Ridge regression for the functional concurrent model

Nadine Hilgert, Tito Manrique Chuquillanqui, Christophe Crambes, André Mas

Ridge Regression for the Functional Concurrent Model

Tito Manrique* [1, 2], Nadine Hilgert [1], Christophe Crambes [2], André Mas [2]

[1] UMR 729 MISTEA, INRA, Montpellier, France
[2] UMR 5149 I3M, Montpellier University, Montpellier, France
*Corresponding author: Tito Manrique
E-mails: tito.manrique@supagro.inra.fr; Nadine.Hilgert@supagro.inra.fr;
christophe.crambes@univ-montp2.fr; andre.mas@univ-montp2.fr

Abstract

The aim of the paper is to propose an estimator of the unknown function in the functional concurrent model. This is a general model in which all functional linear models can be reduced. We follow a strictly functional approach and extend the ridge regression method developed in the classical linear case to the functional data framework. We establish asymptotic statistical properties of the proposed estimator and present some simulations which show its high accuracy in fitting the unknown function, despite a low signal-to-noise ratio.

Keywords: ridge regression; functional data; concurrent model; varying coefficient model.

1 Introduction

Functional Data Analysis (FDA) proposes very good tools to handle data that are functions of some covariate (e.g. time, when dealing with longitudinal data). These tools may allow a better modelling of complex relationships than classical multivariate data analysis would do, as noticed by Ramsay and Silverman [2005, Ch. 1], Yao et al. [2005a,b], among others.

There are several models in FDA to study the relationship between two variables. In particular in this paper we are interested in the Functional Concurrent Model (FCM) because, as stated by Ramsay and Silverman [2005, p. 220], all functional linear models can be reduced to this form. This model can be defined as follows

$$Y(t) = \beta(t)\, X(t) + \epsilon(t), \tag{1}$$

where $t \in \mathbb{R}$, β is an unknown function to be estimated, X, Y are random functions and ϵ is a noise random function.

Some related models have already been discussed by several authors. For instance West et al. [1985] defined a similar model called 'dynamic generalized linear model' which is written in the next equation over time

$$\eta_t = \beta_0(t) + X_1(t)\, \beta_1(t) + \cdots + X_p(t)\, \beta_p(t).$$

Hastie and Tibshirani [1993] themselves proposed a generalization of FCM called 'varying coefficient model'. Afterwards many people studied this model trying to estimate the unknown smooth regression functions β_i, for instance by local maximum likelihood estimation (Dreesman and Tutz [2001]; Cai et al. [2000a,b]), by kernel smoothing (Wu et al. [1998]), or by local polynomial smoothing (Zhang et al. [2000]; Fan et al. [2003]; Zhang et al. [2002]).

As far as we know, despite the abundant literature related to FCM, there is no paper providing a strictly functional approach (i.e. with random functions defined inside normed functional spaces) as noticed by Ramsay and Silverman [2005, p. 259], who said that all these methods come more from a multivariate data analysis approach rather than from a functional one. This may cause a loss of information because these approaches, as noticed by Müller and Sentürk [2010, p. 1257], "do not take full advantage of the functional nature of the underlying data".

The goal of this paper is to extend the ridge regression method developed in the classical linear case to the functional data framework. We establish asymptotic statistical properties of the proposed estimator and present some simulation trials which show its high accuracy in fitting the unknown function, despite a low signal-to-noise ratio.

2 General Hypotheses and Estimator

The space of the real valued continuous functions vanishing at infinity is denoted $C_0(\mathbb{R})$. In this space we use the supremum norm, that is $\|f\|_{C_0} := \sup_{x\in\mathbb{R}} |f(x)|$ for some $f \in C_0(\mathbb{R})$. In the same way, for a compact $K \subset \mathbb{R}$, $C^0(K)$ is the space of real valued continuous functions defined on K, with the supremum norm $\|f\|_{C^0(K)} := \sup_{x\in K} |f(x)|$. Here are the general hypotheses made on the FCM (1) throughout this paper.

