Covariance estimation for aggregated functional data

Guilherme Ludwig[1]
Advisor: Nancy Lopes Garcia[2]
Department of Statistics – IMECC – UNICAMP

1 Introduction

The motivation for this project comes from the necessity of saving resources through an efficient electrical energy distribution. A perfect system of energy distribution would have every power plant and power transformers with a constant load during the entire day, every day of the year. However, this is an impossible scenario, as there are peaks of energy consumption during certain intervals of the day (e.g. during working hours at factories, or at evening due to use of electric showers in residential zones). In order to avoid overloads, energy distribution companies work around this problem by overdimensioning the energy distribution without considering its specific market profiles. An efficient way to optimize the use of existing plants and transformers is identify the consumer profile, and redistribute the consumers in order to reach a more homogeneous demand during the day.

Usual types of consumers include residential, commercial, industrial, among others. Each consumer has a different typical curve, called *typology*. One way to estimate mean and covariance curves is to employ functional data analysis (Ramsay & Silverman, 2005). However, usual techniques require observations made of individual curves and to obtain such a sample is not only costly, but also extremely variable, with a very low signal-to-noise ratio. To overcome this problem we consider a data set consisting of sums of individual curves at each *trafo* (the power transformer), hence the aggregated nature of the model.

Dias, Garcia & Martarelli (2009) obtained estimatives for the typologies based on B-Splines basis expansion and a model of aggregated data regression; also a nonparametric estimative of the covariance function based on the tensor product of B-Splines and and the work by Hall, Fisher & Hoffman (1994).

The goal of this project is to improve the estimation of the covariance function, and develop new methods based on either its spectral decomposition, further B-Spline tensorial product analysis or a nonparametric framework; and compare these methodologies via simulation and application to the real data set.

[1] guilhermeludwig@gmail.com

[2] nancy@ime.unicamp.br

2 Regression Model for Aggregated Data

Consider a population split into C subpopulations which you want to estimate the mean curve (their typology) and the covariance function, for each subpopulation. Consider we have a random sample of J days of aggregated curves observed at T points. For each trafo i, its respective aggregated curve is composed by the sum of $N_{1,i} + \ldots + N_{C,i}$ subpopulation curves, where $N_{c,i}$ is the number of individuals of the cth subpopulation in the ith trafo. Let $Y_{ij}(t)$ denote the curve i from day j observed at time t; our model can then be written as

$$Y_{ij}(t) = \sum_{c=1}^{C} \sum_{n_c=1}^{N_{c,i}} W_{c,j,n_c,i}(t) \tag{1}$$

for $t \in [0, 24]$, $i = 1, \ldots, I$, $j = 1, \ldots, J$ where $W_{c,j,n_c,i}(t)$ represents the n_cth individual curve from subpopulation c, on the jth day of the ith trafo.

We shall assume a nonparametric regression model for the curve of each individual consumer, depending solely on which subpopulation it belongs to. That means that for a type c individual there exists a *typology* $\alpha_c(t)$ such that

$$W_{c,j,n_c,i}(t) = \alpha_c(t) + \varepsilon_{c,j,n_c,i}(t)$$

where $\varepsilon_{c,j,n_c,i}(t)$ is a zero mean Gaussian process. Moreover, we assume $\varepsilon_{c,j,n_c,i}(t)$ to be independent and identically distributed for a given c. Therefore

$$Y_{i,j}(t) = \sum_{c=1}^{C} N_{c,i}\alpha_c(t) + \boldsymbol{\varepsilon}_{i,j}(t) \tag{2}$$

where

$$\boldsymbol{\varepsilon}_{i,j}(t) = \sum_{c=1}^{C} \sum_{n_c=1}^{N_{c,i}} \varepsilon_{c,j,n_c,i}(t) \tag{3}$$

