathcal{I}_2^*$.
 then achieve the STEP 1 for each subinterval (updating $\mathcal{I}^* = \mathcal{I}_j^*$);
 else goto STEP 3
 STEP 3 – **Stop**

1.3 Spectrometric example

In this section the methodology illustrated in the previous section is applied to a well-known benchmark data-set, the Tecator dataset (see: lib.stat.cmu.edu/datasets/tecator). The reader who is interested in a extensive comparative study, where some alternative regression methods are applied to these data, can refer to [9].

The data. The Tecator dataset consists in 215 spectra in the near infrared (NIR) wavelength range from 852 to 1050 nm, discretized on a mesh of 100 equispaced measures, corresponding to as many finely chopped pork samples. The aim is to predict the fat content Y_i , obtained by chemical analysis, from the spectrometric curve. To avoid the well-known “calibration problem”, due to the presence of shifts in the curves that cause noises, it is conventional to take as regressor χ_i the second derivatives of spectrometric curves instead of the original ones (see [8]). These functional data and their second derivatives are represented in Figure 1.1.

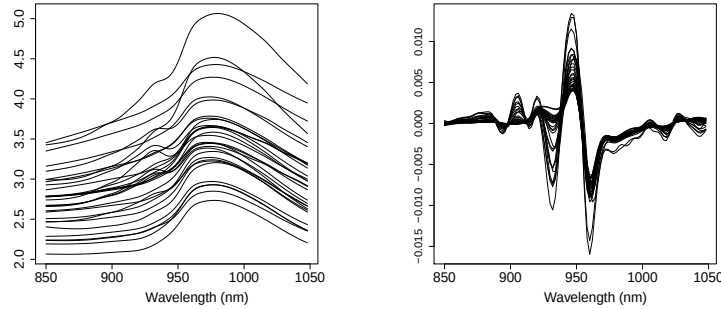


Fig. 1.1 A random selection of original spectrometric data (left panel) and their second derivatives (right panel).

The estimation procedure. We applied the regression methodology described in Section 1.2 to a learning-sample formed by the first 160 couples (x_i, y_i) , and we evaluated the goodness-of-fit over a test-set containing the remaining 55. The first step was the identification of the breaking-point(s) using the algorithm described before and whose partitioning tree is illustrated in Figure 1.2.

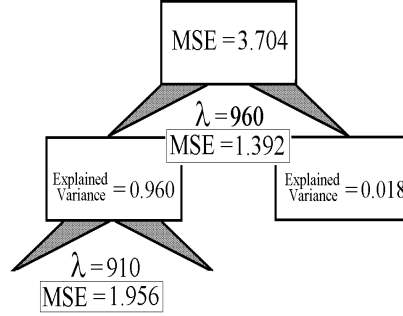


Fig. 1.2 Partitioning tree.

At the first step the minimum for $CV(\lambda)$ was achieved at 960 with a square prediction error $MSE = \frac{1}{55} \sum_{i=1}^{55} (y_i^{out} - \hat{y}_i)^2$ equal to 1.392 (whereas for the single index model without partitioning the MSE was equal to 3.704). Since the variance explained was 0.96 for the first term and 0.02 for the second one, we tried again to partition the interval $582 - 960 \text{ nm}$ in two parts, obtaining the break point 910. Because this partition does not improve the prediction abilities (with the MSE equal to 1.956), we stopped the algorithm at $s = 2$.

Interpretation of the outputs. To provide an interpretation of estimated model, we analyzed the estimated additive terms. As observed previously, the empirical variance explained from the first one is 0.96 against 0.02 for the second one: it emerges that the second part of the spectrum, corresponding to wavelengths longer than 960 nm , is in practice negligible in explaining the fat content. Concentrating our attention on the relevant part of the spectrum and looking at the estimated directions over this region (see Figure 1.3), it appears that the wavelengths between 850 and 890 nm seem not significant, whereas the ones in the range $890 - 950$ are the most important. This is coherent with the results on selection of variables in [5] where such interval appears the most interesting.

References

1. A. Ait-Saïdi, F. Ferraty, R. Kassa, P. Vieu (2008). Cross-validated estimations in the single-functional index model. *Statistics*, 42, 475–94.

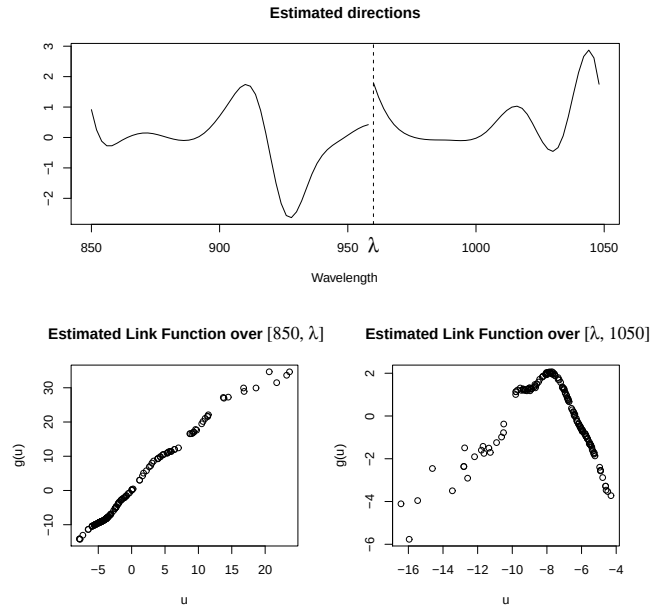


Fig. 1.3 Estimated directions θ_j and link functions g_j for spectrometric data.

2. U. Amato, A. Antoniadis, I. De Feis (2006). Dimension reduction in functional regression with applications. *Comput. Statist. Data Anal.*, 50, 2422–2446.
3. D. Chen, P. Hall, H.G. Müller (2011). Single and multiple index functional regression models with nonparametric link. *Ann. Statist.*, 39(3), 1720–1747.
4. F. Ferraty, A. Goia, E. Salinelli, P. Vieu (2013). Functional projection pursuit regression. *Test*, 22, 293–320.
5. F. Ferraty, P. Hall, P. Vieu (2010). Most-predictive design points for functional data predictors. *Biometrika*, 97, 807–824.
6. F. Ferraty, J. Park, P. Vieu (2011). Estimation of a functional single index model. In: *Recent advances in functional data analysis and related topics*, 111–116, *Contrib Statist*, Physica-Verlag/Springer, Heidelberg.
7. F. Ferraty, A. Peuch, P. Vieu (2003). Modèle à indice fonctionnel simple. *Comptes Rendus Math Académie Sciences Paris*, 336, 1025–8.
8. F. Ferraty, P. Vieu (2006). *Nonparametric functional data analysis*. Springer Series in Statistics. Springer-Verlag, New York.
9. F. Ferraty, P. Vieu (2011). Richesse et complexité des données fonctionnelles. *Revue de Modélad*, 43, 25–43.
10. A. Goia, P. Vieu (2013). A partitioned single functional index model. Submitted for publication.
11. W. Härdle, M. Müller, S. Sperlich, A. Werwatz (2004). *Nonparametric and semiparametric models*. Springer Series in Statistics. Springer-Verlag, New York.
12. H.G. Müller, Y. Wu, F. Yao (2013). Continuously additive models for nonlinear functional regression. *Biometrika*, 100, 607–622.
13. H.G. Müller, F. Yao (2008). Functional additive models. *J. Amer. Statist. Assoc.*, 103, 153–1544.
14. J.O. Ramsay, B.W. Silverman (2005) *Functional data analysis*, 2nd edn., Springer, New York.