Confidence bands for Horvitz-Thompson estimators using sampled noisy functional data

Cardot, Hervé
Institut de Mathématiques de Bourgogne, UMR 5584 CNRS
9 Avenue Alain Savary
21078 Dijon, France
E-mail: herve.cardot@u-bourgogne.fr

Degras, David
Statistical and Applied Mathematical Sciences Institute
19 T.W. Alexander Drive, P.O. Box 14006
Research Triangle Park, NC 27709, USA
E-mail: ddegras@samsi.info

Josserand, Etienne
Institut de Mathématiques de Bourgogne, UMR 5584 CNRS
9 Avenue Alain Savary
21078 Dijon, France
E-mail: etienne.josserand@u-bourgogne.fr

1 Introduction

The recent development of automated sensors has given access to very large collections of signals sampled at fine time scales. However, exhaustive transmission, storage, and analysis of such massive functional data may incur very large investments. In this context, when the goal is to assess a global indicator like the mean temporal signal, survey sampling techniques are appealing solutions as they offer a good trade-off between statistical accuracy and global cost of the analysis. They are in particular competitive with signal compression techniques (Chiky and Hébrail, 2008). Focusing on sampling schemes, Cardot and Josserand (2011) estimate the mean electricity consumption curve in a population of about 19,000 customers whose electricity meters were read every 30 minutes during one week. Assuming exact measurements, they first perform a linear interpolation of the discretized signals and then consider a functional version of the Horvitz-Thompson estimator. They show that estimation can be greatly improved by utilizing stratified sampling over simple random sampling and they extend the Neyman optimal allocation rule (see *e.g.* Fuller (2009)) to the functional setup. As a first contribution, the present work generalizes the framework of Cardot and Josserand (2011) to noisy functional data. Assuming data are observed with errors that may be correlated over time, we replace the interpolation step in their procedure by a local polynomial smoothing step. This sensibly improves the estimation when the noise level is moderate to high.

In relation to mean function estimation, a key statistical task is to build confidence regions. Following the ideas in Degras (2010), we build confidence bands in the finite population setting. Specifically we derive a CLT for the mean function estimator in the space of continuous functions, and then show that the supremum of the limiting process can be approximated in distribution by simulating Gaussian processes conditional on the estimated covariance function (Cardot *et al.*, 2011). The bands attain nominal coverage and are easy and quick to implement.

Finally, the implementation of our mean function estimator requires to select a bandwidth in the data smoothing step. Objective, data-driven methods are desirable for this purpose. As explained by Opsomer and Miller (2005), bandwidth selection in the survey estimation context poses specific prob-

lems that make usual cross-validation or mean square error optimization methods inadequate. In view of the model-assisted survey estimation of a population total, these authors propose a cross-validation method that aims at minimizing the variance of the estimator, the bias component being negligible in their setting. In our functional and design-based framework, the bias is no longer negligible. We therefore devise a weighted cross-validation criterion based on weighted least squares, with weights proportional to the sampling weights. In the case of simple random sampling without replacement, this criterion reduces to the cross-validation technique of Rice and Silverman (1991).

2 Notations and estimators

Consider a finite population $U_N = \{1, \ldots, N\}$ of size N and suppose that to each unit $k \in U_N$ corresponds a real function X_k on $[0, T]$, with $T < \infty$. We assume that each trajectory X_k belongs to the space of continuous functions $C([0, T])$. Our target is the mean trajectory $\mu_N(t),\ t \in [0, T]$, defined as follows:

$$(1) \quad \mu_N(t) \ = \ \frac{1}{N} \sum_{k \in U} X_k(t).$$

We consider a random sample s drawn from U_N without replacement according to a fixed-size sampling design $p_N(s)$, where $p_N(s)$ is the probability of drawing the sample s. The size n_N of s is nonrandom and we suppose that the first and second order inclusion probabilities satisfy $\pi_k := \mathbb{P}(k \in s) > 0$ for all $k \in U_N$, and $\pi_{kl} := \mathbb{P}(k \& l \in s) > 0$ for all $k, l \in U_N$, so that each unit and each pair of units can be drawn with a non null probability from the population. Note that for simplicity of notation the subscript N has been omitted. Also, by convention, we write $\pi_{kk} = \pi_k$ for all $k \in U_N$.

