

Agnostic Model-Assisted Estimation with Machine Learning for Survey Data

Ziming AN

Department of Mathematics and Statistics, University of Ottawa
and

Mehdi DAGDOUG

Department of Mathematics and Statistics, McGill University
and

David HAZIZA

Department of Mathematics and Statistics, University of Ottawa
and

Yves TILLÉ

Institute of Statistics, University of Neuchâtel

Abstract

Model-assisted estimation uses prediction rules to improve the efficiency of estimators of finite population parameters while retaining design-based inference. Although flexible prediction methods have been considered, existing theoretical results are largely method-specific. We develop a learner-agnostic framework that replaces separate analyses for individual learners with general conditions on the sampling design and prediction error. We connect design-aware and design-agnostic cross-fitting and characterize the sampling designs under which they yield conditional independence across folds. Under suitable conditions, conditional weighting gives exact design-unbiasedness. We establish first-order equivalence to oracle estimators, leading to design consistency and asymptotic normality, and clarify when conditional and original inclusion probabilities yield the same first-order behavior. We propose consistent variance estimators based on cross-fitted residuals and construct asymptotically valid confidence intervals. Under additional model and regularity conditions, we establish asymptotic optimality through attainment of the Godambe–Joshi lower bound. Simulations show that cross-fitting substantially reduces finite-sample bias and improves variance estimation and coverage with adaptive learners.

Keywords: Auxiliary information; cross-fitting; design-based inference; finite-population sampling; Neyman orthogonality; variance estimation.

1 Introduction

The main goal in survey sampling is to estimate finite population characteristics, such as totals or means, from a sample. In the classical design-based framework, inference relies on the sampling design, while the population values are treated as fixed. The Horvitz–Thompson estimator is design-unbiased under very general conditions but, apart from auxiliary information incorporated into the sampling design through the inclusion probabilities, it does not use auxiliary information at the estimation stage. It can therefore be inefficient when the survey variable is strongly related to available auxiliary variables (Särndal et al., 1992; Breidt and Opsomer, 2017).

Model-assisted estimation uses a prediction model to exploit this relationship and improve efficiency, while inference remains based on the sampling design. Thus, the model is used to improve precision, not to justify the validity of the methodology. This approach originates in classical regression estimation; see, among others, Cochran (1942), Cassel et al. (1976), Robinson and Särndal (1983), and the monograph of Särndal et al. (1992). A modern review is provided by Breidt and Opsomer (2017).

Much of the early literature focused on parametric settings, particularly linear working models, leading to the generalized regression (GREG) estimator. Because linear models may be too restrictive, nonparametric and semiparametric model-assisted methods have been developed to capture more complex relationships. Examples include local polynomial regression (Breidt and Opsomer, 2000), B-splines (Goga, 2005), penalized splines (Breidt et al., 2005), additive models (Opsomer et al., 2007), neural networks (Montanari and Ranalli, 2005), tree-based methods (Toth and Eltinge, 2011; McConville and Toth, 2019), and random forests (Dagdoug et al., 2023).

Classical linear model-assisted estimators require only the auxiliary values for sampled units and their population totals. In contrast, nonparametric and machine learning-based estimators generally require auxiliary information for every population unit. Although stronger, this requirement is increasingly realistic in applications involving administrative records, registers, or linked data. Such information makes it possible to use flexible prediction methods and potentially achieve substantial efficiency gains (Breidt and Opsomer, 2017).

These methods also introduce new inferential challenges. Because the prediction function is estimated from the same sample used to construct the final estimator, the fitted predictions depend on the sampling indicators and cannot be treated as fixed in the theoretical analysis. Estimating the prediction function may also compromise variance estimation: flexible learners can track the sampled observations too closely, producing artificially small residuals and hence underestimated variances. These problems are particularly relevant for highly adaptive learners,

which may converge slowly and generally do not admit simple linearizations (Opsomer and Miller, 2005).

Recent work on debiased and double machine learning shows that orthogonal estimating equations and cross-fitting can restore valid large-sample inference in the presence of complex nuisance estimation (Chernozhukov et al., 2018). These ideas are increasingly relevant in survey sampling. In particular, Seaman (2025) considers data integration involving probability and non-probability samples, while Dagdoug and Haziza (2026) study item nonresponse through imputation.

This paper develops a general design-based theory of model-assisted estimation and cross-fitting that is not tied to a particular prediction method. The main contribution is to replace learner-specific analyses with common conditions on the sampling design and prediction error. This allows the same estimation and inference results to be used across learners, without repeating the design-based analysis for each method or prescribing a convergence rate for the prediction rule. A central difficulty is that, under a design-based framework and dependent sampling, splitting the population does not generally make the training and evaluation samples independent. We give a necessary and sufficient characterization of the sampling designs that yield conditional independence across folds under uniform population splitting, and establish the connection with design-aware splitting. These results identify when cross-fitting provides the independence needed for design-based inference. Under suitable conditions, we establish exact design-unbiasedness with conditional inclusion probabilities, first-order equivalence to oracle estimators, design consistency, and asymptotic normality. We also determine when retaining the original inclusion probabilities preserves the first-order behavior of conditional weighting. Finally, we establish consistent variance estimation using cross-fitted residuals and give conditions for asymptotic attainment of the Godambe–Joshi lower bound. Together, these results provide a common foundation for using flexible prediction methods in survey estimation while retaining design-based guarantees.

The remainder of the paper is organized as follows. Section 2 introduces the setup and motivates the problem. Section 3 presents conditional cross-fitting schemes and establishes their properties. Section 4 develops a cross-fitted model-assisted estimator based on conditional inclusion probabilities, whereas Section 5 studies its counterpart based on unconditional inclusion probabilities. Section 6 compares the two estimators, and Section 7 reports the simulation results. Section 8 concludes with open questions and extensions to finite-population parameters defined through estimating equations. The appendices contain the proofs and supplementary results.

2 Basic setup and motivation

Consider a finite population $U = \{1, \dots, N\}$ of size N . We are interested in estimating the population mean $\mu = N^{-1} \sum_{k \in U} y_k$, where y_k is the value of the study variable Y for unit $k \in U$. A sample $S \subset U$ of size n is selected according to a sampling design $p(\cdot)$. Let $\{I_k\}_{k \in U}$ denote the sample inclusion indicators, where $I_k = 1$ if $k \in S$ and $I_k = 0$ otherwise. The first-order and second-order inclusion probabilities are given by $\pi_k = \mathbb{P}(I_k = 1)$ and $\pi_{kl} = \mathbb{P}(I_k = 1, I_l = 1)$. Provided that $\pi_k > 0$ for all $k \in U$, the Horvitz–Thompson estimator

$$\hat{\mu}_\pi = \frac{1}{N} \sum_{k \in S} \frac{y_k}{\pi_k}$$

is design-unbiased for μ , that is, $\mathbb{E}_p[\hat{\mu}_\pi] = \mu$, where $\mathbb{E}_p[\cdot]$ denotes the expectation operator with respect to the sampling design. We assume that, for each $k \in U$, the vector $\mathbf{x}_k = (x_{k1}, \dots, x_{kp})^\top$ of p auxiliary variables is known. To exploit this information, consider the working relationship

$$y_k = m(\mathbf{x}_k) + \epsilon_k, \quad k \in U, \quad (1)$$

where $m : \mathbb{R}^p \rightarrow \mathbb{R}$ is an unknown function and $\mathbb{E}(\epsilon_k \mid \mathbf{x}_k) = 0$.

2.1 Revisiting model-based estimation with machine learning

A long-standing approach to using auxiliary information in finite population inference is model-based, or prediction-based, estimation. This perspective goes back at least to [Brewer \(1963\)](#) and [Royall \(1970\)](#); comprehensive accounts are given by [Valliant et al. \(2000\)](#) and [Chambers and Clark \(2012\)](#). The finite population values are viewed as a realization from a superpopulation model, and the unobserved values for nonsampled units are predicted from the observed sample.

Let $\hat{m}(\cdot)$ be a prediction rule fitted using $\{(\mathbf{x}_k, y_k) : k \in S\}$. The resulting model-based estimator is

$$\hat{\mu}_{mb}(\hat{m}) = \frac{1}{N} \left\{ \sum_{k \in S} y_k + \sum_{k \in U \setminus S} \hat{m}(\mathbf{x}_k) \right\}.$$

In the parametric case, when the assumed model is appropriate, this estimator can be highly efficient. Its validity, however, depends on the model: misspecification or informative sampling may lead to biased predictions and invalid measures of uncertainty. This issue was central to the classical debate between design-based and model-based inference; see, for example, [Hansen et al. \(1983\)](#).

Model-based estimators have been studied extensively under parametric models, particularly linear regression models. Nonparametric contributions include regression, kernel, and spline-based methods; see [Dorfman \(1992\)](#), [Chambers et al. \(1993\)](#), [Dorfman and Hall \(1993\)](#), [Kuk \(1993\)](#), [Zheng and Little \(2003\)](#), and [Rueda \(2009\)](#). Their theoretical properties are generally established using assumptions and arguments tailored to the particular prediction method.

A fundamental difficulty with machine learning is that the model-based estimator is not Neyman-orthogonal with respect to the prediction rule. Indeed, for any reference rule m ,

$$\hat{\mu}_{mb}(\hat{m}) - \hat{\mu}_{mb}(m) = \frac{1}{N} \sum_{k \in U \setminus S} \{\hat{m}(\mathbf{x}_k) - m(\mathbf{x}_k)\}.$$

Thus, prediction error enters the estimator at first order. Equivalently, the design expectation of the estimating equation defining $\hat{\mu}_{mb}$ has derivative $N^{-1} \sum_{k \in U} (1 - \pi_k) h(\mathbf{x}_k)$ in a direction h , which is generally nonzero. Without additional structure, the convergence rate of the estimator may therefore inherit the potentially slower rate of the learner, making root- n inference difficult. An elementary example illustrating this phenomenon is given in Section [A.1](#) of the Supplementary Material. See also [Chernozhukov et al. \(2018\)](#) for a formal discussion of Neyman orthogonality.

This observation motivates the model-assisted approach developed below. Unlike the direct model-based estimator, the model-assisted estimator is Neyman-orthogonal, allowing flexible and slowly converging learners to be used without requiring parametric rates. However, using the same sample to fit the prediction rule and construct the estimator creates an additional dependence that must be accounted for. We address this issue using cross-fitting.

2.2 Model-assisted estimation: Advantages and obstacles to a general theory

Let $\tilde{m}(\cdot)$ denote a deterministic target associated with the chosen learning strategy. It is treated as fixed under the sampling design and need not be the true regression function or the rule obtained by fitting the learner to the full finite population. Rather, it represents a stable target toward which the fitted predictions converge in a finite-population average sense, as formalized later through an L_2 stability condition. In classical settings such as linear regression, it may coincide with the usual population-level fit.

If $\tilde{m}(\mathbf{x}_k)$ were known for every $k \in U$, the oracle model-assisted estimator, also known as the difference estimator, would be

$$\hat{\mu}_{ma}(\tilde{m}, \pi) = \frac{1}{N} \left\{ \sum_{k \in U} \tilde{m}(\mathbf{x}_k) + \sum_{k \in S} \frac{y_k - \tilde{m}(\mathbf{x}_k)}{\pi_k} \right\}. \quad (2)$$

Because $\tilde{m}(\cdot)$ is fixed under the sampling design, $\mathbb{E}_p\{\hat{\mu}_{\text{ma}}(\tilde{m}, \pi)\} = \mu$, regardless of its predictive accuracy. Thus, the target need not correctly describe the conditional mean of Y . Although infeasible, (2) provides a useful benchmark for studying the effect of estimating the prediction rule. In the linear case, a natural target is $\tilde{m}(\mathbf{x}_k) = \mathbf{x}_k^\top \mathbf{B}$, where $\mathbf{B}$ is the finite population least-squares coefficient, and (2) reduces to the classical difference estimator studied by [Cassel et al. \(1976\)](#).

In practice, $\tilde{m}(\cdot)$ is replaced by a prediction rule $\hat{m}(\cdot)$ fitted using $\{(\mathbf{x}_k, y_k) : k \in S\}$. The feasible model-assisted estimator is

$$\hat{\mu}_{\text{ma}}(\hat{m}, \pi) = \frac{1}{N} \left\{ \sum_{k \in U} \hat{m}(\mathbf{x}_k) + \sum_{k \in S} \frac{y_k - \hat{m}(\mathbf{x}_k)}{\pi_k} \right\}. \quad (3)$$

Under a linear working model, $\hat{m}(\mathbf{x}_k) = \mathbf{x}_k^\top \hat{\beta}$, with $\hat{\beta}$ estimated by weighted least squares, leading to the generalized regression (GREG) estimator. Unlike more flexible estimators, the GREG estimator requires only the population totals of the auxiliary variables.

In a seminal paper, [Breidt and Opsomer \(2000\)](#) established the $\sqrt{n}$ -consistency and asymptotic normality of a local polynomial estimator, proposed a consistent variance estimator, and showed that the estimator attains the Godambe–Joshi lower bound. As with the other model-assisted methods cited in the introduction, these results do not require the prediction rule to converge at a parametric rate. This contrasts with direct model-based estimation and reflects the Neyman orthogonality of the model-assisted estimator.

A remaining difficulty is that the same sample is used to fit $\hat{m}(\cdot)$ and to construct (3). Consequently, the fitted predictions depend on the sampling indicators that also enter the estimator. Writing them as $\hat{m}(\mathbf{x}_k; \mathbf{I})$, where $\mathbf{I} = (I_1, \dots, I_N)^\top$, gives

$$\hat{\mu}_{\text{ma}}(\hat{m}, \pi) - \mu = \frac{1}{N} \sum_{k \in U} \left(\frac{I_k}{\pi_k} - 1 \right) \{y_k - \hat{m}(\mathbf{x}_k; \mathbf{I})\}.$$

If the prediction rule were fixed under the sampling design, the expectation of the right-hand side would be zero. With sample-fitted predictions, however,

$$\mathbb{E}_p \{ \hat{\mu}_{\text{ma}}(\hat{m}, \pi) - \mu \} = -\frac{1}{N} \sum_{k \in U} \text{Cov}_p \left\{ \frac{I_k}{\pi_k}, \hat{m}(\mathbf{x}_k; \mathbf{I}) \right\}. \quad (4)$$

Thus, the design bias is driven by the dependence between the fitted predictions and the sampling indicators. These covariance terms may be non-negligible in finite samples and need not vanish asymptotically without suitable restrictions on the learner. A simple counterexample showing that the feasible estimator may fail to be design-consistent is given in Section A.2 of the Supplementary Material.

The same dependence affects variance estimation. When the prediction rule is fitted and evaluated on the same observations, the sample residuals, $(y_k - \hat{m}(\mathbf{x}_k))_{k \in S}$, may be artificially small, particularly for highly adaptive learners. Plug-in variance estimators that treat $\hat{m}(\cdot)$ as fixed may then underestimate the sampling variance. A general theory for model-assisted estimation with arbitrary learners must therefore account for the dependence created by data reuse.

2.3 Asymptotic framework

We consider the asymptotic framework of [Isaki and Fuller \(1982\)](#). Let $\{U_v\}_{v \in \mathbb{N}}$ be an increasing sequence of populations, that is, $U_1 \subset U_2 \subset \dots$, with corresponding sizes $\{N_v\}_{v \in \mathbb{N}}$. In each population U_v , a sample S_v is drawn according to a sampling design in order to estimate the population mean μ . In the asymptotic results, $n_v := \mathbb{E}_p(|S_v|) = \sum_{k \in U_v} \pi_k$ denotes the expected sample size; it equals the sample size for fixed-size designs. All subsequent steps (e.g., fitting a prediction rule, prediction, and variance estimation) are carried out at each $v \in \mathbb{N}$. The number of covariates p is assumed to remain fixed as v increases. For simplicity, we omit the index v whenever there is no ambiguity. Below, we state a set of assumptions related to the sampling design.

(D1) The expected sampling fraction satisfies

$$\lim_{v \rightarrow \infty} \frac{n_v}{N_v} = \pi_\star > 0.$$

(D2) There exist positive constants $\lambda > 0$ and $\lambda^\star > 0$ such that, for all $v \in \mathbb{N}$,

$$\min_{k \in U_v} \pi_k \geq \lambda, \quad \text{and} \quad \min_{k, l \in U_v} \pi_{kl} \geq \lambda^\star.$$

(D3) Let $\Delta_{kl} := \pi_{kl} - \pi_k \pi_l$. There exists a positive constant $C > 0$ such that

$$\max_{k \neq l \in U_v} |\Delta_{kl}| \leq \frac{C}{n_v}.$$

(D4) Let D_{t, N_v} denote the set of distinct t -tuples from U_v . Then

$$\lim_{v \rightarrow \infty} \max_{(i, j, k, l) \in D_{4, N_v}} |\mathbb{E}_p[(I_i I_j - \pi_i \pi_j)(I_k I_l - \pi_k \pi_l)]| = 0.$$

These regularity conditions are standard in the survey sampling literature. Assumption (D1) states that the expected sample size n_v grows at the same rate as the population size N_v . Assumption (D2) ensures that the first- and second-order inclusion probabilities are bounded away from zero. Assumption (D3) requires that the sampling covariances decrease fast enough to zero. Finally, Assumption (D4) is commonly used to establish the consistency of the Horvitz–Thompson variance estimator; see [Boistard et al. \(2017\)](#) for a discussion.

3 Cross-fitting in survey sampling

As discussed above, the standard model-assisted estimator suffers from the dependence created by using the same observations to fit the prediction rule and construct the estimator. Cross-fitting addresses this issue through honest predictions, obtained from data not reused at the evaluation stage.

Implementing cross-fitting under design-based inference requires care because the sampling indicators $\{I_k\}_{k \in U}$ are generally dependent. For certain sampling designs, [Lu et al. \(2025\)](#) propose a construction based on conditional rather than unconditional independence. We present this *design-aware* approach in Section 3.1. Section 3.2 introduces a *design-agnostic* alternative based on uniformly randomized folds. We then relate the two approaches and discuss their implementation under more complex designs.

3.1 The design-aware cross-fitting method

We first describe the approach of [Lu et al. \(2025\)](#), independently of any specific estimator. The goal is to partition the finite population into folds satisfying conditional independence properties that separate the data used to train a prediction rule from those on which it is evaluated.

Let $\mathcal{P} = \{U_1, \dots, U_M\}$ be a random partition of U into M disjoint folds, so that $U = \bigcup_{j=1}^M U_j$ and $U_j \cap U_q = \emptyset$ for $j \neq q$. Let $N_j = |U_j|$ and, for a given sample S , define $S_j := S \cap U_j$. Finally, let $\delta_{k,j} = 1$ if $k \in U_j$ and $\delta_{k,j} = 0$ otherwise, and let $\boldsymbol{\delta} = \{\delta_{k,j} : k \in U, j = 1, \dots, M\}$ denote the collection of partition indicators. The construction of [Lu et al. \(2025\)](#) aims to make the sampling indicators conditionally independent across folds, given the partition. More precisely,

$$(\text{Ind-}\boldsymbol{\delta}) : \quad \{I_k\}_{k \in U_j} \perp\!\!\!\perp \{I_k\}_{k \in U \setminus U_j} \mid \boldsymbol{\delta}, \quad j = 1, \dots, M. \quad (5)$$

Thus, conditional on the partition, the sampling indicators in one fold provide no additional information about those in the remaining folds. Examples of the fold constructions proposed by [Lu et al. \(2025\)](#) for Bernoulli sampling, simple random sampling without replacement (SRSWOR), and stratified SRSWOR are given in Section B.1 of the Supplementary Material.

Definition 1. An M -fold cross-fitting scheme is a pair $(\boldsymbol{\delta}, \mathcal{A})$ where $\boldsymbol{\delta}$ is a random partition of U into folds $U_1, \dots, U_M$, and $\mathcal{A}$ is a random element satisfying $\sigma(\boldsymbol{\delta}) \subseteq \sigma(\mathcal{A})$. Let

$$\pi_{k|\mathcal{A}} = \mathbb{P}(k \in S \mid \mathcal{A}), \quad \pi_{kl|\mathcal{A}} = \mathbb{P}(k, l \in S \mid \mathcal{A}), \quad \Delta_{kl|\mathcal{A}} = \pi_{kl|\mathcal{A}} - \pi_{k|\mathcal{A}}\pi_{l|\mathcal{A}}$$

be the first- and second-order inclusion probabilities and design covariances of elements $k, l \in U$.

In Definition 1, the random element $\mathcal{A}$ specifies what is conditioned on to obtain independence across folds. In the examples of Lu et al. (2025) presented in Section B.1 of the Supplementary Material, $\mathcal{A} = \delta$; however, as shown in the next section, $\mathcal{A}$ need not coincide with δ : the required conditional independence may hold after conditioning on a different random element.

Definition 2. Let $(\delta, \mathcal{A})$ be a cross-fitting scheme. We say that $(\delta, \mathcal{A})$ is *honest* if

$$(\text{Ind-}\mathcal{A}) : \quad \{I_k\}_{k \in U_j} \perp\!\!\!\perp \{I_k\}_{k \in U \setminus U_j} \mid \mathcal{A}, \quad j = 1, \dots, M. \quad (6)$$

3.2 A design-agnostic cross-fitting procedure

We now describe a simple cross-fitting procedure, implemented at the estimation stage, after the sample has been selected. The key idea is to uniformly at random partition the population U , and therefore the observed sample S , into folds, and to construct prediction rules using observations outside each fold. Let $\mathcal{P} = \{U_1, \dots, U_M\}$ be a random partition of the population with fixed respective sizes $N_1, N_2, \dots, N_M$, characterized by δ . For each fold $j = 1, \dots, M$, define $S_j = S \cap U_j$, and let $n_j = \sum_{k \in U_j} I_k$ denote the number of sampled units in U_j . We write $\mathbf{n} = (n_1, \dots, n_M)$ for the vector of fold-specific sample sizes. This cross-fitting scheme is summarized in the following algorithm.

Algorithm 1 Uniform cross-fitting scheme.

Input: Population U , sample S , and positive integer fold sizes $N_1, \dots, N_M$.

Do:

1. Independently of S , uniformly partition U into $U_1, \dots, U_M$, where $|U_j| = N_j$.
2. For $j = 1, \dots, M$, set $S_j = S \cap U_j$ and $n_j = |S_j|$.

Output: The population partition $\mathcal{P} = (U_1, \dots, U_M)$.

Unlike the construction in Section 3.1, this procedure does not, in general, satisfy the conditional independence property (Ind- δ) in (5). However, under an important class of sampling designs, another form of conditional independence holds. Specifically, conditional on $(\mathbf{n}, \delta)$, we have

$$(\text{Ind-}\{\mathbf{n}, \delta\}) : \quad \{I_k\}_{k \in U_j} \perp\!\!\!\perp \{I_k\}_{k \in U \setminus U_j} \mid (\mathbf{n}, \delta), \quad j = 1, \dots, M. \quad (7)$$

We now ask under which sampling designs the uniform cross-fitting scheme satisfies condition (Ind- $\{\mathbf{n}, \delta\}$). To answer this question, we recall two definitions from Tillé (2006). Let $\mathcal{P}(U)$ denote the power set of U , and let $\mathcal{Q} := \{s \in \mathcal{P}(U) : p(s) > 0\}$ be the support of the sampling

design p . The support $\mathcal{Q}$ is *symmetric* if, for every $s \in \mathcal{Q}$, all subsets $r \subseteq U$ satisfying $|r| = |s|$ also belong to $\mathcal{Q}$. A sampling design is *exponential* on $\mathcal{Q}$ if, for some $\boldsymbol{\lambda} \in \mathbb{R}^N$,

$$p(s; \mathcal{Q}, \boldsymbol{\lambda}) = \frac{\exp(\sum_{k \in s} \lambda_k)}{\sum_{r \in \mathcal{Q}} \exp\left(\sum_{l \in r} \lambda_l\right)}, \quad s \in \mathcal{Q}.$$

An overview of exponential sampling designs is provided in Appendix D. A more comprehensive discussion is presented in Tillé (2006). Below, we provide additional results regarding these designs, which will be useful for the rest of the article.

Theorem 1. *Let p be a sampling design on a symmetric support $\mathcal{Q}$. Let $M \geq 2$ and let $N_1, \dots, N_M \geq 2$ satisfy $\sum_{j=1}^M N_j = N$. Define*

$$\mathcal{D}_N := \left\{ (U_1, \dots, U_M) : \biguplus_{j=1}^M U_j = U, |U_j| = N_j, \text{ for } j = 1, \dots, M \right\}.$$

Let $\boldsymbol{\delta}$ be a random partition, independent of S , whose distribution has full support on $\mathcal{D}_N$, that is, $\mathbb{P}(\boldsymbol{\delta} = \mathbf{d}) > 0$ for every $\mathbf{d} \in \mathcal{D}_N$. The following statements are equivalent:

- (i) *Property (Ind- $\{\mathbf{n}, \boldsymbol{\delta}\}$) in (7) holds.*
- (ii) *For every sample size n such that $\mathbb{P}(|S| = n) > 0$, the conditional design given $|S| = n$ is an exponential design on $\mathcal{S}_n := \{s \in \mathcal{P}(U) : |s| = n\}$.*

Proof. See Appendix E.1. □

Theorem 1 gives an exact characterization: under the uniform cross-fitting scheme, condition (Ind- $\{\mathbf{n}, \boldsymbol{\delta}\}$) holds if and only if, conditional on each attainable sample size n , the sampling design is exponential on $\mathcal{S}_n$. This class includes SRSWOR, Bernoulli sampling, Poisson sampling, and conditional Poisson sampling. Thus, for these designs, a uniformly generated partition provides the required conditional independence.

3.3 Relationship between design-aware and design-agnostic cross-fitting schemes

Although it is sufficient to obtain conditional independence in an interesting class of sampling designs, the design-agnostic construction has the drawback of not controlling the fold sample sizes. For example, it is possible to obtain a partition $\mathcal{P}$ such that $n_j = 0$ for some fold U_j . On the other hand, the design-aware cross-fitting philosophy, such as the construction proposed for SRSWOR in Lu et al. (2025), allows one to choose how many sampled elements should be in each fold. We restate this procedure for a general fixed-size design.

Algorithm 2 Design-aware cross-fitting scheme.

Input: Population U , sample S , positive integer fold sizes $N_1, \dots, N_M$, and prescribed integer sample fold sizes $n_1, \dots, n_M$.

Do:

1. Partition S uniformly at random into $S_1, \dots, S_M$, where $|S_j| = n_j$.
2. Independently of the first partition, partition $U \setminus S$ uniformly at random into $R_1, \dots, R_M$, where $|R_j| = N_j - n_j$.
3. For $j = 1, \dots, M$, set $U_j = S_j \cup R_j$.

Output: The population partition $\mathcal{P} = (U_1, \dots, U_M)$.

Algorithm 2 happens to be exactly the cross-fitting scheme proposed by Lu et al. (2025) for the particular case of completely randomized trials (corresponding to SRSWOR). It happens that, in fact, this procedure produces valid folds satisfying conditional independence in a wider variety of designs, as the next theorem shows.

Theorem 2. Let δ_{unif} and δ_{aware} denote the partitions obtained from Algorithm 1 and Algorithm 2, respectively, using the same fold sizes $N_1, \dots, N_M$. Fix target fold counts $\mathbf{n} = (n_1, \dots, n_M)$, and suppose that Algorithm 2 uses these counts. For $s \in \mathcal{P}(U)$ and a partition $\mathbf{d}$, define $\mathbf{n}(s, \mathbf{d}) := [|s \cap U_1(\mathbf{d})|, \dots, |s \cap U_M(\mathbf{d})|]^\top$.

