Survey sampling targeted inference

Antoine Chambaz[1,2,3], Emilien Joly[1], and Xavier Mary[1]

[1] Modal'X, Université Paris Nanterre
[2] Division of Biostatistics, School of Public Health, UC Berkeley
[3] MAP5 (UMR CMRS 8145), Université Paris Descartes

1 Introduction

Consider the following situation: we wish to build a confidence interval (CI) for a real-valued pathwise differentiable parameter Ψ evaluated at a law P_0, $\psi_0 \equiv \Psi(P_0)$, from a data set $O_1, \ldots, O_N$ of independent random variables drawn from P_0 but, as is often the case nowadays, N is so large that we will not be able to use all data. To overcome this computational hurdle, we decide *(i)* to select n among N observations randomly with unequal probabilities and *(ii)* to adapt targeted minimum loss inference from the smaller data set that results from the selection. First explored in (Bertail et al., 2016a), our approach is an alternative to the so called "online targeted learning" developed by van der Laan and Lendle (2014).

The selection of n among N observations will be the random outcome of a survey sampling design. From now on, we assume that each observation O_i is summarized by V_i, a low-dimensional random variable, and that $V_1, \ldots, V_N$ are all observed. We will draw advantage from $V_1, \ldots, V_N$ to adjust the probability that each O_i be sampled. We will develop two examples of survey sampling designs: Sampford's and determinantal sampling designs. Also known as rejective sampling design based on Poisson sampling with unequal inclusion probabilities, Sampford's sampling design is a particular case of sampling without replacement (Hanif and Brewer, 1980). It has been thoroughly studied since the publication of the seminal articles (Hajek, 1964; Sampford, 1967). Recently introduced in sampling theory by Loonis and Mary (2015), determinantal sampling design benefits from a rich literature on determinantal point processes (Macchi, 1975; Lyons, 2003; Hough et al., 2006).

The manuscript is organized as follows. Section 2 presents the general template for targeted inference from large data sets by survey sampling. Section 3 introduces and discusses the two examples of survey sampling designs mentioned above. Section 4 addresses their optimization in terms of minimization of the asymptotic variance of the targeted estimator resulting from their use. Section 5 develops an example, that of the inference of a variable importance measure of a continuous exposure. Section 6 presents a simulation study illustrating the implementation of the general template in the example. Finally, Section 7 gathers some elements of proof.

2 Template for targeted inference from large data sets by survey sampling

This section presents a template for carrying out targeted inference from large data sets by survey sampling. Section 2.1 formalizes survey sampling. Section 2.2 quickly describes the construction of the targeted minimum loss estimator (TMLE, which also stands for targeted minimum loss estima*tion*) based on a survey sample and states a central limit theorem (CLT) which enables the construction of CIs of given asymptotic level. The CLT relies on general assumptions typical of empirical processes theory. Section 3 discusses them in the contexts of the Sampford's and determinantal survey sampling designs.

Throughout the manuscript, we denote $\mu f \equiv \int f d\mu$ and $\|f\|_{2,\mu} \equiv (\mu f^2)^{1/2}$ for any measure μ and function f (measurable and integrable with respect to μ).

2.1 Retrieving the observations by survey sampling

As explained in introduction, the first step of the inference procedure is the random selection without replacement of n among N observations. The survey sample size n is set beforehand. Down to earth computational considerations (how many data can the package handle?; how much time are we willing to wait for the results of inference?) typically drive its choice.

Our analysis is asymptotic: we assume that N goes to infinity and that n goes to infinity as N does, in such a way that the ratio n/N go to 0. How n depends on N may or may not need to be described more precisely. The results of this manuscript could be extended to the case that n is random and satisfies these two conditions almost surely (with respect to the law of the sampling design; more details to follow).

The random selection of n among N observations takes the form of a vector $\eta \equiv (\eta_1, \ldots, \eta_N)$ of binary random variables where, for each $1 \leq i \leq N$, O_i is selected if and only if η_i equals 1. The conditional joint distribution of η given $O_1, \ldots, O_N$ is the survey sampling design. By construction, it coincides with the conditional joint distribution of η given the summary measures $V_1, \ldots, V_n$ which, contrary to $O_1, \ldots, O_N$, are all observed at the beginning of the study.

We denote P^s a generic conditional joint distribution of η given $O_1, \ldots, O_n$ (the superscript "s" stands for "survey"). The first order inclusion probabilities are the (conditional marginal) probabilities $\pi_i \equiv P^s(\eta_i = 1)$ for $1 \leq i \leq N$. In case they are equal, the sampling design is said equally weighted. The second order inclusion probabilities are the (conditional joint) probabilities $\pi_{ij} \equiv P^s(\eta_i = 1, \eta_j = 1) = P^s(\eta_i \eta_j = 1)$ for $1 \leq i \neq j \leq N$. The Horvitz-Thompson empirical measure

$$P_n^{\mathrm{HT}} \equiv \frac{1}{N} \sum_{i=1}^{N} \frac{\eta_i}{\pi_i} \mathrm{Dirac}(O_i) \tag{0.1}$$

takes up the role that the empirical measure $P_N \equiv N^{-1} \sum_{i=1}^{N} \mathrm{Dirac}(O_i)$ would play if we had access to it. The former is an unbiased estimator of the latter in the sense that, for any function f of O drawn from P_0,

$$\mathrm{E}_{P^s}\left[P_n^{\mathrm{HT}} f\right] = P_N f.$$

A CLT may hold for $\sqrt{n}(P_n^{\mathrm{HT}} - P_N)f$ (conditionally on $O_1, \ldots, O_N$). Whether it does or not notably depends on the asymptotic behavior of $\mathrm{Var}_{P^s}\left[P_n^{\mathrm{HT}} f\right]$. In general, it holds that

$$\mathrm{Var}_{P^s}\left[P_n^{\mathrm{HT}} f\right] = \frac{1}{N^2} \sum_{i=1}^{N} \left(\frac{1}{\pi_i} - 1\right) f^2(O_i) + \frac{1}{N^2} \sum_{1 \leq i \neq j \leq N} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1\right) f(O_i) f(O_j). \tag{0.2}$$

Sections 3 and 4 focus on the specific examples of the Sampford and determinantal survey sampling designs. The choice of the sampling design affects the limit variance of the TMLE. For the time being, we do not characterize further the survey sampling design.

2.2 CLT on the TMLE and resulting confidence intervals

Constructing the TMLE. We begin with a minor twist. We assume that the real-valued pathwise differentiable parameter of interest Ψ, originally seen as a real-valued mapping on the set of probability measures on $\mathcal{O}$ equipped with the Borel σ-field, can be extended to a real-valued mapping on the set $\mathcal{M}$ of finite measures on $\mathcal{O}$ equipped with the Borel σ-field. Moreover, we assume that the extension can be performed in such a way that the pathwise differentiability of Ψ is preserved: for each $P \in \mathcal{M}$ (not necessarily a probability measure), there exists an influence function $D(P) \in L_0^2(P)$ (the set of measurable functions f on $\mathcal{O}$

satisfying $Pf = 0$ and Pf^2 finite) such that, for all bounded $s \in L_0^2(P)$ with $\|s\|_\infty > 0$, if we characterize $P_\epsilon \in \mathcal{M}$ by setting $dP_\epsilon/dP = 1 + \epsilon s$ (all $\epsilon \in \mathbb{R}$ with $|\epsilon| < \|s\|_\infty^{-1}$), then $\epsilon \mapsto \Psi(P_\epsilon)$ is differentiable at $\epsilon = 0$ with a derivative equal to $PsD(P)$. This is not asking much, as the example developed in Section 5 illustrates.

Suppose that we have constructed $P_n^* \in \mathcal{M}$ targeted to ψ_0 in the sense that

$$P_n^{\mathrm{HT}} D(P_n^*) = o_P(1/\sqrt{n}). \tag{0.3}$$

The TMLE is the substitution estimator $\psi_n^* \equiv \Psi(P_n^*)$.

Central limit theorem and CIs. The CLT hinges on three assumptions. We suppose the existence of $\mathcal{F} \subset \{D(P) : P \in \mathcal{M}\}$ such that

A1. The empirical process $\sqrt{n}(P_n^{\mathrm{HT}} - P_0)$ converges in law in $\ell^\infty(\mathcal{F})$ to a zero-mean Gaussian process with covariance function Σ_h.

A2. With P_0-probability tending to one, $D(P_n^*) \in \mathcal{F}$, and there exists $f_1 \in \mathcal{F}$ such that $\|D(P_n^*) - f_1\|_{2,P_0} = o_P(1)$. Moreover, one knows a conservative estimator σ_n^2 of $\sigma_1^2 \equiv \Sigma_h(f_1, f_1)$.