Assume that noisy measurements of the sampled curves are available at $d = d_N$ fixed discretization points $0 = t_1 < t_2 < \ldots < t_d = T$. For all unit $k \in s$, we observe

$$(2) \quad Y_{jk} = X_k(t_j) + \epsilon_{jk}$$

where the measurement errors ϵ_{jk} are centered random variables that are independent across the index k (units) but not necessarily across j (possible temporal dependence). It is also assumed that the random sample s is independent of the noise ϵ_{jk} and the trajectories $X_k(t), t \in [0, T]$ are deterministic.

Our goal is to estimate μ_N as accurately as possible and to build asymptotic confidence bands, as in Degras (2010) and Cardot and Josserand (2011). For this, we must have a uniformly consistent estimator of its covariance function.

2.1 Linear smoothers and the Horvitz-Thompson estimator

For each (potentially observed) unit $k \in U_N$, we aim at recovering the curve X_k by smoothing the corresponding discretized trajectory $(Y_{1k}, \ldots, Y_{dk})$ with a linear smoother (e.g. spline, kernel, or local polynomial):

$$(3) \quad \widehat{X}_k(t) = \sum_{j=1}^{d} W_j(t) Y_{jk}.$$

Note that the reconstruction can only be performed for the observed units $k \in s$. Here we use local linear smoothers (see *e.g.* Fan and Gijbels (1997)) because of their wide popularity, good statistical

properties, and mathematical convenience. The weight functions $W_j(t)$ express as

$$W_j(t) = \frac{\frac{1}{dh}\left\{s_2(t) - (t_j - t)s_1(t)\right\} K\left(\frac{t_j - t}{h}\right)}{s_2(t)s_0(t) - s_1^2(t)}, \quad j = 1, \ldots, d, \tag{4}$$

where K is a kernel function, $h > 0$ is a bandwidth, and

$$s_l(x) = \frac{1}{dh}\sum_{j=1}^{d}(t_j - t)^l K\left(\frac{t_j - t}{h}\right), \quad l = 0, 1, 2. \tag{5}$$

We suppose that the kernel K is nonnegative, has compact support, satisfies $K(0) > 0$ and $|K(s) - K(t)| \leq C|s - t|$ for some finite constant C and for all $s, t \in [0, T]$.

The classical Horvitz-Thompson estimator (see *e.g.* Fuller (2009)) of the mean curve is

$$\widehat{\mu}_N(t) = \frac{1}{N}\sum_{k\in s}\frac{\widehat{X}_k(t)}{\pi_k} = \frac{1}{N}\sum_{k\in U}\frac{\widehat{X}_k(t)}{\pi_k}I_k, \tag{6}$$

where I_k is the sample membership indicator ($I_k = 1$ if $k \in s$ and $I_k = 0$ otherwise). It holds that $\mathbb{E}(I_k) = \pi_k$ and $\mathbb{E}(I_k I_l) = \pi_{kl}$.

2.2 Covariance estimation

The covariance function of $\widehat{\mu}_N$ can be written as

$$\mathrm{Cov}\left(\widehat{\mu}_N(s), \widehat{\mu}_N(t)\right) = \frac{1}{N}\gamma_N(s, t) \tag{7}$$

for all $s, t \in [0, T]$, where

$$\gamma_N(s, t) = \frac{1}{N}\sum_{k,l\in U}\Delta_{kl}\frac{\tilde{X}_k(s)}{\pi_k}\frac{\tilde{X}_l(t)}{\pi_l} + \frac{1}{N}\sum_{k\in U}\frac{1}{\pi_k}\mathbb{E}\left(\tilde{\epsilon}_k(s)\tilde{\epsilon}_k(t)\right) \tag{8}$$

with

$$\begin{cases} \tilde{X}_k(t) &= \sum_{j=1}^{d} W_j(t)X_k(t_j), \\ \tilde{\epsilon}_k(t) &= \sum_{j=1}^{d} W_j(t)\epsilon_{kj}, \\ \Delta_{kl} &= \mathrm{Cov}(I_k, I_l) = \pi_{kl} - \pi_k\pi_l. \end{cases} \tag{9}$$