In order to estimate the typologies α_c we expand these functions in a linear combination of K basis functions. In this case, cubic B-Splines, recursively defined as, given a sequence of knots $\tau_0, \ldots \tau_L$,

$$B_{k,0}(t) := \begin{cases} 1 & \text{if} \quad \tau_k \leq t < \tau_{k+1} \\ 0 & \quad c.c. \end{cases}$$

$$B_{k,m}(t) := \frac{t - \tau_k}{\tau_{k+m} - \tau_k} B_{k,m-1}(t) + \frac{\tau_{k+m+1} - t}{\tau_{k+m+1} - \tau_{k+1}} B_{k+1,m-1}(t), \quad m = 1, 2, 3$$

Thus we can write

$$\alpha_c(t) \;=\; \sum_{k=1}^{K} \beta_{c,k} B_k(t)$$

where $B_k(t)$, $k = 1, \ldots, K$ are cubic B-splines.

The expression

$$Y_{i,j}(t) = \sum_{c=1}^{C} \sum_{k=1}^{K} N_{c,i} \beta_{c,k} B_k(t) + \varepsilon_{i,j}(t),$$

can be used to estimate the coefficients $\beta_{c,k}$, $c = 1, \ldots, C$ and $k = 1, \ldots, K$.

3 Estimating the covariance function

Given the aggregated nature of the data set, obtaining an estimative for the covariance function of each trafo doesn't provide us with estimates for the consumer type. Denote the covariance for the individual of subpopulation c by

$$\sigma_c(s,t) \;:=\; \mathrm{Cov}\left(\epsilon_{c,j,n_c,i}(s), \epsilon_{c,j,n_c,i}(t)\right)$$

then the covariance function for the ith trafo is

$$\Sigma_i(s,t) \;:=\; \mathrm{Cov}\left(\epsilon_{i,j}(s), \epsilon_{i,j}(t)\right)$$

holding true the relationship:

$$\Sigma_i(s,t) \;=\; \sum_{c=1}^{C} N_{c,i} \sigma_c(s,t)$$

Spectral Decomposition Using the spectral decomposition of the covariance operator, and assuming that all consumers have the same eigenfunctions but different eigenvalues, we have:

$$\sigma_c(s,t) \;:=\; \mathrm{Cov}\left(\epsilon_{c,j,n_c,i}(s), \epsilon_{c,j,n_c,i}(t)\right) \;=\; \sum_{\nu=1}^{\infty} \gamma_{\nu,c} \phi_\nu(t) \phi_\nu(s).$$

and

$$\Sigma_i(s,t) \;=\; \sum_{\nu} \gamma_\nu^{(i)} \phi_\nu(s) \phi_\nu(t)$$

Furthermore, if we add all the covariances

$$\Sigma \;=\; \sum_{i=1}^{I} \Sigma_i$$

we still have the spectral decomposition of the sum of the curves as

$$\Sigma(s,t) \;=\; \sum_{\nu} \gamma_\nu \phi_\nu(s) \phi_\nu(t).$$

Then we can adapt a procedure given by Lee, Zhang & Song (2002), based on smoothing B-Splines of the eigenfunctions of the sampling covariance matrices.

Tensorial product of B-splines We shall assume that there exists a positive integer K and a sequence of knots ξ such that

$$\sigma_c(t,s) = \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} \beta_{c,k_1,k_2} B_{k_1}(t) B_{k_2}(s) = \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} \beta_{c,k_1,k_2} B_{k_1,k_2}(t,s),$$

wher $B_k(t)$, $k = 1, \ldots, K$ are cubic B-splines. In this case, we have the estimative

$$\hat{\sigma}_c(t,s) = \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} \hat{b}_{c,k_1,k_2} B_{k_1}(t) B_{k_2}(s) = \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} \hat{b}_{c,k_1,k_2} B_{k_1,k_2}(t,s),$$

for $c = 1, 2, \ldots, C$, where $\hat{b}_{c,k_1,k_2}$ is the least squares solution of

$$\hat{b}_{c,k_1,k_2} = \arg\min_{b_{c,k_1,k_2}} \sum_{i=1}^{M} \left(\hat{Z}_i(t,s) - \sum_{c=1}^{C} \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} N_{c,i} b_{c,k_1,k_2} B_{k_1,k_2}(t,s) \right)^2$$

$\hat{\sigma}_c$ might not be a covariance function because it might not be positive definite, positiva-definida,

$$\int \int \sigma_c(s,t) h(s) h(t)\, ds\, dt \geq 0$$

for all integrable functions h.