Then, for every $s \in \mathcal{P}(U)$ and every partition $\mathbf{d}$ with fold sizes $N_1, \dots, N_M$ for which $\mathbb{P}(\mathbf{n}(S, \mathbf{d}) = \mathbf{n}) > 0$,

$$\mathbb{P}(S = s | \delta_{\text{aware}} = \mathbf{d}) = \mathbb{P}(S = s | \delta_{\text{unif}} = \mathbf{d}, \mathbf{n}(S, \mathbf{d}) = \mathbf{n}).$$

Consequently, Algorithm 2 satisfies (Ind- δ) for every choice of $\mathbf{n}$ if and only if Algorithm 1 satisfies (Ind- $\{\mathbf{n}, \delta\}$), and both schemes yield the same first- and second-order conditional inclusion probabilities.

Proof. See Appendix E.2. □

For fixed-size sampling designs, Theorems 1 and 2 show that whenever the uniform scheme satisfies (Ind- $\{\mathbf{n}, \delta\}$), Algorithm 2 satisfies (Ind- δ) while allowing the sample fold sizes to be fixed in advance. The construction does not, however, cover conditional independence arising through a different conditioning variable $\mathcal{A}$.

3.4 Assumptions on the cross-fitting scheme

In order to analyze the asymptotic behavior of model-assisted estimators based on cross-fitting, we will require some regularity conditions on the asymptotic behavior of the sequence of cross-fitting schemes as $v \rightarrow \infty$. Throughout, the number of folds M is fixed as $v \rightarrow \infty$. All probabilities and

expectations involving the cross-fitting scheme are taken with respect to the joint distribution of the sample and the random partition.

(CF1) There exist constants $0 < c_1 \leq c_2 < 1$ such that, for $j = 1, \dots, M$,

$$\frac{N_{j,v}}{N_v} \xrightarrow[v \rightarrow \infty]{\mathbb{P}} \alpha_j \in [c_1, c_2].$$

(CF2) Consider the following three versions.

(CF2,a) There exist positive constants $\lambda_{cf} > 0$ and $\lambda_{cf}^* > 0$ such that

$$\lim_{v \rightarrow \infty} \mathbb{P} \left(\min_{k \in U_v} \pi_{k|\mathcal{A}} \geq \lambda_{cf} \right) = 1, \quad \lim_{v \rightarrow \infty} \mathbb{P} \left(\min_{k \neq l \in U_v} \pi_{kl|\mathcal{A}} \geq \lambda_{cf}^* \right) = 1.$$

(CF2,b) First-order conditional inclusion probabilities satisfy

$$\mathbb{P} \left(\min_{k \in U_v} \pi_{k|\mathcal{A}} > 0 \right) = 1 - o(n_v^{-1}).$$

(CF2,c) There exists a constant $\lambda_c > 0$ such that, for all $v \in \mathbb{N}$, almost surely,

$$\min_{k \in U_v} \pi_{k|\mathcal{A}} \geq \lambda_c.$$

(CF3) There exists a positive constant $C_1 > 0$ such that, for all $v \in \mathbb{N}$,

$$\mathbb{E}_p \left[\max_{k \in U_v} (\pi_{k|\mathcal{A}} - \pi_k)^2 \right] \leq \frac{C_1}{n_v}.$$

(CF4) Define $\Delta_{kl|\mathcal{A}} = \pi_{kl|\mathcal{A}} - \pi_{k|\mathcal{A}}\pi_{l|\mathcal{A}}$ for $k, l \in U_v$. There exists a constant $C_{\Delta 2} > 0$ such that, for all $v \in \mathbb{N}$,

$$n_v \max_{k \neq l \in U_v} |\Delta_{kl|\mathcal{A}}| \leq C_{\Delta 2},$$

almost surely.

Assumption (CF1) requires that the fold proportions converge to positive limits below one. This must be ensured when choosing the fold sizes in our constructions. Conditions (CF2,a) and (CF3) require the conditional inclusion probabilities to inherit the good behavior of the unconditional ones. Assumption (CF4) guarantees that the conditional sampling covariances decrease to zero at a rate of at least $\mathcal{O}(n_v^{-1})$, and is used in the analysis of the consistency of the point estimator. This is the conditional analog of what is classically used; see, for example, [Breidt and Opsomer \(2000\)](#).

Some additional assumptions will also be needed for variance estimation. These are the conditional analogs of classical assumptions in the literature; see, e.g., [Breidt and Opsomer \(2000\)](#).

(CF5) Assume that, almost surely,

$$\max_{(i,j,k,l) \in D_{4,N_v}} |\mathbb{E}_p \left[\left(I_i I_j - \pi_{ij|\mathcal{A}} \right) \left(I_k I_l - \pi_{kl|\mathcal{A}} \right) | \mathcal{A} \right]| = o(1).$$

(CF6) Let $\pi_+ := \min_{k \in U_v: \pi_{k|\mathcal{A}} > 0} \pi_{k|\mathcal{A}}$. Set $\pi_+ = 1$ if all the conditional inclusion probabilities are zero. Assume that there exists $\delta > 0$ such that

$$\sup_{v \in \mathbb{N}} \mathbb{E}_p \left[\frac{1}{\pi_+^{2+\delta}} \right] < \infty.$$

The following proposition verifies all these assumptions under SRSWOR.

Proposition 1. *Suppose that p is SRSWOR satisfying (D1). Use Algorithm 2 with deterministic fold sizes and sample counts such that $n_{j,v} \geq 1$ for every j, v , and*

$$\frac{N_{j,v}}{N_v} \xrightarrow{v \rightarrow \infty} \alpha_j \in (0, 1), \quad \max_{1 \leq j \leq M} \left| \frac{n_{j,v}}{N_{j,v}} - \frac{n_v}{N_v} \right| = \mathcal{O}(n_v^{-1/2}).$$

Then the scheme with $\mathcal{A} = (\mathbf{n}, \boldsymbol{\delta})$ is honest and satisfies (CF1)–(CF6), including (CF2,a)–(CF2,c).

Proof. See Appendix E.3. □

The allocation condition holds, for example, when $n_{j,v} = n_v N_{j,v} / N_v + \mathcal{O}(1)$. Similar results are expected for Poisson, conditional Poisson, and stratified sampling under suitable conditions on the design and the cross-fitting scheme.

4 Conditional cross-fitted model-assisted estimation

4.1 Point estimation

We now define model-assisted estimators based on the cross-fitting procedures introduced above. For each fold $j = 1, \dots, M$, a prediction rule $\hat{m}_j$ is fitted using the observations $\{(\mathbf{x}_k, y_k) : k \in S \setminus S_j\}$ outside the fold, and a model-assisted estimator is computed using the units in U_j . Any tuning or variable selection based on the survey variable must also use only the training observations in $S \setminus S_j$. The fold-specific estimators are then aggregated. The prediction rule is assumed to return finite values, including when the training sample is empty. Any randomness used to fit the prediction rules is treated as fixed.

We first introduce the *conditional model-assisted estimator*, which uses the conditional inclusion probabilities $\pi_{k|\mathcal{A}} := \mathbb{P}(I_k = 1 | \mathcal{A})$ for $k \in U$. For each j with $N_j > 0$, define

$$\hat{\mu}_j(\hat{m}_j, \pi_{|\mathcal{A}}) = \frac{1}{N_j} \left\{ \sum_{k \in U_j} \hat{m}_j(\mathbf{x}_k) + \sum_{k \in S_j} \frac{y_k - \hat{m}_j(\mathbf{x}_k)}{\pi_{k|\mathcal{A}}} \right\}.$$

If $N_j = 0$, set $\hat{\mu}_j(\hat{m}_j, \pi_{|\mathcal{A}}) = 0$. The corresponding cross-fitted model-assisted estimator is

$$\hat{\mu}_{cf}(\hat{m}, \pi_{|\mathcal{A}}) = \sum_{j=1}^M \frac{N_j}{N} \hat{\mu}_j(\hat{m}_j, \pi_{|\mathcal{A}}). \quad (8)$$

Remark 1. The estimator in (8) is well defined with probability one because $\mathbb{P}(\pi_{k|\mathcal{A}} > 0 \text{ for every } k \in S) = 1$. This does not imply that $\pi_{k|\mathcal{A}} > 0$ for every $k \in U$. For example, under SRSWOR with the uniform cross-fitting scheme of Algorithm 1, $\pi_{k|\mathcal{A}} = n_j/N_j$ for $k \in U_j$. A fold may contain no sampled unit, in which case $n_j = 0$ and $\pi_{k|\mathcal{A}} = 0$ throughout that fold. Nevertheless, no division by zero occurs in (8), because $S_j = \emptyset$ whenever $n_j = 0$.

Although (8) is defined as a weighted combination of fold-specific estimators, it can also be written in the usual model-assisted form. Let $\hat{m}^{(-k)}$ denote the prediction rule fitted without using the fold containing unit k . Then

$$\hat{\mu}_{cf}(\hat{m}, \pi_{|\mathcal{A}}) = \frac{1}{N} \left\{ \sum_{k \in U} \hat{m}^{(-k)}(\mathbf{x}_k) + \sum_{k \in S} \frac{y_k - \hat{m}^{(-k)}(\mathbf{x}_k)}{\pi_{k|\mathcal{A}}} \right\}. \quad (9)$$

Proposition 2. Assume that the cross-fitting scheme is honest and that $\pi_{k|\mathcal{A}} > 0$ for every $k \in U$ almost surely. Then, $\mathbb{E}_p \left\{ \hat{\mu}_{cf}(\hat{m}, \pi_{|\mathcal{A}}) \right\} = \mu$.

Proof. See Appendix F.1. □

Proposition 2 establishes exact design unbiasedness under (Ind- $\mathcal{A}$) and positivity of the conditional inclusion probabilities, regardless of the machine learning method used to construct the predictions. Thus, flexible prediction methods can be combined with exact finite-sample design validity. Traditional model-assisted estimators, including the GREG estimator, generally have a small-sample bias because the prediction rule is estimated from the same sample. Under the stated conditions, sample splitting removes this source of bias.

For ease of exposition, we will generally assume below that the first-order conditional inclusion probabilities are positive. Some of our theoretical results also cover cases where conditional inclusion probabilities may be equal to zero, although this requires additional technical arguments. In practice, cross-fitting schemes ensuring positivity should be preferred.

We next study the asymptotic properties of the conditional model-assisted estimator when the fitted prediction rule $\hat{m}$ converges to a deterministic target $\tilde{m}$. This target often corresponds to the prediction rule fitted at the population level, although this interpretation is not required. We impose the following conditions.

(M1) The fitted prediction rule $\hat{m}$ converges to its oracle target $\tilde{m}$ in the following finite-population L_2 sense:

$$\lim_{v \rightarrow \infty} \frac{1}{N_v} \sum_{k \in U_v} \mathbb{E}_p \left[\left\{ \hat{m}^{(-k)}(\mathbf{x}_k) - \tilde{m}(\mathbf{x}_k) \right\}^2 \right] = 0.$$

(M2) Let $\tilde{e}_k := y_k - \tilde{m}(\mathbf{x}_k)$ for $k \in U_v$, and consider the following two moment conditions:

(M2,a) The oracle predictor $\tilde{m}$ satisfies

$$\limsup_{v \rightarrow \infty} \frac{1}{N_v} \sum_{k \in U_v} \tilde{e}_k^2 < \infty.$$

(M2,b) The oracle predictor $\tilde{m}$ satisfies

$$\limsup_{v \rightarrow \infty} \frac{1}{N_v} \sum_{k \in U_v} \tilde{e}_k^4 < \infty.$$

Write $\hat{\mu}_{\text{ma}}(\tilde{m}, \pi_{|\mathcal{A}})$ for the estimator in (9) with every fitted prediction replaced by $\tilde{m}$. Theorem 3 establishes its first-order equivalence with the fitted estimator.

Theorem 3. Assume (Ind- $\mathcal{A}$), (D1), (CF2,a), (CF4), and (M1). Then,

$$\sqrt{n_v} \left\{ \hat{\mu}_{cf}(\hat{m}, \pi_{|\mathcal{A}}) - \hat{\mu}_{\text{ma}}(\tilde{m}, \pi_{|\mathcal{A}}) \right\} = o_{\mathbb{P}}(1).$$

Proof. See Appendix F.2. □

The next result shows that the oracle estimator is $\sqrt{n_v}$ -consistent for μ . Together, the two results imply the $\sqrt{n_v}$ -consistency of $\hat{\mu}_{cf}(\hat{m}, \pi_{|\mathcal{A}})$.

Theorem 4. Assume (D1), (CF2,a), (CF4), and (M2,a). Then,

$$\sqrt{n_v} \left\{ \hat{\mu}_{\text{ma}}(\tilde{m}, \pi_{|\mathcal{A}}) - \mu \right\} = \mathcal{O}_{\mathbb{P}}(1).$$

Proof. See Appendix F.3. □

4.2 Variance estimation

Consider the oracle estimator $\hat{\mu}_{\text{ma}}(\tilde{m}, \pi_{|\mathcal{A}})$, and set $\pi_{kk|\mathcal{A}} = \pi_{k|\mathcal{A}}$. If $\pi_{k|\mathcal{A}} > 0$ for every $k \in U_v$ almost surely, then the estimator is conditionally design-unbiased given $\mathcal{A}$. The law of total variance gives

$$\mathbb{V}_p \left\{ \hat{\mu}_{\text{ma}}(\tilde{m}, \pi_{|\mathcal{A}}) \right\} = \mathbb{E}_p \left[\mathbb{V}_p \left\{ \hat{\mu}_{\text{ma}}(\tilde{m}, \pi_{|\mathcal{A}}) \mid \mathcal{A} \right\} \right] = \mathbb{E}_p \left[\frac{1}{N_v^2} \sum_{k \in U_v} \sum_{l \in U_v} \Delta_{kl|\mathcal{A}} \frac{\tilde{e}_k}{\pi_{k|\mathcal{A}}} \frac{\tilde{e}_l}{\pi_{l|\mathcal{A}}} \right],$$

where $\tilde{e}_k = y_k - \tilde{m}(\mathbf{x}_k)$. If, in addition, $\pi_{kl|\mathcal{A}} > 0$ for every $k, l \in U_v$ almost surely, a design-unbiased estimator of this variance would be

$$\frac{1}{N_v^2} \sum_{k \in S_v} \sum_{l \in S_v} \frac{\Delta_{kl|\mathcal{A}}}{\pi_{kl|\mathcal{A}}} \frac{\tilde{e}_k}{\pi_{k|\mathcal{A}}} \frac{\tilde{e}_l}{\pi_{l|\mathcal{A}}}.$$

Because the oracle residuals are unknown, we replace them with the cross-fitted residuals $\hat{e}_{(-k)} := y_k - \hat{m}^{(-k)}(\mathbf{x}_k)$. These residuals are computed using prediction rules fitted without the fold containing unit k . Unlike ordinary in-sample residuals, they are therefore not artificially reduced by overfitting. This leads to

$$\hat{V}_{cf}(\hat{m}, \pi_{|\mathcal{A}}) := \frac{1}{N_v^2} \sum_{k \in S_v} \sum_{l \in S_v} \frac{\Delta_{kl|\mathcal{A}}}{\pi_{kl|\mathcal{A}}} \frac{\hat{e}_{(-k)}}{\pi_{k|\mathcal{A}}} \frac{\hat{e}_{(-l)}}{\pi_{l|\mathcal{A}}}.$$

No zero denominator occurs in this sum almost surely, since $I_k I_l = 1$ implies $\pi_{kl|\mathcal{A}} > 0$, and hence $\pi_{k|\mathcal{A}} > 0$ and $\pi_{l|\mathcal{A}} > 0$. This estimator is generally not design-unbiased for either the variance of the feasible estimator or that of the oracle estimator. The next theorem allows the conditional inclusion probabilities to be zero on events whose probabilities decrease sufficiently fast, and establishes consistency for $\mathbb{V}_p\{\hat{\mu}_{\text{ma}}(\tilde{m}, \pi_{|\mathcal{A}})\}$.

Theorem 5. Assume (CF2,a), (CF2,b), (CF4), (CF5), (CF6), (M1), and (M2,b). Suppose, moreover, that there exists a constant $c > 0$ such that

$$n_v \mathbb{V}_p \left\{ \hat{\mu}_{\text{ma}}(\tilde{m}, \pi_{|\mathcal{A}}) \mid \mathcal{A} \right\} \xrightarrow[v \rightarrow \infty]{\mathbb{P}} c.$$

Then,

$$n_v \left[\hat{V}_{cf}(\hat{m}, \pi_{|\mathcal{A}}) - \mathbb{V}_p \left\{ \hat{\mu}_{\text{ma}}(\tilde{m}, \pi_{|\mathcal{A}}) \right\} \right] = o_{\mathbb{P}}(1).$$

Proof. See Appendix F.4. □

The preceding results can now be combined to establish the asymptotic distribution of the feasible estimator. If the oracle estimator is asymptotically normal, as expected under standard conditions, then the feasible estimator is asymptotically normal as well. Studentized statistics and confidence intervals below are defined arbitrarily when the estimated variance is nonpositive.

Corollary 1. Suppose that the conditions of Theorems 3, 4, and 5 hold. If

$$\frac{\hat{\mu}_{\text{ma}}(\tilde{m}, \pi_{|\mathcal{A}}) - \mu}{\sqrt{\mathbb{V}_p \{ \hat{\mu}_{\text{ma}}(\tilde{m}, \pi_{|\mathcal{A}}) \}}} \xrightarrow[v \rightarrow \infty]{\mathcal{L}} \mathcal{N}(0, 1),$$

then

$$\frac{\hat{\mu}_{\text{cf}}(\hat{m}, \pi_{|\mathcal{A}}) - \mu}{\sqrt{\hat{V}_{\text{cf}}(\hat{m}, \pi_{|\mathcal{A}})}} \xrightarrow[v \rightarrow \infty]{\mathcal{L}} \mathcal{N}(0, 1).$$

Proof. See Appendix F.5. □

It follows from Corollary 1 that an asymptotically valid $100(1 - \alpha)\%$ confidence interval for μ is

$$\text{CI}_{1-\alpha} \left(\hat{\mu}_{\text{cf}}(\hat{m}, \pi_{|\mathcal{A}}), \hat{V}_{\text{cf}}(\hat{m}, \pi_{|\mathcal{A}}) \right) := \hat{\mu}_{\text{cf}}(\hat{m}, \pi_{|\mathcal{A}}) \pm z_{1-\alpha/2} \sqrt{\hat{V}_{\text{cf}}(\hat{m}, \pi_{|\mathcal{A}})}.$$

5 Unconditional cross-fitted model-assisted estimation

5.1 Point estimation

The preceding section provides a complete inferential framework based on the conditional inclusion probabilities. Under SRSWOR and conditional Poisson sampling, which are exponential sampling designs with symmetric support, the first- and second-order conditional inclusion probabilities can be obtained. Their calculation and implementation nevertheless require additional work, particularly for variance estimation. It is therefore natural to ask whether the original, unconditional inclusion probabilities $\pi_k = \mathbb{P}(I_k = 1)$ can be used instead. Retaining the original inclusion probabilities avoids computing $\pi_{k|\mathcal{A}}$ and $\pi_{kl|\mathcal{A}}$ altogether and leads to the estimator

$$\hat{\mu}_{\text{cf}}(\hat{m}, \pi) = \frac{1}{N} \left\{ \sum_{k \in U} \hat{m}^{(-k)}(\mathbf{x}_k) + \sum_{k \in S} \frac{y_k - \hat{m}^{(-k)}(\mathbf{x}_k)}{\pi_k} \right\}. \quad (10)$$

Using unconditional inclusion probabilities generally sacrifices exact design unbiasedness. Unbiasedness is retained if $\pi_{k|\mathcal{A}} = \pi_k$ for every $k \in U_v$ almost surely, in which case the conditional and unconditional estimators coincide. This occurs, for example, under Poisson sampling with a uniform partition and $\mathcal{A} = \delta$. More generally, under $(\text{Ind-}\mathcal{A})$, the design bias of the unconditional estimator is

$$\text{Bias}_p \{ \hat{\mu}_{\text{cf}}(\hat{m}, \pi) \} = -\frac{1}{N_v} \sum_{k \in U_v} \frac{1}{\pi_k} \text{Cov}_p \{ \hat{m}^{(-k)}(\mathbf{x}_k), \pi_{k|\mathcal{A}} \}.$$

By comparison, (4) shows that the bias of the non-cross-fitted estimator depends on $\text{Cov}_p \{ \hat{m}(\mathbf{x}_k), I_k \}$. The distinction is important: $\hat{m}(\mathbf{x}_k)$ is fitted using the same sample and therefore depends directly on I_k , whereas $\hat{m}^{(-k)}(\mathbf{x}_k)$ is conditionally independent of I_k given

$\mathcal{A}$. Its remaining dependence on I_k is mediated by $\mathcal{A}$ and captured by its covariance with $\pi_{k|\mathcal{A}}$. Thus, although cross-fitting with unconditional inclusion probabilities does not eliminate the bias exactly, it can be expected to make it small in many applications.

To bound the bias, we first write

$$\text{Cov}_p \left\{ \hat{m}^{(-k)}(\mathbf{x}_k), \pi_{k|\mathcal{A}} \right\} = \text{Cov}_p \left\{ \hat{m}^{(-k)}(\mathbf{x}_k) - \tilde{m}(\mathbf{x}_k), \pi_{k|\mathcal{A}} - \pi_k \right\},$$

and then apply the Cauchy–Schwarz inequality twice under (D2) to obtain

$$\begin{aligned} |\text{Bias}_p \{ \hat{\mu}_{cf}(\hat{m}, \pi) \}| &\leq \frac{1}{\lambda} \left\{ \max_{k \in U_v} \mathbb{E}_p \left[(\pi_{k|\mathcal{A}} - \pi_k)^2 \right] \right\}^{1/2} \\ &\quad \times \left\{ \frac{1}{N_v} \sum_{k \in U_v} \mathbb{E}_p \left[\left\{ \hat{m}^{(-k)}(\mathbf{x}_k) - \tilde{m}(\mathbf{x}_k) \right\}^2 \right] \right\}^{1/2}. \end{aligned}$$

The first term on the right-hand side is $\mathcal{O}(n_v^{-1/2})$ by (CF3) and, under (M1), the second term is $o(1)$; this shows directly that the bias is negligible at the first-order scale.

We next show that the unconditional cross-fitted estimator is first-order equivalent to its oracle, $\hat{\mu}_{ma}(\tilde{m}, \pi)$. Thus, estimating the prediction rule has no first-order effect. Unlike the corresponding result for the conditionally weighted estimator, this result requires the proximity condition (CF3).

Theorem 6. Assume (Ind- $\mathcal{A}$), (D1), (D2), (CF3), (CF4), and (M1). Then,

$$\sqrt{n_v} \{ \hat{\mu}_{cf}(\hat{m}, \pi) - \hat{\mu}_{ma}(\tilde{m}, \pi) \} = o_{\mathbb{P}}(1).$$

Proof. See Appendix G.1. □

Remark 2. The proof reveals a trade-off between the proximity of the conditional and unconditional inclusion probabilities and the convergence of the prediction rule. More precisely, the part involving these probability differences is controlled by

$$\left\{ \sqrt{n_v} \max_{k \in U_v} |\pi_{k|\mathcal{A}} - \pi_k| \right\} \left\{ \frac{1}{N_v} \sum_{k \in U_v} \left[\hat{m}^{(-k)}(\mathbf{x}_k) - \tilde{m}(\mathbf{x}_k) \right]^2 \right\}^{1/2}.$$

Consequently, the proximity condition (CF3) may be weakened when the prediction rule converges sufficiently fast.

Because $\tilde{m}$ is deterministic, $\hat{\mu}_{ma}(\tilde{m}, \pi)$ is a classical difference estimator. Its first-order properties therefore follow from standard design-based arguments: it is $\sqrt{n_v}$ -consistent and, under many sampling designs, asymptotically normal.

Theorem 7. Assume (D1)–(D3) and (M2,a). Then,

$$\hat{\mu}_{\text{ma}}(\tilde{m}, \pi) - \mu = \mathcal{O}_{\mathbb{P}}\left(\frac{1}{\sqrt{N_v}}\right).$$

Proof. See Appendix G.2. □

Theorems 6 and 7 together establish the $\sqrt{n_v}$ -consistency of $\hat{\mu}_{cf}(\hat{m}, \pi)$ without imposing a specific convergence rate on $\hat{m}$ toward $\tilde{m}$. When the oracle central limit theorem holds, the predictive quality of $\tilde{m}$ affects the first-order distribution only through the asymptotic variance. Optimality results are given and established in Appendix C.

5.2 Variance estimation

Section 4 introduced the variance estimator $\hat{V}_{cf}(\hat{m}, \pi_{|\mathcal{A}})$ for the conditional estimator $\hat{\mu}_{cf}(\hat{m}, \pi_{|\mathcal{A}})$. The conditional and unconditional point estimators are first-order equivalent to different oracle difference estimators and may therefore have different asymptotic variances. We now construct a variance estimator for $\hat{\mu}_{cf}(\hat{m}, \pi)$. Conditions under which the two point estimators are asymptotically equivalent are given in Section 6. Set $\pi_{kk} = \pi_k$, so that $\Delta_{kk} = \pi_k(1 - \pi_k)$. Whenever $\pi_{kl} > 0$ for every $k, l \in U_v$, the natural unconditional analogue of $\hat{V}_{cf}(\hat{m}, \pi_{|\mathcal{A}})$ is

$$\hat{V}_{cf}(\hat{m}, \pi) := \frac{1}{N_v^2} \sum_{k \in S_v} \sum_{l \in S_v} \frac{\Delta_{kl}}{\pi_{kl}} \frac{\hat{e}_{(-k)}}{\pi_k} \frac{\hat{e}_{(-l)}}{\pi_l}.$$

This estimator is closely related to that proposed by Dagdoug et al. (2023). Under leave-one-out cross-fitting, it reduces to the estimator of Opsomer and Miller (2005), who established its properties for local polynomial regression. The leave-one-out version was also recently studied by Bouhadra and Dagdoug (2026) for high-dimensional linear models. For general cross-fitting, however, Dagdoug et al. (2023) investigated only its empirical performance and did not establish consistency. The next result fills this gap for generic prediction rules and an arbitrary number of folds.

Theorem 8. Assume (D1)–(D4), (M1), and (M2,b). Then,

$$n_v \left| \hat{V}_{cf}(\hat{m}, \pi) - \mathbb{V}_p \{ \hat{\mu}_{\text{ma}}(\tilde{m}, \pi) \} \right| = o_{\mathbb{P}}(1).$$

Proof. See Appendix G.3. □

The following corollary establishes valid inference based on $\{\hat{\mu}_{cf}(\hat{m}, \pi), \hat{V}_{cf}(\hat{m}, \pi)\}$ without requiring first- or second-order conditional inclusion probabilities.