General Hypotheses of FCM

$(H1_{FCM})$ X, ϵ are independent $C_0(\mathbb{R})$ valued random functions, $\mathbb{E}(\epsilon) = \mathbb{E}(X) = 0$, $\mathbb{E}[\|\epsilon\|_{C_0}] < +\infty$ and $\mathbb{E}[\|X\|_{C_0}] < +\infty$.

$(H2_{FCM})$ $\beta \in C_0(\mathbb{R})$.

$(H3_{FCM})$ $\mathbb{E}[\, \|X\|_{C_0}^2] < +\infty$.

The Estimator

The definition of the estimator of β is inspired by the estimator introduced by Hoerl [1962] used in the ridge regularization method that deal with ill-posed problems in the classical linear regression.
Let $(X_i, Y_i)_{i=1,\cdots,n}$ be an i.i.d sample of FCM (1) and $\lambda_n > 0$. We define the estimator of β as follows

$$\hat{\beta}_n = \frac{\frac{1}{n}\sum_{i=1}^n Y_i\, X_i}{\frac{1}{n}\sum_{i=1}^n |X_i|^2 \ +\ \frac{\lambda_n}{n}}. \tag{2}$$

In the classical linear regression case, Hoerl and Kennard [1970, p. 62] proved that there is always a regularization parameter for which the ridge estimator is better than the Ordinary Linear Squares (OLS) estimator. Huh and Olkin [1995] made a study of some asymptotic properties of the ridge estimator in this context.

3 Asymptotic Properties

Theorem 3.1. *Let $(X_i, Y_i)_{i=1,\cdots,n}$ be an i.i.d. sample of FCM (1). Then under the following hypotheses*

(A1) The sequence of positive numbers $(\lambda_n)_{n\geq 1} \subset \mathbb{R}^+$ is such that $\frac{\lambda_n}{n} \to 0$ and $\frac{\sqrt{n}}{\lambda_n} \to 0$,

(A2) $0 < \mathbb{E}[\,|X(t)|^2] < +\infty$, for all $t \in \mathbb{R}$,

(A3) There exists a sequence of positive numbers $(D_k)_{k\geq 1} \subset \mathbb{R}^+$ such that

$$\lim_{k\to\infty} \left[\frac{\lambda_k}{k} \cdot \frac{1}{\inf_{t\in[-D_k, D_k]} \mathbb{E}[\,|X(t)|^2]}\right] = 0,$$

and for every $t \in \mathbb{R}$, if $|t| > D_k$ then $|\beta(t)| \leq \frac{1}{k}$,

we obtain

$$\|\hat{\beta}_n - \beta\|_{C_0} \xrightarrow{\mathrm{P}} 0. \tag{3}$$

Proof. *(sketch)* Given that $Y_i = \beta\, X_i + \epsilon_i$, for each $i = 1, \cdots, n$ we can decompose $\hat{\beta}_n$ as follows

$$\hat{\beta}_n = \beta - \frac{\lambda_n}{n}\left(\frac{\beta}{\frac{1}{n}\sum_{i=1}^n |X_i|^2 \ +\ \frac{\lambda_n}{n}}\right) + \left(\frac{\frac{1}{n}\sum_{j=1}^n \epsilon_j\, X_j}{\frac{1}{n}\sum_{i=1}^n |X_i|^2 \ +\ \frac{\lambda_n}{n}}\right). \tag{4}$$

Then the hypothesis *(A2)* and the Strong Law of Large Numbers (SLLN) in the separable Banach space $C_0(\mathbb{R})$ are used to show that

$$\left\| \frac{\frac{1}{n}\sum_{j=1}^n \epsilon_j\, X_j}{\frac{1}{n}\sum_{i=1}^n |X_i|^2 \ +\ \frac{\lambda_n}{n}} \right\|_{C_0} \xrightarrow{\mathrm{P}} 0.$$

Finally *(A3)* and SLLN are used to prove that

$$\left\| \frac{\lambda_n}{n} \left(\frac{\beta}{\frac{1}{n}\sum_{i=1}^{n} |X_i|^2 \; + \; \frac{\lambda_n}{n}} \right) \right\|_{C_0} \xrightarrow{\text{a.s.}} 0,$$

which implies (3) by the triangle inequality in (4). □

3.1 Comments about the Hypotheses

Hypothesis *(A1)***:** This hypothesis is about how fast λ_n has to go to infinity, it must be slower than n but faster than $\sqrt{n}$.

