Semiparametric models with functional responses in a model assisted survey sampling setting

Hervé Cardot[1], Alain Dessertaine[2], and Etienne Josserand[1]

[1]Institut de Mathématiques de Bourgogne, UMR 5584 CNRS
herve.cardot@u-bourgogne.fr, etienne.josserand@u-bourgogne.fr
[2]EDF, R&D, ICAME - SOAD
alain.dessertaine@edf.fr

Computational Statistics - Paris - August 23rd 2010

Outline

Introduction

Survey sampling and curve estimation

Estimation with auxiliary information

Application to electricity consumption curves

Sampling survey on curves

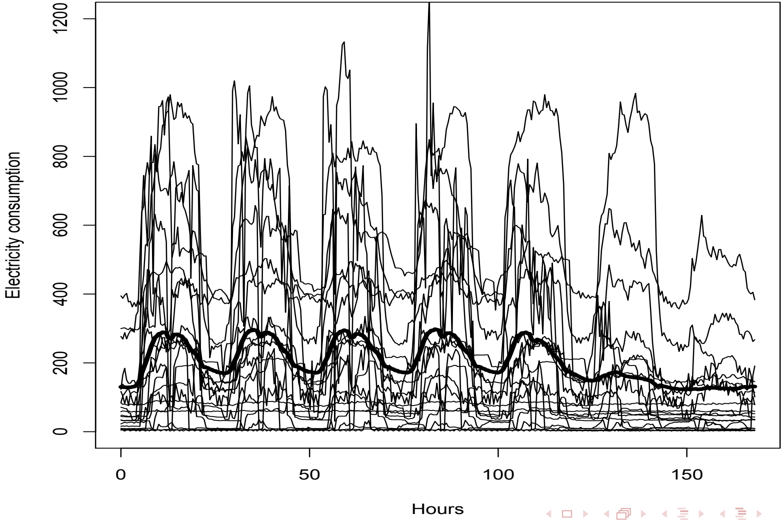
A new subject in statistic boundaries between functional data analysis and survey sampling theory.

EDF problematic :

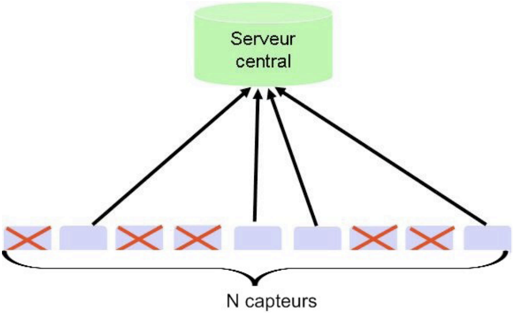
- EDF does not know what their clients consume at each time !
- EDF plans to install electricity meters which will be able to send individual electricity consumptions at very fine time scales.
- Collecting, saving and analysing all this information would be very expensive ($\approx$ 30 millions of electricity meters).
- How to estimate as precisely as possible the mean consumption curve in France or a part of this (particular region, type of clients, ...) ?

Consumption curves

A sample of individual electricity consumption curves measured every half hour during one week.



Survey sampling in large databases of functional data



Chiky 2009 (these, ENST) : survey sampling procedures on the sensors, which allow a trade off between limited storage capacities and accuracy of the data, can be relevant approaches compared to signal compression in order to get accurate approximations to simple estimates such as mean or total trajectories.

Sampling design and mean curve estimation

A population $U = \{1, \dots, k, \dots, N\}$ with finite size N.
At each individual (statistic unit) k of the population U, we associate a deterministic curve

$$Y_k \; = \; (Y_k(t))_{t\in[0,T]} \in C[0,T].$$

Let $\mu \in C[0,T]$, the *mean* of Y_k in the population

$$\mu(t) = \frac{1}{N}\sum_{k\in U} Y_k(t), \quad t \in [0,T].$$

A sample s, *i.e.* a part $s \subset U$, with known size n,
and p a probability law on the set of parts on U,

- $\pi_k = \Pr(k \in s) > 0$ for all $k \in U$,
- $\pi_{kl} = \Pr(k \,\&\, l \in s) > 0$ for all $k, l \in U$, $k \neq l$.

