SEMIPARAMETRIC MODELS WITH FUNCTIONAL RESPONSES IN A MODEL ASSISTED SURVEY SAMPLING SETTING Submitted to COMPSTAT 2010

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

[1] Institut de Mathématiques de Bourgogne, UMR 5584 CNRS,
Université de Bourgogne, 9, Av. A. Savary - B.P. 47 870, 21078 Dijon, France
herve.cardot@u-bourgogne.fr, etienne.josserand@u-bourgogne.fr
[2] EDF, R&D, ICAME - SOAD
1, Av. du Général de Gaulle, 92141 Clamart , France
alain.dessertaine@edf.fr

Abstract. This work adopts a survey sampling point of view to estimate the mean curve of large databases of functional data. When storage capacities are limited, selecting, with survey techniques a small fraction of the observations is an interesting alternative to signal compression techniques. We propose here to take account of real or multivariate auxiliary information available at a low cost for the whole population, with semiparametric model assisted approaches, in order to improve the accuracy of Horvitz-Thompson estimators of the mean curve. We first estimate the functional principal components with a design based point of view in order to reduce the dimension of the signals and then propose semiparametric models to get estimations of the curves that are not observed. This technique is shown to be really effective on a real dataset of 18902 electricity meters measuring every half an hour electricity consumption during two weeks.

Keywords: Design-based estimation, Functional Principal Components, Electricity consumption, Horvitz-Thompson estimator

1 Introduction

With the development of distributed sensors one can have access of potentially huge databases of signals evolving along fine time scales. Collecting in an exhaustive way such data would require very high investments both for transmission of the signals through networks as well as for storage. As noticed in Chiky and Hébrail (2009) 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. Our study is motivated, in such a context of distributed data streams, by the estimation of the temporal evolution of electricity consumption curves. The French operator EDF has planned to install in a few years

more than 30 millions electricity meters, in each firm and household, that will be able to send individual electricity consumptions at very fine time scales. Collecting, saving and analysing all this information which can be seen as functional would be very expensive and survey sampling strategies are interesting to get accurate estimations at reasonable costs (Dessertaine, 2006). It is well known that consumption profiles strongly depend on covariates such as past consumptions, meteorological characteristics (temperature, nebulosity, *etc*) or geographical information (altitude, latitude and longitude). Taking this information into account at an individual level (*i.e* for each electricity meter) is not trivial.

We have a test population of $N = 18902$ electricity meters that have collected electricity consumptions every half an hour during a period of two weeks, so that we have $d = 336$ time points. We are interested in estimating the mean consumption curve during the second week and we suppose that we know the mean consumption, $\bar{Y}_k = \frac{1}{336}\sum_{j=1}^{336} Y_k(t_j)$, for each meter k of the population during the first week. This mean consumption will play the role of auxiliary information. Note that meteorological variables are not available in this preliminary study.

One way to achieve this consists in reducing first the high dimension of the data by performing a functional principal components analysis in a survey sampling framework with a design based approach (Cardot *et al.*, 2010). It is then possible to build models, parametric or nonparametric, on the principal component scores in order to incorporate the auxiliary variables effects and correct our estimator with model assisted approaches (Särndal *et al.*, 1992). Note that this strategy based on modeling the principal components instead of the original signal has already been proposed, with a frequentist point of view, by Chiou *et al.* (2003) with singel index models and Müller and Yao (2008) with additive models.

We present in section 2 the Horvitz-Thomposon estimator of the mean consumption profile as well as the functional principal components analysis. We develop, in section 3, model assisted approaches based on statistical modeling of the principal components scores and derive an approximated variance that can be useful to build global confidence bands. Finally, we illustrate, in section 4, the effectiveness of this methodology which allows to improve significantly more basic approaches on a population of 18902 electricity consumption curves measured every half an hour during one week.