Conservative estimation of σ_1^2 is not as easy as one might think at first sight. For instance, it is not guaranteed in general that the substitution estimator

$$\hat{\sigma}_n^2 \equiv P_n^{\mathrm{HT}} D(P_n^*)^2 h^{-1} \tag{0.4}$$

estimates conservatively σ_1^2. Relying on the non-parametric bootstrap is not a solution either in general. The third assumption guarantees that a second-order term in an expansion of $\psi_n^* = \Psi(P_n^*)$ around P_0 is indeed of second order:

A3. There exists a real-valued random variable γ_n converging in probability to $\gamma_1 \neq 1$ and such that $\gamma_n(\psi_n^* - \psi_0) + [\psi_0 - \psi_n^* - P_0 D(P_n^*)] = o_P(1/\sqrt{n})$. Moreover, one knows an estimator Γ_n such that $\Gamma_n - \gamma_n = o_P(1)$.

The introduction of the term $\gamma_n(\psi_n^* - \psi_0)$ in **A3** gives some flexibility. This proves useful sometimes, as for instance in the example developed in Section 5. If $\gamma_n = 0$, then the main condition in **A3** reduces to the classical $\psi_0 - \psi_n^* - P_0 D(P_n^*) = o_P(1/\sqrt{n})$.

Assumptions **A1**, **A2** and **A3** are variations on the assumptions typically made in the asymptotic analysis of TMLEs. They allow to derive the following CLT, whose proof is sketched in Section 7.1. Set $\alpha \in (0, 1)$ and denote $\xi_{1-\alpha/2}$ the $(1 - \alpha/2)$-quantile of the standard normal distribution.

Proposition 0.1. *Under* **A1**, **A2** *and* **A3**, $(1 - \gamma_n)\sqrt{n}(\psi_n^* - \psi_0)$ *converges in law to the centered Gaussian distribution with variance* σ_1^2. *Consequently,*

$$\left[\psi_n^* \pm \frac{\xi_{1-\alpha/2}\sqrt{\sigma_n^2}}{(1 - \Gamma_n)\sqrt{n}}\right] \tag{0.5}$$

is a confidence interval for ψ_0 *with asymptotic coverage no less than* $(1 - \alpha)$.

Resorting to survey sampling thus makes it possible to construct a CI for ψ_0. It is through σ_n^2 that the width of CI (0.5) depends on the survey sampling design and more precisely on the covariance function in **A1**. In this regard, Sections 3 and 4 will show that all survey sampling designs are not equal. In particular, simple random sampling (selecting n among N observations without replacement and with equal weights) is suboptimal.

3 Survey sampling designs and assumption A1

This section introduces two examples of survey sampling designs and discusses **A1** in their respective contexts.

3.1 Sampford's survey sampling design

Denote $\mathcal{V}$ the space where V drawn from P_0 takes its values and let h be a (measurable) function mapping $\mathcal{V}$ to $\mathbb{R}_+$, chosen by us in such a way that h be bounded away from 0 and $P_0 h = 1$. For each $1 \leq i \leq N$, define

$$p_i \equiv \frac{nh(V_i)}{N}.$$

Let $P^{s\mathrm{P}}$ be characterized by the fact that, under $P^{s\mathrm{P}}$, η is distributed from the conditional law of $(\varepsilon_1, \ldots, \varepsilon_N)$ given $\sum_{i=1}^N \varepsilon_i = n$ when $\varepsilon_1, \ldots, \varepsilon_N$ are independently sampled from Bernoulli laws with parameters $p_1, \ldots, p_N$, respectively (we recall that this statement is conditional on $O_1, \ldots, O_N$). This survey sampling design is an instance of Sampford's survey sampling design. It is also called rejective sampling design based on Poisson (hence the superscript "P" in $P^{s\mathrm{P}}$) sampling with unequal inclusion probabilities (unequal as soon as h is not constant).

By (Bertail et al., 2016a, Theorem 2), which builds upon (Bertail et al., 2016b), assumption **A1** is met when using of Sampford's survey sampling design $P^{s\mathrm{P}}$ provided that $\mathcal{F}$, the class introduced in Section 2.2, is not too complex: this is the message of Proposition 0.2 below.

Proposition 0.2. *Assume that $\mathcal{F}$ is separable (for instance, countable), that it admits an envelope function such that the corresponding uniform entropy integral be finite (see Condition (2.1.7) in van der Vaart and Wellner, 1996), and that $P_0 f^2 h^{-1}$ is finite for all $f \in \mathcal{F}$. Then* **A1** *holds when using Sampford's survey sampling design $P^{s\mathrm{P}}$ with a covariance function Σ_h^{P} given by $\Sigma_h^{\mathrm{P}}(f, g) \equiv P_0 f g h^{-1}$.*

The conclusions of Proposition 0.2 still hold under the same conditions when substituting

$$\frac{1}{N}\sum_{i=1}^N \frac{\eta_i}{p_i}\mathrm{Dirac}(O_i) = \frac{1}{n}\sum_{i=1}^N \frac{\eta_i}{h(V_i)}\mathrm{Dirac}(O_i) \tag{0.6}$$

for P_n^{HT}. It is thus unnecessary to compute the first order inclusion probabilities of Sampford's survey sampling design, which differ from $p_1, \ldots, p_N$ when h is not constant, and the targeting of $P_n^* \in \mathcal{M}$ to ψ_0 can be achieved by ensuring $n^{-1}\sum_{i=1}^N \eta_i D(P_n^*)(O_i) h^{-1}(V_i) = o_P(1/\sqrt{n})$ instead of (0.3).

3.2 Determinantal survey sampling design

A minimalist introduction. Determinantal survey sampling designs are built upon determinantal point processes. Let K be a $N \times N$ Hermitian matrix whose eigenvalues belong to $[0, 1]$. It happens that the set of equalities: for all $I \subset \{1, \ldots, N\}$,

$$\sum_{I' \supset I} P^s(I') = \det(K_{|I}), \tag{0.7}$$

uniquely characterizes the determinantal survey sampling design P^{sK}, a probability measure on the powerset of $\{1, \ldots, N\}$. Here, $K_{|I}$ denotes the Hermitian matrix derived from K by keeping only its rows and columns indexed by the elements of I.

The first and second order inclusion probabilities of P^{sK} characterized by (0.7) are easily derived from the entries of K: for all $1 \leq i \neq j \leq N$, if holds that $\pi_i = \det(K_{|\{i\}}) = K_{ii}$ and $\pi_{ij} = \det(K_{|\{i,j\}}) = K_{ii} \times K_{jj} - |K_{ij}|^2$. Furthermore, draws from P^{sK} are of fixed size if and only if the eigenvalues of K belong to $\{0, 1\}$, in which case K is a projection matrix and the fixed size equals the trace of K. From now on, we focus on this case.

If the first order inclusion probabilities are all positive then, for any bounded function f of O drawn from P_0, $P_n^{\mathrm{HT}} f - \mathrm{E}_{P^{sK}}\left[P_n^{\mathrm{HT}} f\right]$ satisfies a concentration inequality (Pemantle and Peres, 2014, Theorem 3.1; see also (0.20) in our Section 7.3). Moreover, if f meets the so called Soshnikov conditions (0.17), (0.18) and (0.19), then $\sqrt{n}(P_n^{\mathrm{HT}} - P_N) f$ satisfies a CLT (Soshnikov, 2000). These two remarkable properties are the building blocks of Proposition 0.3 below. Let $\mathcal{F}'$ be defined as $\mathcal{F}$ deprived of its elements which depend on O through V only.

Proposition 0.3. *Assume that $\mathcal{F}'$ is countable, uniformly bounded, and that its bracketing entropy with respect to the supremum norm is finite (see the condition preceding Condition (2.1.7) in van der Vaart and Wellner, 1996). Assume moreover that, for every $f \in \mathcal{F}'$, $n \operatorname{Var}_{P^{sK}}\left[P_n^{\mathrm{HT}} f\right] > 0$ converges in P_0-probability to a positive number and f meets the Soshnikov conditions (0.17), (0.18), (0.19) P_0-almost surely. Then* **A1** *holds with $\mathcal{F}'$ substituted for $\mathcal{F}$ when using any fixed-size determinantal survey sampling design P^{sK}, provided that its first order inclusion probabilities are bounded away from 0 uniformly in N. The covariance function is defined as a limit with no closed-form expression in general.*

The message of Proposition 0.3 is the following: if the $\mathcal{F}'$ is not too complex, and if n/N goes to 0 sufficiently slowly, then **A1** is met with $\mathcal{F}'$ substituted for $\mathcal{F}$ when using most determinantal survey sampling designs P^{sK}. The proof of Proposition 0.3 is sketched in Section 7.3. The condition on the ratio n/N is included implicitly in the assumption that the elements of $\mathcal{F}'$ satisfy the Soshnikov conditions P_0-almost surely. We elaborate further on this issue in Proposition 0.4 below.