Hypothesis *(A2)***:** We use *(A2)* because, if for some $t \in \mathbb{R}$, we have $\mathbb{E}[\,|X(t)|^2] = 0$ then almost surely $X(t) = 0$ and thus $\hat{\beta}_n(t) = 0$ also. Therefore when $\beta(t) \neq 0$ and $\mathbb{E}[\,|X(t)|^2] = 0$, $\hat{\beta}_n$ cannot estimate β at the point t.

Hypothesis *(A3)***:** Finally *(A3)* says that β must decrease faster than $\mathbb{E}[\,|X(t)|^2]$. In this sense this hypothesis may be interpreted as an assumption about the decreasing rate of the function β with respect to that of X, as we can see in the following proposition where K_1 can be understood as the decreasing rate of β and K_2 that of $\mathbb{E}[\,|X|^2]$.

Proposition 3.2. *If $\beta(t) = \frac{1}{e^{K_1|t|}}$ and $\mathbb{E}[|X(t)|^2] = \frac{1}{e^{K_2|t|}}$ in such a way that $K_1 > 2K_2 > 0$, then the hypothesis (A3) is satisfied when we take $\lambda_n = \sqrt{n}\log n$ which satisfies (A1) in Theorem 3.1.*

Proof. (sketch) We define $D_k := \frac{\log k}{K_1} > 0$ for each $k \geq 1$ and use the fact that β and $\mathbb{E}[|X|^2]$ are strictly decreasing functions. □

It is possible to get a similar proposition for polynomial decreasing rates.

Proposition 3.3. *If $\beta(t) = \frac{1}{|t|^r}$ and $\mathbb{E}[|X(t)|^2] = \frac{1}{|t|^s}$ in such a way that $r > 2s > 0$ with $r, s \in \mathbb{N}$ then the hypothesis (A3) is satisfied when we take $\lambda_n = \sqrt{n}\log n$ which satisfies (A1) in the Theorem 3.1.*

3.2 Further Results

We can prove the next corollaries by using similar ideas.

Corollary 3.4. *Let $(X_i, Y_i)_{i=1,\cdots,n}$ be an i.i.d. sample of FCM (1), then under (A1) and the following hypotheses*

(A2bis) $\inf_{t \in \overline{supp(\beta)}} \mathbb{E}[\,|X(t)|^2] > 0$,

(A3bis) *$\overline{supp(\beta)}$ is bounded, where $\overline{supp(\beta)}$ is the closure of the support of β, $supp(\beta)$,*

we obtain

$$\|\hat{\beta}_n - \beta\|_{C_0} \xrightarrow{\text{P}} 0. \tag{5}$$

In the following corollary we establish a similar result as that of Theorem 3.1 in the space $C^0(K)$.

Corollary 3.5. *Let $(X_i, Y_i)_{i=1,\cdots,n}$ be an i.i.d. sample of FCM (1), then under (A1), (A2bis) and the following hypothesis*