The Horvitz-Thompson estimator of the mean curve is

$$\widehat{\mu}(t) \; = \; \frac{1}{N}\sum_{k\in s}\frac{Y_k(t)}{\pi_k} = \frac{1}{N}\sum_{k\in U}\frac{Y_k(t)}{\pi_k}\mathbb{1}_{k\in s}, \quad t\in[0,T].$$

Two classical sampling designs

- The simple random sampling without replacement with size n
  - $\pi_k = \frac{n}{N}$for all $k \in U$
  - $\pi_{kl} = \frac{n(n-1)}{N(N-1)}$ for all $k, l \in U$, $k \neq l$

We find again the *common* mean estimator

$$\widehat{\mu}(t) = \frac{1}{n} \sum_{k \in s} Y_k(t).$$

- Stratified sampling with size n.

The population U is stratified in H stratum $\bigcup_{h=1}^{H} U_h = U$, with size N_h

- $\pi_k = \frac{n_h}{N_h}$ for all $k \in U_h$
- $\pi_{kl} = \frac{n_h(n_h-1)}{N_h(N_h-1)}$ for all $k, l \in U_h$, $k \neq l$
- $\pi_{kl} = \frac{n_h n_\ell}{N_h N_\ell}$ for all $k \in U_h$, $l \in U_\ell$, $h \neq \ell$

So

$$\widehat{\mu}(t) = \frac{1}{N} \sum_{h \in H} N_h \frac{1}{n_h} \sum_{k \in s_h} Y_k(t) \; = \; \frac{1}{N} \sum_{h \in H} N_h \, \widehat{\mu}_h(t).$$

Utilization of auxiliary information

Considering information given by m auxiliary variables

- meteorological : temperature, cloud covering , ...
- geographical : altitude, longitude, latitude, ...
- behavioral : past mean consumption, ...

would be able to improve the estimator accuracy of the mean curve.

This requires modeling the behavior of individual electricity meters that are not in the sample :

$$Y_k(t) = \mu(t) + f(x_{k1}, \ldots, x_{km}, t) + error$$

▷ Not much hope to obtain directly an accurate and flexible estimator of the function f which depends on time t and covariables $X_1, \ldots, X_m$.

- Reducing the dimension of data seems to be an interesting way.

Dimension reduction in finite population

▷ The best linear approximation, with quadratic error, of functions Y_k in a functional space of fixed dimension q, $q < N$, generated by q orthonormal functions $\phi_1, \ldots, \phi_q$:

$$Y_k(t) \quad = \quad \mu(t) + \sum_{j=1}^{q} \langle Y_k - \mu, \phi_j \rangle \phi_j(t) + R_{qk}(t)$$

The mean rest with the norm $L^2[0, T]$ satisfies

$$\frac{1}{N} \sum_{k \in U} \|R_{qk}\|^2 = \frac{1}{N} \sum_{k \in U} \|Y_k - \mu\|^2 - \sum_{j=1}^{q} \langle \Gamma \phi_j, \phi_j \rangle$$

where the covariance operator Γ is associated with the covariance function

$$\gamma(s, t) = \frac{1}{N} \sum_{k \in U} (Y_k(t) - \mu(t)) (Y_k(s) - \mu(s)),$$

where for all $f \in L^2[0, T]$, $\Gamma f(s) = \int_0^T \gamma(s, t) f(t) dt \ , \quad s \in [0, T]$.

To minimize against $\phi_1, \ldots, \phi_q$, the mean rest $\frac{1}{N} \sum_{k \in U} \|R_{qk}\|^2$ is the same to find eigen vectors of Γ.

Model on principal components

Property *The rest is minimal for $\phi_1 = v_1, \ldots, \phi_q = v_q$, where*

$$\Gamma v_j(t) \quad = \quad \lambda_j \; v_j(t), \quad t \in [0, T],$$

the functions v_j constitute an orthonormal system in $L^2[0, T]$
the eigen values are sorted, $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_N \geq 0$.

