Monografías del Seminario Matemático García de Galdeano. **27:** 241–248, (2003).

Extending asymptotic least-squares method to fit a reduced rank model to functional data

Simplice Dossou-Gbété

Laboratoire de Mathématiques Appliquées,
Université de Pau et des Pays de l'Adour, BP 1155, 64013 Pau Cedex, France.
e-mail: simplice.dossou-gbete@univ-pau.fr

Abstract

Asymptotic least-squares is a well known estimation method in the framework of parametric generalized linear models. The aim of this paper is to extend this method to the functional data analysis setting. After the presentation of the method, the data motivating the study will be analyzed as an example.

Keywords: asymptotic least-squares, functional data analysis, quasi-likelihood, reduced rank model.

AMS Classification:75F2359

1 Introduction

Asymptotic least-squares is a well known estimation method in the framework of parametric statistics models, especially the generalized linear models. It was considered by Taylor W.F. [8] in the framework of the parametric logistic regression fit (Berkson's and minimum Khi-square methods). More recently, Baccini et al. [1] took Taylors' arguments back to justify the use of least-squares as an alternative for the maximum likelihood method for fitting association models to contingency tables analysis. Parametric generalized linear models can be fitted consistently by asymptotic least-squares which appear as an alternative method to maximum likelihood method, particularly when an efficient computation algorithm is not available (Gouriéroux & Monfort [4], pp.301-314). This is achieved in a two-step procedure combining least-squares with a consistent estimation of the mean. The aim of this study is to extend the asymptotic least-squares method to the functional data analysis framework (Green & Silverman [5]).

2 Functional regression model with time-varying covariate

2.1 Model with additive interaction component

Let Y denotes a numeric response variable; let Z be a J-levels factor covariate and X denotes a numeric valued covariate. We consider a functional version of the generalized linear model as follows

$$\begin{aligned} E\left(Y \backslash Z=j,\, X=x\right) &= \mu_j\left(x\right) \\ Var\left(Y \backslash Z=j,\, X=x\right) &= \sigma_j^2 V\left(\mu_j\left(x\right)\right) \end{aligned}$$

where V is some known function, μ_j is a functional parameter and σ_j is a numeric parameter, and both are unknown. Such a model is closely related to the varying-coefficients models (Hastie & Tibshirani [6])
Let us model the regression function μ_j additively through a link function g as follows: $\eta_j\left(x\right) = g\left(\mu_j\left(x\right)\right) = \alpha_j + \beta\left(x\right) + \gamma_j\left(x\right)$.

2.2 Saturated model vs reduced rank model

The model stated above will be called a saturated model if the parameters (the sequence $\{\alpha_j, j=1,\cdots,J\}$, and the functions β and γ_j $(j=1,\cdots,J)$) are unrestricted. The model is a reduced rank additive model if there is an integer $r \geq 1$ such that $\gamma_j = \sum_{k=1}^{r} c_{kj}\phi_k$, where the parameters $\{c_{kj}\}$ are numeric while $\{\phi_k\}$ are smooth functions, both unknown. The functions ϕ_k may be interpreted as interaction components we can plot to improve data analysis, while the parameters c_{kj} may be considered as loadings for the level j of the factor covariate.

2.3 Identifiability constraints

Neither the saturated model or the reduced rank model is identifiable, since it is not guaranted that the equations below have unique solution

$$\eta_j = \alpha_j + \beta + \gamma_j; \qquad \eta_j = \alpha_j + \beta + \sum_{k=1}^{r} c_{kj}\phi_k$$

So, additional constraints on the model parameters are needed to achieve identifiability. Such constraints can be stated as follows:

$$\begin{aligned} \sum_{j=1}^{J} \pi_j \alpha_j = \alpha_0, \qquad \sum_{j=1}^{J} \pi_j c_{kj} = 0, \\ \int_0^1 \tau\left(x\right)\beta\left(x\right)dx = \beta_0, \qquad \int_0^1 \tau\left(x\right)\phi_k\left(x\right)\phi_l\left(x\right)dx = \delta_{kl}, \end{aligned}$$

where: π_j $(j = 1, \cdots, J)$ and τ are specified positive weights, while α_0 and β_0 are numeric unknown parameters to be estimated.