(A3ter) *There exists a compact $K \subset \mathbb{R}$ such that $supp(\beta) \subset K$, and almost surely $supp(X)$, $supp(\epsilon) \subset K$,*

we obtain

$$\|\hat{\beta}_n - \beta\|_{C^0(K)} \xrightarrow{\text{P}} 0. \tag{6}$$

4 Simulations

The accuracy of the estimator $\hat{\beta}_n$ is illustrated for two choices of the function β. In all experiments, X is a Brownian Motion (BM) on the interval $[0,1]$ and ϵ is a BM on $[0,1]$ too, which is independent of X. To simulate the BM we used the Karhunen-Loève decomposition with the first 100 eigenfunctions. All the functions are observed on 100 evenly spaced points in $[0,1]$.
We calculate the signal-to-noise ratio (SNR) as $\mathrm{SNR} = \frac{\mathrm{Var}[\|Y\|_{C_0}]}{\mathrm{Var}[\|\epsilon\|_{C_0}]}$.
We set two sample sizes $n = 100$ and $n = 200$ and fix three values for λ. For each set of parameter (n, λ), 100 trials were run to estimate the mean and the standard deviation of the relative estimation error $\frac{\|\hat{\beta}_n - \beta\|_{C_0}}{\|\beta\|_{C_0}}$.

Simulation 1 β is defined as follows $\beta(t) = \sqrt{2}\sin((8-1/2)\pi t)$. The accuracy of the estimation of β shall be appreciated on the example given in the Figure 1,

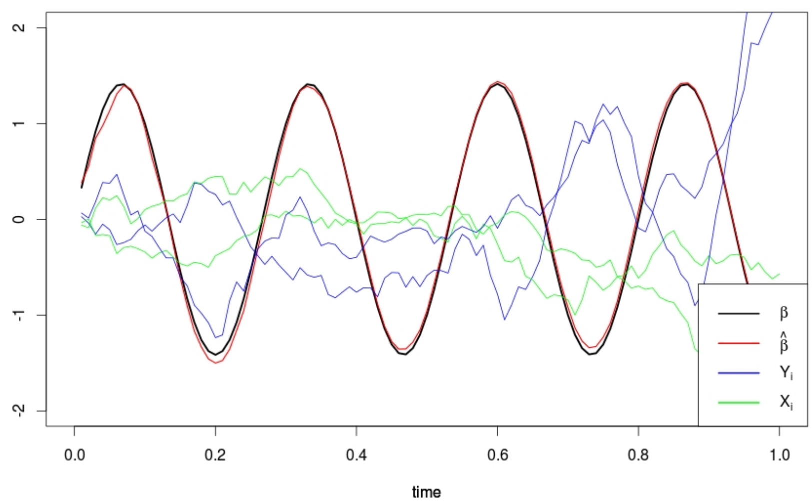

Figure 1: β and its estimator $\hat{\beta}$ with a SNR ≈ 2.937 (around 33% noise).

as well as in the following tables:

$n = 100$

λ	mean	sd
10^{-1}	0.1386390	0.04017871
10^{-2}	0.1391908	0.03942906
10^{-3}	0.1394522	0.03932101

$n = 200$

λ	mean	sd
10^{-1}	0.1003225	0.02600884
10^{-2}	0.1000697	0.02628544
10^{-3}	0.1000806	0.02634147

Simulation 2 β is defined as follows $\beta(t) = 4(t^6 - 1/2t^7 + 2t^3 + 2t^2 - 4t + 1)\sin(22\pi t)$. One example of the estimation is given in the Figure 2, and the relative estimation error is illustrated in the following tables:

$n = 100$

λ	mean	sd
10^{-1}	0.05175795	0.01389744
10^{-2}	0.03762299	0.01079769
10^{-3}	0.03760215	0.01061630

$n = 200$

λ	mean	sd
10^{-1}	0.03233332	0.009219947
10^{-2}	0.02701885	0.007106225
10^{-3}	0.02698448	0.007102699