A natural estimator of $\gamma_N(s, t)$ (see *e.g.* Fuller (2009)) is given by

$$\widehat{\gamma}_N(s, t) = \frac{1}{N}\sum_{k,l\in U}\frac{\Delta_{kl}}{\pi_{kl}}\left(\frac{I_k}{\pi_k}\frac{I_l}{\pi_l}\right)\widehat{X}_k(s)\widehat{X}_l(t). \tag{10}$$

It is unbiased and uniformly consistency in mean square (Cardot *et al.*, 2011).

3 Global confidence bands

In this section we build global confidence bands for μ_N of the form

$$\left\{\left[\widehat{\mu}_N(t) \pm c\,\frac{\widehat{\sigma}_N(t)}{N^{1/2}}\right],\ t \in [0, T]\right\}, \tag{11}$$

where c is a suitable number and $\widehat{\sigma}_N(t) = \widehat{\gamma}_N(t, t)^{1/2}$. More precisely, given a confidence level $1 - \alpha \in (0, 1)$, we seek $c = c_\alpha$ that approximately satisfies

$$\mathbb{P}\left(|G(t)| \leq c\,\sigma(t),\ \forall t \in [0, T]\right) = 1 - \alpha, \tag{12}$$

where G is a Gaussian process with mean zero and covariance function $\gamma = \lim_{N\to\infty} \gamma_N$, and where $\sigma(t) = \gamma(t,t)^{1/2}$. (The convergence of $\widehat{\mu}_N$ to G in the space $(C([0,T]), \|\cdot\|_\infty)$ is proved in Cardot *et al.* (2011).) Computing accurate and as explicit as possible bounds in a general setting is a difficult issue. Under some technical conditions, Cardot *et al.* (2011) prove that it is possible to estimate the threshold c in (12) via simulations: conditionally on $\widehat{\gamma}_N$ defined in (10), one can simulate a large number of sample paths of the Gaussian process $\widehat{G}_N$ with mean zero and covariance $\widehat{\gamma}_N$ and compute their supremum norms and it suffices to set c as the quantile of order $(1-\alpha)$ of this distribution:

$$\mathbb{P}\left(|\widehat{G}_N(t)| \leq c\,\widehat{\sigma}_N(t),\ \forall t \in [0,T] \,\middle|\, \widehat{\gamma}_N\right) = 1-\alpha. \quad (13)$$

4 A simulation study

In this section, we evaluate the performances of the mean curve estimator as well as the coverage and the width of the confidence bands for different bandwidth selection criteria and different levels of noise.

4.1 Simulated data and sampling designs

We have generated a population of $N = 20000$ curves discretized at $d = 200$ and $d = 400$ equidistant instants of time in $[0,1]$. The curves of the population are generated so that they have approximately the same distribution as the electricity consumption curves analyzed in Cardot & Josserand (2011) and each individual curve X_k, for $k \in U$, is simulated as follows

$$X_k(t) = \mu(t) + \sum_{\ell=1}^{3} Z_\ell\ v_\ell(t), \quad t \in [0,1], \quad (14)$$

where μ is the mean function and the random variables Z_ℓ are independent realizations of a centered Gaussian random variable with variance σ_ℓ^2. The three basis function v_1, v_2 and v_3 are orthonormal functions which represent the main mode of variation of the signals. Thus, the covariance function of the population $\gamma(s,t)$ is simply

$$\gamma(s,t) \quad = \quad \sum_{\ell=1}^{3} \sigma_\ell^2\ v_\ell(s) v_\ell(t). \quad (15)$$

To select the samples, we have considered two probabilistic selection procedures, with fixed sample size, $n = 1000$,

