Statistica Sinica Preprint No: SS-2024-0359	
Title	A Maximin Optimal Approach for Sampling Designs in Two-phase Studies
Manuscript ID	SS-2024-0359
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202024.0359
Complete List of Authors	Ruoyu Wang, Qihua Wang and Wang Miao
Corresponding Authors	Qihua Wang
E-mails	qhwang@amss.ac.cn

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0359

A MAXIMIN OPTIMAL APPROACH FOR SAMPLING DESIGNS IN TWO-PHASE STUDIES

Ruoyu Wang[1], Qihua Wang[2*] and Wang Miao[3]

[1]*Department of Biostatistics, Harvard T.H. Chan School of Public Health*

[2]*State Key Laboratory of Mathematical Sciences, Academy of Mathematics and Systems Science,* Chinese Academy of Sciences

and [3] *Department of Probability and Statistics, Peking University*

Abstract: Data collection costs can vary widely across variables in data science tasks. Two-phase designs can be employed to save data collection costs. This paper considers the two-phase studies where inexpensive variables are collected for all subjects in the first phase, and expensive variables are measured for a subsample of subjects in the second phase based on a predetermined sampling rule. The estimation efficiency under two-phase designs relies heavily on the sampling rule. Existing literature primarily focuses on designing sampling rules for estimating a scalar parameter in some parametric models or specific estimating problems. However, real-world scenarios are usually model-unknown and involve two-phase designs for model-free estimation of a scalar or multi-dimensional parameter. This paper proposes a maximin criterion to design an optimal sampling rule based on semiparametric efficiency bounds. The proposed method is model-free and applicable to general estimating problems. The resulting sampling rule can minimize the semiparametric efficiency bound when the parameter is scalar and improve the bound for every component when the parameter is multi-dimensional. Simulation studies demonstrate that the proposed designs reduce the variance of the resulting estimator in various settings. The implementation of the proposed design is illustrated in a real data analysis.

* Corresponding author

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0359

Key words and phrases: Cost-effective sampling; Efficient influence function; Incomplete data; Semiparametric efficiency, Subsample.

1. Introduction

The research of modern data science often involves variables with different collection costs. For example, the research objective in medical and genetic studies often involves analyzing the associations between disease status and biomarkers while adjusting for demographic factors such as age, gender, or race. Measuring the biomarker can be fairly expensive compared to disease status and demographic factors that can be obtained through questionnaires or electronic health records. The two-phase design is commonly employed to achieve cost-effectiveness in ecological, epidemiological and genetic researches (Fattorini et al., 2017; Nab et al., 2021; Lin et al., 2013), especially in large-scale studies such as the National Wilms' Tumor Study (Green et al., 1998), the Caribbean, Central, and South America network for HIV Epidemiology (McGowan et al., 2007), and the National Heart, Lung, and Blood Institute Exome Sequencing Project (Lin et al., 2013).

In two-phase studies considered in e.g., McNamee (2002); McIsaac and Cook (2014, 2015) and Tao et al. (2020), inexpensive variables are collected for all subjects in the first phase, while expensive variables are measured only on a subsample drawn according to some sampling rule in the second phase. The parameter of interest is estimated using observations from the two phases. The design of the second phase sampling rule is crucial to the efficiency of the resulting estimator. Many existing works develop sampling rules tailored for parameter estimation in conditional density models. Zhou et al. (2014) propose to fit a

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0359

model to predict the expensive variables of different subjects and draw a sample from the subjects with extreme predicted values in the second phase. The proposed sampling rule performs well in the empirical studies in (Zhou et al., 2014). However, there is no theoretical guarantee for the optimality of the proposed sampling rules. Recently, researchers have developed optimal sampling rules that minimize the asymptotic variances of some specific estimators or semiparametric efficiency bound under the conditional density model. McIsaac and Cook (2015) investigate the optimal sampling rule for the mean score estimator with discrete inexpensive variables. Chen and Lumley (2022) derive the optimal sampling rule for the inverse-probability weighted estimator and generalized raking estimator for parameter estimation in conditional density models. Tao et al. (2020) propose the optimal sampling rule that minimizes the semiparametric efficiency bound in a general conditional density estimation problem which may contain continuous inexpensive variables and nuisance parameters. A few existing works also consider other estimation problems than the conditional density model. McNamee (2002) proposes the optimal rule for estimating sensitivity, specificity, and positive predictive value of a test with categorical test results and no covariate. Gilbert et al. (2014) study the optimal design that minimized the semiparametric efficiency bound for mean or mean difference estimation with discrete inexpensive variables and proposed to stratify the continuous inexpensive variables. Zhang et al. (2024) explore a problem similar to Gilbert et al. (2014) while extending the scope by accounting for the selection bias in the first-phase data.