3 Fitting the reduced rank additive model by asymptotic least-squares

As in the parametric setting, our proposal to fit a functional additive model to data by an asymptotic least-squares method consists of a two-step procedure combining quasi-likelihood and least-squares. A two-step estimation procedure has been proposed by Fan & Zhang [2] in the gaussian functional linear model setting. Taking into account the gaussian framework, the first step of their method uses least-squares to calculate raw estimates of the coefficient functions for each value of the numeric covariate X. The second step consists in smoothing the raw estimates. They applied the method to the estimation of some functional ANOVA models. This may need a great amount of computation if the data design involves a great number of distinct values of the covariate X. A contrario, taking into account that the covariate Z is a factor, our method starts by smoothing the mean of the response for each level of the covariate Z. Reduced rank model parameters are obtained at the second stage by applyng least-squares to the first step estimates. Compared to Fan & Zhang, our method may involve less computation.

3.1 Estimating the mean functions by local polynomial quasi-likelihood

The first step of the model fitting procedure consists of a consistent nonparametric estimation of $\eta_{jx} = g(\mu_{jx})$. It is achieved through the estimates of the functional parameters η_j which are calculated by local polynomial quasi-likelihood maximization. Let $\widehat{\eta}_j$ denote these estimates which are such that $\widehat{\eta}_j(x)$ converge in probability to the coefficient $\eta_j(x)$ (Loader[6]).

3.2 Least-squares criterion

The weights π_j and τ may depend on data and if such is the case, additional asymptotic assumptions on π_j and τ should be required to obtain consistent estimates of the model parameters.

Let f, f_1 and f_2 be numeric functions belonging to the same linear space of smooth functions. Fix: $(f_1 \mid f_2)_\tau = \int_0^1 \tau(x) f_1(x) f_2(x)\, dx$,$\|f\|_\tau^2 = \int_0^1 \tau(x) f^2(x)\, dx$ and then

consider the least-squares criterion

$$Q = \sum_{j=1}^{J} \pi_j \left\| \widehat{\eta}_j - \alpha_j - \beta - \sum_{k=1}^{r} c_{kj}\phi_k \right\|_\tau^2 .$$

Let

$$\widehat{\alpha}_0 = \widehat{\beta}_0 = \frac{1}{2}\sum_{j=1}^{J} \pi_j \int_0^1 \tau(x)\,\widehat{\eta}_j(x)\,dx; \quad \widehat{\alpha}_j = \textstyle\int_0^1 \tau(x)\,\widehat{\eta}_j(x)\,dx - \widehat{\beta}_0;$$

$$\widehat{\beta} = \sum_{j=1}^{J} \pi_j \widehat{\eta}_j - \widehat{\alpha}_0; \qquad \widetilde{\eta}_j = \widehat{\eta}_j - \widehat{\alpha}_j - \widehat{\beta};$$

$$V = \sum_{j=1}^{J} \pi_j \widetilde{\eta}_j \otimes \widetilde{\eta}_j.$$

Assuming that the following constraints apply

$$\sum_{j=1}^{J} \pi_j \alpha_j = \alpha_0; \qquad \sum_{j=1}^{J} \pi_j \gamma_j = 0;$$