Corollary 2. *Suppose that the conditions of Theorems 6 and 8 hold and that*

$$\liminf_{v \rightarrow \infty} n_v \mathbb{V}_p \{ \hat{\mu}_{\text{ma}}(\tilde{m}, \pi) \} > 0. \quad (11)$$

If

$$\frac{\hat{\mu}_{\text{ma}}(\tilde{m}, \pi) - \mu}{\sqrt{\mathbb{V}_p \{ \hat{\mu}_{\text{ma}}(\tilde{m}, \pi) \}}} \xrightarrow[v \rightarrow \infty]{\mathcal{L}} \mathcal{N}(0, 1),$$

then

$$\frac{\hat{\mu}_{cf}(\hat{m}, \pi) - \mu}{\sqrt{\hat{V}_{cf}(\hat{m}, \pi)}} \xrightarrow[v \rightarrow \infty]{\mathcal{L}} \mathcal{N}(0, 1).$$

Proof. See Appendix G.4. □

6 Conditional versus unconditional cross-fitting

Sections 4 and 5 introduced the conditional estimator $\hat{\mu}_{cf}(\hat{m}, \pi_{|\mathcal{A}})$ in (9) and the unconditional estimator $\hat{\mu}_{cf}(\hat{m}, \pi)$ in (10). They use the same out-of-fold predictions and differ only in the inclusion probabilities used in the residual correction. Consequently, they are first-order equivalent to different oracle estimators. We now compare their asymptotic properties.

Let

$$B_{\mathcal{A}} := \mathbb{E}_p [\hat{\mu}_{\text{ma}}(\tilde{m}, \pi) - \mu \mid \mathcal{A}]$$

denote the conditional bias of the unconditional oracle estimator. Since this estimator is unconditionally design-unbiased, $\mathbb{E}_p(B_{\mathcal{A}}) = 0$. The magnitude of its fluctuations determines whether the conditional and unconditional estimators are first-order equivalent.

Theorem 9. *Suppose that (M2,a) and the conditions of Theorems 3 and 6 hold. Then,*

$$\sqrt{n_v} \left\{ \hat{\mu}_{cf}(\hat{m}, \pi) - \hat{\mu}_{cf}(\hat{m}, \pi_{|\mathcal{A}}) \right\} = \sqrt{n_v} B_{\mathcal{A}} + o_{\mathbb{P}}(1).$$

Consequently, the two estimators are first-order asymptotically equivalent if and only if $\sqrt{n_v} B_{\mathcal{A}} = o_{\mathbb{P}}(1)$.

Proof. See Appendix H.1. □

Condition (CF3) controls the discrepancy between the conditional and unconditional inclusion probabilities and is used in the expansion in Theorem 9. It does not, however, imply that $\sqrt{n_v} B_{\mathcal{A}} = o_{\mathbb{P}}(1)$. Supplementary Section H.3 gives an SRSWOR example for which $\sqrt{n_v} B_{\mathcal{A}} \xrightarrow[v \rightarrow \infty]{\mathcal{L}} \mathcal{N}(0, 1 - \pi_*)$, showing that the two estimators need not be first-order equivalent.

When equivalence fails, the following result provides a general comparison of their first-order variances.

Theorem 10. Assume (D1)–(D2), (CF2,c), (CF3), (CF4), and (M2,a). Then,

$$n_v \left[\mathbb{V}_p \{ \hat{\mu}_{\text{ma}}(\tilde{m}, \pi) \} - \mathbb{V}_p \{ \hat{\mu}_{\text{ma}}(\tilde{m}, \pi_{|\mathcal{A}}) \} \right] = n_v \mathbb{E}_p \left[\left\{ \hat{\mu}_{\text{ma}}(\tilde{m}, \pi_{|\mathcal{A}}) - \hat{\mu}_{\text{ma}}(\tilde{m}, \pi) \right\}^2 \right] + o(1),$$

where the leading term is $\mathcal{O}(1)$.

Proof. See Appendix H.2. □

Since the leading term is nonnegative, Theorem 10 implies

$$\liminf_{v \rightarrow \infty} n_v \left[\mathbb{V}_p \{ \hat{\mu}_{\text{ma}}(\tilde{m}, \pi) \} - \mathbb{V}_p \{ \hat{\mu}_{\text{ma}}(\tilde{m}, \pi_{|\mathcal{A}}) \} \right] \geq 0.$$

Thus, the conditional oracle is at least as efficient as the unconditional oracle to first order. Under the preceding oracle-equivalence conditions and suitable uniform integrability conditions, this comparison extends to the feasible estimators, and their first-order variances coincide when they are first-order equivalent. Supplementary Section H.4 gives a condition for equivalence under SRSWOR. This first-order comparison does not give a second-order ordering. Appendix C gives conditions under which the proposed estimators asymptotically attain the Godambe–Joshi lower bound.

7 Simulation study

We evaluate the finite-sample performance of $\hat{\mu}_{\text{ma}}(\hat{m}, \pi)$, $\hat{\mu}_{\text{cf}}(\hat{m}, \pi^{(\delta)})$, $\hat{\mu}_{\text{cf}}(\hat{m}, \pi_{|\mathcal{A}})$, and $\hat{\mu}_{\text{cf}}(\hat{m}, \pi)$, together with their variance estimators. Point estimators are evaluated in terms of relative bias and efficiency, and variance estimators in terms of relative bias and coverage of Wald confidence intervals.

We generate finite populations of sizes $N \in \{1000, 2500, 5000\}$. For each unit, the $p = 20$ components of $\mathbf{x}_k = (x_{k1}, \dots, x_{kp})^\top$ are generated independently from a normal distribution with mean 5 and variance 2. The survey variable is generated under either the *linear model* $y_k = 4x_{k1} + 3x_{k2} + 3x_{k3} + 2x_{k4} + 5x_{k5} + 7x_{k6} + \varepsilon_k$, with $\varepsilon_k \sim \mathcal{N}(0, 225)$, or the *nonlinear model* $y_k = 10 + 2I(x_{k1} > 5) + (x_{k2} + x_{k3} - 10)^2 + \varepsilon_k$, with $\varepsilon_k \sim \mathcal{N}(0, 9)$. Each population is held fixed throughout the Monte Carlo experiment.

From each population, we draw $R = 20,000$ independent SRSWOR samples of size $n = 0.2N$. Four prediction methods are considered: linear regression; random forests with $B = 200$ trees and

minimum terminal node size $n_0 = 5$; K -nearest neighbors with $K = 5$ and Euclidean distance; and regression trees with minimum terminal node size $n_0 = 5$.

For each sample, we compute four model-assisted estimators. First, $\hat{\mu}_{ma}(\hat{m}, \pi)$ in (3) is the standard estimator without cross-fitting, based on the original inclusion probabilities. Second, $\hat{\mu}_{cf}(\hat{m}, \pi^{(\delta)})$ in (8) uses the design-aware procedure of Section 3.1 and the conditional inclusion probabilities $\pi^{(\delta)}$. We use $M = 2$ population folds with $(N_1, N_2) = N(0.5, 0.5)$ and consider both $(n_1, n_2) = n(0.5, 0.5)$ and $(n_1, n_2) = n(0.8, 0.2)$. Third, $\hat{\mu}_{cf}(\hat{m}, \pi_{|\mathcal{A}})$ in (9) uses the design-agnostic procedure of Section 3.2 (Algorithm 1), with $M = 2$ folds and the conditional inclusion probabilities $\pi_{|\mathcal{A}}$ with $\mathcal{A} = (\mathbf{n}, \delta)$. Finally, $\hat{\mu}_{cf}(\hat{m}, \pi)$ in (10) uses the same design-agnostic procedure but retains the original inclusion probabilities π .

For a generic estimator $\hat{\mu}_m$ among the four above, let $\hat{\mu}_m^{(r)}$ denote its value at replication r . We consider the Monte Carlo percent relative bias RB, relative efficiency RE using the Horvitz–Thompson estimator as benchmark, percent relative bias of the variance estimator RB-Var, and coverage probability CP:

$$\begin{aligned} \text{RB}(\hat{\mu}_m) &= 100\% \times \frac{1}{R} \sum_{r=1}^R \frac{\hat{\mu}_m^{(r)} - \mu}{\mu}, \\ \text{RE}(\hat{\mu}_m) &= 100\% \times \frac{\mathbb{E}_{MC} [(\hat{\mu}_m - \mu)^2]}{\mathbb{E}_{MC} [(\hat{\mu}_\pi - \mu)^2]}, \\ \text{RB-Var}(\hat{\mu}_m) &= 100\% \times \frac{\mathbb{E}_{MC} [\hat{\mathbb{V}}(\hat{\mu}_m)] - \mathbb{V}_{MC}(\hat{\mu}_m)}{\mathbb{V}_{MC}(\hat{\mu}_m)}, \\ \text{CP}(\hat{\mu}_m) &= 100\% \times \mathbb{E}_{MC} \left[I \left\{ \hat{\mu}_m \pm 1.96 \sqrt{\hat{\mathbb{V}}(\hat{\mu}_m)} \ni \mu \right\} \right]. \end{aligned}$$

Table 1 shows that all estimators are essentially unbiased under the linear relationship. Under the nonlinear relationship, the standard model-assisted estimator has small but persistent biases with flexible learners, whereas the cross-fitted estimators remain nearly unbiased.

When the linear model is correctly specified, cross-fitting with balanced folds entails a modest efficiency loss that decreases as the sample size increases. Under the nonlinear relationship, cross-fitting can improve efficiency with flexible learners in larger samples. For random forests with $n = 1000$, the relative efficiency decreases from 31.2% for the standard estimator to 25.2% for design-aware cross-fitting with equal folds. With KNN, persistent bias makes the standard estimator less efficient as n increases, whereas the cross-fitted estimators remain stable. Thus, even a small persistent bias can dominate the mean squared error in large samples.

Balanced sample fold sizes are important for the design-aware procedure. For random forests under the linear relationship with $n = 200$, moving from equal to unequal folds increases the relative efficiency from 68.3% to 111.3%. Apart from this effect, the design-aware and

Table 1. Relative bias and relative efficiency of the estimators under the linear and nonlinear relationships.

Model	Method	n	Linear relationship			Nonlinear relationship		
			200	500	1000	200	500	1000
Linear Regression	$\hat{\mu}_{ma}(\hat{m}, \pi)$	RB	0.0	0.0	0.0	-0.4	-0.2	-0.1
		RE	39.3	34.2	34.2	112.2	105.2	102.5
	$\hat{\mu}_{cf}(\hat{m}, \pi^{(\delta)})$ (equal)	RB	0.0	0.0	0.0	0.0	0.0	0.0
		RE	44.7	35.9	34.8	129.9	110.1	105.6
	$\hat{\mu}_{cf}(\hat{m}, \pi^{(\delta)})$ (unequal)	RB	0.0	0.0	0.0	0.0	0.0	0.0
		RE	78.1	60.8	59.6	224.6	188.5	179.9
	$\hat{\mu}_{cf}(\hat{m}, \pi)$	RB	0.0	0.0	0.0	-0.1	0.0	0.0
		RE	44.4	35.1	34.8	133.4	110.5	105.9
	$\hat{\mu}_{cf}(\hat{m}, \pi_{ \mathcal{A}})$	RB	0.0	0.0	0.0	-0.1	0.0	0.0
		RE	44.1	35.2	34.9	133.2	110.6	106.0
Random Forests	$\hat{\mu}_{ma}(\hat{m}, \pi)$	RB	0.0	0.0	0.0	-1.1	-0.9	-0.7
		RE	60.8	49.6	47.7	45.4	34.1	31.2
	$\hat{\mu}_{cf}(\hat{m}, \pi^{(\delta)})$ (equal)	RB	0.0	0.0	0.0	0.0	0.0	0.0
		RE	68.3	54.6	51.4	59.7	36.7	25.2
	$\hat{\mu}_{cf}(\hat{m}, \pi^{(\delta)})$ (unequal)	RB	0.0	0.0	0.0	0.0	0.0	0.0
		RE	111.3	91.5	86.2	87.4	54.1	39.2
	$\hat{\mu}_{cf}(\hat{m}, \pi)$	RB	0.0	0.0	0.0	0.0	0.0	0.0
		RE	68.8	54.3	51.6	59.5	37.0	25.0
	$\hat{\mu}_{cf}(\hat{m}, \pi_{ \mathcal{A}})$	RB	0.0	0.0	0.0	0.0	0.0	0.0
		RE	68.9	54.3	51.7	59.4	37.0	25.0
KNN	$\hat{\mu}_{ma}(\hat{m}, \pi)$	RB	0.0	0.0	0.0	-1.7	-2.1	-1.7
		RE	72.0	64.5	62.5	89.7	134.6	161.4
	$\hat{\mu}_{cf}(\hat{m}, \pi^{(\delta)})$ (equal)	RB	0.0	0.0	0.0	0.0	0.0	0.0
		RE	77.4	69.1	65.5	88.9	84.6	80.7
	$\hat{\mu}_{cf}(\hat{m}, \pi^{(\delta)})$ (unequal)	RB	0.0	0.0	0.0	0.0	0.0	0.0
		RE	130.4	115.7	111.5	148.9	140.4	135.8
	$\hat{\mu}_{cf}(\hat{m}, \pi)$	RB	0.0	0.0	0.0	-0.1	0.0	0.0
		RE	78.1	68.3	65.7	89.4	84.4	80.6
	$\hat{\mu}_{cf}(\hat{m}, \pi_{ \mathcal{A}})$	RB	0.0	0.0	0.0	-0.1	0.0	0.0
		RE	77.6	68.3	65.6	89.5	84.4	80.8
Regression Trees	$\hat{\mu}_{ma}(\hat{m}, \pi)$	RB	0.0	0.0	0.0	-1.4	-1.3	-1.0
		RE	92.3	78.4	74.9	72.3	65.9	60.0
	$\hat{\mu}_{cf}(\hat{m}, \pi^{(\delta)})$ (equal)	RB	0.0	0.0	0.0	0.0	0.0	0.0
		RE	103.3	87.0	82.5	86.8	57.0	42.5
	$\hat{\mu}_{cf}(\hat{m}, \pi^{(\delta)})$ (unequal)	RB	0.0	0.0	0.0	0.0	0.0	0.0
		RE	169.6	143.5	138.4	130.7	85.3	66.8
	$\hat{\mu}_{cf}(\hat{m}, \pi)$	RB	0.0	0.0	0.0	0.0	0.0	0.0
		RE	103.6	84.3	82.3	88.9	58.2	42.1
	$\hat{\mu}_{cf}(\hat{m}, \pi_{ \mathcal{A}})$	RB	0.0	0.0	0.0	0.0	0.0	0.0
		RE	103.9	84.4	82.5	88.8	58.2	42.2

Table 2. Relative bias of the variance estimators and coverage probabilities under the linear and nonlinear relationships.

Model	Method	n	Linear relationship			Nonlinear relationship		
			200	500	1000	200	500	1000
Linear Regression	$\hat{\mu}_{ma}(\hat{m}, \pi)$	RB-Var	-19.1	-7.1	-4.6	-20.2	-8.9	-5.0
		CP	92.1	94.1	94.4	90.8	93.4	94.1
	$\hat{\mu}_{cf}(\hat{m}, \pi^{(\delta)})$ (equal)	RB-Var	-0.2	0.5	-0.5	0.2	-0.3	-1.0
		CP	94.9	95.0	94.8	94.5	94.7	94.6
	$\hat{\mu}_{cf}(\hat{m}, \pi^{(\delta)})$ (unequal)	RB-Var	0.2	0.7	-1.1	0.4	-0.5	-1.6
		CP	94.6	94.9	94.8	93.0	94.1	94.2
Random Forests	$\hat{\mu}_{cf}(\hat{m}, \pi)$	RB-Var	1.2	1.7	0.1	0.6	-0.4	-0.6
		CP	95.0	95.4	94.9	94.3	94.8	94.7
	$\hat{\mu}_{cf}(\hat{m}, \pi _{\mathcal{A}})$	RB-Var	0.2	1.4	-0.3	0.0	-0.7	-0.9
		CP	94.8	95.2	94.9	94.2	94.7	95.8
	$\hat{\mu}_{ma}(\hat{m}, \pi)$	RB-Var	-66.9	-67.3	-68.1	-53.0	-56.7	-60.1
		CP	74.1	73.4	72.9	75.9	68.4	59.7
KNN	$\hat{\mu}_{cf}(\hat{m}, \pi^{(\delta)})$ (equal)	RB-Var	0.2	0.7	-0.8	-1.4	-2.7	-3.1
		CP	94.9	95.1	94.7	94.1	94.2	94.3
	$\hat{\mu}_{cf}(\hat{m}, \pi^{(\delta)})$ (unequal)	RB-Var	0.8	-0.5	-1.6	0.3	-0.7	-2.4
		CP	94.6	94.9	94.8	93.0	94.1	94.2
	$\hat{\mu}_{cf}(\hat{m}, \pi)$	RB-Var	1.0	0.8	-0.1	-0.8	-2.8	-2.1
		CP	95.0	95.1	94.9	94.1	94.2	94.5
Regression Trees	$\hat{\mu}_{cf}(\hat{m}, \pi _{\mathcal{A}})$	RB-Var	-0.2	0.5	-0.5	-1.7	-3.1	-2.3
		CP	94.9	95.0	94.9	94.0	94.2	94.6
	$\hat{\mu}_{ma}(\hat{m}, \pi)$	RB-Var	-32.0	-31.8	-32.5	-24.6	-26.4	-26.8
		CP	89.3	89.5	89.1	85.3	74.7	68.1
	$\hat{\mu}_{cf}(\hat{m}, \pi^{(\delta)})$ (equal)	RB-Var	0.2	0.3	0.6	1.2	0.1	-0.2
		CP	94.7	94.8	95.0	94.4	94.9	94.9
Regression Trees	$\hat{\mu}_{cf}(\hat{m}, \pi^{(\delta)})$ (unequal)	RB-Var	-0.7	0.3	-0.9	0.5	0.3	-1.1
		CP	94.5	94.8	94.6	93.2	94.2	94.3
	$\hat{\mu}_{cf}(\hat{m}, \pi)$	RB-Var	1.3	1.2	0.8	1.9	0.2	0.6
		CP	95.0	95.0	95.1	94.7	94.8	94.9
	$\hat{\mu}_{cf}(\hat{m}, \pi _{\mathcal{A}})$	RB-Var	0.5	0.8	0.6	1.5	0.1	0.4
		CP	94.5	95.0	95.0	94.6	94.7	94.9
Regression Trees	$\hat{\mu}_{ma}(\hat{m}, \pi)$	RB-Var	-80.0	-80.6	-80.9	-59.0	-65.6	-69.8
		CP	61.4	61.5	61.1	71.7	58.0	49.7
	$\hat{\mu}_{cf}(\hat{m}, \pi^{(\delta)})$ (equal)	RB-Var	-0.8	-0.7	-1.1	-1.1	-2.5	-2.2
		CP	94.8	94.9	94.8	94.2	94.2	94.6
	$\hat{\mu}_{cf}(\hat{m}, \pi^{(\delta)})$ (unequal)	RB-Var	0.7	0.5	-1.5	0.3	0.9	-2.1
		CP	94.5	94.8	94.7	94.1	94.7	94.5
Regression Trees	$\hat{\mu}_{cf}(\hat{m}, \pi)$	RB-Var	0.2	2.2	-0.2	-3.4	-2.7	-0.8
		CP	94.8	95.2	94.8	93.8	94.4	94.7
	$\hat{\mu}_{cf}(\hat{m}, \pi _{\mathcal{A}})$	RB-Var	-0.6	1.9	-0.5	-4.1	-2.9	-1.0
		CP	94.8	95.2	94.8	93.7	94.4	94.8

design-agnostic procedures perform similarly, and the choice between π and $\pi|_{\mathcal{A}}$ is essentially immaterial.

Table 2 shows that the standard model-assisted variance estimator can be severely biased downward. Under the linear relationship, this bias decreases with sample size for linear regression but remains large for random forests and regression trees. The resulting confidence intervals undercover substantially: under the nonlinear relationship with $n = 1000$, coverage is 59.7% for random forests and 49.7% for regression trees.

Cross-fitting largely eliminates this downward variance bias and yields coverage generally around 94%–95%. Unequal folds can slightly reduce coverage for the design-aware procedure, but the effect is much smaller than for point estimation. The design-aware and design-agnostic procedures again perform similarly.

8 Final remarks

In this paper, we developed an agnostic framework for model-assisted estimation with modern machine learning methods. The proposed cross-fitting procedures accommodate flexible prediction rules while ensuring design consistency, asymptotic normality, and valid variance estimation under general conditions. The simulations show negligible bias, competitive or improved efficiency, and variance estimators with small bias, yielding Wald confidence intervals with coverage close to the nominal level, particularly for highly adaptive learners.

Although our development focuses on the population mean, the framework extends naturally to finite-population parameters defined through estimating equations. Let θ_N solve the census estimating equation

$$\Psi(\theta) = \frac{1}{N} \sum_{k \in U} \psi(y_k, \theta) = 0,$$

where $\psi : \mathbb{R} \times \Theta \rightarrow \mathbb{R}$ is known. For example, $\psi(y, \theta) = y - \theta$ yields the population mean, whereas $\psi(y, \theta) = \mathbf{1}(y \leq t) - \theta$ yields the finite-population distribution function at t .

Let $\hat{g}^{(-k)}(\mathbf{x}_k, \theta)$ be a cross-fitted predictor of $\psi(y_k, \theta)$ constructed without observations from the fold containing unit k . A cross-fitted model-assisted estimator $\hat{\theta}_{cf}$ can then be defined as a solution to $\hat{\Psi}_{cf}(\hat{\theta}_{cf}) = 0$, where

$$\hat{\Psi}_{cf}(\theta) = \frac{1}{N} \left\{ \sum_{k \in U} \hat{g}^{(-k)}(\mathbf{x}_k, \theta) + \sum_{k \in S} \frac{\psi(y_k, \theta) - \hat{g}^{(-k)}(\mathbf{x}_k, \theta)}{\pi_k} \right\}.$$

Model-assisted estimation of finite-population distribution functions using auxiliary information was studied by Rao et al. (1990) under simple regression models, including ratio and difference-

type estimators. The formulation above accommodates general prediction methods and addresses data reuse through cross-fitting. Encouraging preliminary simulations motivate a detailed study of these extensions in future work.

References

- Boistard, H., Lopuhaä, H. P. and Ruiz-Gazen, A. (2017), ‘Functional central limit theorems for single-stage sampling designs’, *Annals of Statistics* **45**(4), 1728–1758.
- Bouhadra, K. and Dagdou, M. (2026), ‘High-dimensional variance estimation for the generalized regression estimator’, *arXiv preprint arXiv:2607.05570*.
- Breidt, F. J., Claeskens, G. and Opsomer, J. D. (2005), ‘Model-assisted estimation for complex surveys using penalized splines’, *Biometrika* **92**(4), 831–846.
- Breidt, F. J. and Opsomer, J. D. (2000), ‘Local polynomial regression estimators in survey sampling’, *The Annals of Statistics* **28**(4), 1026–1053.
- Breidt, F. J. and Opsomer, J. D. (2017), ‘Model-assisted survey estimation with modern prediction techniques’, *Statistical Science* **32**(2), 190–205.
- Brewer, K. R. W. (1963), ‘Ratio estimation and finite populations: Some results deducible from the assumption of an underlying stochastic process’, *Australian Journal of Statistics* **5**(3), 93–105.
- Cassel, C.-M., Särndal, C.-E. and Wretman, J. H. (1976), ‘Some results on generalized difference estimation and generalized regression estimation for finite populations’, *Biometrika* **63**(3), 615–620.
- Chambers, R. L. and Clark, R. G. (2012), *An Introduction to Model-Based Survey Sampling with Applications*, Oxford University Press, Oxford.
- Chambers, R. L., Dorfman, A. H. and Wehrly, T. E. (1993), ‘Bias robust estimation in finite populations using nonparametric calibration’, *Journal of the American Statistical Association* **88**(421), 268–277.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and Robins, J. (2018), ‘Double/debiased machine learning for treatment and structural parameters’, *The Econometrics Journal* **21**(1), C1–C68.
- Cochran, W. G. (1942), ‘Sampling theory when the sampling units are of unequal sizes’, *Journal of the American Statistical Association* **37**(218), 199–212.
- Dagdoug, M., Goga, C. and Haziza, D. (2023), ‘Model-assisted estimation through random forests in finite population sampling’, *Journal of the American Statistical Association* **118**(542), 1234–1251.
- Dagdoug, M. and Haziza, D. (2026), ‘Machine learning methods for finite population parameter estimation in survey sampling’, *arXiv preprint arXiv:2604.01160*.
- Dorfman, A. H. (1992), Non-parametric regression for estimating totals in finite populations, in ‘Proceedings of the Section on Survey Research Methods’, American Statistical Association, pp. 622–625.

- Dorfman, A. H. and Hall, P. (1993), ‘Estimators of the finite population distribution function using nonparametric regression’, *The Annals of Statistics* **21**(3), 1452–1475.
- Goga, C. (2005), ‘Réduction de la variance dans les sondages en présence d’information auxiliaire: Une approche non paramétrique par splines de régression’, *Canadian Journal of Statistics* **33**(2), 163–180.
- Hansen, M. H., Madow, W. G. and Tepping, B. J. (1983), ‘An evaluation of model-dependent and probability-sampling inferences in sample surveys’, *Journal of the American Statistical Association* **78**(384), 776–793.
- Isaki, C. T. and Fuller, W. A. (1982), ‘Survey design under the regression superpopulation model’, *Journal of the American Statistical Association* **77**(377), 89–96.
- Kuk, A. Y. C. (1993), ‘A kernel method for estimating finite population functions using auxiliary information’, *Biometrika* **80**(2), 385–392.
- Lu, X., Shi, L., Liu, H. and Ding, P. (2025), ‘Conditional cross-fitting for unbiased machine-learning-assisted covariate adjustment in randomized experiments’, *arXiv preprint arXiv:2508.15664*.
- McConville, K. S. and Toth, D. (2019), ‘Automated selection of post-strata using a model-assisted regression tree estimator’, *Scandinavian Journal of Statistics* **46**(2), 389–413.
- Montanari, G. E. and Ranalli, M. G. (2005), ‘Nonparametric model calibration estimation in survey sampling’, *Journal of the American Statistical Association* **100**(472), 1429–1442.
- Opsomer, J. D., Breidt, F. J., Moisen, G. G. and Kauermann, G. (2007), ‘Model-assisted estimation of forest resources with generalized additive models’, *Journal of the American Statistical Association* **102**(478), 400–409.
- Opsomer, J. and Miller, C. (2005), ‘Selecting the amount of smoothing in nonparametric regression estimation for complex surveys’, *Nonparametric Statistics* **17**(5), 593–611.
- Rao, J. N. K., Kovar, J. G. and Mantel, H. J. (1990), ‘On estimating distribution functions and quantiles from survey data using auxiliary information’, *Biometrika* **77**(2), 365–375.
- Robinson, P. M. and Särndal, C.-E. (1983), ‘Asymptotic properties of the generalized regression estimator in probability sampling’, *Sankhyā, Series B* **45**, 240–248.
- Royall, R. M. (1970), ‘On finite population sampling theory under certain linear regression models’, *Biometrika* **57**(2), 377–387.
- Rueda, M. (2009), ‘A predictive estimator of finite population mean using nonparametric regression’, *Computational Statistics* **24**(1), 1–14.
- Särndal, C.-E., Swensson, B. and Wretman, J. (1992), *Model Assisted Survey Sampling*, Springer.
- Seaman, S. (2025), ‘Debiased machine learning for combining probability and non-probability survey data’, *arXiv preprint arXiv:2508.08948*.
- Tillé, Y. (2006), *Sampling Algorithms*, Springer Series in Statistics, Springer, New York.
- Toth, D. and Eltinge, J. L. (2011), ‘Building consistent regression trees from complex sample data’, *Journal of the American Statistical Association* **106**(496), 1626–1636.
- Valliant, R., Dorfman, A. H. and Royall, R. M. (2000), *Finite Population Sampling and Inference: A Prediction Approach*, John Wiley & Sons, New York.