Despite the fruitful works on sampling rule designs in the second phase, there still exist several challenges in practice. Firstly, current research on optimal design mainly focuses on

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0359

parameter estimation in certain specific models, leaving many important problems unconsidered. Secondly, in many applications, there is a multi-dimensional parameter of interest; however, existing optimal designs work best for scalar parameters, and optimal designs for a multi-dimensional parameter vector are complicated and remain unstudied in the literature. The optimal sampling rule for one component does not necessarily lead to good estimation efficiency for other components. For instance, this issue arises when one estimate sensitivity, specificity, and positive predictive value simultaneously as illustrated by (McNamee, 2002). Therefore, it is crucial to strike a balance between estimation efficiency for different components to achieve efficiency gains for every component.

This paper proposes a model-free design approach for estimating a generic population parameter of interest, which can be either scalar or multi-dimensional. The proposed method applies to many important but understudied problems in two-phase sampling design, such as the estimation of quantiles, Pearson correlation coefficients, and average treatment effects. When designing the sampling rule, the proposed method focuses on optimizing the resulting semiparametric efficiency bound, which represents the smallest asymptotic variance that one can achieve for estimating the parameter of interest under regularity conditions. Once a sampling rule is established, various existing semiparametric methods can be employed to construct an efficient estimator whose asymptotic variance achieves the semiparametric efficiency bound. These methods include the targeted maximum likelihood method (van der Laan and Rubin, 2006), the estimating equation method (Tsiatis, 2007), and the one-step estimation method (Bickel, 1982). Compared to existing works, our main contributions are as follows:

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0359

- When the parameter is scalar, we derive the optimal sampling rule that minimizes the semiparametric efficiency bound subject to a budget constraint. In contrast to many existing methods, our approach does not make any parametric model assumptions and allows the inexpensive variables to be continuous. Obtaining the optimal sampling rule is challenging in this case because it involves an infinite-dimensional constrained optimization problem.

- We study the problem of two-phase sampling design with a multi-dimensional parameter of interest, which remains largely unexplored in the literature to our knowledge. The problem becomes more challenging when the parameter is multi-dimensional. Specifically, under the optimal sampling rule for one given component, the semiparametric efficiency bounds of other components may be larger than those under the uniform sampling rule that includes each subject with a constant probability. To resolve this issue, we define an objective function based on the semiparametric efficiency bound for each component and design the sampling rule by maximizing the objective function, which leads to a *maximin optimal sampling rule*. The maximin optimal sampling rule outperforms the uniform rule regarding the semiparametric efficiency bound for every component under mild conditions. Nevertheless, it involves an intractable infinite-dimensional constrained maximin problem to obtain the proposed sampling rule. To tackle this, we design a novel finite-dimensional optimization problem that shares the same solution as the original infinite-dimensional problem, and obtain the solution by routine optimization algorithms.

The proposed sampling rules depend on unknown quantities determined by the data

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0359

generating process. We adopt the standard "pilot sample" approach to obtain the proposed sampling rules (Gilbert et al., 2014; Tao et al., 2020). To be specific, we select a simple random sample, referred to as a "pilot sample", at the beginning of the second phase and then use the pilot sample to estimate the unknown quantities. The remaining subjects are then selected based on the estimated sampling rule. We establish the optimality of the proposed sampling rules and the consistency of their estimators. Simulation studies demonstrate that the proposed designs can achieve substantial efficiency gains compared to the uniform rule under various settings. The implementation of the proposed design is further illustrated in a real data example.

The proposed sampling rules are derived based on the efficient influence function of an estimation problem. The method to derive the efficient influence function is standard and thoroughly investigated in literature of semiparametric statistics. See, e.g., Bickel (1982); Tsiatis (2007); van der Laan and Robins (2012). This implies that the proposed method is versatile and can be applied to a wide range of estimation problems as long as the corresponding efficient influence function is available. For instance, it can be employed in two-phase designs for data validation in electronic health records to address issues related to missing data and measurement errors (Lotspeich et al., 2022; Zhang et al., 2024). Additionally, this method offers optimal designs for important but underexplored tasks, such as the selection of validation data in observational studies with missing confounders (Lin and Chen, 2014; Yang and Ding, 2020), and quantitative trait analysis with multiple traits of interest (Lin et al., 2013).

