Variance estimation and asymptotic confidence bands for the mean estimator of sampled functional data with high entropy unequal probability sampling designs

Hervé Cardot[(a)], Camelia Goga[(a)] and Pauline Lardin[(a,b)]
(a) Université de Bourgogne, Institut de Mathématiques de Bourgogne,
9 av. Alain Savary, 21078 DIJON, FRANCE
(b) EDF, R&D, ICAME-SOAD, 1 av. du Général de Gaulle,
92141 CLAMART, FRANCE

June 28, 2013

Abstract

For fixed size sampling designs with high entropy it is well known that the variance of the Horvitz-Thompson estimator can be approximated by the Hájek formula. The interest of this asymptotic variance approximation is that it only involves the first order inclusion probabilities of the statistical units. We extend this variance formula when the variable under study is functional and we prove, under general conditions on the regularity of the individual trajectories and the sampling design, that we can get a uniformly convergent estimator of the variance function of the Horvitz-Thompson estimator of the mean function. Rates of convergence to the true variance function are given for the rejective sampling. We deduce, under conditions on the entropy of the sampling design, that it is possible to build confidence bands whose coverage is asymptotically the desired one via simulation of Gaussian processes with variance function given by the Hájek formula. Finally, the accuracy of the proposed variance estimator is evaluated on samples of electricity consumption data measured every half an hour over a period of one week.

Keywords : covariance function, finite population, first order inclusion probabilities, Hájek approximation, Horvitz-Thompson estimator, Kullback-Leibler divergence, rejective sampling, unequal probability sampling without replacement.

1 Introduction

Computing the variance of the Horvitz-Thompson estimator for unequal probability sampling designs can be difficult because the variance formula involves second order probability inclusions which are not always known. The Hájek variance formula, derived in Hájek (1964) for rejective sampling is an asymptotic approximation which only requires the knowledge of the first order inclusion probabilities and is easy to compute. It is shown in Hájek (1964) and Chen et al. (1994) that, for given first order inclusion probabilities, the rejective sampling is the fixed size sampling design with the highest entropy. The validity of this approximation is closely related to the value of the entropy of the considered sampling design. Hájek (1981) proves that this approximation is also valid for the Sampford-Durbin sampling whereas Berger (1998a) gives general conditions on the relative entropy of the sampling design, also called Kullback-Leibler divergence, which justify the use of this approximated variance formula. Variants and refinements of the Hájek variance formula as well as variance estimators are proposed in Deville and Tillé (2005). Matei and Tillé (2005) show on simulations that these approximations to the variance of Horvitz-Thompson estimators are effective, even for moderate sample sizes, provided that the entropy of the underlying sampling design is high enough. Recently Deville and Tillé (2005) and Fuller (2009) consider balanced, or approximately balanced, sampling algorithms which can be useful to build designs with fixed size and given inclusion probabilities. They relate these sampling designs to the rejective sampling, so that the Hájek variance approximation can be used. Note also that there exist other ways to get an approximation to the variance of the Horvitz-Thompson estimator which do not require the knowledge of the second order inclusion probabilities (see *e.g.* Shahbaz and Hanif (2003)). These approaches do not rely on asymptotic developments and are not considered in this work.

When the aim is to build confidence intervals, the asymptotic distribution of the Horvitz-Thompson estimator is required. The Central Limit Theorem has been checked by Erdös and Rényi (1959) and Hájek (1960) for the simple random sampling without replacement, by Hájek (1964) for the rejective sampling and by Víšek (1979) for the Sampford sampling. Berger (1998b) states that the Kullback Leibler divergence of the considered sampling design, with respect to the rejective sampling, should tend to zero when the sample size gets larger for the Horvitz-Thompson estimator to be asymptotically Gaussian.

In recent studies in survey sampling the target was not a mean real value or a mean vector but a mean function (see Cardot and Josserand (2011) and Cardot et al. (2013b) for the estimation of electricity consumption curves) and one important issue was how to build confidence bands when using πps sampling designs. A rapid technique that is well adapted

for large samples has been studied in Degras (2011) and Cardot et al. (2013a). It consists in first estimating the covariance function of the mean estimator and then simulating a Gaussian process, whose covariance function is the estimated covariance function, in order to determine the distribution of its supremum. This strategy which has been employed successfully in Cardot et al. (2013b) to build confidence bands necessitates to have an effective estimator of the variance function of the Horvitz-Thompson estimator. The aim of this work is to prove that under general assumptions on the sampling design and on the regularity of the trajectories, the Hájek formula provides a uniformly consistent estimator of the variance function. So, it is possible to assess rigorously confidence bands built by using the procedure described previously.

The paper is organized as follows. The notations and our estimators are presented in Section 2. In Section 3, we state our main result, namely the uniform convergence of the variance function estimator obtained under broad assumptions on the regularity of the trajectories and the sampling design. We deduce that if the Horvitz-Thompson estimator of the mean curve is pointwise asymptotically Gaussian, then it also satisfies, under the same conditions, a functional central limit theorem. The confidence bands obtained by the Gaussian process simulation techniques have asymptotically the desired coverage. In section 4, we evaluate the performance of the covariance function estimator on samples drawn from a test population of $N = 15055$ electricity consumption curves measured every half an hour over a one-week period. Note there are many ways of drawing samples with high entropy sampling distribution and with given first order inclusion probabilities (see *e.g.* Brewer and Hanif (1983), Tillé (2006), Bondesson et al. (2006) and Bondesson (2010)). Because of our large population and large sample context, we use the vast version of the Cube algorithm (Deville and Tillé (2004)) developed in Chauvet and Tillé (2006) for dealing with very large populations (e.g., of millions of units). Finally, Section 6 contains some concluding remarks. The proofs are gathered in an Appendix.

2 Variance estimation and the Hájek formula

Let us consider a finite population $U = \{1, ..., N\}$ of known size N, and suppose that, for each unit k of the population U, we can observe a deterministic curve $Y_k = (Y_k(t))_{t\in[0,T]}$. We want to estimate the mean trajectory $\mu_N(t)$, $t \in [0, T]$, defined as follows:

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

We consider a sample s, with fixed size n, drawn from U according to a sampling design $p_N(s)$, where $p_N(s)$ is the probability of drawing the sample s. The mean curve $\mu_N(t)$ is

estimated by the Horvitz-Thompson estimator,

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

where $\mathbb{1}_k$ is the sample membership indicator, $\mathbb{1}_k = 1$ if $k \in s$ and $\mathbb{1}_k = 0$ otherwise. We denote by $\pi_k = \mathbb{E}_p(\mathbb{1}_k)$ the first order inclusion probability of unit k with respect to the sampling design $p_N(s)$ and we suppose that $\pi_k > 0$, for all units k in U. It is well known that, for each value of $t \in [0,T]$, $\widehat{\mu}(t)$ is a design-unbiased estimator of $\mu_N(t)$, *i.e.* $\mathbb{E}_p(\widehat{\mu}(t)) = \mu_N(t)$. We denote by $\pi_{kl} = \mathbb{E}_p(\mathbb{1}_{kl})$ with $\mathbb{1}_{kl} = \mathbb{1}_k\mathbb{1}_l$, the second order inclusion probabilities and we suppose that $\pi_{kl} > 0$ for all $k, l \in U$.

Since the sample size is fixed, the variance $\gamma_p(t,t)$ for each instant t of the estimator $\widehat{\mu}(t)$ is given by the Yates and Grundy formula (see Yates and Grundy (1953) and Sen (1953)),

$$\gamma_p(t,t) = -\frac{1}{2}\frac{1}{N^2}\sum_{k\in U}\sum_{l\in U, l\neq k}(\pi_{kl}-\pi_k\pi_l)\left(\frac{Y_k(t)}{\pi_k}-\frac{Y_l(t)}{\pi_l}\right)^2, \tag{2}$$

and it is straightforward to express the covariance $\gamma_p(r,t)$ of $\widehat{\mu}$ between two instants r and t, as follows

$$\gamma_p(r,t) = -\frac{1}{2}\frac{1}{N^2}\sum_{k\in U}\sum_{l\in U, l\neq k}(\pi_{kl}-\pi_k\pi_l)\left(\frac{Y_k(r)}{\pi_k}-\frac{Y_l(r)}{\pi_l}\right)\left(\frac{Y_k(t)}{\pi_k}-\frac{Y_l(t)}{\pi_l}\right). \tag{3}$$

The variance formula (2) clearly indicates that if the first order inclusion probabilities are chosen to be approximately proportional to $Y_k(t)$, the variance of the estimator $\widehat{\mu}(t)$ will be small. In practice, we can consider a non-functional auxiliary variable X of values x_k supposed to be positive and known for all the units $k \in U$. If X is nearly proportional to the variable of interest, it can be very interesting to consider a sampling design whose first order inclusion probabilities are given by

$$\pi_k = n\frac{x_k}{\sum_{k\in U}x_k}.$$

There are many ways of building sampling designs with given first order inclusion probabilities (see *e.g* Brewer and Hanif (1983) and Tillé (2006)) and we focus here on the designs with high entropy, where the entropy of a sampling design p_N (a discrete probability distribution on U) is defined by

$$H(p_N) = -\sum_{k\in s}p_N(s)\ln(p_N(s))$$

with the convention $0\ln 0 = 0$. It has been proven (see Hájek (1981) and Chen et al. (1994)) that, for given first order inclusion probabilities, the rejective sampling, or conditional Poisson sampling, is the fixed size sampling design with the highest entropy. Then, a key result is the following uniform approximation to the second order inclusion probabilities, for $k \neq l$,

$$\pi_{kl} = \pi_k\pi_l\left\{1-\frac{(1-\pi_k)(1-\pi_l)}{d(\pi)}[1+o(1)]\right\} \tag{4}$$

where $d(\pi) = \sum_{k\in U} \pi_k(1-\pi_k)$ is supposed to tend to infinity. Note that this implies that n, N and $N-n$ tend to infinity. This asymptotic approximation is satisfied for the rejective sampling and the Sampford-Durbin sampling which is also a high entropy sampling design (see Hájek (1981)).