- Simple random sampling without replacement (SRSWR).
- Stratified sampling with SRSWR in all strata. The population U is divided into a fixed number of $G = 5$ strata built by considering the quantiles $q_{0.5}, q_{0.7}, q_{0.85}$ and $q_{0.95}$ of the total consumption $\int_0^1 X_k(t)dt$ for all units $k \in U$. For example, the first strata contains all the units k such that $\int_0^1 X_k(t)dt \leq q_{0.5}$, and thus its size is half of the population size N. The sample size n_g in stratum g is determined by a Neyman-like allocation, as suggested in Cardot and Josserand (2011), in order to get a Horvitz-Thompson estimator of the mean trajectory whose variance is as small as possible. The sizes of the different strata, which are optimal according to this mean variance criterion, are reported in Table 1.

We suppose we observe, for each unit k in the sample s, the discretized trajectories, at d equispaced points, $0 = t_1 < \ldots < t_d = 1$,

$$Y_{jk} = X_k(t_j) + \delta\epsilon_{jk} \quad (16)$$

where the $\epsilon_{jk} \sim N(0, \gamma(t_j,t_j))$ are independent random variables and the parameter δ allows to control the noise level.

Stratum number	1	2	3	4	5
Stratum size	10000	4000	3000	2000	1000
Allocation	655	132	98	68	47

Table 1: Strata sizes and optimal allocations.

4.2 Weighted cross-validation for bandwidth selection

Assuming we can access the exact trajectories X_k, $k \in s$, (which is the case in simulations) we consider the oracle-type estimator

$$(17) \quad \widehat{\mu}_s = \sum_{k \in s} \frac{X_k}{\pi_k} \,,$$

which will be a benchmark in our numerical study. We compare different interpolation and smoothing strategies for estimating the X_k, $k \in s$:

- Linear interpolation of the Y_{jk} as in Cardot and Josserand (2011).
- Local linear smoothing of the Y_{jk} with bandwidth h as in (3).

The crucial element here is h. To evaluate the interest of smoothing and the performances of data-driven bandwidth selection criteria, we consider an error measure that compares the oracle $\widehat{\mu}_s$ to any estimator $\widehat{\mu}$ based on the noisy data Y_{jk}, $k \in s$, $j = 1, \dots d$:

$$(18) \quad L(\widehat{\mu}) = \int_0^T \left(\widehat{\mu}_s(t) - \widehat{\mu}(t)\right)^2 dt \,.$$

Considering the estimator defined in (6), we denote by h_{oracle} the bandwidth h that minimizes (18) and call smooth oracle the corresponding estimator.

When $\sum_{k \in s} \pi_k^{-1} = N$, as in SRSWR and stratified sampling, it can be easily checked that $\widehat{\mu}_s$ is the minimum argument of the weighted least squares functional

$$(19) \quad \sum_{k \in s} w_k \int_0^T \left(X_k(t) - \mu(t)\right)^2 dt$$

with respect to $\mu \in L^2([0,T])$, where the weights are $w_k = (N\pi_k)^{-1}$. Then, a simple and natural way to select bandwidth h is to consider the following design-based cross validation

$$(20) \quad \text{WCV}(h) = \sum_{k \in s} w_k \sum_{j=1}^{d} \left(Y_{jk} - \widehat{\mu}_N^{-k}(t_j)\right)^2 ,$$

where $\widehat{\mu}_N^{-k}(t) = \sum_{\ell \in s, \ell \neq k} \widetilde{w}_\ell \; \widehat{X}_\ell(t)$, with new weights $\widetilde{w}_\ell$.