Dias, Garcia & Martarelli (2009) propose the following adaptation of the algorithm developed by Hall, Fisher & Hoffman (1994) that guarantees it to be a positive definite function.

Note the spectral representation of σ_c is given by:

$$\sigma_c(s,t) \;=\; \int_{\mathbb{R}} \int_{\mathbb{R}} \exp\{i(s\omega_1 - t\omega_2)\} d^2 F_c(\omega_1, \omega_2),$$

where $F_c(\omega_1, \omega_2)$ is the spectral function.

If F has a densitity with respect to the Lebesgue measure, this spectral density f_c is again the Fourier transform of the covariance function, that is

$$f_c(\omega_1, \omega_2) \;=\; \frac{1}{(2\pi)^2} \int_0^T \int_0^T \exp\{-i(s\omega_1 - t\omega_2)\} \sigma_c(s,t)\, ds\, dt.$$

The proposed algorithm consists in:

Step 1: Find the estimative $\hat{\sigma}_c$.

Step 2: Find the Fourier transform, $\hat{f}_c$, of $\hat{\sigma}_c$.

Step 3: Obtain $\tilde{f}_c$ by truncating $\hat{f}_c$ at 0, and if necessary smooth it.

Step 4: Invert $\tilde{f}_c$ to obtain the final estimative $\tilde{\sigma}_c$.

Nonparametric model Dias, Garcia, Schmidt (2010) use a nonparametric model for the covariance function in the Bayesian framework. That is, assume that the covariance function has form

$$\sigma_c(t,s) = \eta_c(t)\eta_c(s)e^{-\phi_c|t-s|} \tag{4}$$

with possibly a expansion in B-spline basis

$$\eta_c(t) \;=\; \sum_{k=1}^{K} \theta_{c,k} B_k(t),$$

where $B_k(t)$, $k = 1, \ldots, K$ are cubic B-splines.

Remark that the proposed model in (4) guarantees that for all functions $\eta(\cdot)$ the function $\sigma_c(\cdot,\cdot)$ is a covariance function, because if $\gamma(t) \sim GP(0, \delta(\cdot,\cdot))$ is a zero mean Gaussian process with covariance function $\delta(t,s) = e^{-\phi_c|t-s|}$, then

$$|\eta_c(t)|\gamma(t) \sim GP(0, \sigma_c(\cdot,\cdot))$$

is a zero mean Gaussian process with covariance function $\sigma_c(t,s) = \eta_c(t)\eta_c(s)e^{-\phi_c|t-s|}$.

References

[1] Dias, R., Garcia, N. L. & Martarelli, A. (2009) Non-parametric estimation for aggregated functional data for electric load monitoring. *Environmetrics*, **20**, p. 111-130, 2009.

[2] Dias, R., Garcia, N. L. & Schmidt, A. (2010) A Bayesian Model for Aggregated Functional Data - Basis expansion. Preprint.

[3] Hall, P., Fisher, N. I. & Hoffman, B. (1994) On the nonparametric estimation of covariance functions. *Ann. Statist.*, **22**(4), 2115–2134.

[4] Lee, S.Y., Zhang, W. & Song, X.Y. (2002) Estimating the covariance function with functional data. *The British Journal of Mathematical and Statistical Psychology*, **55**, 247–261.

[5] Ramsay, J. O. & Silverman, B. W. (2005) Functional Data Analysis, Second Edition, *Springer.*