Remark 1. *Formula (4) can be seen rather strange and we give an intuitive and simple interpretation in terms of conditional covariance. Note that this is not a proof. Consider a Poisson sampling design with inclusion probabilities $p_1, \ldots, p_N$ such that $\sum_{k\in U} p_k = n$ and $\mathbb{E}_p(\mathbb{1}_k|\#s = n) = \pi_k$ where $\#s$ denotes the (random) sample size and $\mathbb{1}_k$ is the indicator membership to the sample s of unit k (see Chen et al. (1994) for the existence of such sampling design). Considering now the covariance given the sample size, $cov(\mathbb{1}_k, \mathbb{1}_l|\#s = n) = \pi_{kl} - \pi_k\pi_l$, we get the following approximation, which is similar to (4), if we use the formula for the conditional variance in a Gaussian framework,*

$$\begin{aligned} cov(\mathbb{1}_k, \mathbb{1}_l|\#s = n) &\approx cov(\mathbb{1}_k, \mathbb{1}_l) - \frac{cov(\mathbb{1}_k, \#s)cov(\mathbb{1}_l, \#s)}{var(\#s)} \\ &\approx 0 - \frac{\pi_k(1-\pi_k)\pi_l(1-\pi_l)}{\sum_{k\in U}\pi_k(1-\pi_k)} \end{aligned}$$

since $cov(\mathbb{1}_k, \mathbb{1}_l) = 0$, $cov(\mathbb{1}_k, \#s) = p_k(1-p_k)$ and $var(\#s) = \sum_{k\in U} p_k(1-p_k)$ for Poisson sampling and, for each unit k, p_k tends to π_k as $d(\pi)$ tends to infinity (see Hájek (1964)).

Then, we obtain, for all $(r,t) \in [0,T]\times[0,T]$, the Hájek approximation $\gamma_H(r,t)$ to the covariance function $cov(\hat{\mu}(t), \hat{\mu}(r))$, by plugging in approximation (4) in (3),

$$\gamma_H(r,t) = \frac{1}{N^2}\left[\sum_{k\in U}\frac{Y_k(t)Y_k(r)}{\pi_k}(1-\pi_k) - \frac{1}{d(\pi)}\left(\sum_{k\in U}(1-\pi_k)Y_k(t)\right)\left(\sum_{l\in U}(1-\pi_l)Y_l(r)\right)\right], \tag{5}$$

and we consider in the following two estimators for the covariance

$$\hat{\gamma}_H(r,t) = \frac{1}{N^2}\frac{\hat{d}(\pi)}{d(\pi)}\left[\sum_{k\in s}\frac{1-\pi_k}{\pi_k^2}Y_k(t)Y_k(r) - \frac{1}{\hat{d}(\pi)}\sum_{k\in s}\left(\frac{1-\pi_k}{\pi_k}Y_k(t)\right)\sum_{l\in s}\left(\frac{1-\pi_l}{\pi_l}Y_l(r)\right)\right], \tag{6}$$

and $\hat{\gamma}_H^*(r,t) = \frac{d(\pi)}{\hat{d}(\pi)}\hat{\gamma}_H(r,t)$, where $\hat{d}(\pi) = \sum_{k\in s}(1-\pi_k)$ is the Horvitz-Thompson estimator of $d(\pi)$. Note that $\hat{\gamma}_H(r,t)$ is a slightly modified functional analogue of the variance estimator proposed by Berger (1998a) in the real case. More exactly, the variance estimator considered by Berger (1998a) is $\hat{\gamma}_H(t,t)$ multiplied by the correction factor $n/(n-1)$ so that the expression is exact for simple random sampling without replacement. The second estimator, $\hat{\gamma}_H^*(r,t)$ is the extension to the functional case of the Deville and Tillé (2005)'s estimator. This latter approximation of the variance has been shown to be effective on simulation studies, even for moderate sample sizes, by Matei and Tillé (2005).

We can easily show the following property.

Proposition 2.1. *If, for all $t \in [0,T]$, there is a constant c_t such that $Y_k(t) = c_t \pi_k$ then $\gamma_H(r,t) = 0$ and $\widehat{\gamma}_H(r,t) = \widehat{\gamma}^*_H(r,t) = 0$.*

With real data, we do not observe $Y_k(t)$ at all instants t in $[0,T]$ but only for a finite set of D measurement times, $0 = t_1 < ... < t_D = T$. In functional data analysis, when the noise level is low and the grid of discretization points is fine, it is usual to perform a linear interpolation or a smoothing of the discretized trajectories in order to obtain approximations of the trajectories at every instant t (see Ramsay and Silverman (2005)). When there are nearly no measurement errors and when the trajectories are regular enough, Cardot and Josserand (2011) showed that linear interpolation can provide sufficiently accurate approximations of the trajectories. Thus, for each unit k in the sample s, we build the interpolated trajectory

$$Y_{k,d}(t) = Y_k(t_i) + \frac{Y_k(t_{i+1}) - Y_k(t_i)}{t_{i+1} - t_i}(t - t_i), \quad t \in [t_i, t_{i+1}],$$

and define the estimator of the mean curve $\mu_N(t)$ based on the discretized observations as follows

$$\widehat{\mu}_d(t) = \frac{1}{N} \sum_{k \in s} \frac{Y_{k,d}(t)}{\pi_k}, \quad t \in [t_i, t_{i+1}]. \tag{7}$$

Its covariance function is then estimated by

$$\hat{\gamma}_{H,d}(r,t) = \frac{1}{N^2} \frac{\hat{d}(\pi)}{d(\pi)} \left[\sum_{k \in s} \frac{1 - \pi_k}{\pi_k^2} Y_{k,d}(t) Y_{k,d}(r) - \frac{1}{\hat{d}(\pi)} \sum_{k \in s} \left(\frac{1 - \pi_k}{\pi_k} Y_{k,d}(t) \right) \sum_{l \in s} \left(\frac{1 - \pi_l}{\pi_l} Y_{l,d}(r) \right) \right], \tag{8}$$

and we show in the next section that it is an uniformly consistent estimator of the variance function. Replacing $Y_k(t)$ by $Y_{k,d}(t)$ in $\hat{\gamma}^*_H(r,t)$, yields the variance estimator $\hat{\gamma}^*_{H,d}(r,t)$ based on the discretized values.

3 Asymptotic properties

All the proof are postponed in an Appendix.

3.1 Assumptions

To demonstrate the asymptotic properties, we must suppose that the sample size and the population size become large. Therefore, we adopt the asymptotic approach of Hájek (1964), assuming that $d(\pi) \to \infty$. Note that this assumption implies that $n \to \infty$ and $N - n \to \infty$. We consider a sequence of growing and nested populations U_N with size N tending to infinity and a sequence of samples s_N of size n_N drawn from U_N according to the sampling design $p_N(s_N)$. The first and second order inclusion probabilities are respectively denoted by π_{kN}

and π_{klN}. For simplicity of notations and when there is no ambiguity, we drop the subscript N. To prove our asymptotic results we need to introduce the following assumptions.

A1. We assume that $\lim_{N\to\infty} \frac{n}{N} = \pi \in (0,1)$.

A2. We assume that $\min_{k\in U} \pi_k \geq \lambda > 0$, $\min_{k\neq l\in U} \pi_{kl} \geq \lambda^* > 0$ and

$$\pi_{kl} = \pi_k \pi_l \left\{ 1 - \frac{(1-\pi_k)(1-\pi_l)}{d(\pi)} [1 + o(1)] \right\}$$

uniformly in k and l.

A3. There are two positive constants C_2 and C_3 and $\beta > 1/2$ such that, for all N and for all $(r,t) \in [0,T] \times [0,T]$,

$$\frac{1}{N} \sum_{k\in U} (Y_k(0))^2 < C_2 \quad \text{and} \quad \frac{1}{N} \sum_{k\in U} (Y_k(t) - Y_k(r))^2 < C_3 |t-r|^{2\beta}.$$

A4. There are two positive constants C_4 and C_5 such that, for all N and for all $(r,t) \in [0,T] \times [0,T]$,

$$\frac{1}{N} \sum_{k\in U} (Y_k(0))^4 < C_4 \quad \text{and} \quad \frac{1}{N} \sum_{k\in U} (Y_k(t) - Y_k(r))^4 < C_5 |t-r|^{4\beta}.$$

A5. We assume that

$$\lim_{N\to\infty} \max_{(k_1,l_1,k_2,l_2)\in D_{4,N}} |\mathbb{E}_p \left[(\mathbb{1}_{k_1 l_1} - \pi_{k_1}\pi_{l_1})(\mathbb{1}_{k_2 l_2} - \pi_{k_2}\pi_{l_2}) \right]| = 0$$

where $\mathbb{1}_{kl}$ is the sample membership of the couple (k,l) and $D_{4,N}$ is the set of all distinct quadruples $(i_1, ..., i_4)$ from U.

Assumptions **A1** and **A2** are classical hypotheses in survey sampling and deal with the first and second order inclusion probabilities. They are satisfied for high entropy sampling designs with fixed size (see for example Hájek (1981)). They directly imply that $cn \leq d(\pi) \leq n$, for some strictly positive constant c. The assumption **A2** implies that $\limsup_{N\to\infty} n \max_{k\neq l\in U} |\pi_{kl} - \pi_k\pi_l| < C_1 < \infty$. It also ensures that the Yates-Grundy variance estimator is always positive since $\pi_{kl} \leq \pi_k \pi_l$.