Zheng, H. and Little, R. J. A. (2003), ‘Penalized spline model-based estimation of the finite population total from probability-proportional-to-size samples’, *Journal of Official Statistics* **19**(2), 99–117.

Contents of the Appendix

A	Illustrative examples for model-based and model-assisted estimation	30
A.1	A simple illustration for the model-based estimator	30
A.2	Failure of design consistency under data reuse	30
B	Additional material on cross-fitting	31
B.1	Design-aware cross-fitting schemes	31
B.2	Extension of cross-fitting schemes for stratified sampling	32
C	Godambe-Joshi asymptotic optimality	34
D	Overview of exponential sampling designs	38
D.1	Notation and definitions	38
D.2	Exponential designs	38
D.3	Conditional inference	41
D.4	Conditional inference on a partition	41
D.5	Examples of conditional designs	42
E	Proofs of Section 3	46
E.1	Proof of Theorem 1	46
E.2	Proof of Theorem 2	50
E.3	Proof of Proposition 1	51
F	Proofs of Section 4	53
F.1	Proof of Proposition 2	54
F.2	Proof of Theorem 3	54
F.3	Proof of Theorem 4	55
F.4	Proof of Theorem 5	56
F.5	Proof of Corollary 1	58
G	Proofs of Section 5	59
G.1	Proof of Theorem 6	59
G.2	Proof of Theorem 7	60
G.3	Proof of Theorem 8	61
G.4	Proof of Corollary 2	62
H	Proofs of Section 6	63
H.1	Proof of Theorem 9	63
H.2	Proof of Theorem 10	64
H.3	Failure of first-order equivalence	66
H.4	First-order equivalence under balanced fold residuals	67

A Illustrative examples for model-based and model-assisted estimation

A.1 A simple illustration for the model-based estimator

This section illustrates how the convergence rate of a model-based estimator may be driven by that of the fitted prediction rule. This issue is particularly relevant for nonparametric and machine learning methods, which typically converge more slowly than the parametric rate $n^{-1/2}$. For example, when the regression function is Lipschitz in p dimensions, a typical minimax rate in root mean squared prediction error is $n^{-1/(p+2)}$, up to constants and logarithmic factors. This gives $n^{-1/3}$ for $p = 1$ and $n^{-1/4}$ for $p = 2$.

Consider the prediction rule

$$\hat{m}(\mathbf{x}) = m(\mathbf{x}) + \frac{\xi}{n^\alpha}, \quad 0 < \alpha < \frac{1}{2}, \quad \xi \sim \mathcal{N}(0, 1).$$

Although $\hat{m}$ converges to m at rate $n^{-\alpha}$, its contribution to the model-based estimator is

$$\hat{\mu}_{mb}(\hat{m}) - \hat{\mu}_{mb}(m) = \frac{1}{N} \sum_{k \in U \setminus S} \frac{\xi}{n^\alpha} = \frac{N-n}{N} \frac{\xi}{n^\alpha}.$$

Consequently,

$$\sqrt{n} \{ \hat{\mu}_{mb}(\hat{m}) - \hat{\mu}_{mb}(m) \} = \frac{N-n}{N} n^{1/2-\alpha} \xi.$$

If the nonsampling fraction $(N-n)/N$ is bounded away from zero, this quantity is not bounded in probability when $\alpha < 1/2$. Thus, even though $\hat{m}$ is consistent for m , the error due to estimating the prediction rule may dominate the usual $n^{-1/2}$ sampling variation. This illustrates the lack of Neyman orthogonality of the model-based estimator: prediction errors enter the estimator at first order.

A.2 Failure of design consistency under data reuse

We give a simple example showing that data reuse may prevent the feasible model-assisted estimator from being design-consistent. Consider a population with distinct auxiliary vectors and $y_k = 1$ for every $k \in U$, so that $\mu = 1$, and define the fitted prediction rule by

$$\hat{m}(\mathbf{x}_k; \mathbf{I}) = I_k.$$

Thus, the prediction is exact for every sampled unit but equals zero for every nonsampled unit. Substituting this rule into the model-assisted estimator gives

$$\begin{aligned}\hat{\mu}_{\text{ma}}(\hat{m}, \pi) &= \frac{1}{N} \left\{ \sum_{k \in U} I_k + \sum_{k \in S} \frac{1 - I_k}{\pi_k} \right\} \\ &= \frac{1}{N} \sum_{k \in U} I_k = \frac{n}{N}.\end{aligned}$$

Therefore, if $n/N \xrightarrow{v \rightarrow \infty} f < 1$, then

$$\hat{\mu}_{\text{ma}}(\hat{m}, \pi) \xrightarrow{v \rightarrow \infty} f \neq \mu.$$

The estimator is consequently not design-consistent unless the sampling fraction converges to one. The failure is caused by the perfect dependence between $\hat{m}(\mathbf{x}_k; \mathbf{I})$ and I_k : the fitted rule interpolates the sampled observations but does not provide meaningful predictions outside the sample. This example illustrates why some control of the dependence induced by data reuse is necessary for a general theory.

B Additional material on cross-fitting

B.1 Design-aware cross-fitting schemes

[Lu et al. \(2025\)](#) propose several cross-fitting schemes satisfying the conditional independence property (Ind- δ). We briefly describe their constructions for two folds, that is, $M = 2$.

Bernoulli sampling. For a Bernoulli randomized experiment, corresponding to Bernoulli sampling in our setting, generate N independent Bernoulli random variables $(L_k)_{k \in U}$ with success probability π , independently of S . The two folds are defined by

$$U_1 = \{k \in U : L_k = 1\}, \quad U_2 = \{k \in U : L_k = 0\}.$$

Simple random sampling without replacement. For a completely randomized experiment, corresponding to simple random sampling without replacement (SRSWOR), first split the sample S uniformly at random into two disjoint subsets S_1 and S_2 of prescribed sizes n_1 and n_2 , where $n_1 + n_2 = n$. Independently of this split, split the nonsampled units $U \setminus S$ uniformly at random into two disjoint subsets R_1 and R_2 of respective sizes $N_1 - n_1$ and $N_2 - n_2$. The folds are then defined by

$$U_1 = S_1 \cup R_1, \quad U_2 = S_2 \cup R_2.$$

Stratified simple random sampling without replacement. For a stratified randomized experiment, corresponding to stratified SRSWOR, [Lu et al. \(2025\)](#) describe two constructions.

The first applies the preceding SRSWOR construction independently within each stratum and combines the resulting subsets across strata. The second partitions the stratum labels $1, \dots, H$ between the two folds, independently of S , assigning all units from a given stratum to the same fold.

Each construction produces a random partition for which the sampling indicators in one fold are conditionally independent of those in the other fold, given the partition. Thus, all three schemes satisfy condition (Ind- δ).

B.2 Extension of cross-fitting schemes for stratified sampling

Assume that the population U is partitioned into H strata $U^{(1)}, \dots, U^{(H)}$. Stratified sampling uses the design $p(s) = \prod_{h=1}^H p_h(s_h)$, where $s_h = s \cap U^{(h)}$ and sampling is independent across strata. One valid cross-fitting scheme consists of partitioning the stratum labels into $F_1, \dots, F_M$, independently of S , and setting $U_j = \bigcup_{h \in F_j} U^{(h)}$. Since each stratum belongs to only one fold, independence across strata gives (Ind- δ), regardless of the within-stratum designs. The drawback is that prediction accuracy may suffer: if the relationships between the covariates and the survey variable differ substantially across strata, the prediction rule must extrapolate to strata absent from its training sample.

To circumvent this drawback, we propose the following construction, which uses within-stratum factorization when available and the above observation otherwise.

Algorithm 3 Cross-fitting under stratified sampling.

Input: Strata $U^{(1)}, \dots, U^{(H)}$, sample S , number of folds M , and a fixed partition of the strata labels

$$\{1, \dots, H\} = \mathcal{H}_{\text{split}} \uplus \mathcal{H}_{\text{full}},$$

where a valid within-stratum cross-fitting scheme is specified for each $h \in \mathcal{H}_{\text{split}}$, and each stratum $h \in \mathcal{H}_{\text{full}}$ is kept intact.

Do:

1. For each $h \in \mathcal{H}_{\text{split}}$, apply the specified cross-fitting scheme using only the sample in $U^{(h)}$ and independent randomization across strata, producing

$$U^{(h)} = U_1^{(h)} \uplus \dots \uplus U_M^{(h)}.$$

2. For $j = 1, \dots, M$, combine the corresponding stratum-level folds by setting

$$U_j = \bigcup_{h \in \mathcal{H}_{\text{split}}} U_j^{(h)}.$$

3. For each $h \in \mathcal{H}_{\text{full}}$, assign the entire stratum $U^{(h)}$ to one of the folds $U_1, \dots, U_M$, independently of S and of the preceding randomizations.

Output: The population partition

$$\mathcal{P} = (U_1, \dots, U_M).$$

Proposition 3. *Let p be a stratified sampling design with independent sampling across strata, and let $\mathcal{P}$ be constructed by Algorithm 3. Let $\mathcal{A}$ collect the within-stratum conditioning variables and the assignments of intact strata. Then, the resulting scheme satisfies (Ind- $\mathcal{A}$).*

Proof. The samples and the randomizations used within different strata are independent. Each within-stratum conditioning variable depends only on that stratum’s sample and randomization. Conditioning on their collection therefore preserves independence across strata. The assignments of intact strata are independent of these variables and of the samples, so conditioning on them also preserves this independence.

Within each split stratum, the conditional sample law factorizes across folds by the assumed validity of its cross-fitting scheme. Each intact stratum contributes sampling indicators to only one fold, so its factors for the other folds equal one. For $s \in \mathcal{P}(U)$, these factorizations give

$$\begin{aligned} \mathbb{P}(S = s \mid \mathcal{A}) &= \prod_{h=1}^H \mathbb{P}(S \cap U^{(h)} = s \cap U^{(h)} \mid \mathcal{A}) \\ &= \prod_{h=1}^H \prod_{j=1}^M \mathbb{P}(S \cap U_j \cap U^{(h)} = s \cap U_j \cap U^{(h)} \mid \mathcal{A}) \\ &= \prod_{j=1}^M \mathbb{P}(S \cap U_j = s \cap U_j \mid \mathcal{A}). \end{aligned}$$

The last equality groups the independent stratum samples by fold. This proves (Ind- $\mathcal{A}$). $\square$

Following [Särndal et al. \(1992\)](#), for strata $h \in \mathcal{H}_{\text{split}}$, several possibilities exist in order to train a regression function estimator $\hat{m}$: (i) stratum-specific training; (ii) pooled training. Stratum-specific training consists in fitting a separate model for each stratum h , that is, fitting a regression function estimator $\hat{m}_{jh}$ on $(\mathbf{x}_k, y_k)_{S^{(h)} \setminus S_j^{(h)}}$ for prediction in $U_j^{(h)}$. Pooled training, on the other hand, would consist of pooling all the data across strata together in folds and proceeding to train a common regression function estimator $\hat{m}_j$ on $(\mathbf{x}_k, y_k)_{S \setminus S_j}$, to predict in the whole fold U_j , which would typically contain different strata. The main advantage of the stratum-specific training is that it allows for relationships between covariates X and the survey variable Y to differ freely in each stratum. However, the associated drawback is that the resulting sample size used to fit the stratum-specific learner $\hat{m}_{jh}$ would typically be much smaller than that which could be used with the pooled learner $\hat{m}_j$. If the sample size within strata is large enough and strata are very heterogeneous, this may be an appropriate strategy. If not, fitting a complex learner like a deep neural network, for example, might be challenging. The pooled version, on the other hand, would not suffer from these drawbacks, at the price of perhaps capturing only global trends in the regression function. An alternative strategy, when the number of strata is small compared to the overall sample size, consists of including the strata memberships as additional covariates. Nonetheless, independently of the learning strategy used, the cross-fitting scheme will remain valid. For strata $h \in \mathcal{H}_{\text{full}}$, however, these strategies are not possible, and only pooled training is available.

C Godambe-Joshi asymptotic optimality

To discuss optimality, suppose that the population data $(\mathbf{x}_k, y_k)_{k \in U_v}$ are independent and identically distributed. We can then write

$$y_k = m(\mathbf{x}_k) + \epsilon_k, \quad k \in U_v,$$

where $m : \mathbf{x} \mapsto \mathbb{E}[Y \mid X = \mathbf{x}]$. Conditional on the population covariates $\mathbf{X}_U$, the errors are independent, with $\mathbb{E}_m[\epsilon_k] = 0$ and $\mathbb{E}_m[\epsilon_k^2] = \sigma^2(\mathbf{x}_k)$. The sampling design is non-informative, and the fold construction and $\mathcal{A}$ use only the sample, the covariates, and independent randomization. Thus, given $\mathbf{X}_U$, $(\mathbf{I}_U, \mathcal{A})$ is independent of $\mathbf{y}_U$; see [Pfeffermann and Sverchkov \(2009\)](#) for non-informative sampling. Throughout this appendix, expectations are conditional on the population covariates: $\mathbb{E}_m$ averages over the responses, and $\mathbb{E}_{mp}$ also averages over sampling and fold construction. We moreover assume that

$$\limsup_{v \rightarrow \infty} \frac{1}{N_v} \sum_{k \in U_v} \sigma^2(\mathbf{x}_k) < \infty.$$

In this setup, [Godambe and Joshi \(1965\)](#) showed that, for any design-unbiased $\hat{\mu}$,

$$\mathbb{E}_{mp} [(\hat{\mu} - \mu)^2] \geq \frac{1}{N_v^2} \sum_{k \in U_v} \frac{1 - \pi_k}{\pi_k} \sigma^2(\mathbf{x}_k).$$

Several model-assisted estimators reach this bound to first order under appropriate assumptions, meaning

$$n_v \mathbb{E}_{mp} [(\hat{\mu}_{ma} - \mu)^2] = \frac{n_v}{N_v^2} \sum_{k \in U_v} \frac{1 - \pi_k}{\pi_k} \sigma^2(\mathbf{x}_k) + o(1) := \text{GJ} + o(1).$$

This holds for linear regression with a well-specified model ([Robinson and Särndal, 1983](#)), local polynomials ([Breidt and Opsomer, 2000](#)), and B -splines ([Goga, 2005](#)), to mention a few. The same argument applies to cross-fitting. When $\tilde{m} = m$, the oracle $\hat{\mu}_{ma}(m, \pi)$ reaches the bound exactly. It therefore suffices to strengthen the first-order equivalence with the feasible estimator to convergence in L^2 .

For the conditional oracle $\hat{\mu}_{ma}(m, \pi_{|\mathcal{A}})$, the joint mean-squared error has an additional nonnegative term. A sufficient condition for this term to vanish at first order is that $\mathbb{E}_p[\max_{k \in U_v} (\pi_{k|\mathcal{A}} - \pi_k)^2] = o(1)$. To make the role of this proximity clear, we do not assume [\(CF3\)](#) in the following theorem and instead state the conditions directly.

Theorem 11. *Assume $(\text{Ind-}\mathcal{A})$, [\(D1\)](#), [\(D2\)](#), [\(CF2,c\)](#), [\(CF4\)](#), and [\(M1\)](#), with [\(M1\)](#) holding almost surely under the model. Suppose that $\tilde{m} = m$ and that the prediction errors $\hat{m}^{(-k)}(\mathbf{x}_k) - m(\mathbf{x}_k)$ are bounded by a fixed constant, uniformly in k and v , almost surely under the joint model and sampling law. Then, the following statements hold.*

(i) *Assume that*

$$\mathbb{E}_p \left[\max_{k \in U_v} (\pi_{k|\mathcal{A}} - \pi_k)^2 \right] = o(1).$$

Then, $\hat{\mu}_{cf}(\hat{m}, \pi_{|\mathcal{A}})$ asymptotically reaches the Godambe-Joshi lower bound.

(ii) Assume that

$$\lim_{v \rightarrow \infty} n_v \mathbb{E}_{mp} \left[\max_{k \in U_v} (\pi_{k|\mathcal{A}} - \pi_k)^2 \times \frac{1}{N_v} \sum_{k \in U_v} (\hat{m}^{(-k)}(\mathbf{x}_k) - m(\mathbf{x}_k))^2 \right] = 0.$$

Then, $\hat{\mu}_{cf}(\hat{m}, \pi)$ asymptotically reaches the Godambe-Joshi lower bound.

Proof. Let $\hat{\mu}$ denote either feasible estimator and $\tilde{\mu}$ its corresponding oracle, $\hat{\mu}_{ma}(m, \pi_{|\mathcal{A}})$ or $\hat{\mu}_{ma}(m, \pi)$. We have

$$\mathbb{E}_{mp} [(\hat{\mu} - \mu)^2] = \mathbb{E}_{mp} [(\hat{\mu} - \tilde{\mu})^2] + \mathbb{E}_{mp} [(\tilde{\mu} - \mu)^2] + R_v,$$

where, by Cauchy–Schwarz,

$$\begin{aligned} |R_v| &= 2 |\mathbb{E}_{mp} [(\hat{\mu} - \tilde{\mu})(\tilde{\mu} - \mu)]| \\ &\leq 2 \left\{ \mathbb{E}_{mp} [(\hat{\mu} - \tilde{\mu})^2] \right\}^{1/2} \left\{ \mathbb{E}_{mp} [(\tilde{\mu} - \mu)^2] \right\}^{1/2}. \end{aligned}$$

Since $n_v/N_v \leq 1$, (D2) and the bound on the average model variance give $GJ = \mathcal{O}(1)$. Thus, it is enough to show that

$$n_v \mathbb{E}_{mp} [(\tilde{\mu} - \mu)^2] = GJ + o(1), \quad n_v \mathbb{E}_{mp} [(\hat{\mu} - \tilde{\mu})^2] = o(1).$$

Indeed, the bound on R_v then gives $n_v |R_v| = o(1)$. To bound the difference between each feasible estimator and its oracle, we will use the mean-square convergence of the prediction errors. By (M1), the inner average below tends to zero for almost every model realization. It is bounded by the square of the fixed bound on the prediction errors. Thus, dominated convergence gives

$$\begin{aligned} &\frac{1}{N_v} \sum_{k \in U_v} \mathbb{E}_{mp} \left[\left(\hat{m}^{(-k)}(\mathbf{x}_k) - m(\mathbf{x}_k) \right)^2 \right] \\ &= \mathbb{E}_m \left[\frac{1}{N_v} \sum_{k \in U_v} \mathbb{E}_p \left[\left(\hat{m}^{(-k)}(\mathbf{x}_k) - m(\mathbf{x}_k) \right)^2 \right] \right] \xrightarrow{v \rightarrow \infty} 0. \end{aligned}$$

Proof of (i). We first compute the mean squared error of the conditional oracle. Its estimation error is

$$\hat{\mu}_{ma}(m, \pi_{|\mathcal{A}}) - \mu = \frac{1}{N_v} \sum_{k \in U_v} \left(\frac{I_k}{\pi_{k|\mathcal{A}}} - 1 \right) \epsilon_k.$$

The model errors are independent and centered, and are independent of sampling and fold construction given the covariates. For the diagonal terms,

$$\mathbb{E}_p \left[\left(\frac{I_k}{\pi_{k|\mathcal{A}}} - 1 \right)^2 \mid \mathcal{A} \right] = \frac{1 - \pi_{k|\mathcal{A}}}{\pi_{k|\mathcal{A}}}.$$

Expanding the square and separating the sampling and model expectations, we obtain

$$\begin{aligned} n_v \mathbb{E}_{mp} \left[\left(\hat{\mu}_{\text{ma}}(m, \pi_{|\mathcal{A}}) - \mu \right)^2 \right] &= \frac{n_v}{N_v^2} \sum_{k \in U_v} \sum_{l \in U_v} \mathbb{E}_p \left[\left(\frac{I_k}{\pi_{k|\mathcal{A}}} - 1 \right) \left(\frac{I_l}{\pi_{l|\mathcal{A}}} - 1 \right) \right] \mathbb{E}_m[\epsilon_k \epsilon_l] \\ &= \frac{n_v}{N_v^2} \sum_{k \in U_v} \mathbb{E}_p \left[\frac{1 - \pi_{k|\mathcal{A}}}{\pi_{k|\mathcal{A}}} \right] \sigma^2(\mathbf{x}_k) \\ &= \text{GJ} + \frac{n_v}{N_v^2} \sum_{k \in U_v} \left(\mathbb{E}_p \left[\frac{1}{\pi_{k|\mathcal{A}}} \right] - \frac{1}{\pi_k} \right) \sigma^2(\mathbf{x}_k). \end{aligned}$$

To bound the additional term, use the identity

$$\frac{1}{\pi_{k|\mathcal{A}}} - \frac{1}{\pi_k} = -\frac{\pi_{k|\mathcal{A}} - \pi_k}{\pi_k^2} + \frac{(\pi_{k|\mathcal{A}} - \pi_k)^2}{\pi_k^2 \pi_{k|\mathcal{A}}}$$

and $\mathbb{E}_p[\pi_{k|\mathcal{A}}] = \pi_k$ to obtain

$$\mathbb{E}_p \left[\frac{1}{\pi_{k|\mathcal{A}}} \right] - \frac{1}{\pi_k} = \mathbb{E}_p \left[\frac{(\pi_{k|\mathcal{A}} - \pi_k)^2}{\pi_k^2 \pi_{k|\mathcal{A}}} \right].$$

By (D2) and (CF2,c), the additional term is nonnegative and at most

$$C \mathbb{E}_p \left[\max_{k \in U_v} (\pi_{k|\mathcal{A}} - \pi_k)^2 \right] \frac{1}{N_v} \sum_{k \in U_v} \sigma^2(\mathbf{x}_k) = o(1),$$

for a constant C independent of v .

We next bound the difference between the feasible and oracle estimators. From their definitions,

$$\hat{\mu}_{cf}(\hat{m}, \pi_{|\mathcal{A}}) - \hat{\mu}_{\text{ma}}(m, \pi_{|\mathcal{A}}) = -\frac{1}{N_v} \sum_{k \in U_v} (I_k - \pi_{k|\mathcal{A}}) \frac{\hat{m}^{(-k)}(\mathbf{x}_k) - m(\mathbf{x}_k)}{\pi_{k|\mathcal{A}}}.$$

Applying Lemma 2 with $d_k = \{\hat{m}^{(-k)}(\mathbf{x}_k) - m(\mathbf{x}_k)\}/\pi_{k|\mathcal{A}}$ gives

$$\begin{aligned} n_v \mathbb{E}_p \left[\left\{ \hat{\mu}_{cf}(\hat{m}, \pi_{|\mathcal{A}}) - \hat{\mu}_{\text{ma}}(m, \pi_{|\mathcal{A}}) \right\}^2 \right] \\ \leq \frac{C}{N_v} \sum_{k \in U_v} \mathbb{E}_p \left[\left\{ \frac{\hat{m}^{(-k)}(\mathbf{x}_k) - m(\mathbf{x}_k)}{\pi_{k|\mathcal{A}}} \right\}^2 \right]. \end{aligned}$$

Averaging over the model and using (CF2,c), we obtain

$$\begin{aligned} n_v \mathbb{E}_{mp} \left[\left\{ \hat{\mu}_{cf}(\hat{m}, \pi_{|\mathcal{A}}) - \hat{\mu}_{\text{ma}}(m, \pi_{|\mathcal{A}}) \right\}^2 \right] \\ \leq \frac{C}{N_v} \sum_{k \in U_v} \mathbb{E}_{mp} \left[\left(\hat{m}^{(-k)}(\mathbf{x}_k) - m(\mathbf{x}_k) \right)^2 \right] = o(1). \end{aligned}$$

This proves (i).

Proof of (ii). For the unconditional oracle, the same model calculation gives

$$n_v \mathbb{E}_{mp} \left[(\hat{\mu}_{ma}(m, \pi) - \mu)^2 \right] = \frac{n_v}{N_v^2} \sum_{k \in U_v} \frac{1 - \pi_k}{\pi_k} \sigma^2(\mathbf{x}_k) = \text{GJ}.$$

For the difference between the feasible and oracle estimators, we have

$$\begin{aligned} & \hat{\mu}_{cf}(\hat{m}, \pi) - \hat{\mu}_{ma}(m, \pi) \\ &= -\frac{1}{N_v} \sum_{k \in U_v} (I_k - \pi_{k|\mathcal{A}}) \frac{\hat{m}^{(-k)}(\mathbf{x}_k) - m(\mathbf{x}_k)}{\pi_k} - \frac{1}{N_v} \sum_{k \in U_v} (\pi_{k|\mathcal{A}} - \pi_k) \frac{\hat{m}^{(-k)}(\mathbf{x}_k) - m(\mathbf{x}_k)}{\pi_k}. \end{aligned}$$

For the first sum, Lemma 2 and (D2), followed by averaging over the model, give

$$\begin{aligned} & n_v \mathbb{E}_{mp} \left[\left\{ \frac{1}{N_v} \sum_{k \in U_v} (I_k - \pi_{k|\mathcal{A}}) \frac{\hat{m}^{(-k)}(\mathbf{x}_k) - m(\mathbf{x}_k)}{\pi_k} \right\}^2 \right] \\ & \leq \frac{C}{N_v} \sum_{k \in U_v} \mathbb{E}_{mp} \left[\left(\hat{m}^{(-k)}(\mathbf{x}_k) - m(\mathbf{x}_k) \right)^2 \right]. \end{aligned}$$

For the second sum, Cauchy-Schwarz and (D2) give

$$\begin{aligned} & n_v \left\{ \frac{1}{N_v} \sum_{k \in U_v} (\pi_{k|\mathcal{A}} - \pi_k) \frac{\hat{m}^{(-k)}(\mathbf{x}_k) - m(\mathbf{x}_k)}{\pi_k} \right\}^2 \\ & \leq n_v \left\{ \frac{1}{N_v} \sum_{k \in U_v} \left(\frac{\pi_{k|\mathcal{A}} - \pi_k}{\pi_k} \right)^2 \right\} \left\{ \frac{1}{N_v} \sum_{k \in U_v} \left(\hat{m}^{(-k)}(\mathbf{x}_k) - m(\mathbf{x}_k) \right)^2 \right\} \\ & \leq \frac{n_v}{\lambda^2} \max_{k \in U_v} (\pi_{k|\mathcal{A}} - \pi_k)^2 \times \frac{1}{N_v} \sum_{k \in U_v} \left(\hat{m}^{(-k)}(\mathbf{x}_k) - m(\mathbf{x}_k) \right)^2. \end{aligned}$$

Taking expectations in the second bound and combining the two bounds using $(a+b)^2 \leq 2a^2 + 2b^2$, we obtain

$$\begin{aligned} & n_v \mathbb{E}_{mp} \left[\{ \hat{\mu}_{cf}(\hat{m}, \pi) - \hat{\mu}_{ma}(m, \pi) \}^2 \right] \\ & \leq \frac{C}{N_v} \sum_{k \in U_v} \mathbb{E}_{mp} \left[\left(\hat{m}^{(-k)}(\mathbf{x}_k) - m(\mathbf{x}_k) \right)^2 \right] \\ & \quad + C n_v \mathbb{E}_{mp} \left[\max_{k \in U_v} (\pi_{k|\mathcal{A}} - \pi_k)^2 \times \frac{1}{N_v} \sum_{k \in U_v} \left(\hat{m}^{(-k)}(\mathbf{x}_k) - m(\mathbf{x}_k) \right)^2 \right]. \end{aligned}$$

The first term tends to zero by the joint mean-square convergence above, and the second by the assumption in (ii). This completes the proof. $\square$

D Overview of exponential sampling designs

D.1 Notation and definitions

Let $U = \{1, \dots, N\}$ be a population from which a sample is to be selected. A sample $s \subset U$ is simply a subset of U . A *support* $\mathcal{Q}$ is a set of samples. We will mainly work with two supports: the set of all samples of U (the power set of U), $\mathcal{S} = \{s \subset U\}$, and the set of all samples of fixed size n , $\mathcal{S}_n = \{s \subset U : |s| = n\}$.