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0359

2. Two-phase design and efficiency bound

Let U denote a vector of expensive variables, such as true disease status or biomarkers, and let V denote a vector of inexpensive variables, such as age, gender, race, or surrogate measures for U. In a two-phase study, V is collected in the first phase for all n subjects in the study. In the second phase, a subsample of subjects is drawn from the n subjects according to a sampling rule that may depend on V, and U is measured for the subsample. Let R be the sampling indicator in the second phase. $R = 1$ if the subject is included in the second stage and U is observed; $R = 0$, otherwise. The inclusion probability for the second stage depends on the first-phase variable vector V, that is, $P(R = 1 \mid V) = \rho(V)$ where $\rho(\cdot)$ is the sampling rule. We refer to V and U as the first-phase and second-phase variables, respectively. The sampling framework considered here is commonly adopted in the literature of two-phase studies (Chatterjee et al., 2003; Gilbert et al., 2014; Tao et al., 2017, 2020), and has been widely used in epidemiological and genetic studies to investigate the effect of various risk factors and biomarkers on different diseases-related traits (Nab et al., 2021; Lin et al., 2013).

Suppose θ_0 is the parameter of interest in the two-phase study, which is possibly a multi-dimensional functional of the joint distribution of (V, U). An important question is how to design the sampling rule $\rho(\cdot)$ in the second phase to minimize the asymptotic variance for estimating θ_0. The optimal sampling rule that minimizes the asymptotic variance is usually model-dependent and estimator-specific (Chen and Lumley, 2022). We here consider the model-free semiparametric efficient estimating problem and develop an optimal sample rule by minimizing the semiparametric efficient bound for estimating θ_0. The semiparametric efficiency bound (Bickel, 1982; Tsiatis, 2007) is an extension of the Cramer-Rao bound to

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0359

the semiparametric and nonparametric setting, which characterizes the smallest asymptotic variance one can achieve for estimating θ_0 under the considered distribution class. Besides, researchers usually intend to control the sampling fraction $E[\rho(V)]$ in the second phase to manage the budget since U is expensive to measure. This further motivates us to search for the sampling rule that minimizes the semiparametric efficiency bound under the constraint that $E[\rho(V)]$ is no larger than a given threshold.

Next, we introduce the efficient influence function and the semiparametric efficiency bound under the two-phase design. The efficient influence function is an important concept in semiparametric theory which is useful in deriving the semiparametric efficiency bound and constructing efficient estimators that attain the bound (Tsiatis, 2007). Suppose $\psi(V, U)$ is the efficient influence function for estimating θ_0 under a full data setting where V and U are observed for every subject. We write $\psi(V, U)$ as ψ for short, wherever it does not cause confusion. Under two-phase designs, both the efficient influence function and the semiparametric efficiency bound are highly dependent on ψ. The expression of ψ is well established in many important problems across survey sampling and modern epidemiological and clinical studies, including the outcome mean estimation, linear regression analysis, and average treatment effect estimation (Tsiatis, 2007). See Section S1.1 in Supplementary Material for the expression of ψ in the exemplified problems.

Based on ψ, we can derive the efficient influence function and the semiparametric efficiency bound under the two-phase design without parametric model assumptions. Assume throughout this paper that the distribution of (V, U) has a density with respect to some dominance measure, e.g., the counting measure or the Lebesgue measure. We can derive semi-

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0359

parametric efficiency bound under the two-phase design in the following lemma by using techniques similar to those in Tsiatis (2007) for missing data problems.

Lemma 1. *Let ψ be the full data efficient influence function for estimating θ_0 and $\rho(\cdot)$ the sampling rule in the second phase. Under the regularity conditions in Section S2.1 in Supplementary Material, the efficient influence function under the two-phase design is*

$$\frac{R\psi}{\rho(V)} - \left\{\frac{R}{\rho(V)} - 1\right\}\Pi(V), \tag{2.1}$$

and the semiparametric efficiency bound is

$$E\left[\frac{\Sigma(V)}{\rho(V)}\right] + \mathrm{Var}[\Pi(V)],$$

where $\Pi(V) = E[\psi \mid V]$ and $\Sigma(V) = \mathrm{Var}[\psi \mid V]$ are the mean and variance of ψ conditional on V, respectively.