Assumption **A3** and **A4** are regularity conditions on the individual trajectories. Even if point-wise consistency, for each fixed value of t, can be proven without any condition on β, these regularity conditions are required to get the uniform convergence of the mean estimator (see Cardot and Josserand (2011)). Note finally that assumption **A5** is true for SRSWOR, stratified sampling and rejective sampling (see Arratia et al. (2005) and Boistard

et al. (2012)). More generally, it also holds for unequal probability designs with large entropy as shown in the following proposition. Let us recall before the definition of the Kullback-Leibler divergence $K(p_N, p_{rej})$,

$$K(p_N, p_{rej}) = \sum_{k \in s} p_N(s) \ln \left(\frac{p_N(s)}{p_{rej}(s)} \right), \tag{9}$$

which measures how a sampling distribution $p_N(s)$ is distant from a reference sampling distribution, chosen here to be the rejective sampling $p_{rej}(s)$ since it is the design with maximum entropy for given first order inclusion probabilities. We can now state the following proposition which gives an upper bound of the rates of convergence to zero of the quantity in **A4** in terms of Kullback-Leibler divergence with respect to the rejective sampling.

Proposition 3.1. *Let p_N be a sampling design with the same first order inclusion probabilities as p_{rej}. If $d(\pi) \to \infty$, then*

$$\max_{(k_1,l_1,k_2,l_2) \in D_{4,N}} |\mathbb{E}_p \left[(\mathbb{1}_{k_1 l_1} - \pi_{k_1}\pi_{l_1})(\mathbb{1}_{k_2 l_2} - \pi_{k_2}\pi_{l_2}) \right]| \leq \frac{C}{d(\pi)} + \sqrt{\frac{K(p_N, p_{rej})}{2}}$$

for some constant C.

A direct consequence of Proposition 3.1 is that assumption **A5** is satisfied for the rejective sampling as well as for the Sampford-Durbin design, whose Kullback-Leibler divergence, with respect to the rejective sampling, tends to zero as the sample size n tends to infinity (see Berger (1998b)). Note also that the Kullback-Leibler divergence has been approximated asymptotically for other sampling designs such as the Pareto sampling in Lundqvist (2007).

3.2 Convergence of the estimated variance

Let us first recall Proposition 3.3 in Cardot and Josserand (2011) which states that the estimator $\hat{\mu}_d$ is asymptotically design unbiased and uniformly convergent under mild assumptions. More precisely, if assumptions (A1)-(A3) hold and if the discretization scheme satisfies $\max_{i \in \{1,..,d_N - 1\}} |t_{i+1} - t_i|^{2\beta} = o(n^{-1})$, then for some constant C

$$\sqrt{n}\mathbb{E}_p \left\{ \sup_{t \in [0,T]} |\hat{\mu}_d(t) - \mu_N(t)| \right\} \leq C.$$

We can now state our first result which indicates that the covariance function estimator $\hat{\gamma}_{H,d}(r,t)$ is pointwise convergent and that the variance function estimator $\hat{\gamma}_{H,d}(t,t)$ is uniformly convergent. Note that additional assumptions on the sampling design are required in order to obtain the convergence rates.

Proposition 3.2. *1. Assume (A1)-(A5) hold and the sequence of discretization schemes satisfies* $\lim_{N\to\infty} \max_{i=\{1,..,d_N-1\}} |t_{i+1} - t_i| = 0$. *When N tends to infinity,*

$$n\mathbb{E}_p \{ \mid \widehat{\gamma}_{H,d}(r,t) - \gamma_p(r,t) \mid \} \to 0 \tag{10}$$

for all $(r,t) \in [0,T] \times [0,T]$ and

$$n\mathbb{E}_p \left\{ \sup_{t\in[0,T]} \mid \widehat{\gamma}_{H,d}(t,t) - \gamma_p(t,t) \mid \right\} \to 0. \tag{11}$$

2. Under the same assumptions, the covariance function estimator $\widehat{\gamma}^*_{H,d}(r,t)$ satisfies (10) and the variance function estimator $\widehat{\gamma}^*_{H,d}(t,t)$ satisfies (11).

A sharper result can be stated for the particular case of rejective sampling for which accurate approximations to the multiple inclusion probabilities are available (see Boistard et al. (2012)).

Proposition 3.3. *Suppose that the sample is selected with the rejective sampling design. Assume (A1)-(A4) hold and the sequence of discretization schemes satisfies* $\max_{i=\{1,..,d_N-1\}} |t_{i+1} - t_i|^{2\beta} = O(n^{-1})$. *Then, for all $(r,t) \in [0,T] \times [0,T]$*

$$n^3 \ \mathbb{E}_p \left[\left(\widehat{\gamma}_{H,d}(r,t) - \gamma_p(r,t) \right)^2 \right] \leq C$$

for some positive constant C.

We can note in the proof, given in the Appendix, that the approximation error to the true variance by the Hájek formula is asymptotically negligible compared to the sampling error.

3.3 Asymptotic normality and confidence bands

Let us assume that the Horvitz-Thompson estimator of the mean curve satisfies a Central Limit Theorem for real valued quantities with new moment conditions

A6. There is some $\delta > 0$, such that $N^{-1} \sum_{k\in U_N} |Y_k(t)|^{2+\delta} < \infty$ for all $t \in [0,T]$, and $\{\gamma_p(t,t)\}^{-1/2} \{\hat{\mu}(t) - \mu(t)\} \to \mathcal{N}(0,1)$ in distribution when N tends to infinity.

This asymptotic normality assumption is satisfied for high entropy sampling designs (see Víšek (1979) and Berger (1998b)). Cardot and Josserand (2011) have shown that under the previous assumptions, the central limit theorem also holds in the space of continuous functions $C[0,T]$. More precisely, if assumptions (A1)-(A3) and (A6) hold and the discretization points satisfy $\lim_{N\to\infty} \max_{i=\{1,..,d_N-1\}} |t_{i+1} - t_i|^{2\beta} = o(n^{-1})$, we have

$$\sqrt{n}(\hat{\mu}_d - \mu) \to Z \text{ in distribution in } C[0,T]$$

where Z is a Gaussian random function taking values in $C[0,T]$ with mean 0 and covariance function $\gamma_Z(r,t) = \lim_{N\to\infty} n\gamma_{p_N}(r,t)$. The reader is referred to Cardot et al. (2013c) for a discussion on the reasons of using the convergence in the space $C[0,T]$. This important result gives a theoretical justification of the confidence bands for μ_N built as follows:

$$\left\{ \left[\widehat{\mu}_d(t) \pm c\, \frac{\widehat{\sigma}(t)}{\sqrt{n}} \right],\ t \in [0,T] \right\}, \tag{12}$$

where c is a suitable number and $\widehat{\sigma}(t) = \sqrt{n\widehat{\gamma}_{H,d}(t,t)}$.

Given a confidence level $1-\alpha \in (0,1)$, one way to build such confidence bands, that is to say one way to find an adequate value for c_α, is to perform simulations of centered Gaussian functions $\widehat{Z}$ defined on $[0,T]$ with mean 0 and covariance function $n\widehat{\gamma}_{H,d}(r,t)$ and then compute the quantile of order $1-\alpha$ of $\sup_{t\in[0,T]} \left| \widehat{Z}(t)/\widehat{\sigma}(t) \right|$. In other words, we look for a cut-off point c_α, which is random since it depends on the estimated covariance function $\widehat{\gamma}_{H,d}$, such that

$$\mathbb{P}\left(|\widehat{Z}(t)| \leq c_\alpha \frac{\widehat{\sigma}(t)}{\sqrt{n}},\ \forall t \in [0,T] \mid \widehat{\gamma}_{H,d} \right) = 1-\alpha. \tag{13}$$

Next proposition provides a rigorous justification for this Monte Carlo technique which can be interpreted as parametric bootstrap:

Proposition 3.4. *Assume (A1)-(A6) hold and the discretization scheme satisfies* $\max_{i=\{1,..,d_N-1\}} |t_{i+1}-t_i|^{2\beta} = o(n^{-1})$.

Let Z be a Gaussian process with mean zero and covariance function γ_Z. Let $(\widehat{Z}_N)$ be a sequence of processes such that for each N, conditionally on $\widehat{\gamma}_{H,d}$ defined in (8), $\widehat{Z}_N$ is Gaussian with mean zero and covariance $n\widehat{\gamma}_{H,d}$. Then for all $c>0$, as $N\to\infty$, the following convergence holds in probability:

$$\mathbb{P}\left(|\widehat{Z}_N(t)| \leq c\,\widehat{\sigma}(t),\ \forall t \in [0,T] \,\middle|\, \widehat{\gamma}_{H,d} \right) \to \mathbb{P}\left(|Z(t)| \leq c\,\sigma(t),\ \forall t \in [0,T] \right),$$

where $\widehat{\sigma}(t) = \sqrt{n\widehat{\gamma}_{H,d}(t,t)}$ and $\sigma(t) = \sqrt{\gamma_Z(t,t)}$.

The proof of Proposition 3.4 is very similar to the proof of Proposition 3.5 in Cardot et al. (2013c) and is thus omitted. As in Cardot et al. (2013a), it is possible to deduce from previous proposition that the chosen value $\widehat{c}_\alpha = c_\alpha(\widehat{\gamma}_{H,d})$ provides asymptotically the desired coverage since it satisfies

$$\lim_{N\to\infty} \mathbb{P}\left(\mu(t) \in \left[\widehat{\mu}_d(t) \pm \widehat{c}_\alpha \frac{\widehat{\sigma}(t)}{\sqrt{n}} \right],\ \forall t \in [0,T] \right) = 1-\alpha.$$

4 Example: variance estimation for electricity consumption curves

In this section, we evaluate the performance of the estimators $\hat{\gamma}^*_{H,d}(r,t)$ and $\hat{\gamma}_{H,d}(r,t)$ of the functional variance $\gamma_p(r,t)$ of $\hat{\mu}_d(t)$. Simulation studies not reported here showed that the estimators $\hat{\gamma}^*_{H,d}(r,t)$ and $\hat{\gamma}_{H,d}(r,t)$ conduct very similarly asymptotically. This is why we only give below the simulation results for $\hat{\gamma}_{H,d}(r,t)$.