A random sample is selected by means of a *sampling design* $p(\cdot)$ defined on a support $\mathcal{Q}$, that is, a function satisfying

$$p(s) > 0, \quad s \in \mathcal{Q}, \quad \sum_{s \in \mathcal{Q}} p(s) = 1.$$

The (first-order) *inclusion probabilities* are obtained from the design by

$$\pi_k = \sum_{\substack{s \in \mathcal{Q} \\ s \ni k}} p(s), \quad k \in U,$$

and the second-order (joint) inclusion probabilities by

$$\pi_{k\ell} = \sum_{\substack{s \in \mathcal{Q} \\ s \supset \{k, \ell\}}} p(s), \quad k, \ell \in U.$$

We write S for the random sample selected according to $p(\cdot)$, so that $\Pr(S = s) = p(s)$ for $s \in \mathcal{Q}$.

First, we will define exponential schemes, and then we will list a number of classical sampling schemes, each of which belongs to the exponential family.

D.2 Exponential designs

D.2.1 Definition

Definition 3. Given a sampling design $p(\cdot)$ on $\mathcal{S}$, its *support* is the set of samples with nonzero probability,

$$\mathcal{Q} = \{s \subset U \mid p(s) > 0\}.$$

Definition 4. A support $\mathcal{Q}$ is *symmetric* if, for every $s \in \mathcal{Q}$, every subset $r \subseteq U$ with $|r| = |s|$ also belongs to $\mathcal{Q}$.

Both $\mathcal{S}$ and $\mathcal{S}_n$ are symmetric supports.

Among all designs on a support $\mathcal{Q}$ with prescribed inclusion probabilities π_k , $k \in U$, the *exponential design* is the one that maximizes the entropy

$$I(p) = - \sum_{s \in \mathcal{Q}} p(s) \log p(s).$$

Maximizing $I(p)$ subject to $\sum_{s \in \mathcal{Q}} p(s) = 1$ and $\sum_{s \in \mathcal{Q}, s \ni k} p(s) = \pi_k$, $k \in U$, by means of a Lagrangian function with multipliers $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_N)^\top$ yields a design of the form

$$p_{\text{exp}}(s; \mathcal{Q}, \boldsymbol{\lambda}) = \frac{\prod_{k \in s} \omega_k}{\sum_{r \in \mathcal{Q}} \prod_{k \in r} \omega_k}, \quad \omega_k = \exp(\lambda_k), k \in U. \quad (12)$$

We use $p_{\text{exp}}(s; \mathcal{Q}, \boldsymbol{\lambda})$ as a shorthand for this exponential design, built on support $\mathcal{Q}$ with parameter $\boldsymbol{\lambda}$ (equivalently, with weights $\omega_k = \exp(\lambda_k)$).

The parameters ω_k , $k \in U$, are determined so that the design (12) reproduces the target inclusion probabilities, i.e. by solving the system of equations

$$\sum_{\substack{s \in \mathcal{Q} \\ s \ni k}} p_{\text{exp}}(s; \mathcal{Q}, \boldsymbol{\lambda}) = \sum_{\substack{s \in \mathcal{Q} \\ s \ni k}} \frac{\prod_{\ell \in s} \omega_\ell}{\sum_{r \in \mathcal{Q}} \prod_{\ell \in r} \omega_\ell} = \pi_k, \quad k \in U. \quad (13)$$

Solving this system is not always straightforward: it has a closed-form solution in some cases (Bernoulli sampling, Poisson sampling) but not in others (conditional Poisson sampling).

Throughout, $\mathbf{1} = (1, \dots, 1)^\top \in \mathbb{R}^N$. We now show that several classical designs are exponential designs, i.e. special cases of (12).

D.2.2 Bernoulli sampling

Bernoulli sampling is the exponential design $p_{\text{exp}}(s, \mathcal{S}, \theta \mathbf{1})$ with a common parameter $\theta \in \mathbb{R}$ (so $\lambda_k = \theta$ for every $k \in U$). We obtain

$$\begin{aligned} p(s) &= \frac{\prod_{k \in s} \exp(\theta)}{\sum_{r \in \mathcal{S}} \prod_{k \in r} \exp(\theta)} = \frac{\exp(\theta|s|)}{\sum_{r \in \mathcal{S}} \exp(\theta|r|)} = \frac{\exp(\theta|s|)}{\sum_{g=0}^N \binom{N}{g} \exp(\theta g)} \\ &= \frac{\exp(\theta|s|)}{\{1 + \exp(\theta)\}^N} = \left\{ \frac{\exp(\theta)}{1 + \exp(\theta)} \right\}^{|s|} \left\{ \frac{1}{1 + \exp(\theta)} \right\}^{N-|s|} = \pi^{|s|} (1 - \pi)^{N-|s|}, \quad s \subset U, \end{aligned}$$

where the third equality uses the binomial theorem, and where the inclusion probabilities are $\pi_k = \exp(\theta)/\{1 + \exp(\theta)\}$ for all $k \in U$.

D.2.3 Simple random sampling

Simple random sampling without replacement, with fixed sample size n , is the exponential design $p_{\text{exp}}(s, \mathcal{S}_n, \theta \mathbf{1})$ with common parameter $\theta \in \mathbb{R}$. For $|s| = n$,

$$p(s) = \frac{\prod_{k \in s} \exp(\theta)}{\sum_{r \in \mathcal{S}_n} \prod_{k \in r} \exp(\theta)} = \frac{\exp(\theta n)}{\sum_{r \in \mathcal{S}_n} \exp(\theta n)} = \binom{N}{n}^{-1}, \quad s \in \mathcal{S}_n.$$

Since every $s \in \mathcal{S}_n$ has the same size n , the factor $\exp(\theta n)$ cancels between the numerator and the denominator: the design does not depend on θ and reduces to the uniform design on $\mathcal{S}_n$. The resulting inclusion probabilities are $\pi_k = n/N$, $k \in U$.

D.2.4 Poisson sampling

Poisson sampling is the exponential design $p_{exp}(s, \mathcal{S}, \boldsymbol{\lambda})$, now with a design-specific parameter λ_k for each unit. For this design, the general system (13), with $\mathcal{Q} = \mathcal{S}$, has the closed-form solution

$$\omega_k = \frac{\pi_k}{1 - \pi_k}, \quad k \in U$$

(see Tillé, 2006). Substituting into (12) and simplifying the sum over $\mathcal{S}$ gives

$$p_{ME}(s; \mathcal{S}) = \frac{\prod_{k \in s} \omega_k}{\sum_{r \in \mathcal{S}} \prod_{k \in r} \omega_k} = \left\{ \prod_{k \in s} \pi_k \right\} \left\{ \prod_{k \in U \setminus s} (1 - \pi_k) \right\}, \quad (14)$$

that is, Poisson sampling selects each unit k independently with probability π_k .

D.2.5 Conditional Poisson sampling

Conditional Poisson sampling (CPS), also called the maximum entropy design with fixed sample size, is the exponential design $p_{exp}(s, \mathcal{S}_n, \boldsymbol{\lambda})$. Here the system (13), restricted to $\mathcal{Q} = \mathcal{S}_n$, becomes

$$\sum_{\substack{s \in \mathcal{S}_n \\ s \ni k}} p(s) = \sum_{\substack{s \in \mathcal{S}_n \\ s \ni k}} \frac{\prod_{\ell \in s} \omega_\ell}{\sum_{r \in \mathcal{S}_n} \prod_{\ell \in r} \omega_\ell} = \pi_k, \quad k \in U,$$

and the relationship between π_k and ω_k is considerably more complex than in the Poisson case: this system cannot be solved analytically, and there is no closed form for ω_k as a function of π_k .

Nevertheless, several methods allow ω_k to be computed from π_k without summing over all samples in $\mathcal{S}_n$; a detailed account is given in Tillé (2006). The design is therefore

$$p_{ME}(s; \mathcal{S}_n) = \frac{\prod_{k \in s} \omega_k}{\sum_{r \in \mathcal{S}_n} \prod_{k \in r} \omega_k}, \quad (15)$$

an expression that cannot be simplified further in general. Once the ω_k have been determined, however, several algorithms are available for drawing a sample from this design (see again Tillé, 2006).

In practice, one first computes the ω_k from the target inclusion probabilities, and then draws the sample, as in the following R code:

```
install.packages("sampling")
pik <- c(0.07, 0.17, 0.41, 0.61, 0.83, 0.91)
n <- sum(pik)
pikt <- sampling::UPMEpiktildefrompik(pik)
w <- pikt / (1 - pikt)
sampling::UPMEsfromq(sampling::UPMEqfromw(w, n))
```

D.3 Conditional inference

Conditional inference in survey sampling has been studied by Rao (1985), Deville (1992), and Tillé (1998, 1999). To perform conditional inference, we first derive a *conditional design*; from it we obtain conditional inclusion probabilities, which are in turn used to build an estimator.

Let $\mathcal{Q}_1$ and $\mathcal{Q}_2$ be two supports with $\mathcal{Q}_2 \subset \mathcal{Q}_1$, and let $p(s; \mathcal{Q}_1)$ be a sampling design on $\mathcal{Q}_1$. The conditional design given $\mathcal{Q}_2$ is

$$p(s; \mathcal{Q}_1 \mid \mathcal{Q}_2) = \frac{p(s; \mathcal{Q}_1)}{\sum_{r \in \mathcal{Q}_2} p(r; \mathcal{Q}_1)}, \quad s \in \mathcal{Q}_2.$$

From this conditional design we obtain the conditional inclusion probabilities $\pi_k(\mathcal{Q}_2)$ and $\pi_{kl}(\mathcal{Q}_2)$.

The support $\mathcal{Q}_2$ may itself depend on the selected sample S : for instance, $\mathcal{Q}_2$ may be the set of samples whose size equals the size of the observed sample S . Given the conditional inclusion probabilities, we estimate a population total by

$$\hat{Y}(\mathcal{Q}_2) = \sum_{k \in S} \frac{y_k}{\pi_k(\mathcal{Q}_2)}.$$

This estimator is conditionally unbiased under $p(s; \mathcal{Q}_1 \mid \mathcal{Q}_2)$ if and only if $\pi_k(\mathcal{Q}_2) > 0$ for all $k \in U$.

D.4 Conditional inference on a partition

Let $U_1, \dots, U_H$ be a partition of U into *folds*, i.e.

$$\bigcup_{h=1}^H U_h = U \quad \text{and} \quad U_h \cap U_i = \emptyset \text{ for } h \neq i.$$

For a sample s , write $n_h = |s \cap U_h|$, $h = 1, \dots, H$, and $n = \sum_{h=1}^H n_h$.

Theorem 12. *If $p_{\text{exp}}(s, \mathcal{Q}, \boldsymbol{\lambda})$ is an exponential design on a symmetric support $\mathcal{Q}$, then the design conditional on $n_1, \dots, n_H$ factorizes as a product of independent exponential designs on $U_1, \dots, U_H$.*

Proof. Write $\omega_k = \exp(\lambda_k)$, $k \in U$, and $s_h = s \cap U_h$. By definition, the conditional design is

$$p_{\text{exp}}(s, \mathcal{Q}, \boldsymbol{\lambda} \mid n_1, \dots, n_H) = \frac{\prod_{k \in s} \omega_k}{\sum_{\substack{r \in \mathcal{Q} \\ |r \cap U_h| = n_h, h=1, \dots, H}} \prod_{k \in r} \omega_k}.$$

Because $\mathcal{Q}$ is symmetric, the set of samples in $\mathcal{Q}$ satisfying the H size constraints coincides with the Cartesian product of all subsets of each fold with the prescribed size:

$$\{r \in \mathcal{Q} : |r \cap U_h| = n_h, h = 1, \dots, H\} = \prod_{h=1}^H \{s_h \subset U_h : |s_h| = n_h\};$$

symmetry of $\mathcal{Q}$ guarantees that every such combination of per-fold samples indeed belongs to $\mathcal{Q}$. Consequently the denominator factorizes:

$$\sum_{\substack{r \in \mathcal{Q} \\ |r \cap U_h| = n_h}} \prod_{k \in r} \omega_k = \sum_{\substack{s_1 \subset U_1, \dots, s_H \subset U_H \\ |s_h| = n_h}} \prod_{h=1}^H \prod_{k \in s_h} \omega_k = \prod_{h=1}^H \sum_{\substack{s_h \subset U_h \\ |s_h| = n_h}} \prod_{k \in s_h} \omega_k.$$

Since the numerator also factorizes as $\prod_{k \in s} \omega_k = \prod_{h=1}^H \prod_{k \in s_h} \omega_k$, we obtain

$$p_{exp}(s, \mathcal{Q}, \boldsymbol{\lambda} \mid n_1, \dots, n_H) = \prod_{h=1}^H \frac{\prod_{k \in s_h} \omega_k}{\sum_{\substack{s_h \subset U_h \\ |s_h| = n_h}} \prod_{k \in s_h} \omega_k},$$

which is a product of H independent exponential designs, one on each fold U_h . $\square$

This result applies in particular to Bernoulli sampling, simple random sampling, Poisson sampling, and conditional Poisson sampling.

D.5 Examples of conditional designs

D.5.1 Example 1: conditioning on the n_h in a Bernoulli design

Under Bernoulli sampling, let $s_h = s \cap U_h$ and $n_h = |s_h|$, $h = 1, \dots, H$. Each n_h has a binomial distribution, $n_h \sim \text{Bin}(N_h, \pi)$.

Since $|s| = \sum_h n_h$ is fixed once $n_1, \dots, n_H$ are fixed, the factor $\pi^{|s|}(1 - \pi)^{N - |s|}$ is the same for every s satisfying the constraints, so it cancels in the conditional design:

$$\begin{aligned} p(s \mid n_1, \dots, n_H) &= \frac{\pi^{|s|}(1 - \pi)^{N - |s|}}{\sum_{\substack{r \\ |r_h| = n_h, h=1, \dots, H}} \pi^{|r|}(1 - \pi)^{N - |r|}} \\ &= \frac{1}{|\{r : |r_h| = n_h, h = 1, \dots, H\}|} = \prod_{h=1}^H \binom{N_h}{n_h}^{-1}, \end{aligned}$$

for every s such that $|s_h| = n_h$, $h = 1, \dots, H$.

Conditionally on $n_1, \dots, n_H$, the design is therefore simple random sampling without replacement within each post-stratum, independently across strata, with conditional inclusion probabilities n_h/N_h for $k \in U_h$.

D.5.2 Example 2: conditioning on the n_h in simple random sampling

Under simple random sampling of fixed size n , let again $s_h = s \cap U_h$ and $n_h = |s_h|$, $h = 1, \dots, H$. Here $(n_1, \dots, n_H)$ has a multivariate hypergeometric distribution.

Since $p(s) = \binom{N}{n}^{-1}$ does not depend on s (only on its size n , which is fixed), the conditional design is again uniform over samples satisfying the constraints:

$$p(s \mid n_1, \dots, n_H) = \frac{\binom{N}{n}^{-1}}{\sum_{|r_h|=n_h, h=1, \dots, H} \binom{N}{n}^{-1}} = \frac{1}{|\{r : |r_h| = n_h, h = 1, \dots, H\}|} = \prod_{h=1}^H \binom{N_h}{n_h}^{-1},$$

for every s such that $|s_h| = n_h$, $h = 1, \dots, H$.

As in Example 1, conditionally on $n_1, \dots, n_H$ the design is simple random sampling without replacement within each post-stratum, independent across strata, with conditional inclusion probabilities n_h/N_h for $k \in U_h$.

D.5.3 Example 3: conditioning on the sample size in a Poisson design

Suppose a sample is selected by Poisson sampling, design (14). The design conditional on a fixed sample size is

$$p_{ME}(s; \mathcal{S} \mid \mathcal{S}_n) = \frac{p_{ME}(s; \mathcal{S})}{\sum_{r \in \mathcal{S}_n} p_{ME}(r; \mathcal{S})} = \frac{\prod_{k \in s} \omega_k}{\sum_{r \in \mathcal{S}_n} \prod_{k \in r} \omega_k},$$

which is exactly the CPS design of Section 1.2.5. Formally:

1. Select a random sample S by Poisson sampling with inclusion probabilities π_k .
2. Compute $n = |S|$.
3. Compute $\omega_k = \pi_k / (1 - \pi_k)$, $k \in U$.
4. Compute the inclusion probabilities $\pi_k(\mathcal{S}_n)$ of the CPS design of size n with parameters ω_k .
5. Use $\pi_k(\mathcal{S}_n)$ in the estimator.

The following R code implements this construction, using the `sampling` package of Tillé and Matei (2025) and the `StratifiedSampling` package of Jauslin et al. (2025).

```
# install.packages("sampling")
# install.packages("StratifiedSampling")
#####
rm(list = ls())

cond_incl_prob <- function(pik, n) {
  N <- length(pik)
  if (n == 0) {
    pikn <- rep(0, N)
  } else if (n == N) {
    pikn <- rep(1, N)
  } else {
    TEST <- pik < 1
    n1 <- n - (length(TEST) - sum(TEST))
    w <- pik[TEST] / (1 - pik[TEST])
    pikn <- pik
```

```

    pikn[TEST] <- StratifiedSampling::pikfromq(StratifiedSampling::qfromw(w, n1))
  }
  pikn
}

#####
strat_from_Poisson <- function(pik, s, STRAT) {
  STRAT <- sampling::cleanstrata(STRAT)
  H <- max(STRAT)
  TEST <- pik < 1
  pikn <- pik
  STRATR <- STRAT[TEST]
  nh <- as.vector(t(s[TEST]) %*% sampling::disjunctive(STRATR))
  for (h in 1:H) {
    pikn[TEST][STRATR == h] <- cond_incl_prob(pik[TEST][STRATR == h], nh[h])
  }
  pikn
}

#####
data(MU284, package = "sampling")
n <- 100
pik <- sampling::inclusionprobabilities(MU284$P75, n)

#####
SIM <- 1000
EHT <- EC <- rep(0, SIM)
y <- MU284$REV84 / 10000

for (i in 1:SIM) {
  s <- sampling::UPpoisson(pik)
  n <- sum(s)
  pikn <- cond_incl_prob(pik, n)
  EHT[i] <- sum(s * y / pik)
  EC[i] <- sum(s * y / pikn)
}

boxplot(EHT, EC, names = c("Horvitz-Thompson", "Conditional design"))
var(EC) / var(EHT)

```

D.5.4 Example 4: conditioning on the n_h in a Poisson design

Suppose again a sample is selected by Poisson sampling, design (14). The design conditional on fixed sample sizes $n_1, \dots, n_H$ in each stratum is

$$p_{ME}(s; \mathcal{S} \mid \mathcal{S}_{\text{strat}(n_1, \dots, n_H)}) = \prod_{h=1}^H \frac{\prod_{k \in (s \cap U_h)} \omega_k}{\sum_{\substack{s_h \subset U_h \\ |s_h|=n_h}} \prod_{k \in s_h} \omega_k},$$

i.e. the maximum entropy stratified design. Formally:

1. Select a random sample S by Poisson sampling with inclusion probabilities π_k .
2. Compute $\omega_k = \pi_k / (1 - \pi_k)$, $k \in U$.
3. Compute $n_h = |U_h \cap S|$, $h = 1, \dots, H$.
4. In each stratum, compute the inclusion probabilities $\pi_k(\mathcal{S} \mid \mathcal{S}_{\text{strat}(n_1, \dots, n_H)})$ of the maximum entropy stratified design of sizes $n_1, \dots, n_H$, using the parameters ω_k .
5. Use $\pi_k(\mathcal{S} \mid \mathcal{S}_{\text{strat}(n_1, \dots, n_H)})$ in the estimator.

Again using the `sampling` package of Tillé and Matei (2025) and the `StratifiedSampling` package of Jauslin et al. (2025):

```
#####
CAT <- cut(y, breaks = quantile(y, probs = seq(0, 1, by = 1/3), na.rm = TRUE),
          include.lowest = TRUE, labels = 1:3)

SIM <- 1000
EHT <- ESTRP <- rep(0, SIM)
y <- MU284$REV84 / 10000
STRAT <- MU284$REG
STRAT <- CAT

for (i in 1:SIM) {
  s <- sampling::UPpoisson(pik)
  pikn <- strat_from_Poisson(pik, s, STRAT)
  EHT[i] <- sum((y / pik)[s == 1])
  ESTRP[i] <- sum((y / pikn)[s == 1])
}

boxplot(EHT, ESTRP, names = c("Horvitz-Thompson", "Conditional design"))
var(ESTRP) / var(EHT)
```

D.5.5 Example 5: conditioning on the n_h in a CPS design

Suppose a sample of fixed size n is selected by CPS, design (15). The design conditional on fixed sample sizes $n_1, \dots, n_H$ in each stratum (necessarily with $n_1 + \dots + n_H = n$) is

$$p_{ME}(s; \mathcal{S}_n \mid \mathcal{S}_{\text{strat}(n_1, \dots, n_H)}) = \prod_{h=1}^H \frac{\prod_{k \in (s \cap U_h)} \omega_k}{\sum_{\substack{s_h \subset U_h \\ |s_h| = n_h}} \prod_{k \in s_h} \omega_k},$$

again the maximum entropy stratified design. Formally:

1. Select a random sample S by CPS with inclusion probabilities π_k .
2. Compute ω_k from π_k using one of the methods in Tillé (2006).
3. Compute $n_h = |U_h \cap S|$, $h = 1, \dots, H$.

4. In each stratum, compute the inclusion probabilities $\pi_k(\mathcal{S} \mid \mathcal{S}_{\text{strat}(n_1, \dots, n_H)})$ of the maximum entropy stratified design of sizes $n_1, \dots, n_H$, using the parameters ω_k .
5. Use $\pi_k(\mathcal{S} \mid \mathcal{S}_{\text{strat}(n_1, \dots, n_H)})$ in the estimator.

Again using the `sampling` package of Tillé and Matei (2025) and the `StratifiedSampling` package of Jauslin et al. (2025):

```
#####
pikt <- pik
TEST <- pik < 1
pikt[TEST] <- StratifiedSampling::piktfropik(pik[TEST])

SIM <- 1000
EHT <- ESTRP <- rep(0, SIM)
y <- MU284$REV84 / 10000
STRAT <- MU284$REG
STRAT <- CAT

for (i in 1:SIM) {
  print(i)
  s <- sampling::UPmaxentropy(pik)
  pikn <- strat_from_Poisson(pikt, s, STRAT)
  EHT[i] <- sum((y / pik)[s == 1])
  ESTRP[i] <- sum((y / pikn)[s == 1])
}

boxplot(EHT, ESTRP, names = c("Horvitz-Thompson", "Conditional stratified"))
var(ESTRP) / var(EHT)
```

E Proofs of Section 3

E.1 Proof of Theorem 1

We prove the two implications separately. First, we show that exponential designs factorize after conditioning on fold sample sizes. Second, we show that conditional independence forces the design to be exponential.

Step 1: Exponential designs factorize

Assume (ii), and write $\omega_k(n) = \exp(\lambda_k(n))$ for the weights of the exponential design at each possible sample size n , with

$$\omega_k(n) = \frac{\omega_k}{(\sum_{r \in S_n} \prod_{k \in r} \omega_k)^{1/n}}.$$

Then, exponential designs factorize as

$$p(s; \mathcal{S}_n) = \prod_{k \in s} \omega_k(n).$$

Let $\mathbf{d}$ be a realization of the partition $\boldsymbol{\delta}$, and let $\mathbf{m}$ satisfy $\mathbb{P}(\mathbf{n} = \mathbf{m}, \boldsymbol{\delta} = \mathbf{d}) > 0$. Set $E_{\mathbf{m}, \mathbf{d}} := \{r \in \mathcal{P}(U) : |r \cap U_j(\mathbf{d})| = m_j, j = 1, \dots, M\}$ be the set of compatible samples. For $s \in E_{\mathbf{m}, \mathbf{d}}$, we first compute the conditional sample probability:

$$\begin{aligned} \mathbb{P}(S = s \mid \mathbf{n} = \mathbf{m}, \boldsymbol{\delta} = \mathbf{d}) &= \frac{\mathbb{P}(S = s, \mathbf{n} = \mathbf{m}, \boldsymbol{\delta} = \mathbf{d})}{\mathbb{P}(\mathbf{n} = \mathbf{m}, \boldsymbol{\delta} = \mathbf{d})} \\ &= \frac{\mathbb{P}(S = s, \boldsymbol{\delta} = \mathbf{d})}{\mathbb{P}(\mathbf{n} = \mathbf{m}, \boldsymbol{\delta} = \mathbf{d})}. \end{aligned}$$

The last equality holds because $S = s$ and $\boldsymbol{\delta} = \mathbf{d}$ give $n_j = |s \cap U_j(\mathbf{d})| = m_j$ for every j . For the denominator, summing over all possible samples gives

$$\begin{aligned} \mathbb{P}(\mathbf{n} = \mathbf{m}, \boldsymbol{\delta} = \mathbf{d}) &= \sum_{r \in \mathcal{P}(U)} \mathbb{P}(S = r, \mathbf{n} = \mathbf{m}, \boldsymbol{\delta} = \mathbf{d}) \\ &= \sum_{r \in E_{\mathbf{m}, \mathbf{d}}} \mathbb{P}(S = r, \boldsymbol{\delta} = \mathbf{d}) \\ &= \mathbb{P}(\boldsymbol{\delta} = \mathbf{d}) \sum_{r \in E_{\mathbf{m}, \mathbf{d}}} p(r). \end{aligned}$$

Here the samples outside $E_{\mathbf{m}, \mathbf{d}}$ contribute zero, and the last equality uses $S \perp\!\!\!\perp \boldsymbol{\delta}$. The same independence gives

$$\mathbb{P}(S = s, \boldsymbol{\delta} = \mathbf{d}) = p(s) \mathbb{P}(\boldsymbol{\delta} = \mathbf{d}).$$

Substituting and cancelling $\mathbb{P}(\boldsymbol{\delta} = \mathbf{d})$, we obtain

$$\begin{aligned} \mathbb{P}(S = s \mid \mathbf{n} = \mathbf{m}, \boldsymbol{\delta} = \mathbf{d}) &= \frac{p(s)}{\sum_{r \in E_{\mathbf{m}, \mathbf{d}}} p(r)} \\ &= \frac{\mathbb{P}(S = s \mid |S| = n)}{\sum_{r \in E_{\mathbf{m}, \mathbf{d}}} \mathbb{P}(S = r \mid |S| = n)} \\ &= \frac{\prod_{k \in s} \omega_k(n)}{\sum_{r \in E_{\mathbf{m}, \mathbf{d}}} \prod_{k \in r} \omega_k(n)}. \end{aligned}$$

Every compatible sample has size n , so the common factor $\mathbb{P}(|S| = n)$ cancels in the second equality. The common normalizing constant of the exponential design cancels in the third.