Several methods can be adopted to obtain estimators for θ_0 that have the influence function (2.1) and hence achieve the semiparametric efficiency bound. These methods include the targeted maximum likelihood method (van der Laan and Rubin, 2006), the estimating equation method (Tsiatis, 2007), and the one-step estimation method (Bickel, 1982). Note that the semiparametric efficiency bound depends on $\rho(\cdot)$. Therefore, the efficiency of the efficient estimator can be further improved by employing a carefully designed sampling rule. We delve into this topic in the subsequent sections.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0359

3. Optimal design in two-phase studies

3.1 Optimal design for a scalar parameter

In this section, we consider the optimal sampling rule when θ_0 is scalar. Denote the variance of ψ conditional on V as $\sigma^2(V)$ and assume $\sigma^2(V) > 0$. The semiparametric efficiency bound for θ_0 is

$$E\left[\frac{\sigma^2(V)}{\rho(V)}\right] + \mathrm{Var}[\Pi(V)]. \tag{3.1}$$

We search for the optimal sampling rule that minimizes the semiparametric efficiency bound (3.1) subject to the constraint

$$E[\rho(V)] \leq \varpi, \tag{3.2}$$

where $\varpi \in (0, 1]$ is the maximal sampling fraction determined by study budgets. The problem is an infinite-dimensional optimization problem that is generally difficult to solve. The constraint that the sampling probability belongs to $[0, 1]$ further complicates the problem. Nonetheless, we find that the optimal sampling rule can be obtained via some constructive arguments. For any positive function $\lambda(\cdot)$ and positive τ, let

$$\rho(V; \lambda, \tau) = 1\{\lambda(V) > \tau\} + \frac{1}{\tau}\lambda(V)1\{\lambda(V) \leq \tau\}.$$

Note that $\rho(V; \sigma, \tau)$ is continuous and strictly decreasing with respect to τ, where $\sigma(\cdot)$ is the conditional variance function of ψ conditional on V. This implies that the equation $E[\rho(V; \sigma, \tau)] = \varpi$ has a unique solution, denoted by τ_s. The form of the optimal sampling rule is obtained in the following theorem.

Theorem 1. *The optimal sampling rule that minimizes the semiparametric efficiency bound*

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0359

(3.1) *under the constraint* (3.2) *is*

$$\rho_s(\cdot) = \rho(\cdot; \sigma, \tau_s) = 1\{\sigma(\cdot) > \tau_s\} + \frac{1}{\tau_s}\sigma(\cdot)1\{\sigma(\cdot) \le \tau_s\},$$

where τ_s *is the unique solution of* $E[\rho(V; \sigma, \tau)] = \varpi$ *with respect to* τ.

Please refer to Section S2.3 in Supplementary Material for the proof of Theorem 1. Let us provide some intuitions for Theorem 1. Note that minimizing the semiparametric efficiency bound (3.1) is equivalent to minimizing $E[\sigma^2(V)/\rho(V)]$. Notice that $E[\rho(V)] \le \varpi$. We have $\varpi E[\sigma^2(V)/\rho(V)] \ge E[\rho(V)]E[\sigma^2(V)/\rho(V)] \ge (E[\sigma(V)])^2$ according to the Cauchy–Schwarz inequality and the equality hold only if $\rho(V) = \varpi\sigma(V)/E[\sigma(V)]$. Let $\tau^* = E[\sigma(V)]/\varpi$. Then one may conjecture that $\sigma(V)/\tau^*$ is the optimal sampling rule that minimizes the semiparametric efficiency bound. However, $\sigma(V)/\tau^*$ may not be a feasible sampling rule as it can be larger than one. Thus we consider the truncated version $\rho(V; \sigma, \tau) = 1\{\sigma(V) > \tau\} + \sigma(V)1\{\sigma(V) \le \tau\}/\tau = \min\{\sigma(V)/\tau, 1\}$ and show that it is the desired optimal rule if we determine τ by solving the equation $E[\rho(V; \sigma, \tau)] = \varpi$.

Theorem 1 gives the optimal sampling rule $\rho_s(\cdot)$ for estimating a general population parameter. The optimal sampling rule $\rho_s(\cdot)$ is determined by $\sigma(\cdot) = \mathrm{var}(\psi \mid V = \cdot)$. Thus, to estimate the optimal sampling rule, one only needs to estimate the conditional variance function, which boils down to a well-studied problem in statistics. The form of $\rho_s(\cdot)$ is analogous to the Neyman allocation (Cochran, 2007) for mean estimation problems in survey sampling which has been extended to scalar mean and regression coefficient estimation problems in two-phase studies (Reilly and Pepe, 1995; Gilbert et al., 2014; McIsaac and Cook, 2014; Chen and Lumley, 2022). Our result further extends the Neyman allocation to the problem of estimating a generic parameter. In addition, existing results often assume that the first-

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0359

phase variables V are discrete and propose to discretize the continuous variables regardless of the information loss caused by discretization (McIsaac and Cook, 2014). When V is a discrete variable, minimizing the semiparametric efficiency bound under the budget constraint is a finite-dimensional optimization problem and can be solved by the Lagrange multiplier method. When V is a continuous variable, it becomes more challenging to find the optimal sampling rule, as it requires optimization over an infinite-dimensional function space. We overcome this difficulty by constructively suggesting the sampling rule $\rho_s(\cdot)$ based on the ideas discussed in the last paragraph.