We use the same data frame as in Cardot et al. (2013b). More exactly, we have a population U of $N = 15055$ electricity consumption curves measured every half an hour during one week, so that there are $\mathcal{D} = 336$ time points. The mean consumption during the previous week for each meter k, denoted x_k, is used as an auxiliary variable. This variable is strongly correlated to the consumption curve $Y_k(t)$ (the pointwise correlation is always larger than 0.80) and is nearly proportional to $Y_k(t)$ at each instant t. It is also inexpensive to transmit.

We select samples s of size n drawn with inclusion probabilities π_k proportional to the past mean electricity consumption. This means that $\pi_k = n\frac{x_k}{\sum_{k\in U} x_k}$. As mentioned in Deville and Tillé (2005), this kind of sampling may be viewed as a balanced sampling with the balancing variable $\boldsymbol{\pi} = (\pi_1, ..., \pi_N)$. Note that by construction, the sample is also balanced on $(x_1, \ldots, x_N)$, *i.e* $\sum_{k\in s} x_k/\pi_k = \sum_{k\in U} x_k$. The sample is drawn using the fast version (see Chauvet and Tillé (2006)) of the cube algorithm (see Deville and Tillé (2004)). As suggested in Chauvet (2007), a random sort of the population is made before the sample selection. The true mean consumption curve observed in the population U and one estimation obtained from a sample s' of size $n = 1500$ are drawn in Figure 1.

The inclusion probabilities π_{kl} being unknown, we have obtained an empirical estimation of the covariance function γ_p via Monte Carlo. We draw $J = 10000$ samples, denoted by s_j, for $j = 1, \ldots, J$ and consider the following Monte Carlo approximation to γ_p,

$$\gamma_{emp}(r,t) = \frac{1}{J-1}\sum_{j=1}^{J}(\hat{\mu}_{d,j}(t) - \hat{\bar{\mu}}_d(t))(\hat{\mu}_{d,j}(r) - \hat{\bar{\mu}}_d(r)), \quad (r,t) \in [0,T]\times[0,T], \qquad (14)$$

with $\hat{\mu}_{d,j}(t) = \frac{1}{N}\sum_{k\in s_j}\frac{Y_{k,d}(t)}{\pi_k}$, $\hat{\bar{\mu}}_d(t) = \frac{1}{J}\sum_{j=1}^{J}\hat{\mu}_{d,j}(t)$. The empirical variance function γ_{emp} (solid line) of estimator $\hat{\mu}_d$, the Hájek approximation γ_H (dotted line) and one estimation $\hat{\gamma}_{H,d}$ (dashed line) obtained from the same sample s' are drawn in Figure 2.

To evaluate the performance of estimator $\hat{\gamma}_{H,d}$, we consider different sample sizes, $n = 250$, $n = 500$ and $n = 1500$. The corresponding values of $d(\pi)$ are $d(\pi) = 241.2$, $d(\pi) = 464.7$ and $d(\pi) = 1202.3$ meaning that our asymptotic point of view is justified in this study.

For each sample size, we draw $I = 10000$ samples and we compute the following quadratic

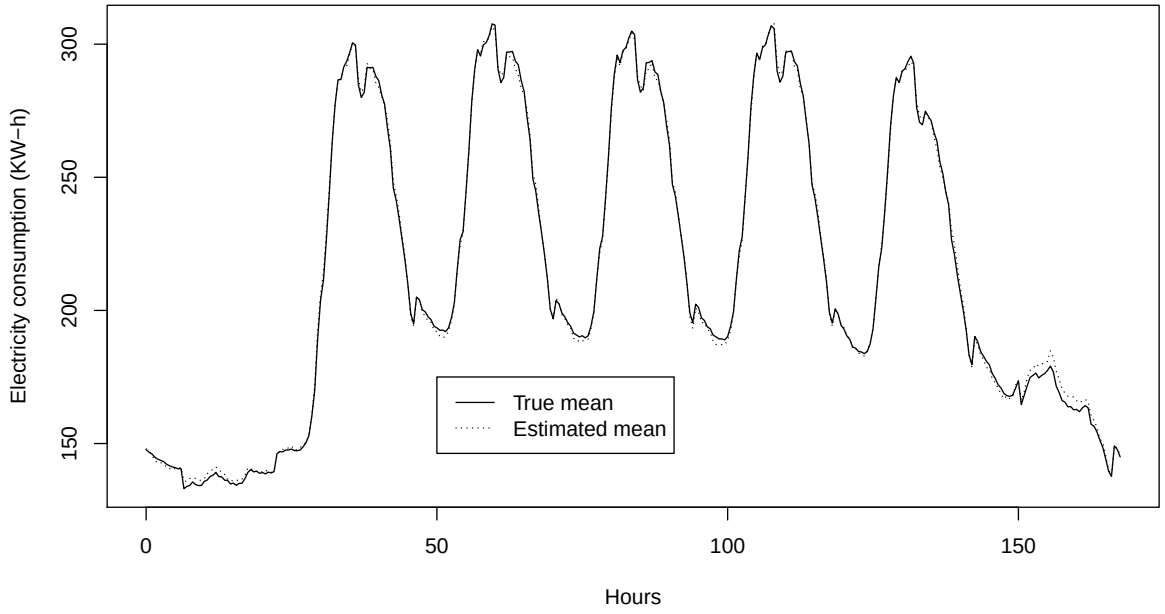

Figure 1: Mean consumption curve and its Horvitz-Thompson estimation obtained from sample s', with $n = 1500$.

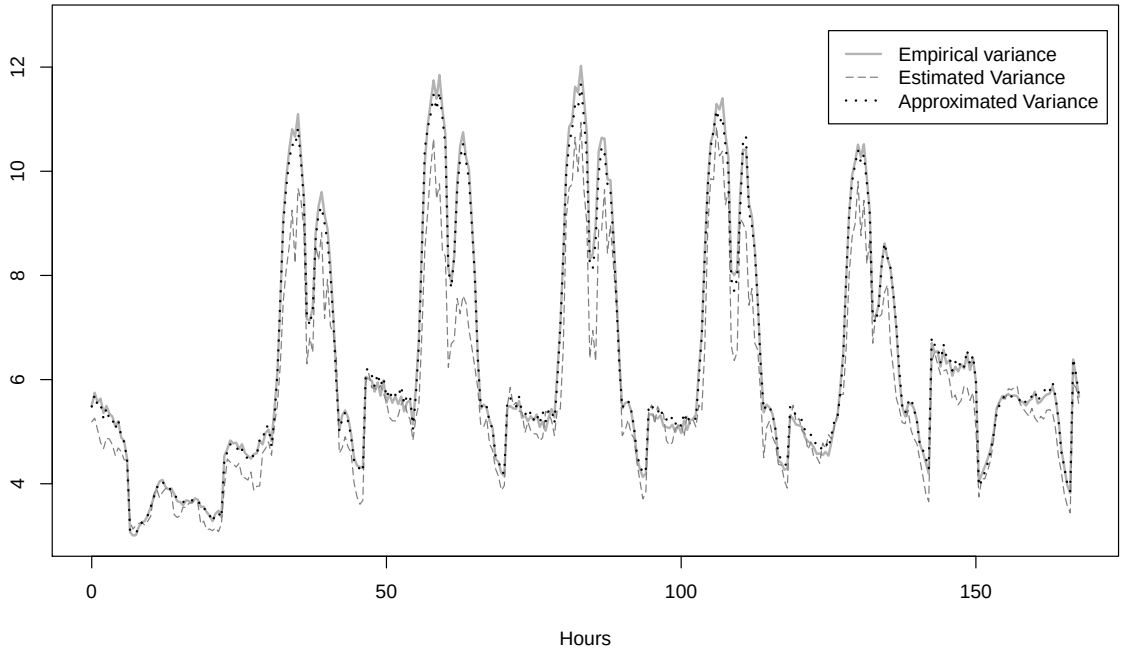

Figure 2: Empirical variance γ_{emp} (solid line), Hájek's approximation γ_H (dotted line) and variance estimation $\hat{\gamma}_{H,d}$ (dashed line) obtained from sample s', with $n = 1500$.

loss criterion

$$R(\hat{\gamma}_{H,d}) = \frac{1}{\mathcal{D}} \sum_{\mathbb{d}=1}^{\mathcal{D}} \frac{|\hat{\gamma}_H(t_{\mathbb{d}}, t_{\mathbb{d}}) - \gamma_{emp}(t_{\mathbb{d}}, t_{\mathbb{d}})|^2}{\gamma_{emp}(t_{\mathbb{d}}, t_{\mathbb{d}})^2}$$
$$\simeq \int \frac{|\hat{\gamma}_H(t,t) - \gamma_{emp}(t,t)|^2}{\gamma_{emp}(t,t)^2} dt. \tag{15}$$

We also compute the relative mean squared error,

$$RMSE = \frac{1}{I} \sum_{i=1}^{I} R(\hat{\gamma}_{H,d}^{(i)})$$
$$= RB^2(\hat{\gamma}_{H,d}) + RV(\hat{\gamma}_{H,d}), \tag{16}$$

where $\hat{\gamma}_{H,d}^{(i)}$ is the value of $\hat{\gamma}_{H,d}$ computed for the ith simulation. It is decomposed as the sum of two terms. The term $RB^2(\hat{\gamma}_{H,d})$ which corresponds to the square relative bias (or approximation error) is defined by

$$RB(\hat{\gamma}_{H,d})^2 = \frac{1}{\mathcal{D}} \sum_{\mathbb{d}=1}^{\mathcal{D}} \left(\frac{\overline{\hat{\gamma}}_{H,d}(t_{\mathbb{d}}, t_{\mathbb{d}}) - \gamma_{emp}(t_{\mathbb{d}}, t_{\mathbb{d}})}{\gamma_{emp}(t_{\mathbb{d}}, t_{\mathbb{d}})} \right)^2$$

where $\overline{\hat{\gamma}}_{H,d}(t_{\mathbb{d}}, t_{\mathbb{d}}) = \sum_{i=1}^{I} \hat{\gamma}_{H,d}^{(i)}(t_{\mathbb{d}}, t_{\mathbb{d}})/I$ and $\hat{\gamma}_{H,d}^{(i)}(t_{\mathbb{d}}, t_{\mathbb{d}})$ is the variance estimation obtained for the ith simulated sample. The second term $RV(\hat{\gamma}_{H,d}) = RMSE - RB^2(\hat{\gamma}_{H,d})$ can be interpreted as the relative variance of estimator $\hat{\gamma}_{H,d}$.