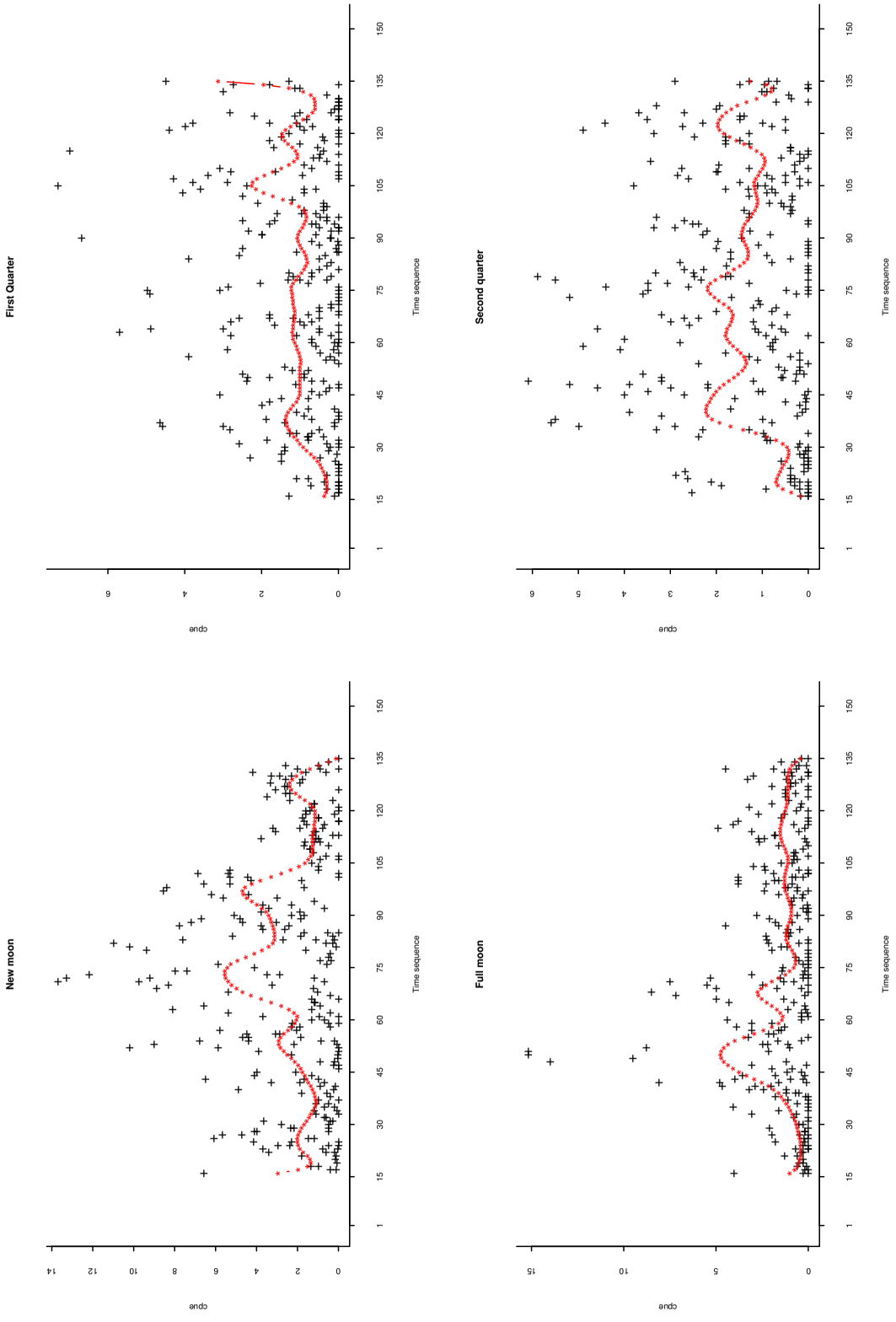
$$\int_0^1 \tau(x)\,\beta(x)\,dx = \beta_0; \qquad \int_0^1 \tau(x)\,\gamma_j(x)\,dx = 0,$$

one can prove, as in Gabriel [3], that the minimum of the criterion Q is attained at the values $\widehat{\alpha}_j$, $\widehat{\beta}$, $\widehat{\phi}_k$ and $\widehat{c}_{kj}$ where: $\left\{\widehat{\phi}_k, k = 1, \cdots, r\right\}$ is an orthonormal sequence (relative to the inner product related to weight function τ) of eigenvectors of the linear kernel operator $V\tau$ associated with its eigenvalues in the decreasing order; $\widehat{c}_{kj} = \int_0^1 \tau(x)\,\widetilde{\eta}_j(x)\,\widehat{\phi}_k(x)\,dx$.

4 Application: Young eels fishing index data

The motivation of our study comes from an analysis of data we have got from IFREMER, a French institution in charge of marine environment studies. The dataset consists in a daily index of the activity of the young eels fishers during an authorized period from November to March devoted to the fishing of young eels in the estuary of the Adour river in the south-west of France. The data analysed were collected during 9 fishing seasons and a time period included between 16th November 1983 and 15th March 1993. Our goal is to point out the effect of the moon on the young eels fishing activity, if any. Thus the factor covariate Z values are the lunar phases at the days corresponding to the fishing indexes in the fishing season.

As shown in the graphs below, the variations of the fishing index do not exhibit the same shape for the different lunar phases. This suggests there is an interaction between the days in a fishing period and the moon.