2 Functional data in a finite population

Let us consider a finite population $U = \{1, \ldots, k, \ldots, N\}$ of size N, and suppose we can observe, for each element k of the population U, a deterministic curve $Y_k = (Y_k(t))_{t\in[0,1]}$ that is supposed to belong to $L^2[0,1]$, the space of square integrable functions defined on the closed interval $[0,1]$ equipped with its usual inner product $\langle\cdot,\cdot\rangle$ and norm denoted by $\|\cdot\|$. Let us define the

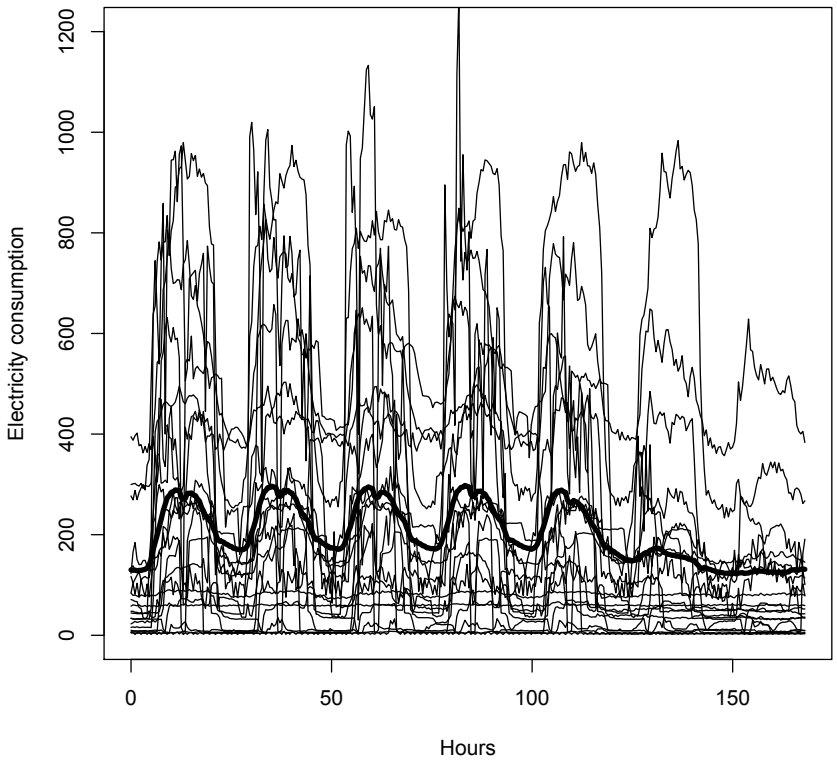

Fig. 1. Mean curve and sample of individual electricity consumption curves.

mean population curve $\mu \in L^2[0,1]$ by

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

Consider now a sample s, *i.e.* a subset $s \subset U$, with known size n, chosen randomly according to a known probability distribution p defined on all the subsets of U. We suppose that all the individuals in the population can be selected, with probabilities that may be unequal, $\pi_k = \Pr(k \in s) > 0$ for all $k \in U$ and $\pi_{kl} = \Pr(k \ \& \ l \in s) > 0$ for all $k, l \in U$, $k \neq l$. The Horvitz-Thompson estimator of the mean curve, which is unbiased, is given by

$$\widehat{\mu}(t) = \frac{1}{N} \sum_{k \in s} \frac{Y_k(t)}{\pi_k} = \frac{1}{N} \sum_{k \in U} \frac{Y_k(t)}{\pi_k} \mathbf{1}_{k \in s}, \quad t \in [0,1]. \tag{2}$$

As in Cardot *et al.* (2010) we would like to describe now the individual variations around the mean function in a functional space whose dimension is as small as possible according to a quadratic criterion. Let us consider a set of q orthonormal functions of $L^2[0,1]$, $\phi_1, \ldots, \phi_q$, and minimize, according to $\phi_1, \ldots, \phi_q$, the remainder $R(q)$ of the projection of the Y_k's onto the space generated by these q functions

$$R(q) = \frac{1}{N} \sum_{k \in U} \|R_{qk}\|^2$$

with $R_{qk}(t) = Y_k(t) - \mu(t) - \sum_{j=1}^{q} \langle Y_k - \mu, \phi_j \rangle \phi_j(t),\ t \in [0,1]$. Introducing now the population covariance function $\gamma(s,t)$,

$$\gamma(s,t) = \frac{1}{N} \sum_{k \in U} \left(Y_k(t) - \mu(t)\right) \left(Y_k(s) - \mu(s)\right), \quad (s,t) \in [0,1] \times [0,1],$$