Sample Size	$RMSE$	$RB^2(\hat{\gamma}_{H,d})$	$R(\hat{\gamma}_{H,d})$				
			5%	1^{st} quartile	median	3^{rd} quartile	95%
250	0.9473	0.0004	0.0188	0.0298	0.0446	0.0748	0.4326
500	0.3428	0.0002	0.0121	0.0191	0.0278	0.0456	0.3510
1500	0.1406	0.0003	0.006	0.0097	0.0144	0.0272	0.0929

Table 1: $RMSE$, $RB^2(\hat{\gamma}_{H,d})$ and estimation errors according to criterion $R(\hat{\gamma}_{H,d})$ for different sample sizes, with $I = 10000$ simulations.

The estimation errors are presented in Table 1 for the three considered sample sizes. We first note that the values of the relative square bias $RB(\hat{\gamma}_{H,d})$ are very low, meaning that the Hájek's formula provides, in our relatively large sample context, a very good approximation to the variance. The median error for $R(\hat{\gamma}_{H,d})$ is slightly larger but remains small (always less than 5%), even for moderate sample sizes (n=250). This means that the most important part of the variance estimation error is due to the sampling error. We have drawn in Figure 3 the approximation error $\gamma_{emp}(t,r) - \gamma_{H,d}(t,r)$ and in Figure 4 the estimation error $\gamma_{emp}(t,r) - \hat{\gamma}_{H,d}(t,r)$ for $t, r \in \{1, \ldots, \mathcal{D}\}$, corresponding to a sample of size $n = 1500$ with an estimation

error close to the median value of the global risk, $R(\hat{\gamma}_{H,d}) = 0.0144$. It appears that the largest estimation errors for the variance occur when the level of consumption is high. We can also observe in these Figures a kind of periodic pattern which can be related to the daily electricity consumption behavior.

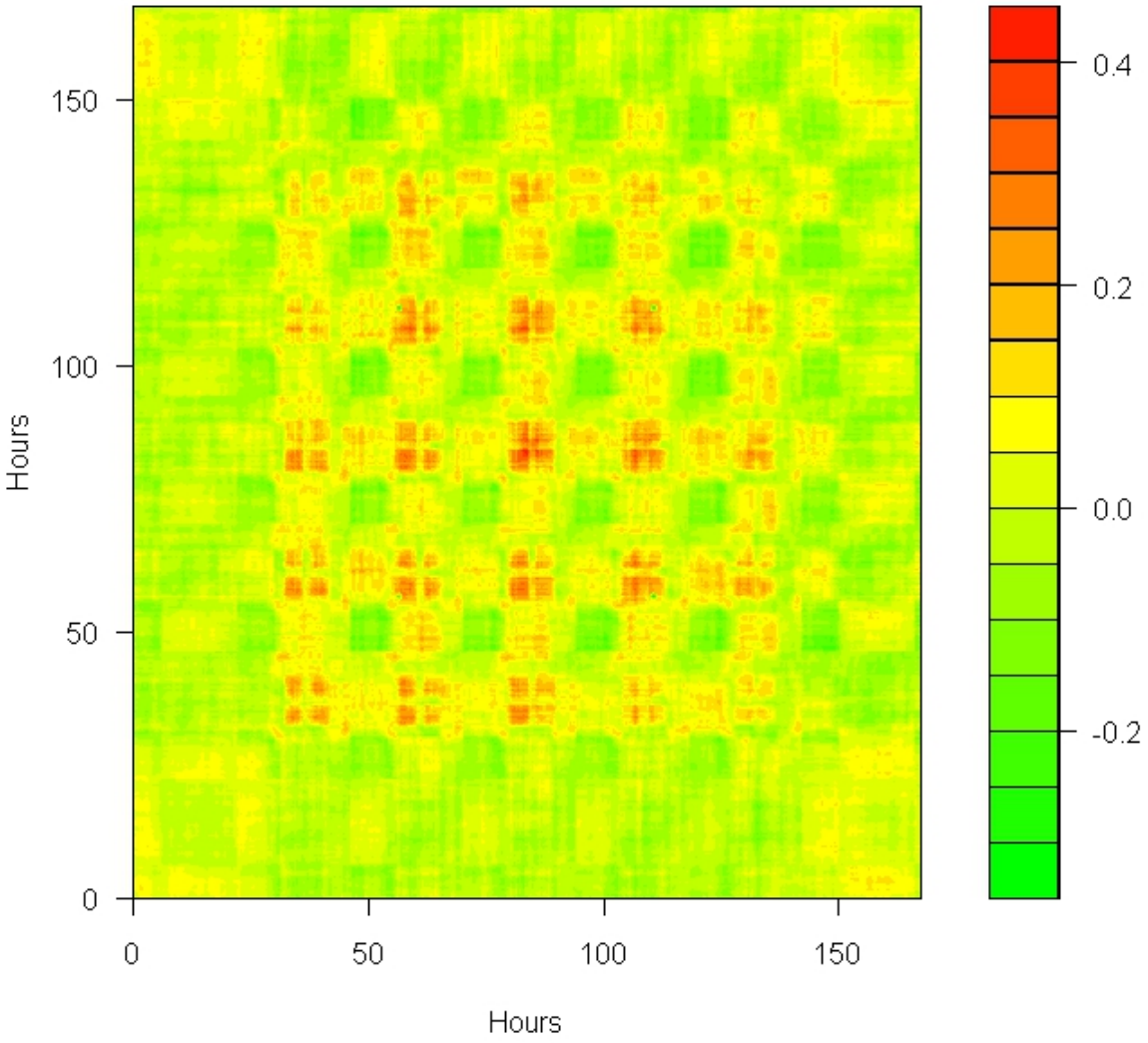

Figure 3: Approximation error $\gamma_{emp} - \gamma_{H,d}$ for a sample of size $n = 1500$.

Nevertheless, we also note that the relative mean squared error $RMSE$, which is approximately equal to the relative variance of the estimator $\hat{\gamma}_{H,d}$, is rather high, especially for small sample sizes ($n = 250$). Looking at the 95 % quantiles of $R(\hat{\gamma}_{H,d})$ in Table 1, we can deduce that bad variance estimations only occur in rare cases but with very large errors. A closer look at the data shows that the bad performance of the variance estimator, in terms of RMSE,