It should be pointed out that Tao et al. (2020) also explore the minimization problem of semiparametric efficiency bound with budget constraint. They focus on the parameter estimation problem under a semiparametric conditional density model. Their approach allows the first-phase variable to be continuous and simplifies the infinite-dimensional optimization problem by invoking the Neyman–Pearson lemma. However, the simplified problem generally remains infinite-dimensional and intractable, except under the logistic model or the linear model with Gaussian error. In contrast, this paper considers the estimation of a generic population parameter. Building on the constructive ideas outlined in previous paragraphs, Theorem 1 effectively converts the sampling rule design problem into a manageable conditional variance estimation problem and is applicable to a wide range of estimation tasks.

3.2 Maximin design for a multi-dimensional parameter

Theorem 1 obtains the optimal design for a scalar parameter. In practice, the parameter of interest θ_0 may well be multi-dimensional, especially when analyzing multiple traits and

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0359

biomarkers or evaluating multi-level treatments. In this section, we study the two-phase design for a multi-dimensional parameter. Let $\theta_0 = (\theta_{01}, \ldots, \theta_{0d})^{\mathrm{T}}$ and $\psi = (\psi_1, \ldots, \psi_d)^{\mathrm{T}}$ be the d-dimensional parameter and the full data influence function, respectively. The semiparametric efficiency bound for the jth component θ_{0j} is

$$E\left[\frac{\sigma_j^2(V)}{\rho(V)}\right] + \mathrm{Var}[\Pi_j(V)],$$

where $\sigma_j(V) = \sqrt{\mathrm{Var}[\psi_j \mid V]}$ and $\Pi_j(V)$ is the jth component of $\Pi(V)$ for $j = 1, \ldots, d$. The optimal sampling rule that minimizes the semiparametric efficiency bound for θ_{0j} can be obtained according to Theorem 1. Let us denote the optimal sampling rule for θ_{0j} by $\rho_j(\cdot)$. Compared to the scalar-parameter case, the main difficulty for the multi-dimensional-parameter problems is that the optimal rules for different components may be different. Thus, there usually does not exist a sampling rule that simultaneously minimizes the semiparametric efficiency bound for different parameter components.

One intuitive approach to determining a sampling rule for a multi-dimensional parameter is to minimize the sum of the semiparametric efficiency bounds for each component

$$\sum_{j=1}^{d} E\left[\frac{\sigma_j^2(V)}{\rho(V)}\right] + \sum_{j=1}^{d} \mathrm{Var}[\Pi_j(V)]. \tag{3.3}$$

This criterion is analogous to the widely-used A-optimality in the experiment design literature (Kiefer, 1959). Arguments similar to those in the proof of Theorem 1 can show that the sampling rule minimizing (3.3) is

$$\begin{aligned}\rho_{\mathrm{sum}}(\cdot) &= \rho(\cdot; \sigma_{\mathrm{sum}}, \tau_{\mathrm{sum}}) \\ &= 1\{\sigma_{\mathrm{sum}}(\cdot) > \tau_{\mathrm{sum}}\} + \frac{1}{\tau_{\mathrm{sum}}}\sigma_{\mathrm{sum}}(\cdot)1\{\sigma_{\mathrm{sum}}(\cdot) \le \tau_{\mathrm{sum}}\},\end{aligned}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0359

where $\sigma_{\text{sum}}(\cdot) = \sqrt{\sum_{j=1}^{d} \sigma_j^2(\cdot)}$ and τ_{sum} denotes the solution of the equation $E[\rho(V; \sigma_{\text{sum}}, \tau)] = \varpi$. However, the sum of semiparametric efficiency bounds may not be suitable in practice when the interest lies on each component. For instance, it could be obscure to sum up the semiparametric efficiency bound for per capita income and sex ratio in a survey study.

The most straightforward sampling rule that satisfies the budget constraint in Equation (3.2) is the uniform sampling rule. Since the uniform rule is intuitive and easy to implement, a designed sampling rule is expected to perform better than the uniform rule. Unfortunately, neither $\rho_{\text{sum}}(\cdot)$ nor $\rho_j(\cdot)$ ($j \in 1, \ldots, d$) fulfills this requirement. Specifically, the semiparametric efficiency bound for some component of θ_0 may be larger under $\rho_{\text{sum}}(\cdot)$ or $\rho_j(\cdot)$ than that under the uniform rule. A simple example illustrating this issue is provided in Section S1.2 in Supplementary Material. Thus, it is important to design a sampling rule that performs better than the uniform rule. This problem, however, remains largely unexplored in the literature to our knowledge.