We wish to follow the same strategy as in Section 3.1, *i.e.*, to define possibly unequal first order inclusion probabilities depending on $V_1, \ldots, V_N$. There exists an algorithm to both construct and sample from a fixed-size determinantal survey sampling design with given first order inclusion probabilities (Loonis and Mary, 2015). Unfortunately, its computational burden is considerable for both tasks in general, especially in the context of large data sets (N large). In addition, the second set of conditions on $\mathcal{F}'$ (and not $\mathcal{F}$) in Proposition 0.3 would typically be very demanding for the yielded determinantal survey sampling design. Moreover, computing the limit variance of the TMLE resulting from its use would be difficult, and its inference would typically be achieved through the use of a very conservative estimator.

These difficulties can be overcome by focusing on V-stratified determinantal survey sampling designs equally weighted on each V-stratum.

V-stratified determinantal sampling equally weighted on each V-stratum. We now consider the case that V drawn from P_0 takes finitely many different values. To alleviate notation, we assume without loss of generality that $\mathcal{V} \equiv \{1, \ldots, \nu\}$ and that $O_1, \ldots, O_N$ are ordered by values of $V_1, \ldots, V_N$.

Let h be a function mapping $\mathcal{V}$ to $\mathbb{R}_+^*$ such that $P_N h = N^{-1} \sum_{i=1}^N h(V_i) = 1$. We will hide and neglect notation-wise the dependency of h on $V_1, \ldots, V_N$ due to the normalization $P_N h = 1$. In the limit, h does not depend on the summary measures anymore: by the strong law of large numbers, $P_N h$ converges P_0-almost surely to $P_0 h$, revealing that condition $P_N h = 1$ is similar to its counterpart $P_0 h = 1$ from Section 3.1. For each $1 \leq i \leq N$, define

$$\pi_i \equiv \frac{n h(V_i)}{N}.$$

Similar to the proportions $p_1, \ldots, p_N$ used in Section 3.1 to characterize a Sampford survey sampling design, $\pi_1, \ldots, \pi_N$ are the exact (as opposed to approximate) first order inclusion probabilities that we choose for our determinantal survey sampling design. Its complete characterization now boils down to elaborating a $N \times N$ Hermitian matrix Π with $\pi_1, \ldots, \pi_N$ as diagonal elements and eigenvalues in $\{0, 1\}$. Since $\sum_{i=1}^N \pi_i = n$, the resulting determinantal survey sampling design will be of fixed size n.

For simplicity, we elaborate Π under the form of a block matrix with zero matrices as off-diagonal blocks and make each of the ν diagonal blocks be a projection matrix featuring the prescribed diagonal elements. This last step is easy provided that $n_v \equiv \sum_{i=1}^N \pi_i \mathbf{1}\{V_i = v\}$ is an integer dividing $N_v \equiv \sum_{i=1}^N \mathbf{1}\{V_i = v\}$. In that case, the projection matrix can be a block matrix consisting of n_v^2 square matrices of size $N_v/n_v \times N_v/n_v$, with zero off-diagonal blocks and diagonal blocks having all their entries equal to n_v^{-1}. Otherwise, we may rely on an algorithm to derive the desired projection matrix.

The determinantal survey sampling design $P^{s\Pi}$ encoded by Π (hence the superscript "Π") is said V-stratified and equally weighted on each V-stratum. It randomly selects a deterministic number n_v of observations from the stratum where $V = v$, for each $1 \leq v \leq \nu$. Sampling from it makes it possible to derive the next result, proven in Section 7.2: for any function f of O drawn from P_0,

$$\mathrm{E}_{P_0}\left[\mathrm{Var}_{P^{s\Pi}}\left[P_n^{\mathrm{HT}} f\right]\right] = \frac{1}{n}\mathrm{E}_{P_0}\left[\mathrm{Var}_{P_0}\left[f(O)|V\right] h^{-1}(V)\right] - \frac{1}{N}\mathrm{E}_{P_0}\left[\mathrm{Var}_{P_0}\left[f(O)|V\right]\right]. \tag{0.8}$$

Equality (0.8) is instrumental in deriving the following corollary to Proposition 0.3, whose proof is sketched in Section 7.3.

Proposition 0.4. *Let us impose that n is chosen in such a way that $N/n = o((N^2/n)^\epsilon)$ for all $\epsilon > 0$. This is the case if $n \equiv N/\log^a(N)$ for some $a > 0$, for instance. Assume that $\mathcal{F}$ is separable (for instance, countable) and that its bracketing entropy with respect to the supremum norm is finite (see the condition preceding Condition (2.1.7) in van der Vaart and Wellner, 1996). Then* **A1** *holds when using the V-stratified and equally weighted on each V-stratum determinantal survey sampling design $P^{s\Pi}$ with a covariance function Σ_h^{Π} given by $\Sigma_h^{\Pi}(f,g) = \mathrm{E}_{P_0}\left[\mathrm{Cov}_{P_0}\left[f(O)g(O)|V\right] h^{-1}(V)\right]$.*

Note that $\Sigma_h^{\Pi}(f,f) = 0$ for every $f \in \mathcal{F}$ which depends on O through V only. Actually, for such a function, $\sqrt{n}(P_n^{\mathrm{HT}} - P_0)f = \sqrt{n}(P_N - P_0)f = O_P(\sqrt{n/N}) = o_P(1)$. Moreover, for every $f \in \mathcal{F}$, combining (0.8) and equality $\mathrm{E}_{P^{s\Pi}}\left[P_n^{\mathrm{HT}} f\right] = P_N f$ readily implies

$$\mathrm{Var}_{P_0 P^{s\Pi}}\left[\sqrt{n}(P_n^{\mathrm{HT}} - P_0)f\right] = \Sigma_h^{\Pi}(f,f) + \frac{n}{N}\Big(\mathrm{Var}_{P_0}\left[f(O)\right] - \mathrm{E}_{P_0}\left[\mathrm{Var}_{P_0}\left[f(O)|V\right]\right]\Big). \tag{0.9}$$

Proved in Section 7.2, (0.9) relates the exact variance of $\sqrt{n}(P_n^{\mathrm{HT}} - P_0)f$ with the limit variance $\Sigma_h^{\Pi}(f,f)$, showing that their difference is upper-bounded by a $O(n/N) = o(1)$-expression.

It is an open question to determine whether or not the extra condition on how n depends on N could be relaxed or even given up by proving directly a functional CLT for $\sqrt{n}(P_n^{\mathrm{HT}} - P_0)$. By "directly", we mean without building up on functional CLTs conditional on the observations, and managing to go around the Soshnikov conditions. This route was followed to prove (Bertail et al., 2016a, Theorem 2).

Sobolev classes are known to have finite bracketing entropy with respect to the supremum norm (van der Vaart, 1998, Example 19.10). The fact that the bracketing entropy is meant relative to the supremum norm instead of the $L^2(P_0)$-norm is a little frustrating, though. Indeed, a bracketing entropy condition relative to the latter would have allowed a larger variety of classes. The supremum norm comes from the concentration inequality (Pemantle and Peres, 2014, Theorem 3.1). Perhaps the aforementioned direct proof might also allow to replace it with the $L^2(P_0)$-norm.

The covariance functions Σ_h^{P} and Σ_h^{Π} in Propositions 0.2 and Proposition 0.4 differ. In particular, for every $f \in \mathcal{F}$,

$$\Sigma_h^{\mathrm{P}}(f,f) = \mathrm{E}_{P_0}\left[\mathrm{E}_{P_0}\left[f^2(O)|V\right] h^{-1}(V)\right] \geq \mathrm{E}_{P_0}\left[\mathrm{Var}_{P_0}\left[f(O)|V\right] h^{-1}(V)\right] = \Sigma_h^{\Pi}(f,f) \tag{0.10}$$

(using the same h on both sides of (0.10) is allowed because, in the limit, condition $P_N h = 1$ is similar to condition $P_0 h = 1$). Consequently, $P^{s\Pi}$ is more efficient than $P^{s\mathrm{P}}$ when $\mathcal{V}$ is finite in the sense that whichever function h^{P} is used to define $P^{s\mathrm{P}}$, it is always possible to choose function h^{Π} to define $P^{s\Pi}$ in such a way that $\Sigma_{h^{\mathrm{P}}}^{\mathrm{P}}(f,f) \geq \Sigma_{h^{\Pi}}^{\Pi}(f,f)$ for every $f \in \mathcal{F}$.

4 Optimizing the survey sampling designs

This section discusses the optimization of functions h^{P} and h^{Π} used to define the first order inclusion probabilities of the survey sampling designs $P^{s\mathrm{P}}$ and $P^{s\Pi}$ that we developed in Sections 3.1 and 3.2. The optimization is relative to the asymptotic variance of the TMLE, $\Sigma_h^{\mathrm{P}}(f_1,f_1)$ or $\Sigma_h^{\Pi}(f_1,f_1)$, respectively.