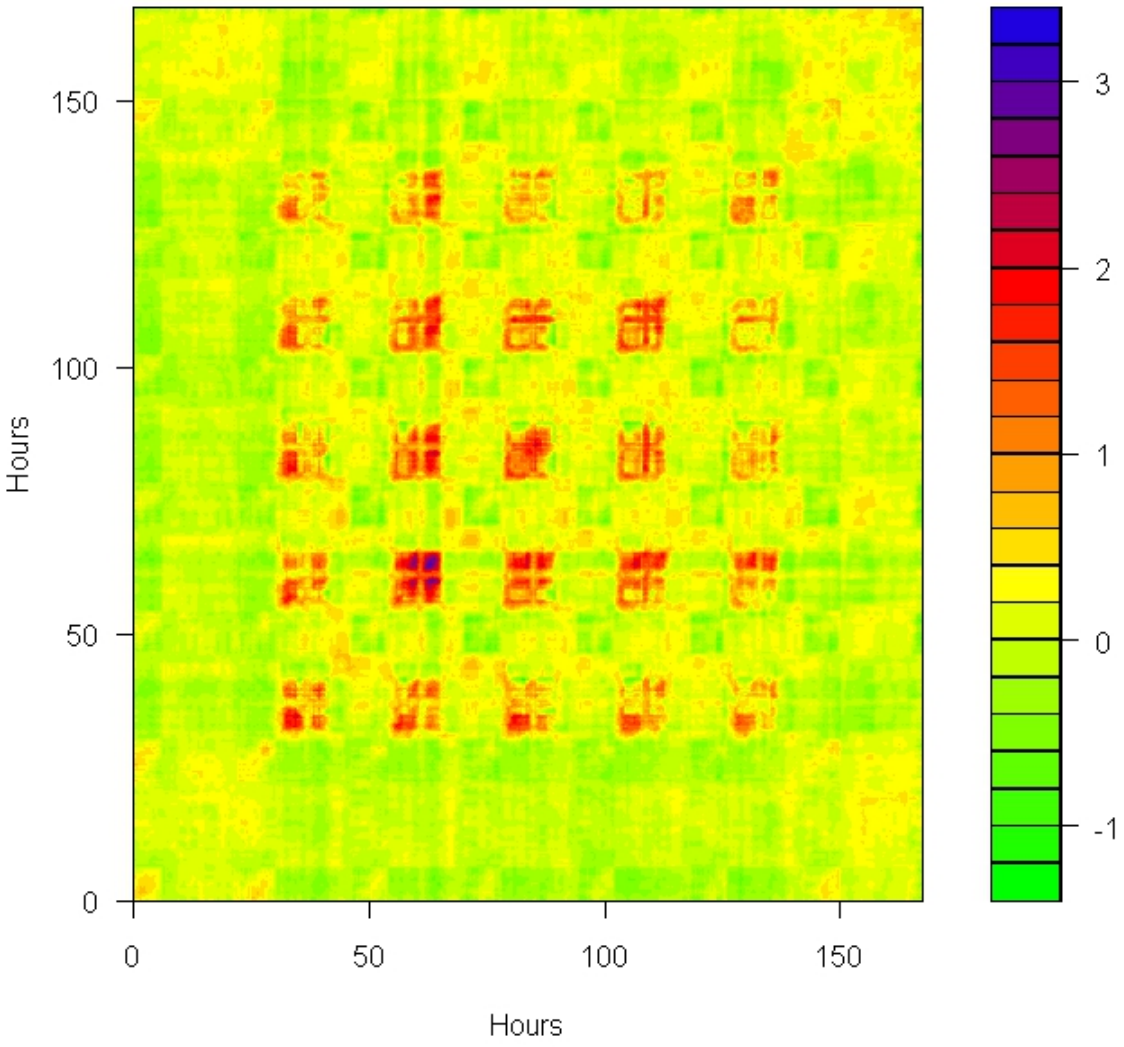

Figure 4: Estimation error $\gamma_{emp} - \hat{\gamma}_{H,d}$ for a a sample of size $n = 1500$.

is in fact due to a few individuals in the population that have both a very small inclusion probability π_k and a consumption level Y_k that can be very important at some instants of the period. Their selection in the sample, which occurs rarely, leads to an overestimation of the mean curve and to a large error $R(\hat{\gamma}_{H,d})$ when estimating the variance at these instants.

5 Concluding remarks

We have studied in this work simple estimators of the covariance function of the Horvitz-Thompson estimator for curve data considering high entropy unequal probability sampling designs. Our variance estimators, which are based on the asymptotic Hájek approximation to the second order inclusion probabilities, are well suited for large samples drawn in large populations. It is shown under reasonable conditions on the regularity of the curves and on the sampling design that we get consistent estimators that can also be used to build confidence bands for the mean, or total, curve by employing an approach based on Gaussian process simulations. The illustration on the estimation of mean electricity consumption curves with πps samples drawn with the Cube algorithm shows that, in most of cases, the estimation error of the covariance function is small. Nevertheless, we have in the population a few very influent observations (about 10 units in a population of $N = 15055$) which are characterized by very small inclusion probabilities and high values of electricity consumption at some instant of the considered period. These influent observations, which can be detected in the sample by considering the extreme values of the real variable $m_k = \sup_{t\in[0,T]} |Y_k(t)|/\pi_k$, completely deteriorate the quality of the variance estimator when they belong to the sample, which rarely occurs.

More robust estimators could be obtained at the sampling stage by preventing the inclusion probabilities from being too close to zero and by introducing a threshold $\delta > 0$ such that

$$\pi_k = n\frac{\max(\delta, x_k)}{\sum_{k\in U}\max(\delta, x_k)}.$$

Even if the resulting Horvitz-Thompson estimator would certainly be a bit less efficient, since the proportionality would not be respected anymore, it would permit to get a more stable estimation by attenuating the eventual effect of influent observations. On the other hand, another possible way to deal with this robustness issue would consist in modifying the weights of the influent observations at the estimation stage by introducing a correction such as winsorization (see *e.g* Beaumont and Rivest (2009) for a review). In our variance estimation functional context, this topic is new and would certainly deserve further investigation.

Acknowledgements. The authors thanks the two anonymous referees as well as an associate editor for their constructive remarks.

A Proofs

Throughout the proofs we use the letter C to denote a generic constant whose value may vary from place to place. Let us also define $\Delta_{kl} = \pi_{kl} - \pi_k\pi_l$ and $\Delta_{kk} = \pi_k(1-\pi_k)$. More detailed proofs can be found in Lardin (2012).

A.1 Proof of proposition 3.1

We first consider the case of the rejective sampling $p_{rej}(s)$ and show that **A5** is true if $d(\boldsymbol{\pi}_N)$ tends to infinity. By Theorem 1 in in Boistard et al. (2012) and hypothesis **A2**, we have

$$\mathbb{E}_p\left(\mathbb{1}_{k_1k_2l_1l_2}\right) - \pi_{k_1}\pi_{k_2}\pi_{l_1}\pi_{l_2} = O(d(\pi)^{-1})$$

uniformly for $(k_1, l_1, k_2, l_2) \in D_{4,N}$. Since $\pi_{k_1}\pi_{k_2} - \pi_{k_1k_2} = O(d(\pi)^{-1})$ and $\pi_{l_1}\pi_{l_2} - \pi_{l_1l_2} = O(d(\pi)^{-1})$ uniformly for $(k_1, l_1, k_2, l_2) \in D_{4,N}$, we directly obtain that, for rejective sampling

$$\max_{(k_1,l_1,k_2,l_2)\in D_{4,N}} \left|\mathbb{E}_p\left[(\mathbb{1}_{k_1l_1} - \pi_{k_1}\pi_{l_1})(\mathbb{1}_{k_2l_2} - \pi_{k_2}\pi_{l_2})\right]\right| \leq \frac{C}{d(\pi)},$$

for some constant C.

If we consider now a different sampling design $p_N(s)$, we have with Pinsker inequality (see Theorem 6.1 in Kemperman (1969)) and the property of the total variation distance,

$$\sup_{A\in\mathcal{A}_\mathcal{N}} |p_N(A) - p_{rej}(A)| \leq \sqrt{K(p_N, p_{rej})/2}$$

where $\mathcal{A}_\mathcal{N}$ is the set of all partitions of U_N. Considering the particular cases $A = \{(k_1, l_1, k_2, l_2) \in D_{4,N}\}$, and denoting by $\pi_{k_1k_2l_1l_2} = p_N(A)$ and by $\pi^{rej}_{k_1k_2l_1l_2} = p_{rej}(A)$, we directly get that

$$\sup_{(k_1,l_1,k_2,l_2)\in D_{4,N}} \left|\pi_{k_1k_2l_1l_2} - \pi^{rej}_{k_1k_2l_1l_2}\right| \leq \sqrt{K(p_N, p_{rej})/2}$$

and the proof is complete.

A.2 Proof of Proposition 3.2 (consistency of the covariance and the variance functions)

The proof follows the same steps as in Cardot et al. (2013c). We show first that for all $t, r \in [0, T]$, the estimator of the covariance function $\widehat{\gamma}_{H,d}(r, t)$ is pointwise convergent for $\gamma_p(r, t)$ and then, that the random variable $n(\widehat{\gamma}_{H,d}(t, t) - \gamma_p(t, t))$ converges in distribution to zero in the space $C([0, T])$. By the definition of the convergence in distribution in $C([0, T])$ and the boundedness and continuity of the sup functional, we then directly obtain the uniformly convergence of the variance function estimator. As in Cardot et al. (2013c), in order to obtain the convergence in distribution of $n(\widehat{\gamma}_{H,d}(t, t) - \gamma_p(t, t))$, we first show the convergence of all finite linear combinations which results easily from the pointwise convergence. Next, we check that the sequence $n(\widehat{\gamma}_{H,d}(t, t) - \gamma_p(t, t))$ is tight.

Step 1. Pointwise convergence

We want to show, that for each $(t, r) \in [0, T] \times [0, T]$, we have

$$n\mathbb{E}_p \{ \mid \widehat{\gamma}_{H,d}(r,t) - \gamma_p(r,t) \mid \} \to 0, \quad \text{when } N \to \infty.$$

Let us decompose

$$n(\widehat{\gamma}_{\mathrm{H},d}(r,t) - \gamma_p(r,t)) = n(\widehat{\gamma}_{\mathrm{H},d}(r,t) - \widehat{\gamma}_H(r,t)) + n(\gamma_H(r,t) - \gamma_p(r,t)) + n(\widehat{\gamma}_H(r,t) - \gamma_H(r,t))$$

and study separately the interpolation, the approximation and the estimation errors.

Interpolation error

We suppose that $t \in [t_i, t_{i+1})$ and $r \in [t_{i'}, t_{i'+1})$. Using the assumptions (**A1**)-(**A3**), we can bound

$$n|\hat{\gamma}_{H,d}(r,t) - \hat{\gamma}_H(r,t)| \leq C_1|t_{i+1} - t_i|^\beta + C_2|t_{i'+1} - t_{i'}|^\beta$$

and the assumption on the grid of discretization points leads to

$$n|\hat{\gamma}_{H,d}(r,t) - \hat{\gamma}_H(r,t)| = o(1). \tag{17}$$

Approximation error

We show that, for each $(r, t) \in [0, T] \times [0, T]$, $n \mid \gamma_H(r,t) - \gamma_p(r,t) \mid = o(1)$. We write the approximation (4) as follows

$$\pi_{kl} - \pi_k\pi_l = -\pi_k\pi_l \frac{(1-\pi_k)(1-\pi_l)}{d(\pi)} + \frac{c_{kl}}{d(\pi)} \tag{18}$$

where $\max_{k \neq l \in U} |c_{kl}| \to 0$ and we use it in the expression of the covariance function given by (3):

$$\begin{aligned}\gamma_p(r,t) &= \frac{1}{2}\frac{1}{d(\pi)N^2} \sum_{k \in U} \sum_{l \neq k \in U} [\pi_k\pi_l(1-\pi_k)(1-\pi_l) - c_{kl}] \left(\frac{Y_k(r)}{\pi_k} - \frac{Y_l(r)}{\pi_l}\right)\left(\frac{Y_k(t)}{\pi_k} - \frac{Y_l(t)}{\pi_l}\right) \\ &= \gamma_H(r,t) - \frac{1}{2}\frac{1}{N^2} \sum_{k \in U} \sum_{l \neq k \in U} \frac{c_{kl}}{d(\pi)} \left(\frac{Y_k(r)}{\pi_k} - \frac{Y_l(r)}{\pi_l}\right)\left(\frac{Y_k(t)}{\pi_k} - \frac{Y_l(t)}{\pi_l}\right).\end{aligned}$$