Cardot *et al.* (2010) have shown that $R(q)$ attains its minimum when $\phi_1, \ldots, \phi_q$ are the eigenfunctions of the covariance operator Γ associated to the largest eigenvalues, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_q \geq 0$,

$$\Gamma \phi_j(t) = \int_0^1 \gamma(s,t) \phi_j(s) ds = \lambda_j \phi_j(t), \quad t \in [0,1], j \geq 1.$$

When observing individuals from a sample s, a simple estimator of the covariance function

$$\widehat{\gamma}(s,t) = \frac{1}{N} \sum_{k \in s} \frac{1}{\pi_k} \left(Y_k(t) - \widehat{\mu}(t)\right) \left(Y_k(s) - \widehat{\mu}(s)\right)\ (s,t) \in [0,1] \times [0,1], \quad (3)$$

allows to derive directly estimators of the eigenvalues $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_q$ and the corresponding eigenfuctions $\widehat{\phi}_1, \ldots, \widehat{\phi}_q$.

Remark: *with real data, one only gets discretized trajectories of the Y_k at d points, $t_1, \ldots, t_d$, so that we observe $\mathbf{Y}_k = (Y_k(t_1), \ldots, Y_k(t_d)) \in \mathbb{R}^d$. When observations are not corrupted by noise, linear interpolation allows to get accurate approximations to the true trajectories,*

$$\widetilde{Y}_k(t) = Y_k(t_j) + \frac{Y_k(t_{j+1}) - Y_k(t_j)}{t_{j+1} - t_j}(t - t_j), \quad t \in [t_j, t_{j+1}]$$

and to build consistent estimates of the mean function provided the grid of time points is dense enough (Cardot and Josserand, 2009).

3 Semiparametric estimation with auxiliary information

Suppose now we have access to m auxiliary variables $X_1, \ldots, X_m$ that are supposed to be linked to the individual curves Y_k and we are able to observe

these variables, at a low cost, for every individual k in the population. Taking this additional information into account would certainly be helpful to improve the accuracy of the basic estimator $\widehat{\mu}$. Going back to the decomposition of the individual trajectories Y_k on the eigenfunctions,

$$Y_k(t) = \mu(t) + \sum_{j=1}^{q} \langle Y_k - \mu, \phi_j \rangle \phi_j(t) + R_{qk}(t), \quad t \in [0,1],$$

and borrowing ideas from Chiou *et al.* (2003) and Müller and Yao (2008), an interesting approach consists in modeling the population principal components scores $\langle Y_k - \mu, \phi_j \rangle$ with respect to auxiliary variables at each level j of the decomposition on the eigenfunctions, $\langle Y_k - \mu, \phi_j \rangle \approx f_j(x_{k1}, \ldots, x_{km})$ where the regression function f_j can be parametric or not and $(x_{k1}, \ldots, x_{km})$ is the vector of observations of the m auxiliary variables for individual k.

It is possible to estimate the principal component scores

$$\widehat{C}_{kj} = \langle Y_k - \widehat{\mu}, \widehat{\phi}_j \rangle,$$

for $j = 1, \ldots, q$ and all $k \in s$. Then, a design based least squares estimator for the functions f_j

$$\widehat{f}_j = \arg\min_{g_j} \sum_{k \in s} \frac{1}{\pi_k} \left(\widehat{C}_{kj} - g_j(x_{k1}, \ldots, x_{km}) \right)^2, \tag{4}$$

is useful to construct the following model-assisted estimator $\widehat{\mu}_X$ of μ,

$$\widehat{\mu}_x(t) = \widehat{\mu}(t) - \frac{1}{N} \left(\sum_{k \in s} \frac{\widehat{Y}_k(t)}{\pi_k} - \sum_{k \in U} \widehat{Y}_k(t) \right) \tag{5}$$

where the predicted curves $\widehat{Y}_k$ are estimated for all the individuals of the population U thanks to the m auxiliary variables,

$$\widehat{Y}_k(t) = \widehat{\mu}(t) + \sum_{j=1}^{q} \widehat{f}_j(x_{k1}, \ldots, x_{km})\ \widehat{v}_j(t),\ t \in [0,1].$$

4 Estimation of electricity consumption curves

We consider now the population consisting in the $N = 18902$ electricity consumption curves measured during the second week very half an hour. We have $d = 336$ time points. Note that meteorological variables are not available in this preliminary study and our auxiliary information is the mean consumption, for each meter k, during the first week.