Since the folds form a partition, each compatible sample is uniquely determined by its intersections with the folds:

$$E_{\mathbf{m}, \mathbf{d}} = \left\{ \bigcup_{j=1}^M r_j : r_j \subseteq U_j(\mathbf{d}), |r_j| = m_j, j = 1, \dots, M \right\}.$$

Thus, the denominator factors as

$$\begin{aligned} \sum_{r \in E_{\mathbf{m}, \mathbf{d}}} \prod_{k \in r} \omega_k(n) &= \sum_{\substack{r_1 \subseteq U_1(\mathbf{d}) \\ |r_1|=m_1}} \cdots \sum_{\substack{r_M \subseteq U_M(\mathbf{d}) \\ |r_M|=m_M}} \prod_{j=1}^M \prod_{k \in r_j} \omega_k(n) \\ &= \prod_{j=1}^M \sum_{\substack{r_j \subseteq U_j(\mathbf{d}) \\ |r_j|=m_j}} \prod_{k \in r_j} \omega_k(n). \end{aligned}$$

The numerator also factors over the folds, giving

$$\mathbb{P}(S = s \mid \mathbf{n} = \mathbf{m}, \boldsymbol{\delta} = \mathbf{d}) = \prod_{j=1}^M \frac{\prod_{k \in s \cap U_j(\mathbf{d})} \omega_k(n)}{\sum_{\substack{r_j \subseteq U_j(\mathbf{d}) \\ |r_j|=m_j}} \prod_{k \in r_j} \omega_k(n)}.$$

This proves (i).

Step 2: Conditional independence given fold sizes holds only for exponential designs

The following lemma characterizes exponential designs through ratios of sample probabilities.

Lemma 1. *Suppose that p has symmetric support, let $2 \leq n \leq N - 2$ be a possible sample size, and set $p_n(s) = \mathbb{P}[S = s \mid |S| = n]$, with $\mathcal{S}_n = \{s \in \mathcal{P}(U) : |s| = n\}$. Then, the following statements are equivalent.*

(i) *For every pair of distinct elements $k, l \in U$, the ratio*

$$\rho_{kl} = \frac{p_n(r \cup \{k\})}{p_n(r \cup \{l\})}$$

does not depend on $r \in \{s \subset U \setminus \{k, l\} : |s| = n - 1\}$.

(ii) *The sampling design p_n is exponential.*

Proof. *Proof of (ii) $\Rightarrow$ (i).* Assume that the design p_n is exponential. Then, for some positive weights $\omega_k(n)$,

$$p_n(s) \propto \prod_{k \in s} \omega_k(n).$$

Hence,

$$\rho_{kl} = \frac{\omega_k(n) \prod_{i \in r} \omega_i(n)}{\omega_l(n) \prod_{i \in r} \omega_i(n)} = \frac{\omega_k(n)}{\omega_l(n)},$$

from which the implication follows.

Proof of (i) $\Rightarrow$ (ii). Assume (i). For distinct $k, l, m \in U$, choose $r \subseteq U \setminus \{k, l, m\}$ with $|r| = n - 1$; this is possible since $n \leq N - 2$. Then,

$$\rho_{km} = \frac{p_n(r \cup \{k\})}{p_n(r \cup \{m\})} = \frac{p_n(r \cup \{k\})}{p_n(r \cup \{l\})} \times \frac{p_n(r \cup \{l\})}{p_n(r \cup \{m\})} = \rho_{kl} \rho_{lm}.$$

We now construct the weights. For an arbitrary $k_0 \in U$, let $\omega_{k_0}(n) = 1$, and $\omega_k(n) = \rho_{kk_0}$ for $k \in U \setminus \{k_0\}$. For distinct $k, l \in U \setminus \{k_0\}$, the identity above gives $\omega_k(n) = \rho_{kl}\rho_{lk_0} = \rho_{kl}\omega_l(n)$, and hence $\rho_{kl} = \omega_k(n)/\omega_l(n)$. The same equality is immediate if k or l equals k_0 .

Now, let $s, t \in \mathcal{S}_n$. Because both are of the same size, the sets $s \setminus t$ and $t \setminus s$ contain the same number of elements $L := n - |s \cap t|$; write $s \setminus t = \{a_1, \dots, a_L\}$ and $t \setminus s = \{b_1, \dots, b_L\}$. Starting from $t_0 = t$, set $t_q = (t_{q-1} \setminus \{b_q\}) \cup \{a_q\}$, so that $t_L = s$. At each exchange, the common set $t_{q-1} \setminus \{b_q\}$ has size $n - 1$ and contains neither a_q nor b_q . Hence,

$$\frac{p_n(t_q)}{p_n(t_{q-1})} = \frac{p_n((t_{q-1} \setminus \{b_q\}) \cup \{a_q\})}{p_n((t_{q-1} \setminus \{b_q\}) \cup \{b_q\})} = \rho_{a_q b_q} = \frac{\omega_{a_q}(n)}{\omega_{b_q}(n)}.$$

Multiplying these ratios gives

$$\frac{p_n(s)}{p_n(t)} = \frac{p_n(t_1)}{p_n(t_0)} \frac{p_n(t_2)}{p_n(t_1)} \dots \frac{p_n(t_L)}{p_n(t_{L-1})} = \prod_{k=1}^L \frac{\omega_{a_k}(n)}{\omega_{b_k}(n)} = \frac{\prod_{k \in s} \omega_k(n)}{\prod_{k \in t} \omega_k(n)}.$$

Thus, $p_n(s)/\prod_{k \in s} \omega_k(n)$ is constant over $\mathcal{S}_n$. Since $\sum_{s \in \mathcal{S}_n} p_n(s) = 1$, we get

$$p_n(s) = \frac{\prod_{k \in s} \omega_k(n)}{\sum_{t \in \mathcal{S}_n} \prod_{k \in t} \omega_k(n)}.$$

Since $\omega_k(n) > 0$, setting $\lambda_k(n) = \log \omega_k(n)$ shows that the design is exponential. $\square$

Proof of Step 2. Assume (i). Fix a possible sample size n and write $p_n(s) = \mathbb{P}(S = s \mid |S| = n)$. If $n = 0$ or $n = N$, then p_n is a point mass and is therefore exponential. If $n = 1$, take $\omega_k(1) = p_1(\{k\})$. If $n = N - 1$, take $\omega_k(N - 1) = 1/p_{N-1}(U \setminus \{k\})$. These choices give the exponential form directly. Thus, suppose that $2 \leq n \leq N - 2$.

Let $k, l \in U$ be distinct and let $r \in \{s \subset U \setminus \{k, l\} : |s| = n - 1\}$. We show that the ratio

$$\rho_{kl}(r) = \frac{p_n(r \cup \{k\})}{p_n(r \cup \{l\})}$$

does not depend on r . We first compare samples that differ by one element, then use a sequence of such exchanges.

Step 2. a. Let $r' \in \{s \subset U \setminus \{k, l\} : |s| = n - 1\}$ differ from r by one unit, so that $r' = (r \setminus \{a\}) \cup \{b\}$ for some elements a, b . The elements k, l, a, b are distinct. Since $N_1, N_2 \geq 2$ and the random partition has full support, we can choose a realization $\mathbf{d}$ such that $k, l \in U_1(\mathbf{d})$ and $a, b \in U_2(\mathbf{d})$. Consider the four samples $s_k = r \cup \{k\}$, $s_l = r \cup \{l\}$, $s'_k = r' \cup \{k\}$, and $s'_l = r' \cup \{l\}$. They have the same fold-count vector, which we denote by $\mathbf{m}$. Moreover, $\mathbb{P}(\mathbf{n} = \mathbf{m}, \boldsymbol{\delta} = \mathbf{d}) > 0$ because symmetry gives p_n full support on $\mathcal{S}_n$, the distribution of $\boldsymbol{\delta}$ has full support, and $\boldsymbol{\delta} \perp\!\!\!\perp S$.

For $s \in E_{\mathbf{m}, \mathbf{d}}$, set

$$Q(s) := \mathbb{P}(S = s \mid \mathbf{n} = \mathbf{m}, \boldsymbol{\delta} = \mathbf{d}).$$

We compare $Q(s_k)/Q(s_l)$ and $Q(s'_k)/Q(s'_l)$. Conditional independence between fold one and its complement gives

$$Q(s) = \mathbb{P}(S \cap U_1(\mathbf{d}) = s \cap U_1(\mathbf{d}) \mid \mathbf{n} = \mathbf{m}, \boldsymbol{\delta} = \mathbf{d}) \times \mathbb{P}(S \setminus U_1(\mathbf{d}) = s \setminus U_1(\mathbf{d}) \mid \mathbf{n} = \mathbf{m}, \boldsymbol{\delta} = \mathbf{d}).$$

In each ratio, the two samples agree outside $U_1(\mathbf{d})$, so the second factor cancels. Since $r' \cap U_1(\mathbf{d}) = r \cap U_1(\mathbf{d})$, the remaining factors give

$$\begin{aligned} \frac{Q(s_k)}{Q(s_l)} &= \frac{\mathbb{P}(S \cap U_1(\mathbf{d}) = (r \cap U_1(\mathbf{d})) \cup \{k\} \mid \mathbf{n} = \mathbf{m}, \boldsymbol{\delta} = \mathbf{d})}{\mathbb{P}(S \cap U_1(\mathbf{d}) = (r \cap U_1(\mathbf{d})) \cup \{l\} \mid \mathbf{n} = \mathbf{m}, \boldsymbol{\delta} = \mathbf{d})} \\ &= \frac{Q(s'_k)}{Q(s'_l)}. \end{aligned}$$

Now, on $E_{\mathbf{m}, \mathbf{d}}$,

$$\begin{aligned} Q(s) &= \frac{p(s) \mathbb{P}[\boldsymbol{\delta} = \mathbf{d} \mid S = s]}{\sum_{r \in E_{\mathbf{m}, \mathbf{d}}} p(r) \mathbb{P}[\boldsymbol{\delta} = \mathbf{d} \mid S = r]} \\ &= \frac{p(s)}{\sum_{r \in E_{\mathbf{m}, \mathbf{d}}} p(r)} \quad (\text{Using that } \boldsymbol{\delta} \perp\!\!\!\perp S) \\ &= \frac{p_n(s)}{\sum_{r \in E_{\mathbf{m}, \mathbf{d}}} p_n(r)}. \quad (\text{Since } p(s) = p_n(s) \mathbb{P}(|S| = n) \text{ for } s \in E_{\mathbf{m}, \mathbf{d}}) \end{aligned}$$

Therefore, $\rho_{kl}(r) = \rho_{kl}(r')$ when r and r' differ by one element.

Step 2. b. Finally, let $r, r' \subset U \setminus \{k, l\}$ have size $n - 1$, and set $L = n - 1 - |r \cap r'|$. We can construct r' from r by a sequence of L one-element swaps, which defines a sequence $r_0 = r, r_1, \dots, r_L = r'$ where

$$\rho_{kl}(r) = \rho_{kl}(r_0) = \rho_{kl}(r_1) = \dots = \rho_{kl}(r_L) = \rho_{kl}(r').$$

This follows by applying Step 2.a. to each one-element swap. Thus, $\rho_{kl}(r)$ does not depend on r . Lemma 1 shows that p_n is exponential. Since n was arbitrary, this proves (ii).

E.2 Proof of Theorem 2

Fix a partition $\mathbf{d}$ such that $\mathbb{P}(\mathbf{n}(S, \mathbf{d}) = \mathbf{n}) > 0$ and a sample $s \in \mathcal{P}(U)$. We show that the two schemes give the same conditional sample probabilities.

For the design-aware scheme, the numbers of ways to partition the sampled and nonsampled units are, respectively,

$$\frac{n!}{n_1! n_2! \dots n_M!} \quad \text{and} \quad \frac{(N - n)!}{(N_1 - n_1)! (N_2 - n_2)! \dots (N_M - n_M)!}.$$

Hence, the total number of possible pairs of partitions in Algorithm 2 is

$$K_{\text{aware}} = \frac{n!}{n_1! n_2! \dots n_M!} \frac{(N - n)!}{(N_1 - n_1)! (N_2 - n_2)! \dots (N_M - n_M)!}.$$

Given $S = s$, Algorithm 2 can produce $\mathbf{d}$ if and only if $\mathbf{n}(s, \mathbf{d}) = \mathbf{n}$. In that case, the sampled and nonsampled parts are uniquely determined by $S_j = s \cap U_j(\mathbf{d})$ and $R_j = U_j(\mathbf{d}) \setminus s$, for $j = 1, \dots, M$. Since all pairs of partitions are equally likely,

$$\mathbb{P}(S = s, \boldsymbol{\delta}_{\text{aware}} = \mathbf{d}) = \frac{p(s) \mathbb{1}_{\mathbf{n}(s, \mathbf{d}) = \mathbf{n}}}{K_{\text{aware}}}.$$

Summing over all samples gives

$$\mathbb{P}(\boldsymbol{\delta}_{\text{aware}} = \mathbf{d}) = \frac{1}{K_{\text{aware}}} \sum_{r \in \mathcal{P}(U)} p(r) \mathbb{1}_{\mathbf{n}(r, \mathbf{d}) = \mathbf{n}}.$$

Therefore,

$$\mathbb{P}(S = s | \boldsymbol{\delta}_{\text{aware}} = \mathbf{d}) = \frac{p(s) \mathbb{1}_{\mathbf{n}(s, \mathbf{d}) = \mathbf{n}}}{\sum_{r \in \mathcal{P}(U)} p(r) \mathbb{1}_{\mathbf{n}(r, \mathbf{d}) = \mathbf{n}}}.$$

For the uniform scheme, we use the independence of $\boldsymbol{\delta}_{\text{unif}}$ and S to obtain

$$\begin{aligned} & \mathbb{P}(S = s | \boldsymbol{\delta}_{\text{unif}} = \mathbf{d}, \mathbf{n}(S, \mathbf{d}) = \mathbf{n}) \\ &= \frac{\mathbb{P}(S = s, \mathbf{n}(S, \mathbf{d}) = \mathbf{n} | \boldsymbol{\delta}_{\text{unif}} = \mathbf{d})}{\mathbb{P}(\mathbf{n}(S, \mathbf{d}) = \mathbf{n} | \boldsymbol{\delta}_{\text{unif}} = \mathbf{d})} \\ &= \frac{\mathbb{P}(S = s | \boldsymbol{\delta}_{\text{unif}} = \mathbf{d}) \mathbb{1}_{\mathbf{n}(s, \mathbf{d}) = \mathbf{n}}}{\sum_{r \in \mathcal{P}(U)} \mathbb{P}(S = r | \boldsymbol{\delta}_{\text{unif}} = \mathbf{d}) \mathbb{1}_{\mathbf{n}(r, \mathbf{d}) = \mathbf{n}}} \\ &= \frac{p(s) \mathbb{1}_{\mathbf{n}(s, \mathbf{d}) = \mathbf{n}}}{\sum_{r \in \mathcal{P}(U)} p(r) \mathbb{1}_{\mathbf{n}(r, \mathbf{d}) = \mathbf{n}}} \\ &= \mathbb{P}(S = s | \boldsymbol{\delta}_{\text{aware}} = \mathbf{d}). \end{aligned}$$

Thus, the two conditional laws are equal. It follows that one law factorizes across folds if and only if the other does. Applying this result to every target vector gives the stated equivalence between the two conditional independence properties. The equality of the first- and second-order conditional inclusion probabilities follows by summing the common probabilities over samples containing k and samples containing both k and l , respectively.

E.3 Proof of Proposition 1

By Theorem 2, under SRSWOR, all samples with the prescribed fold counts have the same conditional probability given $\mathcal{A}$. There are $\prod_{j=1}^M \binom{N_{j,v}}{n_{j,v}}$ such samples, since the units in each fold can be chosen separately. Thus, for every compatible s ,

$$\mathbb{P}(S_v = s | \mathcal{A}) = \frac{1}{\prod_{j=1}^M \binom{N_{j,v}}{n_{j,v}}} = \prod_{j=1}^M \frac{1}{\binom{N_{j,v}}{n_{j,v}}}.$$

Each factor is the probability of an SRSWOR sample in its fold. Hence, the fold samples are conditionally independent, the scheme is honest, and

$$\pi_{k|\mathcal{A}} = \frac{n_{j,v}}{N_{j,v}}, \quad k \in U_{j,v}.$$

For distinct $k, l \in U_{j,v}$,

$$\pi_{kl|\mathcal{A}} = \frac{n_{j,v}(n_{j,v} - 1)}{N_{j,v}(N_{j,v} - 1)}, \quad \Delta_{kl|\mathcal{A}} = -\frac{(n_{j,v}/N_{j,v})(1 - n_{j,v}/N_{j,v})}{N_{j,v} - 1}.$$

For units in different folds, $\pi_{kl|\mathcal{A}} = \pi_{k|\mathcal{A}}\pi_{l|\mathcal{A}}$ and $\Delta_{kl|\mathcal{A}} = 0$.

The fold-size condition gives (CF1). By (D1) and the allocation condition, $n_{j,v}/N_{j,v} \xrightarrow{v \rightarrow \infty} \pi_\star > 0$, uniformly over the fixed number of folds. Since $n_{j,v} \geq 1$ for every j, v , these ratios are bounded below by a positive constant over all j, v . This proves (CF2,b), (CF2,c), and (CF6). The second-order probabilities for distinct units also converge uniformly to $\pi_\star^2$, which gives (CF2,a).

For (CF3), use $\pi_k = n_v/N_v$ and the allocation condition to obtain

$$\mathbb{E}_p \left[\max_{k \in U_v} (\pi_{k|\mathcal{A}} - \pi_k)^2 \right] = \max_{1 \leq j \leq M} \left(\frac{n_{j,v}}{N_{j,v}} - \frac{n_v}{N_v} \right)^2 = \mathcal{O}(n_v^{-1}).$$

For (CF4), the covariance formula gives, for all sufficiently large v ,

$$n_v \max_{k \neq l \in U_v} |\Delta_{kl|\mathcal{A}}| \leq \frac{n_v}{\min_j N_{j,v} - 1} = \mathcal{O}(1),$$

since each $N_{j,v}$ is of order N_v and $n_v \leq N_v$. The constants in both bounds can be enlarged to cover the finitely many initial populations.

It remains to verify (CF5). For $1 \leq r \leq 4$ distinct units $k_1, \dots, k_r$ in the same fold, and all sufficiently large v ,

$$\mathbb{E}_p \left[\prod_{t=1}^r I_{k_t} \mid \mathcal{A} \right] = \prod_{t=0}^{r-1} \frac{n_{j,v} - t}{N_{j,v} - t} = \left(\frac{n_{j,v}}{N_{j,v}} \right)^r + \mathcal{O}(N_{j,v}^{-1}).$$

For the last bound, all factors are in $[0, 1]$, so replacing them one at a time gives

$$\left| \prod_{t=0}^{r-1} \frac{n_{j,v} - t}{N_{j,v} - t} - \left(\frac{n_{j,v}}{N_{j,v}} \right)^r \right| \leq \sum_{t=0}^{r-1} \left| \frac{n_{j,v} - t}{N_{j,v} - t} - \frac{n_{j,v}}{N_{j,v}} \right| \leq \sum_{t=0}^{r-1} \frac{t}{N_{j,v} - t} = \mathcal{O}(N_{j,v}^{-1}).$$

By independence across folds, both $\mathbb{E}_p[I_i I_j I_k I_l \mid \mathcal{A}]$ and $\pi_{ij|\mathcal{A}} \pi_{kl|\mathcal{A}}$ differ from $\pi_{i|\mathcal{A}} \pi_{j|\mathcal{A}} \pi_{k|\mathcal{A}} \pi_{l|\mathcal{A}}$ by $\mathcal{O}(N_v^{-1})$, since each $N_{j,v}$ is of order N_v . Subtracting gives, uniformly over four distinct units,

$$\mathbb{E}_p \left[(I_i I_j - \pi_{ij|\mathcal{A}})(I_k I_l - \pi_{kl|\mathcal{A}}) \mid \mathcal{A} \right] = \mathcal{O}(N_v^{-1}) = o(1).$$

This proves (CF5) and completes the proof.

F Proofs of Section 4

We first give a second-moment bound for centered sums within the folds.

Lemma 2. Assume (Ind- $\mathcal{A}$) and (CF4). Let $(d_k)_{k \in U_v}$ be square-integrable random variables such that, for each j , $(d_k)_{k \in U_{j,v}}$ is measurable with respect to $\mathcal{A}$ and $\{I_q\}_{q \in U_v \setminus U_{j,v}}$. Then,

$$n_v \mathbb{E}_p \left[\left\{ \frac{1}{N_v} \sum_{k \in U_v} (I_k - \pi_{k|\mathcal{A}}) d_k \right\}^2 \right] \leq \frac{C}{N_v} \sum_{k \in U_v} \mathbb{E}_p(d_k^2),$$

where C depends only on M and $C_{\Delta 2}$.

Proof. Conditional on $\mathcal{A}$ and the sampling indicators outside $U_{j,v}$, the coefficients d_k , $k \in U_{j,v}$, are fixed. By (Ind- $\mathcal{A}$),

$$\begin{aligned} & \mathbb{E}_p \left[\left\{ \sum_{k \in U_{j,v}} (I_k - \pi_{k|\mathcal{A}}) d_k \right\}^2 \mid \mathcal{A}, \{I_q\}_{q \in U_v \setminus U_{j,v}} \right] \\ &= \sum_{k, l \in U_{j,v}} d_k d_l \mathbb{E}_p \left[(I_k - \pi_{k|\mathcal{A}})(I_l - \pi_{l|\mathcal{A}}) \mid \mathcal{A}, \{I_q\}_{q \in U_v \setminus U_{j,v}} \right] \\ &= \sum_{k, l \in U_{j,v}} \Delta_{kl|\mathcal{A}} d_k d_l \\ &\leq \frac{1}{4} \sum_{k \in U_{j,v}} d_k^2 + \frac{C_{\Delta 2}}{n_v} \left(\sum_{k \in U_{j,v}} |d_k| \right)^2 \\ &\leq \left(\frac{1}{4} + \frac{C_{\Delta 2} N_{j,v}}{n_v} \right) \sum_{k \in U_{j,v}} d_k^2, \end{aligned}$$

where we used $\Delta_{kk|\mathcal{A}} \leq 1/4$, (CF4), and the Cauchy–Schwarz inequality. For the sum over all folds, Cauchy–Schwarz and the preceding bound give

$$\begin{aligned} & n_v \mathbb{E}_p \left[\left\{ \frac{1}{N_v} \sum_{k \in U_v} (I_k - \pi_{k|\mathcal{A}}) d_k \right\}^2 \right] \\ &\leq \frac{n_v M}{N_v^2} \sum_{j=1}^M \mathbb{E}_p \left[\left\{ \sum_{k \in U_{j,v}} (I_k - \pi_{k|\mathcal{A}}) d_k \right\}^2 \right] \\ &\leq \frac{M}{N_v^2} \sum_{j=1}^M \mathbb{E}_p \left[\left(\frac{n_v}{4} + C_{\Delta 2} N_{j,v} \right) \sum_{k \in U_{j,v}} d_k^2 \right] \\ &\leq \frac{C}{N_v} \sum_{k \in U_v} \mathbb{E}_p(d_k^2), \end{aligned}$$

where we used $n_v \leq N_v$ and $N_{j,v} \leq N_v$. □

F.1 Proof of Proposition 2

By the law of total expectation, it is enough to show that $\mathbb{E}_p[\hat{\mu}_{cf}(\hat{m}, \pi_{|\mathcal{A}}) \mid \mathcal{A}] = \mu$. Since $\hat{m}_j$ is fitted using only the observations outside U_j , it is measurable with respect to $\mathcal{A}$ and $\{I_l\}_{l \in U \setminus U_j}$. Moreover, by (Ind- $\mathcal{A}$),

$$\mathbb{E}_p \left[I_k \mid \mathcal{A}, \{I_l\}_{l \in U \setminus U_j} \right] = \mathbb{E}_p[I_k \mid \mathcal{A}] = \pi_{k|\mathcal{A}}, \quad k \in U_j.$$

Consequently, for $k \in U_j$,

$$\begin{aligned} \mathbb{E}_p \left[\hat{m}_j(\mathbf{x}_k) + I_k \frac{y_k - \hat{m}_j(\mathbf{x}_k)}{\pi_{k|\mathcal{A}}} \mid \mathcal{A}, \{I_l\}_{l \in U \setminus U_j} \right] \\ = \hat{m}_j(\mathbf{x}_k) + \pi_{k|\mathcal{A}} \frac{y_k - \hat{m}_j(\mathbf{x}_k)}{\pi_{k|\mathcal{A}}} = y_k. \end{aligned}$$

Therefore, by the tower property,

$$\begin{aligned} \mathbb{E}_p \left[\hat{\mu}_{cf}(\hat{m}, \pi_{|\mathcal{A}}) \mid \mathcal{A} \right] &= \frac{1}{N} \sum_{j=1}^M \sum_{k \in U_j} \mathbb{E}_p \left[\hat{m}_j(\mathbf{x}_k) + I_k \frac{y_k - \hat{m}_j(\mathbf{x}_k)}{\pi_{k|\mathcal{A}}} \mid \mathcal{A} \right] \\ &= \frac{1}{N} \sum_{j=1}^M \sum_{k \in U_j} y_k = \mu. \end{aligned}$$

F.2 Proof of Theorem 3

Let $\mathcal{E}_v := \{\min_{k \in U_v} \pi_{k|\mathcal{A}} \geq \lambda_{cf}\}$ and let $\epsilon > 0$. By Markov's inequality,

$$\begin{aligned} \mathbb{P} \left(\sqrt{n_v} \left| \hat{\mu}_{cf}(\hat{m}, \pi_{|\mathcal{A}}) - \hat{\mu}_{ma}(\tilde{m}, \pi_{|\mathcal{A}}) \right| > \epsilon \right) \\ \leq \mathbb{P}(\mathcal{E}_v^c) + \frac{n_v}{\epsilon^2} \mathbb{E}_p \left[\left\{ \hat{\mu}_{cf}(\hat{m}, \pi_{|\mathcal{A}}) - \hat{\mu}_{ma}(\tilde{m}, \pi_{|\mathcal{A}}) \right\}^2 \mathbf{1}_{\mathcal{E}_v} \right]. \end{aligned}$$

Assumption (CF2,a) gives $\mathbb{P}(\mathcal{E}_v^c) \xrightarrow{v \rightarrow \infty} 0$. It remains to control the expectation.

For $k \in U_v$, set $\rho_k := \hat{m}^{(-k)}(\mathbf{x}_k) - \tilde{m}(\mathbf{x}_k)$. On $\mathcal{E}_v$,

$$\hat{\mu}_{cf}(\hat{m}, \pi_{|\mathcal{A}}) - \hat{\mu}_{ma}(\tilde{m}, \pi_{|\mathcal{A}}) = -\frac{1}{N_v} \sum_{k \in U_v} (I_k - \pi_{k|\mathcal{A}}) \frac{\rho_k}{\pi_{k|\mathcal{A}}}.$$

Apply Lemma 2 with $d_k = \rho_k / \pi_{k|\mathcal{A}}$ on $\mathcal{E}_v$ and $d_k = 0$ otherwise. The required measurability follows because ρ_k uses only observations outside its fold, while $\mathcal{E}_v$ is determined by $\mathcal{A}$. Since $d_k^2 \leq \rho_k^2 / \lambda_{cf}^2$, the lemma gives

$$n_v \mathbb{E}_p \left[\left\{ \hat{\mu}_{cf}(\hat{m}, \pi_{|\mathcal{A}}) - \hat{\mu}_{ma}(\tilde{m}, \pi_{|\mathcal{A}}) \right\}^2 \mathbf{1}_{\mathcal{E}_v} \right] \leq \frac{C}{N_v} \sum_{k \in U_v} \mathbb{E}_p(\rho_k^2),$$

where C is independent of v . The right-hand side converges to zero by (M1). This proves the result.