In light of (0.10), let f_2^{P} and f_2^{Π} be the functions from $\mathcal{V}$ to $\mathbb{R}_+$ given by

$$f_2^{\mathrm{P}}(V) \equiv \sqrt{\mathrm{E}_{P_0}\left[f_1^2(O)|V\right]} \quad \text{and} \quad f_2^{\Pi}(V) \equiv \sqrt{\mathrm{Var}_{P_0}\left[f_1(O)|V\right]}. \tag{0.11}$$

Then (0.10) shows in particular that $\Sigma_h^{\mathrm{P}}(f_1,f_1) = P_0(f_2^{\mathrm{P}})^2 h^{-1}$ is always larger than $\Sigma_h^{\Pi}(f_1,f_1) = P_0(f_2^{\Pi})^2 h^{-1}$. Now, with f_2 equal to either f_2^{P} or f_2^{Π}, the Cauchy-Schwarz inequality yields that

$$P_0(f_2)^2 h^{-1} \times P_0 h \geq (P_0 f_2)^2 ,$$

where equality holds if and only if h is proportional to f_2.

In the case of $P^{s\mathrm{P}}$, h satisfies $P_0 h = 1$. Therefore, the optimal h and corresponding optimal asymptotic variance of the TMLE are

$$h^{\mathrm{P}} \equiv f_2^{\mathrm{P}} / P_0 f_2^{\mathrm{P}} \quad \text{and} \quad \Sigma^{\mathrm{P}}_{h^{\mathrm{P}}}(f_1, f_1) = \left(P_0 f_2^{\mathrm{P}}\right)^2 . \tag{0.12}$$

In the case of $P^{s\Pi}$, h satisfies $P_N h = 1$ and $P_0 h = 1$ in the limit. By analogy with (0.12), the optimal h and corresponding optimal asymptotic variance of the TMLE are

$$h^{\Pi} \equiv f_2^{\Pi} / P_0 f_2^{\Pi} \quad \text{and} \quad \Sigma^{\Pi}_{h^{\Pi}}(f_1, f_1) = \left(P_0 f_2^{\Pi}\right)^2 . \tag{0.13}$$

5 Example: variable importance measure of a continuous exposure

We illustrate our template for survey sampling targeted learning with the inference of a variable importance measure of a continuous exposure. In this example, the ith observation O_i writes $(W_i, A_i, Y_i) \in \mathcal{O} \equiv \mathcal{W} \times \mathcal{A} \times [0,1]$. Here, $W_i \in \mathcal{W}$ is the ith context, $A_i \in \mathcal{A}$ is the ith exposure and $Y_i \in [0,1]$ is the ith outcome. Exposures take their values in $\mathcal{A} \ni 0$, a bounded subset of $\mathbb{R}$ containing 0, which serves as a reference level of exposure. Typically, in biostatistics or epidemiology, W_i could be the baseline covariate describing the ith subject, A_i could describe her assignment (*e.g.*, a dose-level) or her level of exposure, and Y_i could quantify her biological response.

Section 5.1 presents and analyzes the parameter of interest. Section 5.2 discusses the construction of the corresponding TMLE.

5.1 Preliminaries

For each finite measure P on $\mathcal{O}$ equipped with the Borel σ-field, we denote P_W, $P_{A|W}$ and $P_{Y|A,W}$ the marginal measures of W and conditional measures of A and Y given W and (A, W), respectively. (The conditional measure $P_{A|W}$ is $P(\mathcal{O})$ times the conditional distribution of A given W under the probability distribution $P/P(\mathcal{O})$. The conditional measure $P_{Y|A,W}$ is defined analogously.) Moreover, for each $(w, a) \in \mathcal{W} \times \mathcal{A}$, we introduce and denote $g_P(0|w) \equiv P_{A|W=w}(\{0\})$ and $Q_P(a, w) \equiv \int_{[0,1]} y dP_{Y|A=a,W=w}(y)$. In particular if $P(\mathcal{O}) = 1$, then $g_P(0|W) = P(A = 0|W)$ is the conditional probability that the exposure equal the reference value 0 and $Q_P(A, W) = \mathrm{E}_P[Y|A, W]$ is the conditional expectation of the response given exposure and context.

We assume that $P_{0,A|W}(A \neq 0|W) > 0$ $P_{0,W}$-almost surely and the existence of a constant $c(P_0) > 0$ such that $g_P(0|W) \geq c(P_0)$ $P_{0,W}$-almost surely. Introduced in (Chambaz et al., 2012; Chambaz and Neuvial, 2015), the true parameter of interest is

$$\begin{aligned} \psi_0^c &\equiv \arg\min_{\beta \in \mathbb{R}} \mathrm{E}_{P_0}\left[\left(Y - \mathrm{E}_{P_0}[Y|A = 0, W] - \beta A\right)^2\right] \\ &= \arg\min_{\beta \in \mathbb{R}} \mathrm{E}_{P_0}\left[\left(\mathrm{E}_{P_0}[Y|A, W] - \mathrm{E}_{P_0}[Y|A = 0, W] - \beta A\right)^2\right] \end{aligned}$$

(the superscript "c" stands for "continuous").

In the context of this example, $\mathcal{M}$ stands for the set of finite measures P on $\mathcal{O}$ equipped with the Borel σ-field such that there exists a constant $c(P) > 0$ guaranteeing that the marginal measure of $\{w \in \mathcal{W} : P_{A|W=w}(\mathcal{A} \setminus \{0\}) > 0 \text{ and } P_{A|W=w}(\{0\}) \geq c(P)\}$ under P_W equals $P(\mathcal{O})$. In particular, $P_0 \in \mathcal{M}$ by the above assumption. We see ψ_0^c as the value at P_0 of the functional Ψ^c characterized over $\mathcal{M}$ by

$$\Psi^c(P) \equiv \arg\min_{\beta \in \mathbb{R}} \int_{\mathcal{A} \times \mathcal{W}} \left(Q_P(a, w) - Q_P(0, w) - \beta a\right)^2 dP_{A|W=w}(a) dP_W(w).$$

By (Chambaz et al., 2012, Proposition 1), for each $P \in \mathcal{M}$,

$$\Psi^c(P) = \frac{\int_{\mathcal{A}\times\mathcal{W}} a(Q_P(a,w) - Q_P(0,w))dP_{A|W=w}(a)dP_W(w)}{\int_{\mathcal{A}\times\mathcal{W}} a^2 dP_{A|W=w}(a)dP_W(w)}.$$

If P is a *distribution*, then

$$\Psi^c(P) = \frac{\mathrm{E}_P\left[A(Q_P(A,W) - Q_P(0,W))\right]}{\mathrm{E}_P\left[A^2\right]}.$$

For clarity, we define $\mu_P(w) \equiv \int_{\mathcal{A}} a dP_{A|W=w}(a)$ and $\zeta^2(P) \equiv \int_{\mathcal{A}\times\mathcal{W}} a^2 dP_{A|W=w}(a)dP_W(w)$ for all $P \in \mathcal{M}$, $(w,a) \in \mathcal{W}\times\mathcal{A}$. If $P(\mathcal{O}) = 1$, then $\mu_P(W) = \mathrm{E}_P[A|W]$ and $\zeta^2(P) = \mathrm{E}_P\left[A^2\right]$. Adapting (Chambaz et al., 2012, Proposition 1) yields that Ψ^c is pathwise differentiable with influence curve $D^c(P) \equiv D_1^c(P) + D_2^c(P) \in L_0^2(P)$,

$$\begin{aligned}
\zeta^2(P)D_1^c(P)(O) &\equiv A\left(Q_P(A,W) - Q_P(0,W) - A\Psi^c(P)\right),\\
\zeta^2(P)D_2^c(P)(O) &\equiv (Y - Q_P(A,W))\left(A - \frac{\mu_P(W)\mathbf{1}\{A=0\}}{g_P(0|W)}\right)
\end{aligned}$$

(all $P \in \mathcal{M}$). Let now $\mathcal{R}^c : \mathcal{M}^2 \to \mathbb{R}$ be given by

$$\begin{aligned}
\mathcal{R}^c(P,P') &\equiv \Psi^c(P') - \Psi^c(P) - (P'-P)D^c(P)\\
&\equiv \Psi^c(P') - \Psi^c(P) - P'D^c(P).
\end{aligned}$$

In light of **A3**, $\psi_0 - \psi_n^* - P_0 D^c(P_n^*) = \mathcal{R}^c(P_n^*, P_0)$. Adapting the last step of the proof of (Chambaz et al., 2012, Proposition 1) yields that, for every $P, P' \in \mathcal{M}$,

$$\begin{aligned}
\mathcal{R}^c(P,P') = &\left(1 - \frac{\zeta^2(P')}{\zeta^2(P)}\right)(\Psi^c(P') - \Psi^c(P))\\
&+ \frac{1}{\zeta^2(P)} P'\left((Q_{P'}(0,\cdot) - Q_P(0,\cdot))\left(\mu_{P'} - \mu_P \frac{g_{P'}(0|\cdot)}{g_P(0|\cdot)}\right)\right). \qquad (0.14)
\end{aligned}$$

We will use this equality to derive an easy to interpret sufficient condition for **A3** to hold.