4.1 Data modelling

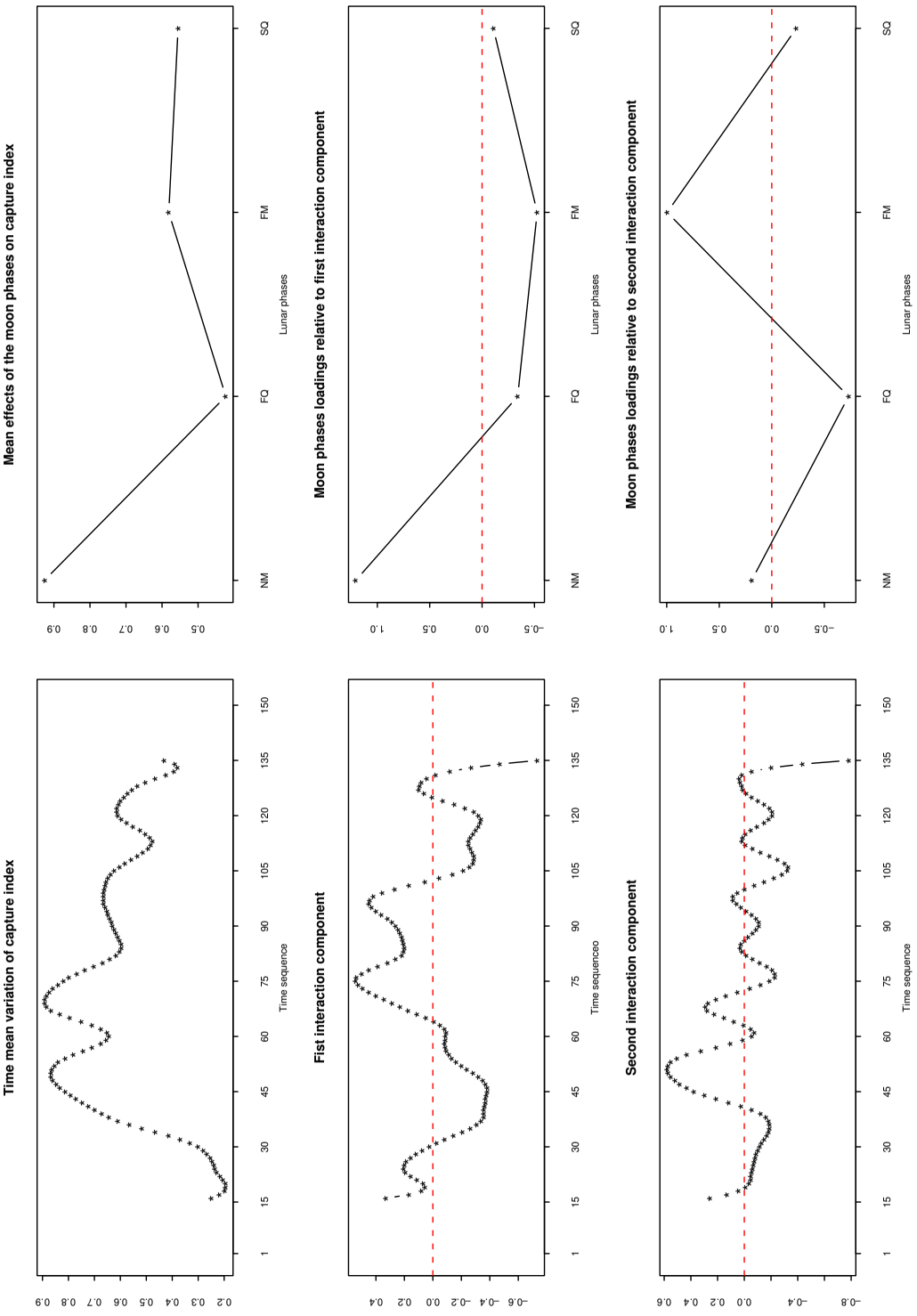
Let $y_{st_{sk}}$ denote the fishing index on the day t_{sk}, $(t_{sk} \in \{1, \cdots, 120\})$ of the fishing season s, $(s = 1, \cdots, n)$ where the moon phase is j $(j = 1, \cdots, 4)$. We consider $y_{st_{sk}}$ as the outcome of