F.3 Proof of Theorem 4

Let $\mathcal{E}_v := \{\min_{k \in U_v} \pi_{k|\mathcal{A}} \geq \lambda_{cf}\}$. On $\mathcal{E}_v$,

$$\hat{\mu}_{\text{ma}}(\tilde{m}, \pi_{|\mathcal{A}}) - \mu = \frac{1}{N_v} \sum_{k \in U_v} \tilde{e}_k \left(\frac{I_k}{\pi_{k|\mathcal{A}}} - 1 \right).$$

Its conditional mean is

$$\mathbb{E}_p \left[\hat{\mu}_{\text{ma}}(\tilde{m}, \pi_{|\mathcal{A}}) - \mu \mid \mathcal{A} \right] = \frac{1}{N_v} \sum_{k \in U_v} \tilde{e}_k \left(\frac{\pi_{k|\mathcal{A}}}{\pi_{k|\mathcal{A}}} - 1 \right) = 0.$$

Therefore,

$$\begin{aligned} n_v \mathbb{E}_p \left[\left(\hat{\mu}_{\text{ma}}(\tilde{m}, \pi_{|\mathcal{A}}) - \mu \right)^2 \mid \mathcal{A} \right] &= n_v \mathbb{V}_p \left(\hat{\mu}_{\text{ma}}(\tilde{m}, \pi_{|\mathcal{A}}) \mid \mathcal{A} \right) \\ &= n_v \left\{ \frac{1}{N_v^2} \sum_{k \in U_v} \Delta_{kk|\mathcal{A}} \frac{\tilde{e}_k^2}{\pi_{k|\mathcal{A}}^2} + \frac{1}{N_v^2} \sum_{k \in U_v} \sum_{\substack{l \in U_v \\ l \neq k}} \Delta_{kl|\mathcal{A}} \frac{\tilde{e}_k}{\pi_{k|\mathcal{A}}} \frac{\tilde{e}_l}{\pi_{l|\mathcal{A}}} \right\} \\ &\leq \frac{1}{N_v \lambda_{cf}^2} \sum_{k \in U_v} \tilde{e}_k^2 \left\{ \frac{n_v}{N_v} + C_{\Delta 2} \right\}. \end{aligned}$$

Here we used (CF4) and

$$\sum_{k \in U_v} \sum_{\substack{l \in U_v \\ l \neq k}} |\tilde{e}_k \tilde{e}_l| \leq \left(\sum_{k \in U_v} |\tilde{e}_k| \right)^2 \leq N_v \sum_{k \in U_v} \tilde{e}_k^2.$$

Multiplying by $\mathbf{1}_{\mathcal{E}_v}$, taking expectations, and using (M2,a), we obtain, for all sufficiently large v ,

$$n_v \mathbb{E}_p \left[\left\{ \hat{\mu}_{\text{ma}}(\tilde{m}, \pi_{|\mathcal{A}}) - \mu \right\}^2 \mathbf{1}_{\mathcal{E}_v} \right] \leq C,$$

for some constant C . Now, Markov's inequality gives, for every $K > 0$,

$$\begin{aligned} &\mathbb{P} \left(\left| \sqrt{n_v} \left\{ \hat{\mu}_{\text{ma}}(\tilde{m}, \pi_{|\mathcal{A}}) - \mu \right\} \right| > K \right) \\ &\leq \mathbb{P}(\mathcal{E}_v^c) + \frac{n_v}{K^2} \mathbb{E}_p \left[\left\{ \hat{\mu}_{\text{ma}}(\tilde{m}, \pi_{|\mathcal{A}}) - \mu \right\}^2 \mathbf{1}_{\mathcal{E}_v} \right] \leq \mathbb{P}(\mathcal{E}_v^c) + \frac{C}{K^2}. \end{aligned}$$

Assumption (CF2,a) gives $\mathbb{P}(\mathcal{E}_v^c) \xrightarrow{v \rightarrow \infty} 0$. The result follows by choosing K large enough.

F.4 Proof of Theorem 5

Write

$$\begin{aligned} & \left| \widehat{V}_{cf}(\widehat{m}, \pi_{|\mathcal{A}}) - \mathbb{V}_p \left\{ \widehat{\mu}_{\text{ma}}(\widetilde{m}, \pi_{|\mathcal{A}}) \right\} \right| \\ & \leq \left| \widehat{V}_{cf}(\widehat{m}, \pi_{|\mathcal{A}}) - \mathbb{V}_p \left\{ \widehat{\mu}_{\text{ma}}(\widetilde{m}, \pi_{|\mathcal{A}}) \mid \mathcal{A} \right\} \right| + \left| \mathbb{V}_p \left\{ \widehat{\mu}_{\text{ma}}(\widetilde{m}, \pi_{|\mathcal{A}}) \mid \mathcal{A} \right\} - \mathbb{V}_p \left\{ \widehat{\mu}_{\text{ma}}(\widetilde{m}, \pi_{|\mathcal{A}}) \right\} \right| \\ & =: T_{1v} + T_{2v}. \end{aligned}$$

We first consider T_{1v} . Let

$$\mathcal{E}_v := \left\{ \min_{k \in U_v} \pi_{k|\mathcal{A}} \geq \lambda_{cf}, \min_{\substack{k, l \in U_v \\ k \neq l}} \pi_{kl|\mathcal{A}} \geq \lambda_{cf}^* \right\}.$$

By (CF2,a), $\mathbb{P}(\mathcal{E}_v^c) \xrightarrow{v \rightarrow \infty} 0$. On $\mathcal{E}_v$, set $c_{kl|\mathcal{A}} := \Delta_{kl|\mathcal{A}} / (\pi_{k|\mathcal{A}} \pi_{l|\mathcal{A}})$. We have

$$\begin{aligned} T_{1v} & \leq \left| \frac{1}{N_v^2} \sum_{k \in U_v} \sum_{l \in U_v} c_{kl|\mathcal{A}} \widetilde{e}_k \widetilde{e}_l \left(\frac{I_k I_l}{\pi_{kl|\mathcal{A}}} - 1 \right) \right| \\ & \quad + \left| \frac{1}{N_v^2} \sum_{k \in S_v} \sum_{l \in S_v} \frac{c_{kl|\mathcal{A}}}{\pi_{kl|\mathcal{A}}} \left(\widehat{e}_{(-k)} \widehat{e}_{(-l)} - \widetilde{e}_k \widetilde{e}_l \right) \right| \\ & =: A_v + B_v. \end{aligned}$$

For A_v , we condition on $\mathcal{A}$ and use the argument in the proof of Theorem 3 of [Breidt and Opsomer \(2000\)](#), with the conditional inclusion probabilities in place of the unconditional ones. On $\mathcal{E}_v$, the required lower bounds follow from (CF2,a), while (CF4), (CF5), and (M2,b) give the remaining bounds. Hence, $n_v A_v = o_{\mathbb{P}}(1)$ on $\mathcal{E}_v$.

We next control B_v . Set $d_k := \widehat{e}_{(-k)} - \widetilde{e}_k = \widetilde{m}(\mathbf{x}_k) - \widehat{m}^{(-k)}(\mathbf{x}_k)$. For $k, l \in S_v$,

$$\widehat{e}_{(-k)} \widehat{e}_{(-l)} - \widetilde{e}_k \widetilde{e}_l = d_k \widetilde{e}_l + d_l \widetilde{e}_k + d_k d_l.$$

On $\mathcal{E}_v$, the diagonal and off-diagonal coefficients satisfy

$$\frac{|c_{kk|\mathcal{A}}|}{\pi_{kk|\mathcal{A}}} = \frac{1 - \pi_{k|\mathcal{A}}}{\pi_{k|\mathcal{A}}^2} \leq \frac{1}{\lambda_{cf}^2}, \quad \max_{k \neq l \in U_v} \frac{|c_{kl|\mathcal{A}}|}{\pi_{kl|\mathcal{A}}} \leq \frac{C_{\Delta 2}}{n_v \lambda_{cf}^2 \lambda_{cf}^*},$$

by (CF4). Taking absolute values and bounding the sample sums by population sums gives

$$\begin{aligned} n_v B_v & \leq \frac{n_v}{N_v^2 \lambda_{cf}^2} \sum_{k \in U_v} \left(2|d_k \widetilde{e}_k| + d_k^2 \right) \\ & \quad + \frac{C_{\Delta 2}}{N_v^2 \lambda_{cf}^2 \lambda_{cf}^*} \left[2 \left(\sum_{k \in U_v} |d_k| \right) \left(\sum_{k \in U_v} |\widetilde{e}_k| \right) + \left(\sum_{k \in U_v} |d_k| \right)^2 \right]. \end{aligned}$$

Cauchy–Schwarz gives

$$\begin{aligned} \frac{1}{N_v} \sum_{k \in U_v} |d_k \tilde{e}_k| &\leq \left\{ \left(\frac{1}{N_v} \sum_{k \in U_v} d_k^2 \right) \left(\frac{1}{N_v} \sum_{k \in U_v} \tilde{e}_k^2 \right) \right\}^{1/2}, \\ \frac{1}{N_v^2} \left(\sum_{k \in U_v} |d_k| \right)^2 &\leq \frac{1}{N_v} \sum_{k \in U_v} d_k^2, \\ \frac{1}{N_v^2} \left(\sum_{k \in U_v} |d_k| \right) \left(\sum_{k \in U_v} |\tilde{e}_k| \right) &\leq \left\{ \left(\frac{1}{N_v} \sum_{k \in U_v} d_k^2 \right) \left(\frac{1}{N_v} \sum_{k \in U_v} \tilde{e}_k^2 \right) \right\}^{1/2}. \end{aligned}$$

Together with $n_v \leq N_v$, these bounds give

$$n_v B_v \leq C_B \left[2 \left\{ \left(\frac{1}{N_v} \sum_{k \in U_v} d_k^2 \right) \left(\frac{1}{N_v} \sum_{k \in U_v} \tilde{e}_k^2 \right) \right\}^{1/2} + \frac{1}{N_v} \sum_{k \in U_v} d_k^2 \right],$$

where C_B is a constant independent of v . By (M1) and Markov's inequality, $N_v^{-1} \sum_k d_k^2 = o_{\mathbb{P}}(1)$. Moreover, by the Cauchy–Schwarz inequality and (M2,b),

$$\left(\frac{1}{N_v} \sum_{k \in U_v} \tilde{e}_k^2 \right)^2 \leq \frac{1}{N_v} \sum_{k \in U_v} \tilde{e}_k^4 = \mathcal{O}(1).$$

Hence, $n_v B_v = o_{\mathbb{P}}(1)$ on $\mathcal{E}_v$. Since $\mathbb{P}(\mathcal{E}_v^c) \xrightarrow{v \rightarrow \infty} 0$, we obtain $n_v T_{1v} = o_{\mathbb{P}}(1)$.

For T_{2v} , set

$$Z_v := n_v \mathbb{V}_p \left\{ \hat{\mu}_{\text{ma}}(\tilde{m}, \pi|_{\mathcal{A}}) \mid \mathcal{A} \right\}.$$

The law of total variance gives

$$n_v T_{2v} \leq |Z_v - \mathbb{E}_p(Z_v)| + n_v \mathbb{V}_p \left\{ \mathbb{E}_p \left[\hat{\mu}_{\text{ma}}(\tilde{m}, \pi|_{\mathcal{A}}) \mid \mathcal{A} \right] \right\}.$$

We first show that $Z_v - \mathbb{E}_p(Z_v) = o_{\mathbb{P}}(1)$. Since $Z_v \xrightarrow{v \rightarrow \infty} c$ by assumption, it is enough to verify uniform integrability. In the sums below, terms for which a first-order conditional inclusion probability is zero are omitted. By (CF4),

$$\begin{aligned} Z_v &\leq \frac{n_v}{N_v^2} \sum_{k \in U_v} \frac{\tilde{e}_k^2}{\pi_+} + \frac{C_{\Delta 2}}{N_v^2 \pi_+^2} \sum_{k \in U_v} \sum_{\substack{l \in U_v \\ l \neq k}} |\tilde{e}_k \tilde{e}_l| \\ &\leq \left\{ \frac{n_v}{N_v \pi_+} + \frac{C_{\Delta 2}}{\pi_+^2} \right\} \frac{1}{N_v} \sum_{k \in U_v} \tilde{e}_k^2. \end{aligned}$$

By (M2,b),

$$C_e := \sup_{v \in \mathbb{N}} \frac{1}{N_v} \sum_{k \in U_v} \tilde{e}_k^2 < \infty.$$

Since $n_v/N_v \leq 1$ and $\pi_+ \leq 1$,

$$Z_v \leq \frac{C_e(1 + C_{\Delta 2})}{\pi_+^2}.$$

It follows from (CF6) that

$$\sup_{v \in \mathbb{N}} \mathbb{E}_p \left(Z_v^{1+\delta/2} \right) \leq \{C_e(1 + C_{\Delta 2})\}^{1+\delta/2} \sup_{v \in \mathbb{N}} \mathbb{E}_p \left(\frac{1}{\pi_+^{2+\delta}} \right) < \infty.$$

Thus, $(Z_v)_{v \in \mathbb{N}}$ is uniformly integrable, so $\mathbb{E}_p(Z_v) \xrightarrow{v \rightarrow \infty} c$, and therefore

$$Z_v - \mathbb{E}_p(Z_v) = o_{\mathbb{P}}(1).$$

It remains to bound the variance of the conditional mean. Let $\mathcal{F}_v := \{\min_{k \in U_v} \pi_{k|\mathcal{A}} > 0\}$. On $\mathcal{F}_v$, the oracle estimator is conditionally unbiased. In general, however,

$$\begin{aligned} & \mathbb{E}_p \left[\hat{\mu}_{\text{ma}}(\tilde{m}, \pi_{|\mathcal{A}}) \mid \mathcal{A} \right] - \mu \\ &= \frac{1}{N_v} \left\{ \sum_{k \in U_v} \tilde{m}(\mathbf{x}_k) + \sum_{\substack{k \in U_v \\ \pi_{k|\mathcal{A}} > 0}} \tilde{e}_k - \sum_{k \in U_v} y_k \right\} \\ &= -\frac{1}{N_v} \sum_{\substack{k \in U_v \\ \pi_{k|\mathcal{A}} = 0}} \tilde{e}_k. \end{aligned}$$

The Cauchy–Schwarz inequality gives

$$\begin{aligned} & \left| \mathbb{E}_p \left[\hat{\mu}_{\text{ma}}(\tilde{m}, \pi_{|\mathcal{A}}) \mid \mathcal{A} \right] - \mu \right| \\ & \leq \left\{ \left(\frac{1}{N_v} \sum_{k \in U_v} \mathbb{1}_{\{\pi_{k|\mathcal{A}} = 0\}} \right) \left(\frac{1}{N_v} \sum_{k \in U_v} \tilde{e}_k^2 \right) \right\}^{1/2} \\ & \leq C_e^{1/2} \mathbb{1}_{\mathcal{F}_v^c}. \end{aligned}$$

The variance is at most the mean squared deviation from μ . Hence, by (CF2,b),

$$\begin{aligned} & \mathbb{V}_p \left\{ \mathbb{E}_p \left[\hat{\mu}_{\text{ma}}(\tilde{m}, \pi_{|\mathcal{A}}) \mid \mathcal{A} \right] \right\} \\ & \leq \mathbb{E}_p \left[\left\{ \mathbb{E}_p \left[\hat{\mu}_{\text{ma}}(\tilde{m}, \pi_{|\mathcal{A}}) \mid \mathcal{A} \right] - \mu \right\}^2 \right] \\ & \leq C_e \mathbb{P}(\mathcal{F}_v^c) = o(n_v^{-1}). \end{aligned}$$

These two bounds give $n_v T_{2v} = o_{\mathbb{P}}(1)$. Combining the bounds for T_{1v} and T_{2v} proves the result.

F.5 Proof of Corollary 1

The proof of Theorem 5 gives

$$n_v \mathbb{V}_p \left\{ \hat{\mu}_{\text{ma}}(\tilde{m}, \pi_{|\mathcal{A}}) \right\} \xrightarrow{v \rightarrow \infty} c > 0.$$

Together with Theorem 5, this implies

$$\frac{\widehat{V}_{cf}(\widehat{m}, \pi_{|\mathcal{A}})}{\mathbb{V}_p\{\widehat{\mu}_{ma}(\widetilde{m}, \pi_{|\mathcal{A}})\}} \xrightarrow[v \rightarrow \infty]{\mathbb{P}} 1.$$

In particular, $\mathbb{P}\{\widehat{V}_{cf}(\widehat{m}, \pi_{|\mathcal{A}}) > 0\} \xrightarrow[v \rightarrow \infty]{} 1$. Moreover, the oracle standard deviation is of order $n_v^{-1/2}$, so Theorem 3 gives

$$\frac{\widehat{\mu}_{cf}(\widehat{m}, \pi_{|\mathcal{A}}) - \widehat{\mu}_{ma}(\widetilde{m}, \pi_{|\mathcal{A}})}{\sqrt{\mathbb{V}_p\{\widehat{\mu}_{ma}(\widetilde{m}, \pi_{|\mathcal{A}})\}}} = o_{\mathbb{P}}(1).$$

On the event that the estimated variance is positive,

$$\begin{aligned} & \frac{\widehat{\mu}_{cf}(\widehat{m}, \pi_{|\mathcal{A}}) - \mu}{\sqrt{\widehat{V}_{cf}(\widehat{m}, \pi_{|\mathcal{A}})}} \\ &= \left[\frac{\widehat{\mu}_{ma}(\widetilde{m}, \pi_{|\mathcal{A}}) - \mu}{\sqrt{\mathbb{V}_p\{\widehat{\mu}_{ma}(\widetilde{m}, \pi_{|\mathcal{A}})\}}} + o_{\mathbb{P}}(1) \right] \left[\frac{\mathbb{V}_p\{\widehat{\mu}_{ma}(\widetilde{m}, \pi_{|\mathcal{A}})\}}{\widehat{V}_{cf}(\widehat{m}, \pi_{|\mathcal{A}})} \right]^{1/2}. \end{aligned}$$

The result follows from the assumed central limit theorem and Slutsky's theorem.

G Proofs of Section 5

G.1 Proof of Theorem 6

We first split the difference between the two estimators into two terms:

$$\begin{aligned} & \sqrt{n_v} \{\widehat{\mu}_{cf}(\widehat{m}, \pi) - \widehat{\mu}_{ma}(\widetilde{m}, \pi)\} \\ &= \frac{\sqrt{n_v}}{N_v} \sum_{j=1}^M \sum_{k \in U_{j,v}} \frac{I_k - \pi_{k|\mathcal{A}}}{\pi_k} \{\widetilde{m}(\mathbf{x}_k) - \widehat{m}^{(-k)}(\mathbf{x}_k)\} \\ &+ \frac{\sqrt{n_v}}{N_v} \sum_{j=1}^M \sum_{k \in U_{j,v}} \left(\frac{\pi_{k|\mathcal{A}}}{\pi_k} - 1 \right) \{\widetilde{m}(\mathbf{x}_k) - \widehat{m}^{(-k)}(\mathbf{x}_k)\} \\ &:= \sum_{j=1}^M \left\{ \sqrt{n_v} T_{j,v}^{(1)} + \sqrt{n_v} T_{j,v}^{(2)} \right\}. \end{aligned}$$

We first control the centered term $\sqrt{n_v} \sum_{j=1}^M T_{j,v}^{(1)}$. For $k \in U_v$, set $d_k := \widetilde{m}(\mathbf{x}_k) - \widehat{m}^{(-k)}(\mathbf{x}_k)$. Apply Lemma 2 to the coefficients d_k/π_k , which depend only on the observations outside the fold containing k and on $\mathcal{A}$. By (D2) and (M1),

$$n_v \mathbb{E}_p \left[\left\{ \sum_{j=1}^M T_{j,v}^{(1)} \right\}^2 \right] \leq \frac{C}{N_v} \sum_{k \in U_v} \mathbb{E}_p(d_k^2) = o(1),$$

where C is independent of v . Thus, Markov's inequality gives $\sqrt{n_v} \sum_{j=1}^M T_{j,v}^{(1)} = o_{\mathbb{P}}(1)$.

For the second term, the Cauchy–Schwarz inequality gives

$$\begin{aligned} n_v |T_{j,v}^{(2)}|^2 &\leq \frac{n_v}{N_v^2 \lambda^2} \left(\sum_{k \in U_{j,v}} |\pi_{k|\mathcal{A}} - \pi_k| |d_k| \right)^2 \\ &\leq \frac{n_v N_{j,v}}{N_v^2 \lambda^2} \max_{k \in U_v} (\pi_{k|\mathcal{A}} - \pi_k)^2 \sum_{k \in U_{j,v}} d_k^2 \\ &\leq \frac{n_v}{\lambda^2} \max_{k \in U_v} (\pi_{k|\mathcal{A}} - \pi_k)^2 \frac{1}{N_v} \sum_{k \in U_v} d_k^2. \end{aligned}$$

For every $\epsilon, K > 0$, the last bound and Markov's inequality give

$$\begin{aligned} \mathbb{P} \left(\sqrt{n_v} |T_{j,v}^{(2)}| > \epsilon \right) &\leq \mathbb{P} \left(n_v \max_{k \in U_v} (\pi_{k|\mathcal{A}} - \pi_k)^2 > K \right) \\ &\quad + \mathbb{P} \left(\frac{1}{N_v} \sum_{k \in U_v} d_k^2 > \frac{\epsilon^2 \lambda^2}{K} \right) \\ &\leq \frac{C_1}{K} + \frac{K}{\epsilon^2 \lambda^2 N_v} \sum_{k \in U_v} \mathbb{E}_p(d_k^2), \end{aligned}$$

where the last line uses (CF3). By (M1), letting $v \rightarrow \infty$ and then $K \rightarrow \infty$ proves that $\sqrt{n_v} T_{j,v}^{(2)} = o_{\mathbb{P}}(1)$. Since M is fixed, the result follows.

G.2 Proof of Theorem 7

Since $\tilde{m}$ is deterministic,

$$\hat{\mu}_{\text{ma}}(\tilde{m}, \pi) - \mu = \frac{1}{N_v} \sum_{k \in U_v} \tilde{e}_k \left(\frac{I_k}{\pi_k} - 1 \right),$$

and this quantity has mean zero. Therefore, by (D2) and (D3),

$$\begin{aligned} \mathbb{E}_p \left[\{ \hat{\mu}_{\text{ma}}(\tilde{m}, \pi) - \mu \}^2 \right] &= \frac{1}{N_v^2} \sum_{k \in U_v} \frac{1 - \pi_k}{\pi_k} \tilde{e}_k^2 + \frac{1}{N_v^2} \sum_{k \in U_v} \sum_{\substack{l \in U_v \\ l \neq k}} \frac{\Delta_{kl}}{\pi_k \pi_l} \tilde{e}_k \tilde{e}_l \\ &\leq \left\{ \frac{1 - \lambda}{N_v \lambda} + \frac{C}{n_v \lambda^2} \right\} \frac{1}{N_v} \sum_{k \in U_v} \tilde{e}_k^2, \end{aligned}$$

where we used

$$\sum_{k \in U_v} \sum_{\substack{l \in U_v \\ l \neq k}} |\tilde{e}_k \tilde{e}_l| \leq \left(\sum_{k \in U_v} |\tilde{e}_k| \right)^2 \leq N_v \sum_{k \in U_v} \tilde{e}_k^2.$$

By (D1), n_v and N_v have the same order, while (M2,a) bounds the average squared residuals. Thus, the second moment is $\mathcal{O}(N_v^{-1})$, and the result follows from Chebyshev's inequality.

G.3 Proof of Theorem 8

The variance estimator based on the fixed residuals is

$$\tilde{V}(\tilde{m}, \pi) = \frac{1}{N_v^2} \sum_{k \in S_v} \sum_{l \in S_v} \frac{\Delta_{kl}}{\pi_{kl}} \frac{\tilde{e}_k}{\pi_k} \frac{\tilde{e}_l}{\pi_l}.$$

We split the variance estimation error as follows:

$$\begin{aligned} & n_v \left| \hat{V}_{cf}(\hat{m}, \pi) - \mathbb{V}_p \{ \hat{\mu}_{ma}(\tilde{m}, \pi) \} \right| \\ & \leq n_v \left| \hat{V}_{cf}(\hat{m}, \pi) - \tilde{V}(\tilde{m}, \pi) \right| + n_v \left| \tilde{V}(\tilde{m}, \pi) - \mathbb{V}_p \{ \hat{\mu}_{ma}(\tilde{m}, \pi) \} \right| \\ & =: A_v + B_v. \end{aligned}$$

We first control A_v . Set $d_k := \hat{e}_{(-k)} - \tilde{e}_k = \tilde{m}(\mathbf{x}_k) - \hat{m}^{(-k)}(\mathbf{x}_k)$. By (D2) and (D3),

$$\frac{|\Delta_{kk}|}{\pi_k^2 \pi_{kk}} = \frac{1 - \pi_k}{\pi_k^2} \leq \frac{1}{\lambda^2}, \quad \max_{k \neq l \in U_v} \frac{|\Delta_{kl}|}{\pi_k \pi_l \pi_{kl}} \leq \frac{C}{n_v \lambda^2 \lambda^*}.$$

We have

$$\hat{e}_{(-k)} \hat{e}_{(-l)} - \tilde{e}_k \tilde{e}_l = d_k d_l + d_k \tilde{e}_l + d_l \tilde{e}_k.$$

Taking absolute values and bounding the sample sums by population sums gives

$$\begin{aligned} A_v & \leq \frac{n_v}{N_v^2 \lambda^2} \sum_{k \in U_v} \left(d_k^2 + 2|d_k \tilde{e}_k| \right) \\ & \quad + \frac{C}{N_v^2 \lambda^2 \lambda^*} \left[\left(\sum_{k \in U_v} |d_k| \right)^2 + 2 \left(\sum_{k \in U_v} |d_k| \right) \left(\sum_{k \in U_v} |\tilde{e}_k| \right) \right]. \end{aligned}$$

Cauchy–Schwarz gives

$$\begin{aligned} \frac{1}{N_v} \sum_{k \in U_v} |d_k \tilde{e}_k| & \leq \left\{ \left(\frac{1}{N_v} \sum_{k \in U_v} d_k^2 \right) \left(\frac{1}{N_v} \sum_{k \in U_v} \tilde{e}_k^2 \right) \right\}^{1/2}, \\ \frac{1}{N_v^2} \left(\sum_{k \in U_v} |d_k| \right)^2 & \leq \frac{1}{N_v} \sum_{k \in U_v} d_k^2, \\ \frac{1}{N_v^2} \left(\sum_{k \in U_v} |d_k| \right) \left(\sum_{k \in U_v} |\tilde{e}_k| \right) & \leq \left\{ \left(\frac{1}{N_v} \sum_{k \in U_v} d_k^2 \right) \left(\frac{1}{N_v} \sum_{k \in U_v} \tilde{e}_k^2 \right) \right\}^{1/2}. \end{aligned}$$

Together with $n_v \leq N_v$, these bounds give

$$A_v \leq C_A \left[\frac{1}{N_v} \sum_{k \in U_v} d_k^2 + 2 \left\{ \left(\frac{1}{N_v} \sum_{k \in U_v} d_k^2 \right) \left(\frac{1}{N_v} \sum_{k \in U_v} \tilde{e}_k^2 \right) \right\}^{1/2} \right],$$

where C_A is a constant independent of v . By (M1) and Markov's inequality, $N_v^{-1} \sum_{k \in U_v} d_k^2 = o_{\mathbb{P}}(1)$. Moreover, (M2,b) and the Cauchy–Schwarz inequality imply that $N_v^{-1} \sum_{k \in U_v} \tilde{e}_k^2$ is bounded. Hence, $A_v = o_{\mathbb{P}}(1)$.