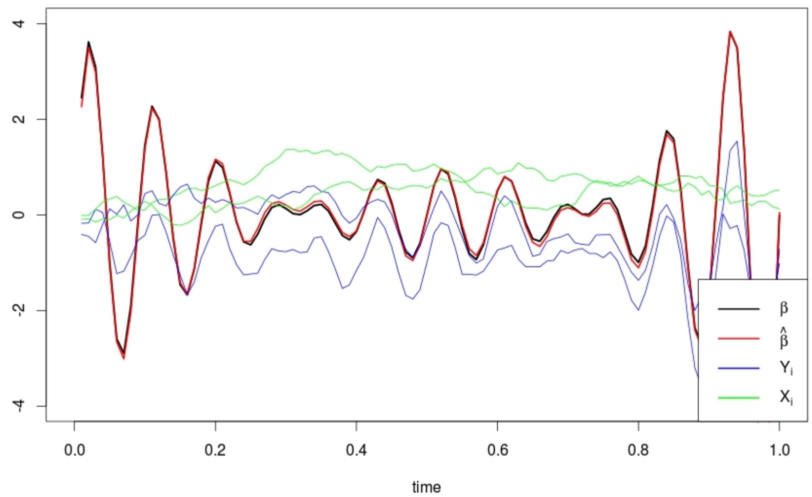

Figure 2: β and its estimator $\hat{\beta}$ with a SNR ≈ 32.743 (around 3% noise).

5 Conclusions

We established the asymptotic convergence of the functional ridge estimator. The simulations showed the good accuracy of the estimator even for a low signal-to-noise ratio (around 3 in Simulation 1).
For further research, some work on the rate of convergence should be considered as well as a discussion about the choice of the regularization parameter λ_n.

6 Acknowledgments

The authors would like to thank the Labex NUMEV (convention ANR-10-LABX-20) for partly funding the PhD thesis of Tito Manrique (under project 2013-1-007).

References

Cai, Z., Fan, J., and Li, R. (2000a). Efficient estimation and inferences for varying-coefficient models. *Journal of the American Statistical Association*, 95(451), 888-902.

Cai, Z., Fan, J., and Yao, Q. (2000b). Functional-coefficient regression models for nonlinear time series. *Journal of the American Statistical Association*, 95(451), 941-956.

Dreesman, J. M., and Tutz, G. (2001). Non Stationary Conditional Models for Spatial Data Based on Varying Coefficients. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 50(1), 1-15.

Fan, J., Yao, Q., and Cai, Z. (2003). Adaptive varying coefficient linear models. *Journal of the Royal Statistical Society: series B (statistical methodology)*, 65(1), 57-80.

Hastie, T., and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 757-796.

Hoerl, A. E. (1962). Application of ridge analysis to regression problems. *Chemical Engineering Progress* 58, 54-59.

Hoerl, A. E., and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55-67.

Huh, M. H., and Olkin, I. (1995). Asymptotic aspects of ordinary ridge regression. *American Journal of Mathematical and Management Sciences*, 15(3-4), 239-254.

Müller, H. G. and Sentürk, D. (2010). Functional varying coefficient models for longitudinal data. *Journal of the American Statistical Association*, 105(491), 1256-1264.

Ramsay J. and Silverman B. (2005) *Functional Data Analysis*, Springer–Verlag, USA.

West, M., Harrison, P. J., and Migon, H. S. (1985). Dynamic generalized linear models and Bayesian forecasting. *Journal of the American Statistical Association*, 80(389), 73-83.

Wu, C. O., Chiang, C. T., and Hoover, D. R. (1998). Asymptotic confidence regions for kernel smoothing of a varying-coefficient model with longitudinal data. *Journal of the American statistical Association*, 93(444), 1388-1402.

Yao, F., Müller, H. G., and Wang, J. L. (2005a). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470), 577-590.

Yao, F., Müller, H. G., and Wang, J. L. (2005b). Functional linear regression analysis for longitudinal data. *The Annals of Statistics*, 33(6), 2873-2903.

Zhang, W., and Lee, S. Y. (2000). Variable bandwidth selection in varying-coefficient models. *Journal of Multivariate Analysis*, 74(1), 116-134.

Zhang, W., Lee, S. Y., and Song, X. (2002). Local polynomial fitting in semivarying coefficient model. *Journal of Multivariate Analysis*, 82(1), 166-188.