Thus, we directly get with assumptions (**A1**)-(**A3**) that

$$d(\pi) \ |\gamma_H(r,t) - \gamma_p(r,t)| = o(1). \tag{19}$$

Sampling error

To establish the convergence of $n(\hat{\gamma}_H(r,t) - \gamma_H(r,t))$ to zero in probability as $N \to \infty$, it is enough to show that, for all $(r,t) \in [0,T] \times [0,T]$,

$$n^2 \mathbb{E}_p \left[(\hat{\gamma}_H(r,t) - \gamma_H(r,t))^2 \right] \to 0, \quad \text{when } N \to \infty.$$

Noting that

$$\begin{aligned} n|\hat{\gamma}_H(r,t) - \gamma_H(r,t)| &\leq \frac{n}{N^2} \left| \sum_{k \in U} \left(\frac{\hat{d}(\pi)}{d(\pi)} - 1 \right) \frac{\mathbb{1}_k}{\pi_k^2} (1-\pi_k) Y_k(t) Y_k(r) \right| \\ &+ \frac{n}{N^2} \left| \sum_{k \in U} \left(\frac{\mathbb{1}_k}{\pi_k} - 1 \right) \frac{1-\pi_k}{\pi_k} Y_k(t) Y_k(r) \right| \\ &+ \frac{n}{N^2} \frac{1}{d(\pi)} \left| \sum_{k \in U} \sum_{l \in U} \left(\frac{\mathbb{1}_{kl}}{\pi_k \pi_l} - 1 \right) (1-\pi_k)(1-\pi_l) Y_k(t) Y_l(r) \right| \\ &:= |B_1(r,t)| + |B_2(r,t)| + |B_3(r,t)|, \end{aligned} \tag{20}$$

we get

$$n^2 \mathbb{E}_p \left[(\hat{\gamma}_H(r,t) - \gamma_H(r,t))^2 \right] \leq 3\mathbb{E}_p(B_1(r,t)^2) + 3\mathbb{E}_p(B_2(r,t)^2) + 3\mathbb{E}_p(B_3(r,t)^2). \tag{21}$$

Let us show now that $\mathbb{E}_p(B_1(r,t)^2) \to 0$ when $N \to \infty$. Let $M = \max_{\pi_k \neq 1} \pi_k$. Under the assumptions (**A1**)-(**A3**) and the inequality $\frac{1}{d(\pi)} \leq \frac{1}{N\lambda(1-M)}$, we have

$$\mathbb{E}_p(B_1(r,t)^2) \leq \frac{n^2}{\lambda^4 d(\pi)^2} \mathbb{E}_p \left[\frac{1}{N^2} (\hat{d}(\pi) - d(\pi))^2 \right] \left[\frac{1}{N} \sum_{k \in U} Y_k^2(t) \right] \left[\frac{1}{N} \sum_{k \in U} Y_k^2(r) \right] \leq \frac{1}{n} C$$

since $\mathbb{E}_p(\frac{1}{N^2}(\hat{d}(\pi) - d(\pi))^2) = O(n^{-1})$. Hence, $E_p(B_1(r,t)^2) \to 0$ when $N \to \infty$. Now,

$$\begin{aligned} \mathbb{E}_p(B_2(r,t)^2) &\leq \frac{n^2}{N^4} \sum_{k \in U} \sum_{l \in U} \frac{|\Delta_{kl}|}{\pi_k \pi_l} \frac{1-\pi_k}{\pi_k} \frac{1-\pi_l}{\pi_l} |Y_k(t) Y_k(r) Y_l(t) Y_l(r)| \\ &\leq \frac{1}{\lambda^3} \frac{1}{N} \left(\frac{n^2}{N^2} + \frac{n^2 \max_{k \neq l \in U} |\Delta_{kl}|}{N\lambda} \right) \left(\frac{1}{N} \sum_{k \in U} |Y_k(t)|^4 \right)^{1/2} \left(\frac{1}{N} \sum_{k \in U} |Y_k(r)|^4 \right)^{1/2} \\ &\leq \frac{1}{N} C \end{aligned}$$

by assumptions (**A1**)-(**A4**). Thus $\mathbb{E}_p(B_2(r,t)^2) \to 0$ when $N \to \infty$. For the third term, we have

$$\begin{aligned}
\mathbb{E}_p(B_3(r,t)^2) = n^2\mathbb{E}_p\Bigg[&\frac{1}{N^4}\frac{1}{d(\pi)^2}\sum_{k,l\in U}\sum_{k',l'\in U}\left(\frac{\mathbb{1}_{kl}}{\pi_k\pi_l}-1\right)\left(\frac{\mathbb{1}_{k'l'}}{\pi_{k'}\pi_{l'}}-1\right)\\
&\cdot(1-\pi_k)(1-\pi_l)(1-\pi_{k'})(1-\pi_{l'})Y_k(t)Y_l(r)Y_{k'}(t)Y_{l'}(r)\Bigg]\\
&\leq \frac{n^2}{N^4}\frac{1}{d(\pi)^2}\sum_{k\in U}\sum_{k'\in U}\left|\mathbb{E}_p\left[\left(\frac{\mathbb{1}_k}{\pi_k^2}-1\right)\left(\frac{\mathbb{1}_{k'}}{\pi_{k'}^2}-1\right)\right]\right||Y_k(t)Y_k(r)Y_{k'}(t)Y_{k'}(r)|\\
&+\frac{2n^2}{N^4}\frac{1}{d(\pi)^2}\sum_{k\in U}\sum_{k'\neq l'\in U}\left|\mathbb{E}_p\left[\left(\frac{\mathbb{1}_k}{\pi_k^2}-1\right)\left(\frac{\mathbb{1}_{k'l'}}{\pi_{k'}\pi_{l'}}-1\right)\right]\right||Y_k(t)Y_k(r)Y_{k'}(t)Y_{l'}(r)|\\
&+\frac{n^2}{N^4}\frac{1}{d(\pi)^2}\sum_{k\neq l\in U}\sum_{k'\neq l'\in U}\left|\mathbb{E}_p\left[\left(\frac{\mathbb{1}_{kl}}{\pi_k\pi_l}-1\right)\left(\frac{\mathbb{1}_{k'l'}}{\pi_{k'}\pi_{l'}}-1\right)\right]\right||Y_k(t)Y_l(r)Y_{k'}(t)Y_{l'}(r)|\\
&:= v_1+v_2+v_3.
\end{aligned}$$

To bound v_1, v_2, v_3, the proof follows the same lines as above. We write each double sum $\sum_{k\in U}\sum_{l\in U}$ as the sum of two terms: the first one is $\sum_{k\in U}$ and is obtained for $k=l$ and the second one is $\sum_{k\in U}\sum_{l\neq k\in U}$. Under assumptions (**A1**), (**A2**) and (**A4**) and the facts that $\pi_{kl} \leq \pi_k\pi_l$ and $d(\pi)\to\infty$, we get that $v_1 \to 0$. Next, we can write

$$v_3 \leq \frac{C}{N} + \frac{n^2}{\lambda^4 d^2(\pi)}\max_{(k,l,k',l')\in D_{4,N}}\left|\mathbb{E}_p\left[\left(\mathbb{1}_{kl}-\pi_k\pi_l\right)\left(\mathbb{1}_{k'l'}-\pi_{k'}\pi_{l'}\right)\right]\right|\left(\frac{1}{N}\sum_{k\in U}Y_k^2(t)\right)\left(\frac{1}{N}\sum_{l\in U}Y_l^2(r)\right),$$

so that $v_3 \to 0$ when $N\to\infty$ and the assumptions (**A1**)-(**A5**) are fulfilled. By the Cauchy-Schwarz inequality, we have $v_2 \to 0$ when $N\to\infty$. Finally, we have that for all $(r,t)\in[0,T]\times[0,T]$, $n|\hat{\gamma}_H(r,t)-\gamma_H(r,t)| \to 0$, when $N\to\infty$. Finally, the proof of step 1 is complete using (17) and (19).

Step 2. Tightness

To check the tightness of $n(\hat{\gamma}_H(t,t)-\gamma_H(t,t))$ in $C[0,T]$, we use the Theorem 12.3 from Billingsley (1968) which requires that the sequence is tight for $t=0$ and that the increments of $n(\hat{\gamma}_H-\gamma_H)$ between two instants t and r satisfy

$$d_\gamma^2(t,r) = n^2\mathbb{E}_p(|\widehat{\gamma}_H(t,t)-\gamma_H(t,t)-\widehat{\gamma}_H(r,r)+\gamma_H(r,r)|^2) \leq C|t-r|^{2\beta}, \quad \beta > 1/2$$

for some positive constant C and all $(r,t)\in[0,T]\times[0,T]$.