We first perform a simple random sampling without replacement (SRSWR) with fixed size of $n = 2000$ electricity meters during the second week

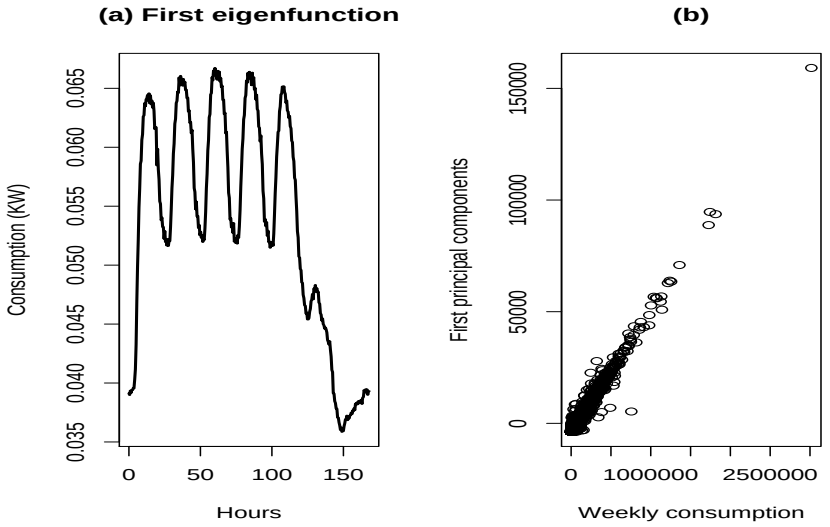

Fig. 2. Mean curve and sample of individual electricity consumption curves.

in order to get $\widehat{\mu}$ and perform the functional principal components analysis (FPCA). The true mean consumption curve $\mu(t)$ during this period is drawn in Figure 1 whereas Figure 2 (a) present the result of the FPCA. The first principal component explains more than 80% of the total variance telling us that there is a strong temporal structure in these data. The associated estimated eigenfunction $\widehat{\phi}_1$ presents strong daily periodicity. Looking now at the relationship between the estimated first principal components and the auxiliary variable, we can notice that there is a strong linear relationship between these two variables and thus considering a linear regression model for estimating f_1 seems to be appropriate.

To evaluate the accuracy of estimator (5) we made 500 replications of the following scheme

- Draw a sample of size $n = 2000$ in population U with SRSWR and estimate $\widehat{\mu}, \widehat{\phi}_1$ and $\widehat{C}_{k1}$, for $k \in s$, during the second week.
- Estimate a linear relationship between X_k and $\widehat{C}_{k1}$, for $k \in s$ where $X_k = \frac{1}{336}\sum_{j=1}^{336} Y_k(t_j)$ is the mean consumption during the first week, and predict the principal component using the estimated relation $\widehat{C}_{k1} \approx \widehat{\beta}_0 + \widehat{\beta}_1 X_k$.
- Estimate $\widehat{\mu}_X$ taking the auxiliary information into account with equation (5).

The following loss criterion $\int |\mu(t) - \widehat{\mu}(t)|dt$ has been considered to evaluate the accuracy of the estimators $\widehat{\mu}$ and $\widehat{\mu}_X$. We also compare the estimation error with an optimal stratification sampling scheme in which strata are built

	SAS	OPTIM	MA1
Mean	4.245	1.687	1.866
Median	3.348	1.613	1.813
First quartile	2.213	1.343	1.499
Third quartile	5.525	1.944	2.097

Table 1. Comparison of mean absolute deviation from the true mean curve for SRSWR, optimal allocation for stratification (OPTIM) and model assisted (MA1) estimation procedures.

on the curves of the population observed during the first week. As in Cardot and Josserand (2009), the population is partitioned into $K = 7$ strata thanks to a k-means algorithm. It is then possible to determine the optimal allocation weights, according to a mean variance criterion, in each stratum for the stratified sampling procedure during the second week.

The estimation errors are presented in Table 1 for the three estimators. We first remark that considering optimal stratification (OPTIM) or model assisted estimators (MA1) lead to a significant improvement compared to the basic SRSWR approach. Secondly, the performances of the stratification and the model assisted approaches are very similar in terms of accuracy but they do not need the same amount of information. The optimal stratification approach necessitates to know the cluster of each individual of the population and the covariance function within each cluster whereas the model assisted estimator only needs the past mean consumption for each element of the population.