5.2 Construction of the TMLE

Let $\mathcal{Q}^w$, $\mathcal{M}^w$ and $\mathcal{G}^w$ be three user-supplied classes of functions mapping $\mathcal{A}\times\mathcal{W}$, $\mathcal{W}$ and $\mathcal{W}$ to $[0,1]$, respectively. We first estimate Q_{P_0}, μ_{P_0} and g_{P_0} with Q_n and μ_n and g_n built upon P_n^{HT}, $\mathcal{Q}^w$, $\mathcal{M}^w$ and $\mathcal{G}^w$. For instance, one could simply minimize (weighted) empirical risks and define

$$\begin{aligned}
Q_n \equiv \operatorname*{argmin}_{Q\in\mathcal{Q}^w} P_n^{\mathrm{HT}}\ell(Y, Q(A,W)), \qquad \mu_n \equiv \operatorname*{argmin}_{\mu\in\mathcal{M}^w} P_n^{\mathrm{HT}}\ell(A,\mu(W)),\\
g_n \equiv \operatorname*{argmin}_{g\in\mathcal{G}^w} P_n^{\mathrm{HT}}\ell(\mathbf{1}\{A=0\}, g(0|W))
\end{aligned}$$

(assuming that the argmins exist). Alternatively, one could prefer minimizing cross-validated (weighted) empirical risks. One then should keep in mind that the observations are dependent, because of the selection process by survey sampling. We also estimate the marginal distribution $P_{0,W}$ of W under P_0 with $P_{n,W}^{\mathrm{HT}}$, defined as in (0.1) with W_i substituted for O_i, and the real-valued parameter $\zeta^2(P_0)$ with $\zeta^2(P_{n,X}^{\mathrm{HT}})$ where $P_{n,X}^{\mathrm{HT}}$ is defined as in (0.1) with X_i substituted for O_i.

Let P_n^0 be a measure such that $Q_{P_n^0} = Q_n$, $\mu_{P_n^0} = \mu_n$, $g_{P_n^0} = g_n$, $\zeta^2(P_n^0) = \zeta^2(P_{n,X}^{\mathrm{HT}})$, $P_{n,W}^0 = P_{n,W}^{\mathrm{HT}}$, *and from which we can sample A conditionally on W*. Picking up such a P_n^0 is an easy technical task,

see (Chambaz et al., 2012, Lemma 5) for a computationally efficient choice. Then the initial estimator $\Psi^b(P_n^0)$ of ψ_0^b can be computed with high accuracy by Monte-Carlo. It suffices to sample a large number B (say $B = 10^7$) of independent $(A^{(b)}, W^{(b)})$ by *(i)* sampling $W^{(b)}$ from $P_{n,W}^0 = P_{n,W}^{\mathrm{HT}}$ then *(ii)* sampling $A^{(b)}$ from the conditional distribution of A given $W = W^{(b)}$ under P_n^0 repeatedly for $b = 1, \ldots, B$ and to make the approximation

$$\Psi^c(P_n^0) \approx \frac{B^{-1}\sum_{b=1}^B A^{(b)}(Q_n(A^{(b)}, W^{(b)}) - Q_n(0, W^{(b)}))}{\zeta^2(P_n^0)}.$$

We now target the inference procedure and bend P_n^0 into P_n^* satisfying (0.3) with D^c substituted for D. We proceed iteratively. Suppose that P_n^k has been constructed for some $k \geq 0$. We fluctuate P_n^k with the one-dimensional parametric model $\{P_n^k(\epsilon) : \epsilon \in \mathbb{R}, \epsilon^2 \leq c(P_n^k)/\|D^c(P_n^k)\|_\infty\}$ characterized by $dP_n^k(\epsilon)/dP_n^k = 1 + \epsilon D^c(P_n^k)$. Lemma 1 in (Chambaz et al., 2012) shows how $Q_{P_n^k(\epsilon)}$, $\mu_{P_n^k(\epsilon)}$, $g_{P_n^k(\epsilon)}$, $\zeta^2(P_n^k(\epsilon))$ and $P_{n,W}^k(\epsilon)$ depart from their counterparts at $\epsilon = 0$. The optimal move along the fluctuation is indexed by

$$\epsilon_n^k \equiv \arg\max_\epsilon P_n^{\mathrm{HT}} \log\left(1 + \epsilon D^c(P_n^k)\right),$$

i.e., the maximum likelihood estimator of ϵ. Note that the random function $\epsilon \mapsto P_n^{\mathrm{HT}} \log(1 + \epsilon D^c(P_n^k))$ is strictly concave. The optimal move results in the $(k+1)$-th update of P_n^0, $P_n^{k+1} \equiv P_n^k(\epsilon_n^k)$.

There is no guarantee that a P_n^{k+1} will coincide with its predecessor P_n^k. We assume that the iterative updating procedure converges (in k) in the sense that, for k_n large enough, $P_n^{\mathrm{HT}} D^c(P_n^{k_n}) = o_P(1/\sqrt{n})$. We set $P_n^* \equiv P_n^{k_n}$. It is actually possible to come up with a one-step updating procedure (*i.e.*, an updating procedure such that $P_n^k = P_n^{k+1}$ for all $k \geq 1$) by relying on universally least favorable models (van der Laan, 2016). We adopt this multi-step updating procedure for simplicity.

We can assume without loss of generality that we can sample A conditionally on W from P_n^*. The final estimator is computed with high accuracy like $\Psi^c(P_n^0)$ previously: with $Q_n^* \equiv Q_{P_n^*}$, we sample B independent $(A^{(b)}, W^{(b)})$ by *(i)* sampling $W^{(b)}$ from $P_{n,W}^*$ then *(ii)* sampling $A^{(b)}$ from the conditional distribution of A given $W = W^{(b)}$ under P_n^* repeatedly for $b = 1, \ldots, B$ and make the approximation

$$\psi_n^* \equiv \Psi^c(P_n^*) \approx \frac{B^{-1}\sum_{b=1}^B A^{(b)}(Q_n^*(A^{(b)}, W^{(b)}) - Q_n^*(0, W^{(b)}))}{\zeta^2(P_n^*)}.$$

To conclude this section, let us use (0.14) to derive an easy to interpret sufficient condition for **A3** to hold. If we introduce

$$\gamma_n \equiv 1 - \frac{\zeta^2(P_0)}{\zeta^2(P_n^*)} \quad \text{and} \quad \Gamma_n \equiv 1 - \frac{\zeta_n^2(P_0)}{\zeta_n^2(P_n^*)}$$

where $\zeta_n^2(P_0)$ and $\zeta_n^2(P_n^*)$ estimate $\zeta^2(P_0)$ and $\zeta^2(P_n^*)$, then **A3** is met provided that $\zeta^2(P_n^*)$ converge in probability to a finite real number such that $\gamma_1 \neq 1$ and

$$\frac{1}{\zeta^2(P_n^*)} P_0\left((Q_{P_0}(0,\cdot) - Q_{P_n^*}(0,\cdot))\left(\mu_{P_0} - \mu_{P_n^*}\frac{g_{P_0}(0|\cdot)}{g_{P_n^*}(0|\cdot)}\right)\right) = o_P(1/\sqrt{n}).$$

Through the product, we draw advantage of the synergistic convergences of $Q_{P_n^*}(0,\cdot)$ to $Q_{P_0}(0,\cdot)$ and $(\mu_{P_n^*}, g_{P_n^*})$ to (μ_{P_0}, g_{P_0}) (by the Cauchy-Schwarz inequality for example).

The next section illustrates further this example of application of our survey sampling targeted learning methodology with a simulation study.

6 Simulation study

Section 6.1 presents the setting of the simulation study, which illustrates the example developed in Section 5 in three related but different scenarios. Section 6.2 summarizes its results and comments on them.

6.1 Setting

We consider three data-generating distributions $P_{0,1}$, $P_{0,2}$ and $P_{0,3}$ of a data-structure $O = (W, A, Y)$. The three distributions differ only in terms of the conditional mean and variance of Y given (A, W). Specifically, $O = (W, A, Y)$ drawn from $P_{0,j}$ $(j = 1, 2, 3)$ is such that

- $W \equiv (V, W_1, W_2)$ with $P_{0,j}(V = 1) = 1/6$, $P_{0,j}(V = 2) = 1/3$, $P_{0,j}(V = 3) = 1/2$ and, conditionally on V, (W_1, W_2) is a Gaussian random vector with mean $(0, 0)$ and variance $\left(\begin{smallmatrix} 1 & -0.2 \\ -0.2 & 1 \end{smallmatrix}\right)$ (if $V = 1$), $(1, 1/2)$ and $\left(\begin{smallmatrix} 0.5 & 0.1 \\ 0.1 & 0.5 \end{smallmatrix}\right)$ (if $V = 2$), $(1/2, 1)$ and $\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ (if $V = 3$);
- conditionally on W, $A = 0$ with probability 80% if $W_1 \geq 1.1$ and $W_2 \geq 0.8$ and 10% otherwise; moreover, conditionally on W and $A \neq 0$, $3A - 1$ is drawn from the χ^2-distribution with 1 degree of freedom and non-centrality parameter $\sqrt{(W_1 - 1.1)^2 + (W_2 - 0.8)^2}$;
- conditionally on (W, A), Y is a Gaussian random variable with mean
  - $A(W_1 + W_2)/6 + W_1 + W_2/4 + \exp((W_1 + W_2)/10)$ for $j = 1, 2$,
  - $A(W_1 + W_2)/6 + W_1 + W_2/4 + \exp((W_1 + W_2)/10) + 3AV$ for $j = 3$,

  and standard deviation
  - 2 (if $V = 1$), 1.5 (if $V = 2$) and 1 (if $V = 3$) for $j = 1$,
  - 9 (if $V = 1$), 4 (if $V = 2$) and 1 (if $V = 3$) for $j = 2, 3$.