The pointwise convergence of $n(\hat{\gamma}_H-\gamma_H)$ implies that $n(\hat{\gamma}_H(0,0)-\gamma_H(0,0))$ is tight. Using (20), we can decompose $d_\gamma^2(t,r)$ into 3 parts,

$$\begin{aligned}
d_\gamma^2(r,t) &\leq 3\Big(\mathbb{E}_p\left([B_1(t,t)-B_1(r,r)]^2\right)+\mathbb{E}_p\left([B_2(t,t)-B_2(r,r)]^2\right)+\mathbb{E}_p\left([B_3(t,t)-B_3(r,r)]^2\right)\Big)\\
&:= 3\left(d_{B_1}^2+d_{B_2}^2+d_{B_3}^2\right).
\end{aligned}$$

Denote by $\phi_{kl}(t,r) = Y_k(t)Y_l(t) - Y_k(r)Y_l(r)$ with $\phi_k(t,r) = Y_k^2(t) - Y_k^2(r)$ for $k = l$. Assuming (**A3**), we get that $\left(\frac{1}{N}\sum_{k\in U}|\phi_k(t,r)|\right)^2 \leq C|t-s|^{2\beta}$ and $\left(\frac{1}{N^2}\sum_{k,l\in U}|\phi_{kl}(t,r)|\right)^2 \leq C|t-s|^{2\beta}$. Moreover, under the assumptions (**A1**) and (**A2**), we have

$$d_{B_1}^2 \leq \frac{n^2}{N^2}\left(\frac{1+\lambda}{\lambda^3}\right)^2\left(\frac{1}{N}\sum_{k\in U}|\phi_k(t,r)|\right)^2 \leq C|t-r|^{2\beta} \tag{22}$$

as well as

$$d_{B_2}^2 \leq \frac{n^2}{N^2}\left(\frac{1+\lambda}{\lambda^2}\right)^2\left(\frac{1}{N}\sum_{k\in U}|\phi_k(t,r)|\right)^2 \leq C|t-r|^{2\beta}. \tag{23}$$

Finally,

$$d_{B_3}^2 \leq \frac{n^2}{d(\pi)^2}\left[\frac{1}{N^2}\sum_{k\in U}\sum_{l\in U}|\phi_{kl}(t,r)|\right]^2 \leq C|t-r|^{2\beta} \tag{24}$$

and with inequalities (22), (23) we deduce that $d_\gamma^2(r,t) \leq C|t-r|^{2\beta}$. The proof is complete.

Proof of Proposition 3.2, point (2): Under the assumptions (**A1**) and (**A2**), it is clear that $\hat{d}(\pi)/d(\pi) = 1 + o_p(1)$. The pointwise convergence of $n\hat{\gamma}_{H,d}^*(r,t)$ is then a direct consequence of Proposition 3.2, point (1) and the fact that $\hat{\gamma}_{H,d}^*(r,t) = \dfrac{d(\pi)}{\hat{d}(\pi)}\hat{\gamma}_{H,d}(r,t)$. Furthermore, we may write

$$n(\hat{\gamma}_{H,d}^* - \gamma_H) = n\frac{d(\pi)}{\hat{d}(\pi)}(\hat{\gamma}_{H,d} - \gamma_H) + n\left(\frac{d(\pi)}{\hat{d}(\pi)} - 1\right)\gamma_H.$$

By Slutsky's theorem, the first term at the righthand-side of previous equation converges in distribution to zero in $C([0,T])$ while the second term goes to zero in probability since $\sup_{(r,t)\in[0,T]\times[0,T]}|n\gamma_H(r,t)| < \infty$ and $\frac{d(\pi)}{\hat{d}(\pi)} - 1 = o_p(1)$. Hence, the sequence $n(\hat{\gamma}_{H,d}^* - \gamma_H)$ converges in distribution to zero in $C([0,T])$.

A.3 Proof of Proposition 3.3

We first note that the interpolation error, bounded in (17), satisfies

$$n^{3/2}|\hat{\gamma}_{H,d}(r,t) - \hat{\gamma}_H(r,t)| = O(1) \tag{25}$$

provided that $\lim_{N\to\infty}\max_{i=\{1,..,d_N-1\}}|t_{i+1}-t_i|^{2\beta} = O(n^{-1})$. We then use the fact (see Theorem 1 in Boistard et al. (2012)) that for rejective sampling the terms c_{kl} defined in (18) satisfy, for some constant C, $\max_{k,l}|c_{kl}| \leq Cd(\pi)^{-1}$. Thus, bound (19) is now $d(\pi)^2\,|\gamma_H(r,t) - \gamma_p(r,t)| = O(1)$. If we examine the sampling error, we can check that the terms B_1 and B_2 are of order n^{-1}. Concerning the term B_3, it is bounded by the sum $v_1 + v_2 + v_3$ with $v_1 = O(d^{-2}(\pi))$ and $v_2 \leq \sqrt{v_1 v_3}$. Thanks to Proposition 3.1, we get that the term v_3 satisfies $v_3 = O(d^{-1}(\pi))$ and consequently, $\mathbb{E}_p(B_3(r,t)^2) = O(n^{-1})$. Thus, $n^2\mathbb{E}_p\left[(\hat{\gamma}_H(r,t) - \gamma_H(r,t))^2\right] = O(n^{-1})$ and the proof is complete.

References

Arratia, R. and Goldstein, L. and Langholz, B. (2005). Local central limit theorems, the high-order correlations of rejective sampling and logistic likelihood asymptotics. *Ann. Statist*, 33:871–891.

Beaumont, J.-F. and Rivest, L.-P. (2009). Dealing with outliers in survey data. In *Sample surveys: design, methods and applications*, volume 29 of *Handbook of Statist.*, pages 247–279. Elsevier/North-Holland, Amsterdam.

Berger, Y. G. (1998a). Rate of convergence for asymptotic variance of the Horvitz-Thompson estimator. *Journal of Statistical Planning and Inference*, 74:149–168.

Berger, Y. G. (1998b). Rate of convergence to normal distribution for the Horvitz-Thompson estimator. *Journal of Statistical Planning and Inference*, 67:209–226.

Billingsley, P. (1968). *Convergence of Probability Measures*. John Wiley and Sons.

Boistard, H., Lopuhaä, H. P., and Ruiz-Gazen, A. (2012). Approximation of rejective sampling inclusion probabilities and application to higher order correlation. *Electron. J. Stat.*, 6:1967–1983.

Bondesson, L. (2010). Conditional and restricted Pareto sampling: two new methods for unequal probability sampling. *Scand. J. Stat.*, 37(3):514–530.

Bondesson, L., Traat, I., and Lundqvist, A. (2006). Pareto sampling versus Sampford and conditional Poisson sampling. *Scand. J. Statist.*, 33(4):699–720.

Brewer, K. R. W. and Hanif, M. (1983). *Sampling with unequal probabilities*, volume 15 of *Lecture Notes in Statistics*. Springer-Verlag, New York.

Cardot, H., Degras, D., and Josserand, E. (2013a). Confidence bands for Horvitz-Thompson estimators using sampled noisy functional data. *Bernoulli*, to appear.

Cardot, H., Dessertaine, A., Goga, C., Josserand, E., and Lardin, P. (2013b). Comparaison de différents plans de sondage et construction de bandes de confiance pour l'estimation de la moyenne de données fonctionnelles : une illustration sur la consommation électrique. *Survey Methodology / Technique d'enquêtes*, to appear.

Cardot, H., Goga, C., and Lardin, P. (2013c). Uniform convergence and asymptotic confidence bands for model-assisted estimators of the mean of sampled functional data. *Electron. J. Stat.*, 6:2535–2562.

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.

Chauvet, G. (2007). *Méthodes de bootstrap en population finie*. PhD thesis, Université de Rennes II.

Chauvet, G. and Tillé, Y. (2006). A fast algorithm of balanced sampling. *Computational Statistics*, 21:53–61.

Chen, X.-H., Dempster, A. P., and Liu, J. S. (1994). Weighting finite population sampling to maximise entropy. *Biometrika*, 81:457–469.

Degras, D. A. (2011). Simultaneous confidence bands for nonparametric regression with functional data. *Statist. Sinica*, 21:1735–1765.

Deville, J.-C. and Tillé, Y. (2004). Efficient balanced sampling: the cube algorithm. *Biometrika*, 91:893–912.

Deville, J.-C. and Tillé, Y. (2005). Variance approximation under balanced sampling. *J. Statist. Plann. Inference*, 128:569–591.

Erdös, P. and Rényi, A. (1959). On the central limit theorem for samples from a finite population. *Publ. Math. Inst. Hungar. Acad.Sci.*, 4:49–61.

Fuller, W. A. (2009). Some design properties of a rejective sampling procedure. *Biometrika*, 96(4):933–944.

Hájek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Publ. Math. Inst. Hungar. Acad.Sci.*, 77:361–374.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics*, 35:1491–1523.

Hájek, J. (1981). *Sampling from a finite population*. Statistics: Textbooks and Monographs. Marcel Dekker, New York.

Kemperman, J. H. B. (1969). On the optimum rate of transmitting information. *Ann. Math. Statist.*, 40:2156–2177.

Lardin, P. (2012). *Estimation de synchrones de consommation électrique par sondage et prise en compte d'information auxiliaire*. PhD thesis, Université de Bourgogne, France.

Lundqvist, A. (2007). On the distance between some πps sampling designs. *Acta Appl. Math.*, 97(1-3):79–97.

Matei, A. and Tillé, Y. (2005). Evaluation of variance approximations and estimators in maximum entropy sampling with unequal probability and fixed sample size. *Journal of Official Statistics*, 21(4):543–570.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional data analysis*. Springer-Verlag, New York, second edition.

Sen, A. (1953). On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 5:119–127.

Shahbaz, M-Q. and Hanif, M. (2003). Variance formulas for Horvitz-Thompson estimator using first order inclusion probabilities. *J. Applied Statistical Science*, 12:201–208.

Tillé, Y. (2006). *Sampling algorithms*. Springer Series in Statistics. Springer, New York.

Víšek, J. Á. (1979). Asymptotic distribution of simple estimate for rejective, Sampford and successive sampling. In *Contributions to statistics*, pages 263–275. Reidel, Dordrecht.

Yates, F. and Grundy, P. (1953). Selection without replacement from within strata with probability proportional to size. *J. Royal Statist. Soc., B*, 15:235–261.