Looking now at the empirical variance, at each instant, of these estimators, we see in Figure (3) that the simple SRSWR has much larger variances, in which we recognize the first eigenfunction of the covariance operator, than the more sophisticated OPTIM and MA1. Among these two estimators the model assisted estimator has a smaller pointwise variance, indicating that it is certainly more reliable.

Acknowledgment. Etienne Josserand thanks the Conseil Régional de Bourgogne, France, for its financial support (FABER PhD grant).

References

CARDOT, H., CHAOUCH, M., GOGA, C. and C. LABRUÈRE (2010). Properties of Design-Based Functional Principal Components Analysis, *J. Statist. Planning and Inference.*, **140**, 75-91.

CARDOT, H., JOSSERAND, E. (2009). Horvitz-Thompson Estimators for Functional Data: Asymptotic Confidence Bands and Optimal Allocation for Stratified Sampling. `http://arxiv.org/abs/0912.3891`.

CHIKY, R., HEBRAIL, G. (2009). Spatio-temporal sampling of distributed data streams. *J. of Computing Science and Engineering*, to appear.

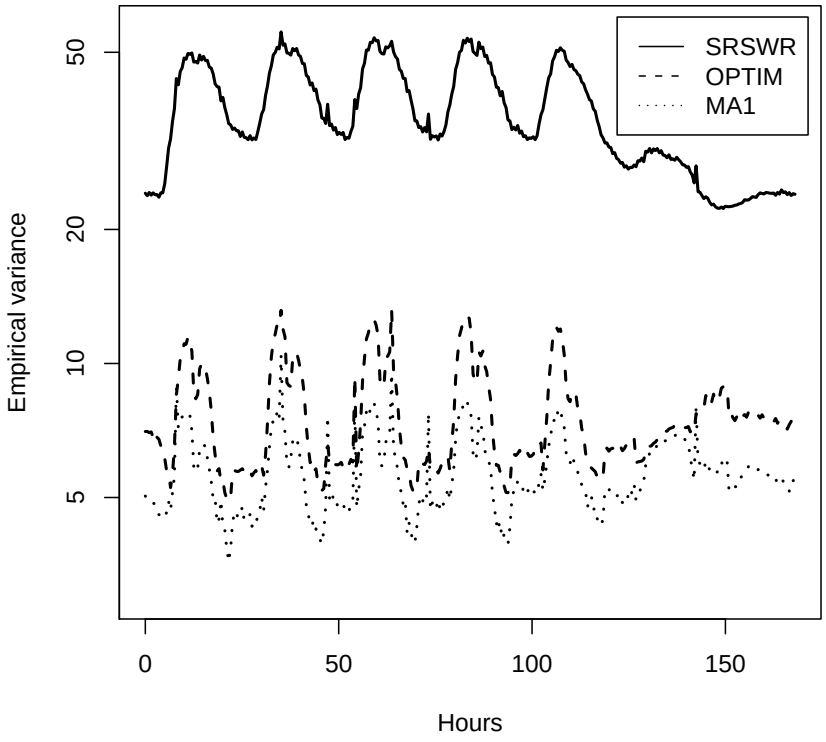

Fig. 3. Comparison of the empirical pointwise variance for SRSWR, optimal allocation for stratification (OPTIM) and model assisted (MA1) estimation procedures.

CHIOU, J-M., MÜLLER, H.G. and WANG, J.L. (2003). Functional quasi-likelihood regression models with smooth random effects. *J.Roy. Statist. Soc., Ser. B*, **65**, 405-423.

DESSERTAINE, A. (2006). Sampling and Data-Stream : some ideas to built balanced sampling using auxiliary hilbertian informations. *56th ISI Conference*, Lisboa, Portugal, 22-29 August 2007.

MÜLLER, H-G., YAO, F. (2008). Functional Additive Model. *J. Am. Statist. Ass.* **103**, 1534-1544.

SÄRNDAL, C.E., SWENSSON, B. and J. WRETMAN, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag.

SKINNER, C.J, HOLMES, D.J, SMITH, T.M.F (1986). The Effect of Sample Design on Principal Components Analysis. *J. Am. Statist. Ass.* **81**, 789-798.