In the same spirit, we consider a more general problem. Given a benchmark sampling rule $\rho_0(\cdot)$ that satisfies $E[\rho_0(V)] = \varpi$, such as the uniform rule, our objective is to search for a sampling rule under which the semiparametric efficiency bound for each parameter is smaller or at least no larger than the corresponding semiparametric efficiency bound under $\rho_0(\cdot)$. Hereafter, a sampling rule $\rho(\cdot)$ is considered to *dominates* $\rho_0(\cdot)$ if the semiparametric efficiency bound for each parameter component under $\rho(\cdot)$ is smaller than that under $\rho_0(\cdot)$.

Let $\xi_j = E[\sigma_j^2(V)/\rho_0(V)]$. Define $b_j = E\left[\sigma_j^2(V)/\rho_0(V)\right] + \text{Var}[\Pi_j(V)]$ as the semiparametric efficiency bound for the jth component under the benchmark sampling rule $\rho_0(\cdot)$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0359

for $j = 1, \ldots, d$. For any $\rho(\cdot)$, the expression

$$b_j^{-1}\left(\xi_j - E\left[\frac{\sigma_j^2(V)}{\rho(V)}\right]\right)$$

measures the relative improvement in the semiparametric efficiency bound for the jth component compared to the benchmark rule $\rho_0(\cdot)$. We focus on relative improvements to provide a normalized measure that accounts for different scales in the efficiency bounds across components. Consider the minimum of the difference over all components

$$M(\rho) = \min_{j=1,\ldots,d}\left\{b_j^{-1}\left(\xi_j - E\left[\frac{\sigma_j^2(V)}{\rho(V)}\right]\right)\right\}.$$

When maximizing $M(\rho)$ over a candidate set of sampling rules $\mathcal{P}$ that contains $\rho_0(\cdot)$, the resulting maximin optimal sampling rule $\rho_*(\cdot)$ satisfies $M(\rho_*) = \max_{\rho\in\mathcal{P}} M(\rho) \geq M(\rho_0) = 0$. This ensures that the semiparametric efficiency bound for every component under $\rho_*(\cdot)$ is no larger than that under $\rho_0(\cdot)$. The idea to derive the sampling rule by solving the maximin problem $\max_{\rho\in\mathcal{P}} M(\rho)$ resembles the minimax regret strategy used in decision theory (Cox and Hinkley, 1979), prediction (Meinshausen and Bühlmann, 2015), experimental design (Dette, 1997), and hypothesis testing (Liu et al., 2022) for various purposes. Here, the strategy is employed to ensure that the resulting sampling rule can improve the benchmark sampling rule in terms of the semiparametric efficiency bound for every parameter component, which is not precedented in the context of two-phase sampling design to our knowledge.

Although guaranteed to be no worse than $\rho_0(\cdot)$, the actual improvement that $\rho_*(\cdot)$ can achieve depends on the choice of $\mathcal{P}$. For example, if $\mathcal{P}$ is the singleton set $\{\rho_0(\cdot)\}$, then $\rho_*(\cdot) = \rho_0(\cdot)$ and no improvement can be achieved. On the other hand, if there exists a sampling rule $\rho(\cdot) \in \mathcal{P}$ that dominates $\rho_0(\cdot)$, then we have $M(\rho_*) \geq M(\rho) > 0$ which

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0359

implies $\rho_*(\cdot)$ also dominates $\rho_0(\cdot)$. Hence, a careful specification of the set $\mathcal{P}$ is necessary to achieve significant efficiency improvement over $\rho_0(\cdot)$.

We provide some sensible candidates for $\mathcal{P}$. Recall that $\rho_j(\cdot)$ is the optimal sampling rule for θ_{0j} for $j = 1, \ldots, d$. A reasonable choice is to take $\mathcal{P} = \mathcal{P}_c := \{\rho_0(\cdot) + \sum_{j=1}^d w_j(\rho_j(\cdot) - \rho_0(\cdot))\}$ to be the set which consists of all convex combinations of the component-wise optimal sampling rule $\rho_j(\cdot)$ and the benchmark rule $\rho_0(\cdot)$. This results in the following constrained maximin problem

$$\max_{w \in \mathcal{W}} \min_{j=1,\ldots,d} \left\{ b_j^{-1} \left(\xi_j - E\left[\frac{\sigma_j^2(V)}{\rho_0(V) + \sum_{k=1}^d w_k\{\rho_k(V) - \rho_0(V)\}} \right] \right) \right\}, \tag{3.4}$$