The true parameters equal approximately $\Psi^c(P_{0,1}) = \Psi^c(P_{0,2}) = 0.1201$ and $\Psi^c(P_{0,3}) = 6.9456$.

For $B = 10^3$ and each $j = 1, 2, 3$, we repeat independently the following steps:

1. simulate a data set of $N = 10^7$ independent observations drawn from $P_{0,j}$;
2. extract $n_0 \equiv 10^3$ observations from the data set by simple random sampling (SRS, which is identical to $P^{s\mathrm{P}}$ with $h_0 \equiv 1$), and based on these observations:
   a) apply the procedure described in Section 5 and retrieve $f_{n_0,1} \equiv D^c(P_{n_0}^{k_{n_0}})$;
   b) regress $f_{n_0,1}(O)$ and $f_{n_0,1}(O)^2$ on V, call $f_{n_0,2}^{\mathrm{P}}$ the square root of the resulting estimate of f_2^{P} and $f_{n_0,2}^{\Pi}$ the square root of the resulting estimate of f_2^{Π}, see (0.11);
   c) estimate the marginal distribution of V, estimate $P_0 f_{n_0,2}^{\mathrm{P}}$ with $\pi_{n_0,2}$ and set $h_{n_0}^{\mathrm{P}} \equiv f_{n_0,2}^{\mathrm{P}}/\pi_{n_0,2}$, $h_{n_0}^{\Pi} \equiv f_{n_0,2}^{\Pi}/P_N f_{n_0,2}^{\Pi}$, see (0.12) and (0.13);
3. for each n in $\{10^3, 5 \times 10^3, 10^4\}$, successively, and for each survey sampling design among SRS, $P^{s\mathrm{P}}$ with $h_{n_0}^{\mathrm{P}}$ and $P^{s\Pi}$ with $P_{n_0}^{\Pi}$, extract a sub-sample of n observations from the data set (deprived of the observations extracted in step 2) and, based on these observations, apply the procedure described in Section 5. We use $\hat{\sigma}_n^2$ given in (0.4) to estimate σ_1^2, although we are not sure in advance that it is a conservative estimator.

We thus obtain $27 \times B$ estimates and their respective CIs.

To give an idea of what are $h_{n_0}^{\mathrm{P}}$ and $h_{n_0}^{\Pi}$ in each scenario, we report their averages across the B simulation studies under $P_{0,1}$, $P_{0,2}$ and $P_{0,3}$:

- under $P_{0,1}$, we expect similar $h_{n_0}^{\mathrm{P}}$ and $h_{n_0}^{\Pi}$, and do get that they are approximately equal (on average) to $(h_1(1), h_1(2), h_1(3)) \approx (2.10, 0.83, 0.75)$;
- under $P_{0,2}$, we also expect similar $h_{n_0}^{\mathrm{P}}$ and $h_{n_0}^{\Pi}$, and do get that they are approximately equal (on average) to $(h_1(1), h_1(2), h_1(3)) \approx (3.39, 0.83, 0.32)$;
- under $P_{0,3}$, we do not expect similar $h_{n_0}^{\mathrm{P}}$ and $h_{n_0}^{\Pi}$, and get that they are approximately equal (on average) to $(h_1(1), h_1(2), h_1(3)) \approx (2.93, 0.66, 0.58)$ and $(h_1(1), h_1(2), h_1(3)) \approx (2.97, 0.68, 0.56)$, respectively (although small, the differences are significant).

Applying the TMLE procedure is straightforward thanks to the `R` package called `tmle.npvi` (Chambaz and Neuvial, 2016, 2015). Note, however, that it is necessary to compute Γ_n and $\hat{\sigma}_n^2$. Specifically, we fine-tune the TMLE procedure by setting `iter` (the maximum number of iterations of the targeting step) to 7 and `stoppingCriteria` to `list(mic=0.01, div=0.001, psi=0.05)`. Moreover, we use the default `flavor` called

`"learning"`, thus notably rely on parametric linear models for the estimation of the infinite-dimensional parameters Q_{P_0}, μ_{P_0} and g_{P_0} and their fluctuation. We refer the interested reader to the package's manual and vignette for details.

The Sampford sampling method (Sampford, 1967) implements $P^{s\mathrm{P}}$. However, when the ratio n/N is close to 0 or 1 (here, when n/N differs from 10^{-3}), this acceptance-rejection algorithm typically takes too much time to succeed. To circumvent the issue, we approximate $P^{s\mathrm{P}}$ with a Pareto sampling (see Algorithm 2 in Bondesson et al., 2006, Section 5). We implement $P^{s\Pi}$ as described in Section 3.2, with minor changes to account for the fact that for some $1 \leq v \leq 3$, $\sum_{i=1}^{N} K_{ii}\mathbf{1}\{V_i = v\}$ may not be an integer or may not divide $\sum_{i=1}^{N} \mathbf{1}\{V_i = v\}$.

6.2 Results

The results are summarized in Table 0.1. We focus on the empirical coverage, empirical variance and mean of the estimated variance of the TMLE.

		SRS			$P^{s\mathrm{P}}$ with $h^{\mathrm{P}}_{n_0}$			$P^{s\Pi}$ with $h^{\Pi}_{n_0}$		
		n			n			n		
		10^3	5×10^3	10^4	10^3	5×10^3	10^4	10^3	5×10^3	10^4
$P_{0,1}$	empirical coverage	96.2%	98.9%	99.2%	98.1%	98.6%	99.4%	97.8%	99.2%	99.3%
	empirical variance	09	08	07	07	06	06	07	06	06
	estimated variance	13	14	14	11	11	11	11	11	11
$P_{0,2}$	empirical coverage	94.0%	98.9%	99.2%	98.9%	99.9%	99.1%	98.4%	99.4%	99.3%
	empirical variance	129	104	102	44	41	44	49	42	42
	estimated variance	171	196	200	85	86	87	86	85	86
$P_{0,3}$	empirical coverage	95.6%	98.8%	97.8%	97.8%	97.9%	97.1%	98.5%	98.3%	96.3%
	empirical variance	157	134	168	85	91	116	81	85	104
	estimated variance	216	242	245	130	133	135	124	128	127

Table 0.1. Summarizing the results of the simulation study. The top, middle and bottom groups of rows correspond to simulations under $P_{0,1}$, $P_{0,2}$ and $P_{0,3}$. Each of them reports the empirical coverage of the CIs ($B^{-1}\sum_{b=1}^{B}\mathbf{1}\{\Psi^c(P_{0,j}) \in I_{n,b}\}$), n times the empirical variance of the estimators ($n[B^{-1}\sum_{b=1}^{B}\psi^{*2}_{n,b} - (B^{-1}\sum_{b=1}^{B}\psi^{*}_{n,b})^2]$) and empirical mean of n times the estimated variance of the estimators ($B^{-1}\sum_{b=1}^{B}\hat{\sigma}^2_{n,b}$), for every sub-sample size n and for each survey sampling design.

All empirical coverages are larger than 95% but one (equal to 94%). In each case, the mean of estimated variances is larger than the corresponding empirical variance, revealing that we achieve the conservative estimation of σ_1^2. Regarding the variances, we observe that $P^{s\mathrm{P}}$ and $P^{s\Pi}$ perform similarly and provide slightly better results than SRS under $P_{0,1}$. This is in line with what was expected, due to the contrast induced by the conditional standard deviation of Y given (A, W) under $P_{0,1}$. Under $P_{0,2}$, we observe that $P^{s\mathrm{P}}$ and $P^{s\Pi}$ perform similarly and provide significantly better results than SRS. This too is in line with what was expected, due to the contrast induced by the conditional standard deviation of Y given (A, W), which is stronger under $P_{0,2}$ than under $P_{0,1}$. Finally, under $P_{0,3}$, we observe that $P^{s\mathrm{P}}$ performs better than SRS and that $P^{s\Pi}$ performs even slightly better than $P^{s\mathrm{P}}$. This again is in line with what was expected, due to the contrast induced by the conditional standard deviation of Y given (A, W) (same as under $P_{0,2}$) and to the different conditional means of Y given (A, W) under $P_{0,3}$ and $P_{0,2}$.