It remains to control B_v . By (D2),

$$\begin{aligned} \mathbb{E}_p \left\{ \tilde{V}(\tilde{m}, \pi) \right\} &= \frac{1}{N_v^2} \sum_{k, l \in U_v} \frac{\mathbb{E}_p(I_k I_l)}{\pi_{kl}} \frac{\Delta_{kl}}{\pi_k \pi_l} \tilde{e}_k \tilde{e}_l \\ &= \frac{1}{N_v^2} \sum_{k, l \in U_v} \frac{\Delta_{kl}}{\pi_k \pi_l} \tilde{e}_k \tilde{e}_l \\ &= \mathbb{V}_p \left\{ \hat{\mu}_{\text{ma}}(\tilde{m}, \pi) \right\}. \end{aligned}$$

The argument used in the proof of Theorem 3 of [Breidt and Opsomer \(2000\)](#) applies to the fixed residuals $(\tilde{e}_k)_{k \in U_v}$. Under (D1)–(D4) and (M2,b), it gives $B_v = o_{\mathbb{P}}(1)$.

G.4 Proof of Corollary 2

We first compare the estimated and oracle variances. By (11), the oracle variance is bounded below by a positive constant times n_v^{-1} . Therefore, Theorem 8 gives

$$\frac{\hat{V}_{cf}(\hat{m}, \pi)}{\mathbb{V}_p \left\{ \hat{\mu}_{\text{ma}}(\tilde{m}, \pi) \right\}} \xrightarrow[v \rightarrow \infty]{\mathbb{P}} 1.$$

In particular, $\mathbb{P}\{\hat{V}_{cf}(\hat{m}, \pi) > 0\} \xrightarrow[v \rightarrow \infty]{} 1$. For the difference between the two numerators, Theorem 6 and (11) give

$$\frac{\hat{\mu}_{cf}(\hat{m}, \pi) - \hat{\mu}_{\text{ma}}(\tilde{m}, \pi)}{\sqrt{\mathbb{V}_p \left\{ \hat{\mu}_{\text{ma}}(\tilde{m}, \pi) \right\}}} = o_{\mathbb{P}}(1).$$

On the event that the estimated variance is positive,

$$\begin{aligned} &\frac{\hat{\mu}_{cf}(\hat{m}, \pi) - \mu}{\sqrt{\hat{V}_{cf}(\hat{m}, \pi)}} \\ &= \left[\frac{\hat{\mu}_{\text{ma}}(\tilde{m}, \pi) - \mu}{\sqrt{\mathbb{V}_p \left\{ \hat{\mu}_{\text{ma}}(\tilde{m}, \pi) \right\}}} + o_{\mathbb{P}}(1) \right] \left[\frac{\mathbb{V}_p \left\{ \hat{\mu}_{\text{ma}}(\tilde{m}, \pi) \right\}}{\hat{V}_{cf}(\hat{m}, \pi)} \right]^{1/2}. \end{aligned}$$

The result follows from the assumed central limit theorem and Slutsky's theorem.

H Proofs of Section 6

H.1 Proof of Theorem 9

We first compare the oracle estimators; the first-order equivalences will then give the result. By definition,

$$B_{\mathcal{A}} = \frac{1}{N_v} \sum_{k \in U_v} \left(\frac{\pi_{k|\mathcal{A}}}{\pi_k} - 1 \right) \tilde{e}_k.$$

Let $\mathcal{E}_v := \{\min_{k \in U_v} \pi_{k|\mathcal{A}} \geq \lambda_{cf}\}$. On $\mathcal{E}_v$, centering the sampling indicators at their conditional means gives

$$\begin{aligned} & \hat{\mu}_{\text{ma}}(\tilde{m}, \pi) - \hat{\mu}_{\text{ma}}(\tilde{m}, \pi_{|\mathcal{A}}) - B_{\mathcal{A}} \\ &= \frac{1}{N_v} \sum_{k \in U_v} \tilde{e}_k (I_k - \pi_{k|\mathcal{A}}) \left(\frac{1}{\pi_k} - \frac{1}{\pi_{k|\mathcal{A}}} \right). \end{aligned}$$

Its conditional mean is zero, since $\mathbb{E}_p(I_k - \pi_{k|\mathcal{A}} \mid \mathcal{A}) = 0$. To bound its conditional second moment, first note that, by (D2),

$$\left| \frac{1}{\pi_k} - \frac{1}{\pi_{k|\mathcal{A}}} \right| \leq \frac{|\pi_{k|\mathcal{A}} - \pi_k|}{\lambda \lambda_{cf}}.$$

Separating the diagonal and off-diagonal terms, and using (CF4) and the Cauchy–Schwarz inequality, we obtain on $\mathcal{E}_v$

$$\begin{aligned} & n_v \mathbb{E}_p \left[\left\{ \hat{\mu}_{\text{ma}}(\tilde{m}, \pi) - \hat{\mu}_{\text{ma}}(\tilde{m}, \pi_{|\mathcal{A}}) - B_{\mathcal{A}} \right\}^2 \mid \mathcal{A} \right] \\ &= \frac{n_v}{N_v^2} \sum_{k, l \in U_v} \Delta_{kl|\mathcal{A}} \tilde{e}_k \tilde{e}_l \left(\frac{1}{\pi_k} - \frac{1}{\pi_{k|\mathcal{A}}} \right) \left(\frac{1}{\pi_l} - \frac{1}{\pi_{l|\mathcal{A}}} \right) \\ &\leq \frac{\max_{k \in U_v} (\pi_{k|\mathcal{A}} - \pi_k)^2}{\lambda^2 \lambda_{cf}^2} \left[\frac{n_v}{4N_v^2} \sum_{k \in U_v} \tilde{e}_k^2 + \frac{C_{\Delta 2}}{N_v^2} \sum_{k \in U_v} \sum_{\substack{l \in U_v \\ l \neq k}} |\tilde{e}_k \tilde{e}_l| \right] \\ &\leq C \left(\frac{1}{N_v} \sum_{k \in U_v} \tilde{e}_k^2 \right) \max_{k \in U_v} (\pi_{k|\mathcal{A}} - \pi_k)^2, \end{aligned}$$

where C is a constant independent of v . Here we used $n_v \leq N_v$ and

$$\sum_{k \in U_v} \sum_{\substack{l \in U_v \\ l \neq k}} |\tilde{e}_k \tilde{e}_l| \leq \left(\sum_{k \in U_v} |\tilde{e}_k| \right)^2 \leq N_v \sum_{k \in U_v} \tilde{e}_k^2.$$

Multiplying by $\mathbb{1}_{\mathcal{E}_v}$ and taking expectations, (CF3) and (M2,a) give

$$\begin{aligned} n_v \mathbb{E}_p \left[\left\{ \hat{\mu}_{\text{ma}}(\tilde{m}, \pi) - \hat{\mu}_{\text{ma}}(\tilde{m}, \pi_{|\mathcal{A}}) - B_{\mathcal{A}} \right\}^2 \mathbb{1}_{\mathcal{E}_v} \right] \\ \leq C \left(\frac{1}{N_v} \sum_{k \in U_v} \tilde{e}_k^2 \right) \mathbb{E}_p \left[\max_{k \in U_v} (\pi_{k|\mathcal{A}} - \pi_k)^2 \right] = \mathcal{O}(n_v^{-1}) = o(1). \end{aligned}$$

Since $\mathbb{P}(\mathcal{E}_v^c) \xrightarrow{v \rightarrow \infty} 0$ by (CF2,a), Markov's inequality gives

$$\sqrt{n_v} \left\{ \hat{\mu}_{\text{ma}}(\tilde{m}, \pi) - \hat{\mu}_{\text{ma}}(\tilde{m}, \pi_{|\mathcal{A}}) - B_{\mathcal{A}} \right\} = o_{\mathbb{P}}(1).$$

The result follows from Theorems 3 and 6.

H.2 Proof of Theorem 10

We first compare the two oracle variances by conditioning on $\mathcal{A}$. By (D2) and (CF2,c),

$$\mathbb{E}_p [\hat{\mu}_{\text{ma}}(\tilde{m}, \pi) \mid \mathcal{A}] = \mu + B_{\mathcal{A}}, \quad \mathbb{E}_p [\hat{\mu}_{\text{ma}}(\tilde{m}, \pi_{|\mathcal{A}}) \mid \mathcal{A}] = \mu.$$

The tower property gives

$$\mathbb{E}_p(B_{\mathcal{A}}) = \frac{1}{N_v} \sum_{k \in U_v} \tilde{e}_k \left\{ \frac{\mathbb{E}_p[\mathbb{E}_p(I_k \mid \mathcal{A})]}{\pi_k} - 1 \right\} = \frac{1}{N_v} \sum_{k \in U_v} \tilde{e}_k \left(\frac{\pi_k}{\pi_k} - 1 \right) = 0.$$

The law of total variance therefore gives

$$\mathbb{V}_p \{ \hat{\mu}_{\text{ma}}(\tilde{m}, \pi) \} = \mathbb{E}_p \left[\frac{1}{N_v^2} \sum_{k, l \in U_v} \Delta_{kl|\mathcal{A}} \frac{\tilde{e}_k \tilde{e}_l}{\pi_k \pi_l} \right] + \mathbb{E}_p(B_{\mathcal{A}}^2) \quad (16)$$

and

$$\mathbb{V}_p \{ \hat{\mu}_{\text{ma}}(\tilde{m}, \pi_{|\mathcal{A}}) \} = \mathbb{E}_p \left[\frac{1}{N_v^2} \sum_{k, l \in U_v} \Delta_{kl|\mathcal{A}} \frac{\tilde{e}_k \tilde{e}_l}{\pi_{k|\mathcal{A}} \pi_{l|\mathcal{A}}} \right]. \quad (17)$$

Subtracting yields

$$\mathbb{V}_p \{ \hat{\mu}_{\text{ma}}(\tilde{m}, \pi) \} - \mathbb{V}_p \{ \hat{\mu}_{\text{ma}}(\tilde{m}, \pi_{|\mathcal{A}}) \} = \mathbb{E}_p(B_{\mathcal{A}}^2) + R_v, \quad (18)$$

where

$$R_v := \mathbb{E}_p \left[\frac{1}{N_v^2} \sum_{k, l \in U_v} \Delta_{kl|\mathcal{A}} \tilde{e}_k \tilde{e}_l \left(\frac{1}{\pi_k \pi_l} - \frac{1}{\pi_{k|\mathcal{A}} \pi_{l|\mathcal{A}}} \right) \right].$$

We next bound R_v . Write

$$\pi_{k|\mathcal{A}} \pi_{l|\mathcal{A}} - \pi_k \pi_l = (\pi_{k|\mathcal{A}} - \pi_k) \pi_{l|\mathcal{A}} + \pi_k (\pi_{l|\mathcal{A}} - \pi_l).$$

Since all inclusion probabilities are at most one, (D2) and (CF2,c) give

$$\left| \frac{1}{\pi_k \pi_l} - \frac{1}{\pi_{k|\mathcal{A}} \pi_{l|\mathcal{A}}} \right| \leq \frac{|\pi_{k|\mathcal{A}} - \pi_k| + |\pi_{l|\mathcal{A}} - \pi_l|}{\lambda^2 \lambda_c^2}.$$

Using $\Delta_{kk|\mathcal{A}} \leq 1/4$, (CF4), and the Cauchy–Schwarz inequality as in the preceding proof, we get

$$\begin{aligned} n_v |R_v| &\leq \frac{2}{\lambda^2 \lambda_c^2} \mathbb{E}_p \left[\max_{k \in U_v} |\pi_{k|\mathcal{A}} - \pi_k| \right] \times \left[\frac{n_v}{4N_v^2} \sum_{k \in U_v} \tilde{e}_k^2 + \frac{C_{\Delta 2}}{N_v^2} \sum_{k \in U_v} \sum_{\substack{l \in U_v \\ l \neq k}} |\tilde{e}_k \tilde{e}_l| \right] \\ &\leq C \left(\frac{1}{N_v} \sum_{k \in U_v} \tilde{e}_k^2 \right) \mathbb{E}_p \left[\max_{k \in U_v} |\pi_{k|\mathcal{A}} - \pi_k| \right], \end{aligned}$$

for a constant C independent of v . By (CF3),

$$\mathbb{E}_p \left[\max_{k \in U_v} |\pi_{k|\mathcal{A}} - \pi_k| \right] \leq \left\{ \mathbb{E}_p \left[\max_{k \in U_v} (\pi_{k|\mathcal{A}} - \pi_k)^2 \right] \right\}^{1/2} = \mathcal{O}(n_v^{-1/2}).$$

Thus, (M2,a) gives $n_v R_v = o(1)$. The leading term $n_v \mathbb{E}_p(B_{\mathcal{A}}^2)$ is bounded, since

$$n_v \mathbb{E}_p(B_{\mathcal{A}}^2) \leq \frac{n_v}{\lambda^2} \left(\frac{1}{N_v} \sum_{k \in U_v} \tilde{e}_k^2 \right) \mathbb{E}_p \left[\max_{k \in U_v} (\pi_{k|\mathcal{A}} - \pi_k)^2 \right] = \mathcal{O}(1).$$

It remains to relate $B_{\mathcal{A}}$ to the difference between the estimators. Under (CF2,c), the second-moment bound in the preceding proof holds on the whole sample space. Hence, by (CF3) and (M2,a),

$$\begin{aligned} n_v \mathbb{E}_p \left[\left\{ \hat{\mu}_{\text{ma}}(\tilde{m}, \pi) - \hat{\mu}_{\text{ma}}(\tilde{m}, \pi_{|\mathcal{A}}) - B_{\mathcal{A}} \right\}^2 \right] \\ \leq C \left(\frac{1}{N_v} \sum_{k \in U_v} \tilde{e}_k^2 \right) \mathbb{E}_p \left[\max_{k \in U_v} (\pi_{k|\mathcal{A}} - \pi_k)^2 \right] = \mathcal{O}(n_v^{-1}) = o(1). \end{aligned}$$

The quantity inside braces has conditional mean zero given $\mathcal{A}$. Since $B_{\mathcal{A}}$ is determined by $\mathcal{A}$, the tower property gives

$$\begin{aligned} \mathbb{E}_p \left[B_{\mathcal{A}} \left\{ \hat{\mu}_{\text{ma}}(\tilde{m}, \pi) - \hat{\mu}_{\text{ma}}(\tilde{m}, \pi_{|\mathcal{A}}) - B_{\mathcal{A}} \right\} \right] \\ = \mathbb{E}_p \left[B_{\mathcal{A}} \mathbb{E}_p \left[\hat{\mu}_{\text{ma}}(\tilde{m}, \pi) - \hat{\mu}_{\text{ma}}(\tilde{m}, \pi_{|\mathcal{A}}) - B_{\mathcal{A}} \mid \mathcal{A} \right] \right] = 0. \end{aligned}$$

Together with (18), the following equality proves the result:

$$\begin{aligned} n_v \mathbb{E}_p \left[\left\{ \hat{\mu}_{\text{ma}}(\tilde{m}, \pi) - \hat{\mu}_{\text{ma}}(\tilde{m}, \pi_{|\mathcal{A}}) \right\}^2 \right] \\ = n_v \mathbb{E}_p(B_{\mathcal{A}}^2) + n_v \mathbb{E}_p \left[\left\{ \hat{\mu}_{\text{ma}}(\tilde{m}, \pi) - \hat{\mu}_{\text{ma}}(\tilde{m}, \pi_{|\mathcal{A}}) - B_{\mathcal{A}} \right\}^2 \right] \\ = n_v \mathbb{E}_p(B_{\mathcal{A}}^2) + o(1). \end{aligned}$$

H.3 Failure of first-order equivalence

Let p be SRSWOR of size n_v , with $n_v/N_v \xrightarrow{v \rightarrow \infty} \pi_\star \in (0, 1)$. Assume that N_v is even and fix $M = 2$ population folds of sizes $N_1 = N_2 = N_v/2$ before sampling. With $\mathcal{A} = (\mathbf{n}, \boldsymbol{\delta})$, we get

$$\pi_{k|\mathcal{A}} = \frac{2n_j}{N_v}, \quad k \in U_j,$$

while $\pi_k = n_v/N_v$. Suppose that $\hat{m}^{(-k)} = \tilde{m} = 0$ for every k , and that $y_k = 1$ on U_1 and $y_k = -1$ on U_2 , so that $\mu = 0$. Then,

$$\hat{\mu}_{cf}(\hat{m}, \pi) = \frac{2n_1 - n_v}{n_v}, \quad \hat{\mu}_{cf}(\hat{m}, \pi_{|\mathcal{A}}) = 0 \quad \text{if } n_1 n_2 > 0.$$

The previous event has probability tending to one. Indeed, for $n_v \leq N_v/2$,

$$\mathbb{P}(n_1 n_2 = 0) = 2 \frac{\binom{N_v/2}{n_v}}{\binom{N_v}{n_v}} = 2 \prod_{r=0}^{n_v-1} \frac{N_v/2 - r}{N_v - r} \leq 2^{1-n_v},$$

and this probability is zero if $n_v > N_v/2$.

Conditional on the fold counts, the samples in the two folds are independent SRSWOR samples. The count n_1 is hypergeometric, with

$$\mathbb{E}_p(n_1) = \frac{n_v}{2}, \quad \mathbb{V}_p(n_1) = \frac{n_v(N_v - n_v)}{4(N_v - 1)}.$$

For every $\epsilon > 0$, Chebyshev's inequality gives

$$\mathbb{P}\left(\left|\frac{n_1}{n_v} - \frac{1}{2}\right| > \epsilon\right) \leq \frac{N_v - n_v}{4n_v(N_v - 1)\epsilon^2} \xrightarrow{v \rightarrow \infty} 0.$$

Since $n_2 = n_v - n_1$, it follows that $2n_j/N_v \xrightarrow{v \rightarrow \infty} \pi_\star$ for $j = 1, 2$. The second-order conditional inclusion probabilities are

$$\frac{n_j(n_j - 1)}{(N_v/2)(N_v/2 - 1)} \quad \text{within fold } j, \quad \frac{4n_1 n_2}{N_v^2} \quad \text{across folds.}$$

Both converge in probability to $\pi_\star^2$, giving (CF2,a). Moreover,

$$\begin{aligned} \mathbb{E}_p \left[\max_{k \in U_v} (\pi_{k|\mathcal{A}} - \pi_k)^2 \right] &= \frac{4}{N_v^2} \mathbb{E}_p \left[\left(n_1 - \frac{n_v}{2} \right)^2 \right] \\ &= \frac{4}{N_v^2} \mathbb{V}_p(n_1) = \frac{n_v(N_v - n_v)}{N_v^2(N_v - 1)} = \mathcal{O}(n_v^{-1}), \end{aligned}$$

which verifies (CF3). For distinct $k, l \in U_j$,

$$\Delta_{kl|\mathcal{A}} = \frac{n_j(n_j - 1)}{(N_v/2)(N_v/2 - 1)} - \frac{4n_j^2}{N_v^2} = -\frac{\pi_{k|\mathcal{A}}(1 - \pi_{k|\mathcal{A}})}{N_v/2 - 1}.$$

Across folds, the conditional covariance is zero. Therefore,

$$n_v \max_{k \neq l \in U_v} |\Delta_{kl|\mathcal{A}}| \leq \frac{n_v}{4(N_v/2 - 1)} = \mathcal{O}(1),$$

which verifies (CF4).

Here $B_{\mathcal{A}} = (2n_1 - n_v)/n_v$, so

$$\sqrt{n_v} B_{\mathcal{A}} = \sqrt{\frac{N_v - n_v}{N_v - 1} \frac{n_1 - n_v/2}{\mathbb{V}_p(n_1)}}.$$

The hypergeometric central limit theorem (Lahiri et al., 2006) and $(N_v - n_v)/(N_v - 1) \xrightarrow{v \rightarrow \infty} 1 - \pi_*$ give

$$\sqrt{n_v} B_{\mathcal{A}} \xrightarrow{v \rightarrow \infty} \mathcal{N}(0, 1 - \pi_*).$$

Thus, first-order equivalence fails. The fold residual means differ.

H.4 First-order equivalence under balanced fold residuals

Suppose that the assumptions of Theorem 10 hold. Consider SRSWOR with $n_v/N_v \xrightarrow{v \rightarrow \infty} \pi_* \in (0, 1)$, fix the population folds before sampling, and let $\mathcal{A} = (\mathbf{n}, \boldsymbol{\delta})$. Define

$$\bar{e} = \frac{1}{N_v} \sum_{k \in U_v} \tilde{e}_k, \quad \bar{e}_j = \frac{1}{N_{j,v}} \sum_{k \in U_{j,v}} \tilde{e}_k, \quad j = 1, \dots, M.$$

Then,

$$\lim_{v \rightarrow \infty} n_v \left\{ \mathbb{V}_p(\hat{\mu}_{\text{ma}}(\tilde{m}, \pi)) - \mathbb{V}_p(\hat{\mu}_{\text{ma}}(\tilde{m}, \pi_{|\mathcal{A}})) \right\} = 0$$

if and only if

$$\lim_{v \rightarrow \infty} |\bar{e}_j - \bar{e}| = 0, \quad j = 1, \dots, M.$$

Proof. By (18) and $n_v R_v = o(1)$, it is enough to determine when $n_v \mathbb{E}_p(B_{\mathcal{A}}^2)$ tends to zero. Since $\pi_{k|\mathcal{A}} = n_j/N_{j,v}$ for $k \in U_{j,v}$, the conditional bias is

$$B_{\mathcal{A}} = \sum_{j=1}^M \left(\frac{n_j}{n_v} - \frac{N_{j,v}}{N_v} \right) \bar{e}_j = \sum_{j=1}^M \left(\frac{n_j}{n_v} - \frac{N_{j,v}}{N_v} \right) (\bar{e}_j - \bar{e}).$$

To compute its second moment, write $n_j = \sum_{k \in U_{j,v}} I_k$. Under SRSWOR,

$$\Delta_{kk} = \frac{n_v(N_v - n_v)}{N_v^2}, \quad \Delta_{kq} = -\frac{n_v(N_v - n_v)}{N_v^2(N_v - 1)}, \quad k \neq q.$$

Summing these covariances over $U_{j,v}$ and $U_{l,v}$ gives

$$\begin{aligned}\text{Cov}_p(n_j, n_l) &= \sum_{k \in U_{j,v}} \sum_{q \in U_{l,v}} \Delta_{kq} \\ &= n_v \frac{N_v - n_v}{N_v - 1} \left\{ \frac{N_{j,v}}{N_v} \mathbb{1}_{\{j=l\}} - \frac{N_{j,v} N_{l,v}}{N_v^2} \right\}.\end{aligned}$$

Also, $\mathbb{E}_p(n_j) = n_v N_{j,v} / N_v$, so $\mathbb{E}_p(B_{\mathcal{A}}) = 0$. Hence,

$$\begin{aligned}n_v \mathbb{E}_p(B_{\mathcal{A}}^2) &= \frac{1}{n_v} \sum_{j=1}^M \sum_{l=1}^M \text{Cov}_p(n_j, n_l) (\bar{e}_j - \bar{e})(\bar{e}_l - \bar{e}) \\ &= \frac{N_v - n_v}{N_v - 1} \left[\sum_{j=1}^M \frac{N_{j,v}}{N_v} (\bar{e}_j - \bar{e})^2 - \left\{ \sum_{j=1}^M \frac{N_{j,v}}{N_v} (\bar{e}_j - \bar{e}) \right\}^2 \right] \\ &= \frac{N_v - n_v}{N_v - 1} \sum_{j=1}^M \frac{N_{j,v}}{N_v} (\bar{e}_j - \bar{e})^2,\end{aligned}$$

where the last equality uses $\sum_{j=1}^M (N_{j,v} / N_v) \bar{e}_j = \bar{e}$. Combining this identity with (18) and $n_v R_v = o(1)$ yields

$$\begin{aligned}&n_v \left\{ \mathbb{V}_p(\hat{\mu}_{\text{ma}}(\tilde{m}, \pi)) - \mathbb{V}_p(\hat{\mu}_{\text{ma}}(\tilde{m}, \pi_{|\mathcal{A}})) \right\} \\ &= \frac{N_v - n_v}{N_v - 1} \sum_{j=1}^M \frac{N_{j,v}}{N_v} (\bar{e}_j - \bar{e})^2 + o(1).\end{aligned}$$

The factoring term tends to $1 - \pi_{\star} > 0$. Moreover, (CF2,c) requires $N_{j,v} > N_v - n_v$ for every fold, since otherwise that fold could contain no sampled unit. Thus, the fold proportions are bounded away from zero. As M is fixed, the result follows. $\square$

References

- Breidt, F. J. and Opsomer, J. D. (2000), ‘Local polynomial regression estimators in survey sampling’, *The Annals of Statistics* **28**(4), 1026–1053.
- Deville, J.-C. (1992), Constrained samples, conditional inference, weighting: Three aspects of the utilisation of auxiliary information, in ‘Proceedings of the Workshop on the Uses of Auxiliary Information in Surveys’, Statistics Sweden, Örebro, pp. 1–18.
- Godambe, V. and Joshi, V. (1965), ‘Admissibility and Bayes estimation in sampling finite populations. I’, *The Annals of Mathematical Statistics* **36**(6), 1707–1722.
- Goga, C. (2005), ‘Réduction de la variance dans les sondages en présence d’information auxiliaire: Une approche non paramétrique par splines de régression’, *Canadian Journal of Statistics* **33**(2), 163–180.
- Jauslin, R., Eustache, E., Panahbehagh, B. and Tillé, Y. (2025), *StratifiedSampling: Different Methods for Stratified Sampling*. R package version 0.4.2.
URL: <https://CRAN.R-project.org/package=StratifiedSampling>

- Lahiri, S. N., Chatterjee, A. and Maiti, T. (2006), ‘A sub-Gaussian Berry-Esseen theorem for the hypergeometric distribution’, *arXiv preprint math/0602276* .
- Lu, X., Shi, L., Liu, H. and Ding, P. (2025), ‘Conditional cross-fitting for unbiased machine-learning-assisted covariate adjustment in randomized experiments’, *arXiv preprint arXiv:2508.15664* .
- Pfeffermann, D. and Sverchkov, M. (2009), Inference under informative sampling, in ‘Handbook of statistics’, Vol. 29, Elsevier, pp. 455–487.
- Rao, J. N. K. (1985), ‘Conditional inference in survey sampling’, *Survey Methodology* **11**, 15–31.
- Robinson, P. M. and Särndal, C.-E. (1983), ‘Asymptotic properties of the generalized regression estimator in probability sampling’, *Sankhyā, Series B* **45**, 240–248.
- Särndal, C.-E., Swensson, B. and Wretman, J. (1992), *Model Assisted Survey Sampling*, Springer.
- Tillé, Y. (2006), *Sampling Algorithms*, Springer Series in Statistics, Springer, New York.
- Tillé, Y. (1998), ‘Estimation in surveys using conditional inclusion probabilities: simple random sampling’, *International Statistical Review* **66**, 303–322.
- Tillé, Y. (1999), ‘Estimation in surveys using conditional inclusion probabilities: complex design’, *Survey Methodology* **25**, 57–66.
- Tillé, Y. and Matei, A. (2025), *sampling: Survey Sampling*. R package version 2.11.
URL: <https://CRAN.R-project.org/package=sampling>