where $\mathcal{W} = \{w = (w_1, \ldots, w_d) : \sum_{j=1}^d w_j \leq 1,\ 0 \leq w_j \leq 1,\ \text{for } j = 1, \ldots, d\}$. In problem (3.4), by taking the feasible set to be the set $\mathcal{P}_c$, we ensure that the optimization problem is a finite-dimensional convex problem and hence allows for effective computation. Let $w_c = (w_{c,1}, \ldots, w_{c,d})$ be the solution of the maximin problem (3.4). Then we obtain the constrained maximin optimal sampling rule

$$\rho_c(\cdot) = \rho_0(\cdot) + \sum_{j=1}^d w_{c,j}\{\rho_j(\cdot) - \rho_0(\cdot)\}$$

that solves the constrained maximin problem (3.4). By using the optimal sampling rule of each component as basis functions, this method takes the accuracy of each component into consideration. By taking a convex combination and solving the maximin problem (3.4), we achieve a balance in the estimation efficiency of different components. The resulting sampling rule $\rho_c(\cdot)$ dominates $\rho_0(\cdot)$ as long as $(0, \ldots, 0)$ is not a solution of problem (3.4).

In problem (3.4), the optimization problem is simplified by constraining the feasible set $\mathcal{P}$ to be the finite-dimensional set $\mathcal{P}_c$. However, the resulting sampling rule $\rho_c(\cdot)$ might

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0359

be suboptimal because many promising sampling rules may not be in $\mathcal{P}_c$. To address this concern, one may consider the problem of maximizing $M(\rho)$ over all sampling rules under the budget constraint, i.e., taking $\mathcal{P} = \mathcal{P}_\varsigma := \{\rho(\cdot) : 0 \le \rho(\cdot) \le 1,\ E[\rho(V)] \le \varpi\}$. Then we arrive at the maximin problem

$$\max_{\rho \in \mathcal{P}_\varsigma} \min_{j=1,\ldots,d} \left\{ b_j^{-1} \left(\xi_j - E\left[\frac{\sigma_j^2(V)}{\rho(V)} \right] \right) \right\}. \tag{3.5}$$

Solving the problem (3.5) directly is challenging due to the functional nature of $\rho(\cdot)$ and the fact that it involves an infinite-dimensional maximin problem. To resolve this issue, by invoking the result in Theorem 1, we design a novel finite-dimensional optimization problem that shares the same solution as (3.5) .

Theorem 2. *The global maximin optimal sampling rule that solves problem* (3.5) *is*

$$\rho_\varsigma(\cdot) = \rho(\cdot; \sigma_{w_\varsigma}, \tau_{w_\varsigma}) = 1\{\sigma_{w_\varsigma}(\cdot) > \tau_{w_\varsigma}\} + \frac{1}{\tau_{w_\varsigma}} \sigma_{w_\varsigma}(\cdot) 1\{\sigma_{w_\varsigma}(\cdot) \le \tau_{w_\varsigma}\},$$

where $w_\varsigma = (w_{\varsigma,1}, \ldots, w_{\varsigma,d})^{\mathrm{T}}$ *is the solution of the problem*

$$\min_{w \in \mathcal{W}^\dagger} \left\{ \sum_{j=1}^d w_j b_j^{-1} \xi_j - E\left[\sigma_w(V) \max\{\sigma_w(V), \tau_w\} \right] \right\}, \tag{3.6}$$

$\mathcal{W}^\dagger = \{w = (w_1, \ldots, w_d) : \sum_{j=1}^d w_j = 1,\ 0 \le w_j \le 1,$ *for* $j = 1, \ldots, d\}$, $\sigma_w(\cdot) = \sqrt{\sum_{j=1}^d w_j b_j^{-1} \sigma_j^2(\cdot)}$ *and* τ_w *is the unique solution of* $E[\rho(V; \sigma_w, \tau)] = \varpi$ *with respect to* τ.

The proof of Theorem 2 can be found in Section S2.4 in Supplementary Material. Similar arguments as in the proof of Theorem 1 can show that ρ_ς actually minimizes a weighted sum of the semiparametric efficiency bound for parameter component

$$\sum_{j=1}^d w_{\varsigma,j} b_j^{-1} \left\{ E\left[\frac{\sigma_j^2(V)}{\rho(V)} \right] + \mathrm{Var}[\Pi_j(V)] \right\}. \tag{3.7}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0359

From the proof of Theorem 2, it can be seen that the objective function in (3.6) is actually a dual problem of (3.5), which can be viewed as a criterion to adaptively determine the weight of the efficiency bound for each parameter component. Then, with w_{ϱ} selected according to (3.6), the optimal rule that minimizes the weighted sum of semiparametric efficiency bounds in (3.7) is also the solution of the maximin problem (3.5). Theorem 2 contributes to the literature by providing a tractable way to solve the complex infinite-dimensional maximin problem (3.5). By applying Theorem 2, the problem (3.5) reduces to the more manageable finite-dimensional minimization problem (3.6). This simplification significantly eases the optimization process and facilities the data-based estimation.