- Obtaining estimations of individual variations on principal components (real)

$$\langle Y_k - \mu, v_j \rangle \approx g_j(x_{k1}, \ldots, x_{km})$$

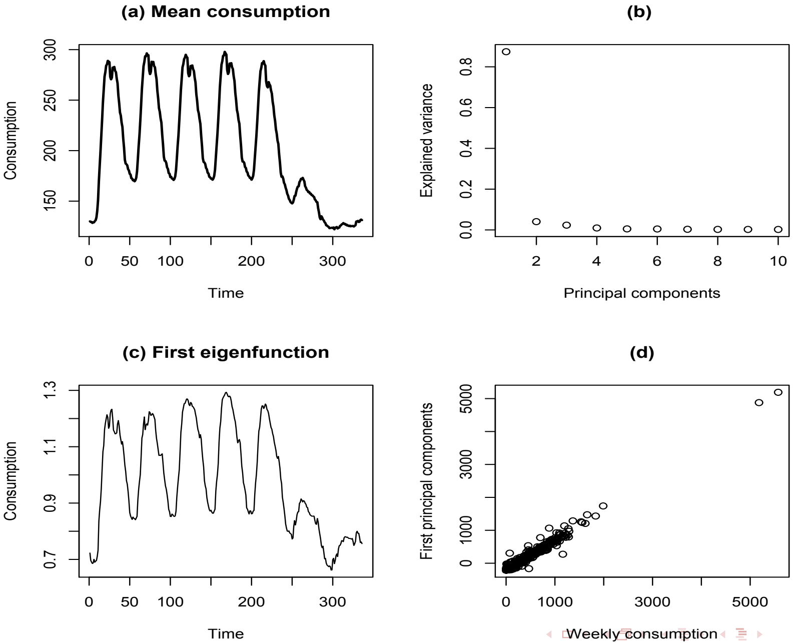
allow the application of model-assisted techniques to build an estimator of μ

$$\widehat{\mu}_x(t) = \widehat{\mu}(t) - \frac{1}{N}\left(\sum_{k \in s} \frac{\widehat{Y}_k(t)}{\pi_k} - \sum_{k \in U} \widehat{Y}_k(t)\right)$$

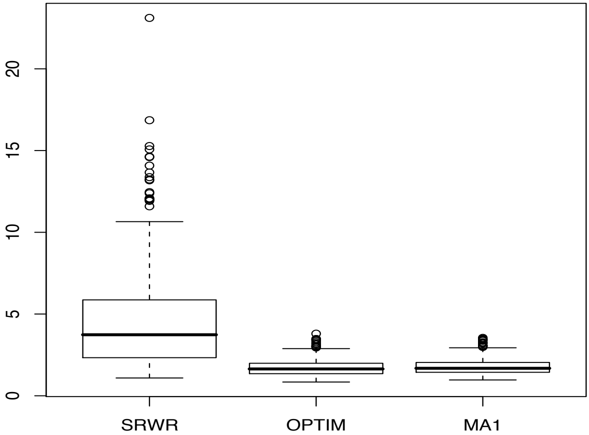
where

$$\widehat{Y}_k(t) = \widehat{\mu}(t) + \sum_{j=1}^{q} \widehat{g}_j(x_{k1}, \ldots, x_{km}) \; \widehat{v}_j(t).$$

An illustration of EDF consumption curves



Error estimation of μ : $\|\widehat{\mu} - \mu\|$

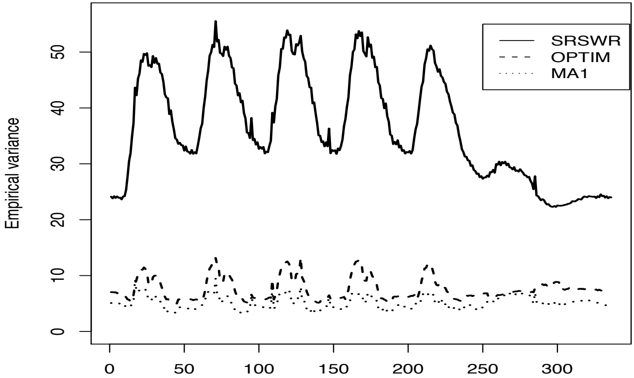


The model (MA1) considered is very simple

$$\widehat{Y}_k(t) = \widehat{\mu}(t) + (\widehat{\beta}_0 + \widehat{\beta}_1 X_k)\ \widehat{v}_1(t)$$

where X_k is the mean consumption of the last week.

Variances comparison $\gamma(t,t)$ of estimators $\widehat{\mu}$



Problem : Lack of explicit formula for variance estimation

- Candidate for asymptotic formula (when $n, N \to \infty$)
- Need a corrected variance which depends on eigen vectors's variances (perturbations) ?
- ...