Acknowledgements.

The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR), under grant ANR-13-BS01-0005 (project SPADRO).

7 Elements of proof

For every $f \in \mathcal{F}$, let $\bar{f}$, $\overline{f^2}$ be given by $\bar{f}(V) \equiv \mathrm{E}_{P_0}[f(O)|V]$, $\overline{f^2}(V) \equiv \mathrm{E}_{P_0}\left[f^2(O)|V\right]$. Note that $\overline{f^2}(V) - \bar{f}^2(V) = \mathrm{Var}_{P_0}[f(O)|V]$.

For every $1 \leq v \leq \nu$, let $\ell_1, \ldots, \ell_\nu$ and $I_1, \ldots, I_\nu$ be given by $\ell_v(V) \equiv \mathbf{1}\{V = v\}$ and $I_v \equiv \{1 \leq i \leq N : V_i = v\}$.

7.1 Proof of Proposition 0.1

Combining (0.3) and **A3** yields that

$$\begin{aligned}(1-\gamma_n)\sqrt{n}(\psi_n^* - \psi_0) &= \sqrt{n}(P_n^{\mathrm{HT}} - P_0)D(P_n^*) + o_P(1) \\ &= \sqrt{n}(P_n^{\mathrm{HT}} - P_0)f_1 + \sqrt{n}(P_n^{\mathrm{HT}} - P_0)(D(P_n^*) - f_1) + o_P(1),\end{aligned}$$

where $f_1 \in \mathcal{F}$ is introduced in **A2**. By **A1**, the first RHS term in the above equation converges in distribution to the centered Gaussian distribution with variance σ_1^2. Moreover, by a classical argument of empirical processes theory (van der Vaart, 1998, Lemma 19.24), **A1** and the convergence of $D(P_n^*)$ to f_1 in **A2** imply that the second RHS term converges to 0 in probability. This completes the sketch of proof.

7.2 Proof of Eq. (0.8) and (0.9)

By construction of $P^{s\Pi}$, the number of observations sampled from each V-stratum is deterministic. In other words, it holds for each $1 \leq v \leq \nu$ that $\mathrm{Var}_{P^{s\Pi}}\left[P_n^{\mathrm{HT}}\ell_v\right] = 0$. In light of (0.2), this is equivalent to

$$\sum_{i \in I_v}\left(\frac{1}{\Pi_{ii}} - 1\right) = \sum_{i \neq j \in I_v}\frac{|\Pi_{ij}|^2}{\Pi_{ii}\Pi_{jj}} \tag{0.15}$$

for each $1 \leq v \leq \nu$.

Now, since $V_i \neq V_j$ implies $\Pi_{ij} = 0$ by construction, (0.2) rewrites

$$\begin{aligned}N^2 \mathrm{Var}_{P^{s\Pi}}\left[P_n^{\mathrm{HT}} f\right] &= \sum_{i=1}^N \left(\frac{1}{\Pi_{ii}} - 1\right) f^2(O_i) - \sum_{1 \leq i \neq j \leq N} |\Pi_{ij}|^2 \frac{f(O_i)}{\Pi_{ii}}\frac{f(O_j)}{\Pi_{jj}} \\ &= \sum_{v=1}^\nu \sum_{i \in I_v}\left(\frac{1}{\Pi_{ii}} - 1\right) f^2(O_i) - \sum_{v=1}^\nu \sum_{i \neq j \in I_v}\frac{|\Pi_{ij}|^2}{\Pi_{ii}\Pi_{jj}} f(O_i)f(O_j).\end{aligned}$$

Because $O_1, \ldots, O_N$ are conditionally independent given $(V_1, \ldots, V_N)$ and since each factor $|\Pi_{ij}|^2/\Pi_{ii}\Pi_{jj}$ is deterministic given $i, j \in I_v$, the previous equality and (0.15) then imply

$$\begin{aligned}N^2 \mathrm{E}_{P_0}\left[\mathrm{Var}_{P^{s\Pi}}\left[P_n^{\mathrm{HT}} f\right]\right] &= \mathrm{E}_{P_0}\left[\sum_{v=1}^\nu \overline{f^2}(v) \sum_{i \in I_v}\left(\frac{1}{\Pi_{ii}} - 1\right) - \sum_{v=1}^\nu \bar{f}^2(v) \sum_{i \neq j \in I_v}\frac{|\Pi_{ij}|^2}{\Pi_{ii}\Pi_{jj}}\right] \\ &= \sum_{v=1}^\nu \left(\overline{f^2}(v) - \bar{f}^2(v)\right) \mathrm{E}_{P_0}\left[\sum_{i \in I_v}\left(\frac{1}{\Pi_{ii}} - 1\right)\right]. \end{aligned} \tag{0.16}$$

For each $1 \leq v \leq \nu$,

$$\mathrm{E}_{P_0}\left[\sum_{i \in I_v}\left(\frac{1}{\Pi_{ii}} - 1\right)\right] = \left(\frac{N}{nh(v)} - 1\right)\mathrm{E}_{P_0}\left[\mathrm{card}(I_v)\right] = \left(\frac{N}{nh(v)} - 1\right) N P_0(V = v).$$

Therefore, (0.16) yields

$$\begin{aligned}\mathrm{E}_{P_0}\left[\mathrm{Var}_{P^{s\Pi}}\left[P_n^{\mathrm{HT}} f\right]\right] &= \frac{1}{n}\sum_{v=1}^{\nu}\left(\overline{f^2}(v) - \bar{f}^2(v)\right) h^{-1}(v) P_0(V=v) - \frac{1}{N}\sum_{v=1}^{\nu}\left(\overline{f^2}(v) - \bar{f}^2(v)\right) P_0(V=v) \\ &= \frac{1}{n}\mathrm{E}_{P_0}\left[\mathrm{Var}_{P_0}\left[f(O)|V\right] h^{-1}(V)\right] - \frac{1}{N}\mathrm{E}_{P_0}\left[\mathrm{Var}_{P_0}\left[f(O)|V\right]\right],\end{aligned}$$

as stated in (0.8).

We now turn to (0.9). Since $\mathrm{E}_{P^{s\Pi}}\left[P_n^{\mathrm{HT}} f\right] = P_N f$, it holds that

$$\begin{aligned}\mathrm{Var}_{P_0 P^{s\Pi}}\left[\sqrt{n}(P_n^{\mathrm{HT}} - P_0) f\right] &= n\mathrm{E}_{P_0}\left[\mathrm{E}_{P^{s\Pi}}\left[(P_n^{\mathrm{HT}} f)^2\right]\right] - n\left(\mathrm{E}_{P_0 P^{s\Pi}}\left[P_n^{\mathrm{HT}} f\right]\right)^2 \\ &= n\mathrm{E}_{P_0}\left[\mathrm{Var}_{P^{s\Pi}}\left[P_n^{\mathrm{HT}} f\right]\right] + n\,\mathrm{Var}_{P_0}\left[P_N f\right] \\ &= n\Sigma_h^{\Pi}(f,f) + \frac{n}{N}\Big(\mathrm{Var}_{P_0}\left[f(O)\right] - \mathrm{E}_{P_0}\left[\mathrm{Var}_{P_0}\left[f(O)|V\right]\right]\Big),\end{aligned}$$

where the last equality follows from (0.8). This completes the proof.

7.3 Proof of Proposition 0.3

Let us first state the so called Soshnikov conditions (Soshnikov, 2000). A function f of O drawn from P_0 meets them if

$$N^2\,\mathrm{Var}_{P^{sK}}\left[P_n^{\mathrm{HT}} f\right] \text{ goes to infinity,} \tag{0.17}$$

$$\max_{1\leq i\leq N} K_{ii}^{-1} f(O_i) = o\left(N^2\,\mathrm{Var}_{P^{sK}}\left[P_n^{\mathrm{HT}} f\right]\right)^{\epsilon} \text{ for all } \epsilon > 0, \tag{0.18}$$

$$N\mathrm{E}_{P^{sK}}\left[P_n^{\mathrm{HT}}|f|\right] = O\left(N^2\,\mathrm{Var}_{P^{sK}}\left[P_n^{\mathrm{HT}} f\right]\right)^{\delta} \text{ for some } \delta > 0. \tag{0.19}$$

Conditions (0.17), (0.18) and (0.19) are expressed conditionally on a trajectory $(O_i)_{i\geq 1}$ of mutually independent random variables drawn from P_0. We denote $\Omega(f)$ the set of trajectories for which they are met. By assumption, $P_0(\Omega(f)) = 1$ for all $f \in \mathcal{F}'$. It is worth emphasizing that this assumption may implicitly require that the ratio n/N go to zero sufficiently slowly, as evident in the sketch of proof of Proposition 0.4. Since $\mathcal{F}'$ is countable, $\Omega \equiv \cap_{f\in\mathcal{F}'}\Omega(f)$ also satisfies $P_0(\Omega) = 1$.