This weighted cross validation criterion is simpler than the cross validation criteria based on the estimated variance proposed in Opsomer and Miller (2005). Indeed, in our case, the bias may be non negligible and focusing only on the variance part of the error leads to too large selected values for the bandwidth. Furthermore, Opsomer and Miller (2005) suggested to consider weights defined as follows $\widetilde{w}_\ell = w_\ell/(1 - w_k)$. For SRSWR, since $w_k = n^{-1}$ one has $\widetilde{w}_k = (n-1)^{-1}$, so that the weighted cross validation criterion defined in (20) is exactly the cross validation criterion introduced by Rice and Silverman (1991) in the independent case. We denote in the following by h_{cv} the bandwidth value minimizing this criterion. For stratified sampling, a better approximation which keeps the design-based properties of the estimator $\widehat{\mu}_N^{-k}$ can be obtained by taking into account the sampling rates in the different strata. We have G strata with sizes N_g, $g = 1, \dots, G$ and we sample n_g observations,

with SRSWR, in each stratum g. If unit k comes from strata g, we have $w_k = N_g(Nn_g)^{-1}$. Thus, we take $\widetilde{w}_\ell = (N_g - 1)\{(N-1)(n_g-1)\}^{-1}$ for all the units $\ell \neq k$ in stratum g and just scale the weights for all the units ℓ' of the sample that do not belong to stratum g, $\widetilde{w}_{\ell'} = N(N-1)^{-1}w_{\ell'}$. We denote by h_{wcv} the bandwidth value minimizing (20).

We can first note that stratified sampling allows to improve much the estimation of the mean curve. We also remark that, for such large samples, linear interpolation performs nearly as well as the smooth oracle estimator, especially when the noise level is low ($\delta \leq 15\%$). As far as bandwidth selection is concerned, we can note that the usual cross validation criterion h_{cv} is not adapted to unequal probability sampling and does not perform as well as linear interpolation for stratified sampling by selecting too large values for the bandwidth. On the other hand, the weighted cross-validation criterion seems to be effective to select good bandwidth values and produce estimators whose estimation errors are very close to the oracle and perform better than the other estimators when the noise level is moderate or high ($\delta \geq 20\%$). We now examine the empirical coverage and the width of the confidence bands, which are built as described in Section 3. For each sample, we estimate the covariance function $\widehat{\gamma}_N$ and draw 10000 realizations of a centered Gaussian process with variance function $\widehat{\gamma}_N$ in order to obtain a suitable coefficient c with a confidence level of $1 - \alpha = 0.95$. The area of the confidence band is then $\int_0^T 2c\sqrt{\widehat{\gamma}(t,t)}\ \mathrm{d}t$. The results highlight now the interest of considering smoothing strategies combined with the weighted cross validation bandwidth selection criterion (20). It appears that linear interpolation, which does not intend to get rid of the noise, always gives larger confidence bands than the smoothed estimators based on h_{wcv}. Moreover, smoothing approaches become more interesting as the number of discretization points and the noise level increase.

As a conclusion of this simulation study, it appears that smoothing is not a crucial aspect when the only target is the estimation of the mean, and that bandwidth values should be chosen by a cross validation criterion that takes the sampling weights into account. When the goal is also to build confidence bands, smoothing with weighted cross validation criteria lead to narrower bands compared to interpolation techniques, without deteriorating the empirical coverage.

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

REFERENCES

- Cardot, H.. Degras, D. and Josserand, E. (2011). Confidence bands for Horvitz-Thompson estimators using sampled noisy functional data. Submitted. `http://arxiv.org/abs/1105.2135.`
- Cardot, H. and Josserand, E. (2011). Horvitz-Thompson estimators for functional data: asymptotic confidence bands and optimal allocation for stratified sampling. *Biometrika*, **98**, 107-118.
- Chiky, R. and Hébrail, G. (2008). Summarizing distributed data streams for storage in data warehouses. In DaWaK 2008, I-Y. Song, J. Eder and T. M. Nguyen, Eds. *Lecture Notes in Computer Science*, Springer, 65-74.
- Degras, D. (2010). Simultaneous confidence bands for nonparametric regression with functional data. Accepted for publication at *Statistica Sinica*. `http://arxiv.org/abs/0908.1980`
- Fuller, W.A. (2009). *Sampling Statistics.* John Wiley and Sons.
- Opsomer, J. D. and Miller, C. P. (2005). Selecting the amount of smoothing in nonparametric regression estimation for complex surveys. *J. Nonparametric Statistics,* **17**, 593-611.
- Rice, J. A. and Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *J. Roy. Statist. Soc. Ser. B*, **53**, 233-243.