a random variable $Y_{st_{sk}}$ with mean $E\left(Y_{st_{sk}} \setminus Z_{st_{sk}} = j,\ X_{st_{sk}} = x_{sk}\right) = \mu_j(x_{sk})$ and variance $Var\left(Y_{st_{sk}} \setminus Z_{st_{sk}} = j,\ X_{st_{sk}} = x_{sk}\right) = \sigma_j^2 \mu_j(x_{sk})$ where $x_{sk} = \frac{t_{sk}-1}{119}$. Furthermore, the random variables $Y_{st_{sk}},\ s = 1, \cdots, n,\ t_{sk} \in \{1, \cdots, 120\}$ are assumed to be independent.
Taking into account the results of the preliminary analysis of the data in the preceding section, the relationship of the mean $\mu_j(x_{sk})$ with the phase j of the moon ($j = 1, \cdots, 4$) and the daily period t_{sk} of a fishing season ($t_{sk} \in \{1, \cdots, 120\}$) is modelled additively through some known link function g as follows: $g\left(\mu_j(x_{sk})\right) = \eta_j(xjh) = \alpha_j + \beta(x_{sk}) + \gamma_j(x_{sk})$. We assume $\beta(x_{sk})$ and $\gamma_j(x_{sk})$ are respectively the values of smooth functions β and γ_j at $x_{sk} = \frac{t_{sk}-1}{119} \in [0, 1]$. Then we are dealing with a functional parameters model which involves functional (nonparametric) estimation procedures. Let $\eta_j = \alpha_j + \beta + \gamma_j$ and $\mu_j = g^{-1}\left(\eta_j\right)$. For this particular example we have chosen g as the square-root function, since the data are modelled as quasi-poisson and the square-root function has a nice variance stabilisation property in this situation. Thus it seems evident to choose the weight $\pi_j = \frac{\bar{\sigma}}{\widehat{\sigma}_j}$,where $\bar{\sigma} = \sum_{j=1}^{4} \widehat{\sigma}_j$ and the weight function τ is the identity function.

4.2 Interpreting the results

The main results of the data analysis are presented in the graphics below. The left figure in the top panel shows the main behaviour of the fishing activity during a fishing season. It is shown that the young eels catch is a maximum between the forty-fifth and the seventy-fifth days of the season. The figure in the right side of this first panel suggests that the new moon is globally more favorable than the other lunar phases to fishing for young eels.
In the middle panel the left figure shows the first interaction component variations in a fishing season while the right figure shows the corresponding loadings for the lunar phases. The interpretation of these loadings is that the first interaction component puts the new moon against the other lunar phases. The loading of the new moon is positive while the other loadings are negative. Moreover the first interaction function is significantly positive from the 70th day to the 100th day of a fishing season. The interpretation of the component is obtained by considering together the interaction function and the loadings. Then one can say that, from the 70th day to the 100th day of a fishing season, the young eels catch increases during the days of the new moon. Furthermore, at the beginning or at the end of a fishing season, the young eels fishing level is reduced from the main level during the days where the new moon happens.
The bottom panel is devoted to the second interaction component which puts the first quarter of the moon against the full moon. The interaction function is positive in the first half of a fishing season and exhibits mainly negative values in the second half of a fishing season. Between the 40th and the 70th days of the fishing season the young eels fishing

decreases in the days where the moon stands in the first quarter while the fishing level is put up the main level when the first quarter happens in the second half of the season.



References

[1] Baccini A.,Caussinus H., de Falguerolles A. (1993). Analysing dependence in largecontingency tables: Dimensionality and patterns in scatter-plots. Multivariateanalysis: future directions 2. Cuadras C.M. an Rao C. R. editors , 245-263.

[2] Fan J. & Zhang J.-T.(2000). Two-step estimation of functional linear model with application to longitudinal data. J. of Royal Statistical Society B, 62, 303-322

[3] Gabriel K. R. (1978). Least-squares approximation of matrices by additive and multiplicative models. J. of Royal Statistical Society B, 40, 186-196.

[4] Gouriéroux Ch. & Monfort A. (1989). Statistique et modèles économétriques,vol. 1. Economica.

[5] Green P.J. & Silverman B.W. (1994). Nonparametric regression and generalized linear models. A rougthness penalty approach. Chapman & Hall.

[6] Hastie T. & Tibshinari R. (1993). Varying-coeficient models. J. Roy. Statist. Soc B vol. 55, 757-796

[7] Loader C. (1999). Local regression and likelihood. Springer.

[8] Taylor W.F. (1953). Distance functions and regular best asymptotically normal estimates. Annals of mathematical statistics, vol. 24, 1, 85-92.