Set $f \in \mathcal{F}'$ and define $Z_N(f) \equiv (\mathrm{Var}_{P^s}\left[P_n^{\mathrm{HT}} f\right])^{-1/2}(P_n^{\mathrm{HT}} - P_0)f$. On Ω, the characteristic function $t \mapsto \mathrm{E}_{P^{sK}}\left[e^{itZ_N(f)}\right]$ converges pointwise to $t \mapsto e^{-t^2/2}$. Therefore, $t \mapsto \mathrm{E}_{P_0}\left[\mathrm{E}_{P^{sK}}\left[e^{itZ_N(f)}\right]\mathbf{1}\{\Omega\}\right]$ also does. Since $P_0(\Omega) = 1$, this implies the convergence in distribution of $Z_N(f)$ to the standard normal law hence, by Slutsky's lemma, that of $\sqrt{n}(P_n^{\mathrm{HT}} - P_0)f$ to the centered Gaussian law with a variance equal to the limit in probability of $n\,\mathrm{Var}_{P^{sK}}\left[P_n^{\mathrm{HT}} f\right]$. The asymptotic tightness of $\sqrt{n}(P_n^{\mathrm{HT}} - P_0)f$ follows. Finally, applying the Cramér-Wold device yields the convergence to a centered multivariate Gaussian law of all marginals $\sqrt{n}(P_n^{\mathrm{HT}} - P_0)(f_1,\ldots,f_M)$ with $f_1,\ldots,f_M \in \mathcal{F}'$.

The second step of the proof hinges on the following concentration inequality (Pemantle and Peres, 2014, Theorem 3.1): if $C(f) \equiv \max_{1\leq i\leq N}|K_{ii}^{-1} f(O_i)|$ then, for all $t > 0$,

$$P^{sK}\left[|(P_n^{\mathrm{HT}} - P_0)f| \geq t\right] \leq 2\exp\left(-nt^2/8C(f)^2\right). \tag{0.20}$$

This statement is conditional on $O_1,\ldots,O_N$. Note that there exists a deterministic upper-bound to all $C(f)$s because $\mathcal{F}'$ is uniformly bounded and because the first order inclusion probabilities are bounded away from 0 uniformly in N. We go from the convergence of all marginals to **A1** by developing a so called chaining argument typical of empirical processes theory. The argument builds upon (0.20) and the assumed finiteness of the bracketing entropy of $\mathcal{F}'$ with respect to the supremum norm. This completes the sketch of the proof.

7.4 Proof of Proposition 0.4

Consider $f \in \mathcal{F}\setminus\mathcal{F}'$, a function of O drawn from P_0 which depends on V only. It holds that

$$P_n^{\mathrm{HT}} f = \frac{1}{n} \sum_{v=1}^{\nu} f(v) h^{-1}(v) n_v,$$

where $n_v = \sum_{i=1}^N \eta_i \ell_v(V_i) = nh(v)N_v/N$ with $N_v \equiv \sum_{i=1}^N \ell_v(V_i)$ (each $1 \leq v \leq \nu$). Therefore, the above display rewrites $P_n^{\mathrm{HT}} f = P_N f$, hence $\mathrm{Var}_{P^{s\Pi}}\left[P_n^{\mathrm{HT}} f\right] = 0$. Moreover, the CLT for bounded, independent and identically distributed observations implies $\sqrt{n}(P_n^{\mathrm{HT}} - P_0)f = \sqrt{n/N} \times \sqrt{N}(P_N - P_0)f = O_P(\sqrt{n/N}) = o_P(1)$.

Consider now $f \in \mathcal{F}'$. We wish to prove that f meets the Soshnikov conditions and that $n \,\mathrm{Var}_{P^{s\Pi}}\left[P_n^{\mathrm{HT}} f\right]$ converges in P_0-probability to $\Sigma_h^{\Pi}(f,f)$, which is positive because $f \in \mathcal{F}'$. When relying on $P^{s\Pi}$, the LHS expression in (0.18) rewrites $\max_{1 \leq i \leq N} Nf(O_i)/nh(V_i)$ and is clearly upper-bounded by a constant times N/n. As for the LHS of (0.19), it equals $\sum_{i=1}^N |f(O_i)|$ and is thus clearly upper-bounded by a constant times N. Let us now turn to $\mathrm{Var}_{P^{s\Pi}}\left[P_n^{HT} f\right]$. By construction of $P^{s\Pi}$, the variance decomposes as the sum of the variances over each V-stratum, each of them being a quadratic form in sub-Gaussian, independent and identically distributed random variables conditionally on $(V_1, \ldots, V_N)$. Because quadratic forms of independent sub-Gaussian random variables are known to concentrate exponentially fast around their expectations (see the Hanson-Wright concentration inequality in Rudelson and Vershynin, 2013), $\mathrm{Var}_{P^{s\Pi}}\left[P_n^{HT} f\right]$ concentrates around its expectation (0.8). Consequently, $N^2 \,\mathrm{Var}_{P^{s\Pi}}\left[P_n^{HT} f\right]$ is of order N^2/n. It is then clear that $N/n = o((N^2/n)^{\epsilon})$ for all $\epsilon > 0$ ensures that f meets the Soshnikov conditions. This holds for instance if $n \equiv N/\log^a(N)$ for some $a > 0$. Finally, the concentration of $\mathrm{Var}_{P^s}\left[P_n^{\mathrm{HT}} f\right]$ around its expectation also yields the convergence of $n \,\mathrm{Var}_{P^s}\left[P_n^{\mathrm{HT}} f\right]$ to $\Sigma_h^{\Pi}(f,f)$ in P_0-probability.

At this point, we have shown that $\sqrt{n}(P_n^{\mathrm{HT}} - P_0)f$ converges in distribution to the centered Gaussian law with variance $\Sigma_h^{\Pi}(f,f)$. The rest of the proof is similar to the end of the proof of Proposition 0.3. This completes the sketch of proof.

References

P. Bertail, A. Chambaz, and E. Joly. Practical targeted learning from large data sets by survey sampling. *arXiv preprint*, 2016a. Submitted.

P. Bertail, E. Chautru, and S. Clémençon. Empirical processes in survey sampling. *Scandinavian Journal of Statistics*, 2016b. To appear.

L. Bondesson, I. Traat, and A. Lundqvist. Pareto sampling versus Sampford and conditional Poisson sampling. *Scandinavian Journal of Statistics. Theory and Applications*, 33(4):699–720, 2006.

A. Chambaz and P. Neuvial. Targeted, integrative search of associations between DNA copy number and gene expression, accounting for DNA methylation. *Bioinformatics*, 31(18):3054–3056, 2015.

A. Chambaz and P. Neuvial. *Targeted Learning of a Non-Parametric Variable Importance Measure of a Continuous Exposure*, 2016. URL http://CRAN.R-project.org/package=tmle.npvi. R package version 0.10.0.

A. Chambaz, P. Neuvial, and M. J. van der Laan. Estimation of a non-parametric variable importance measure of a continuous exposure. *Electron. J. Stat.*, 6:1059–1099, 2012.

J. Hajek. Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, 35(4):1491–1523, 12 1964.

M. Hanif and K. R. W. Brewer. Sampling with unequal probabilities without replacement: a review. *International Statistical Review/Revue Internationale de Statistique*, pages 317–335, 1980.

J. B. Hough, M. Krishnapur, Y. Peres, and B. Virág. Determinantal processes and independence. *Probability Surveys*, 3:206–229, 2006.

V. Loonis and X. Mary. Determinantal sampling designs. *arXiv preprint arXiv:1510.06618*, 2015. Submitted.

R. Lyons. Determinantal probability measures. *Publications Mathématiques de l'Institut des Hautes Études Scientifiques*, 98:167–212, 2003.

O. Macchi. The coincidence approach to stochastic point processes. *Advances in Applied Probability*, 7: 83–122, 1975.

R. Pemantle and Y. Peres. Concentration of Lipschitz functionals of determinantal and other strong Rayleigh measures. *Combinatorics, Probability and Computing*, 23(1):140–160, 2014.

M. Rudelson and R. Vershynin. Hanson-Wright inequality and subgaussian concentration. *Electron. Commun. Probab.*, 18(82):1–9, 2013.

M. R. Sampford. On sampling without replacement with unequal probabilities of selection. *Biometrika*, 54 (3-4):499–513, 1967.

A. Soshnikov. Gaussian limit for determinantal random point fields. *Ann. Probab.*, 30(1):171–187, 2000.

M. J. van der Laan. One-step targeted minimum loss-based estimation based on universal least favorable one-dimensional submodels. *International Journal of Biostatistics*, 2016. To appear.

M. J. van der Laan and S. Lendle. Online targeted learning. Technical Report 330, Division of Biostatistics, University of California, Berkeley, 2014.

A. W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer, Berlin Heidelberg New York, 1996.