Both $\rho_{c}(\cdot)$ and $\rho_{\varrho}(\cdot)$ reduce to the optimal sampling rule $\rho_{\mathcal{S}}(\cdot)$ in Theorem 1 when $d = 1$, but they differ when $d > 1$. As $\rho_{\varrho}(\cdot)$ solves the maximin problem (3.5), it dominates $\rho_0(\cdot)$ whenever there exists a sampling rule that dominates $\rho_0(\cdot)$. The sampling rule $\rho_{\varrho}(\cdot)$ maximizes the minimal relative improvement compared to the benchmark sampling rule, which makes it promising for estimating a multi-dimensional parameter.

In contrast to sampling rules based on other optimality criteria such as A-optimality or D-optimality (Kiefer, 1959), the maximin sampling rules $\rho_{c}(\cdot)$ and $\rho_{\varrho}(\cdot)$ ensure improved efficiency for each parameter component relative to the benchmark rule $\rho_0(\cdot)$. This property is particularly desirable when it is hard to determine the importance of the estimation efficiency for each parameter component at the design stage, making $\rho_{c}(\cdot)$ and $\rho_{\varrho}(\cdot)$ appealing choices in epidemiological and genetic studies with multiple biomarkers, traits, or a multi-level treatment.

Remark 1. The formulation $M(\rho)$ and (3.5) the case where all the parameters are equally

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0359

important. In practice, different parameters can have different levels of importance or interest. The proposed framework can be extended to take the different importance into consideration. A method is to add weights to the objective function (3.6). That is, we consider the optimization problem

$$\max_{\rho\in\mathcal{P}_{\mathcal{G}}}\min_{j=1,\ldots,d}\left\{a_j^{-1}b_j^{-1}\left(\xi_j - E\left[\frac{\sigma_j^2(V)}{\rho(V)}\right]\right)\right\} \tag{3.8}$$

instead of (3.6), where $a_j > 0$ is a specified weight to reflect the importance of the jth parameter component for $j = 1, \ldots, d$. Note that the optimization problem is unchanged if we multiply each of $\{a_j\}_j^d$ by a positive constant. Thus, without loss of generality, we can assume that $\sum_{j=1}^d a_j = 1$. Let

$$\mathcal{P}_{\mathrm{NL}} = \left\{\begin{array}{l}\rho(\cdot) : 0 \le \rho(\cdot) \le 1,\ E[\rho(V)] \le \varpi,\ \text{the semiparametric efficiency bound for} \\ \\ \text{every parameter component under } \rho(\cdot) \text{ is no larger than that under } \rho_0(\cdot)\end{array}\right\}.$$

Then, it is not hard to verify that the solution of (3.8) belongs to $\mathcal{P}_{\mathrm{NL}}$, which is a desirable property. Moreover, this implies that the solutions of (3.8) and

$$\max_{\rho\in\mathcal{P}_{\mathrm{NL}}}\min_{j=1,\ldots,d}\left\{a_j^{-1}b_j^{-1}\left(\xi_j - E\left[\frac{\sigma_j^2(V)}{\rho(V)}\right]\right)\right\} \tag{3.9}$$

are identical. For any $\rho(\cdot) \in \mathcal{P}_{\mathrm{NL}}$, a larger a_j makes it more likely that

$$\min_{j'=1,\ldots,d}\left\{a_{j'}^{-1}b_{j'}^{-1}\left(\xi_{j'} - E\left[\frac{\sigma_{j'}^2(V)}{\rho(V)}\right]\right)\right\} = a_j^{-1}b_j^{-1}\left(\xi_j - E\left[\frac{\sigma_j^2(V)}{\rho(V)}\right]\right)$$

and hence prioritizes the efficiency for estimating the jth parameter component in the maximin optimization (3.9). Thus, if the jth parameter component is important in a research, one can set a_j to be large to prioritize the efficiency of the estimator for the jth parameter component. According to such intuition, users have the flexibility to set the parameters $\{a_j\}_{j=1}^d$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0359

to prioritize some parameter components based on their research interest. Similar arguments as in the proof of Theorem 2 can show that the optimal sampling rule that minimizes (3.8) is

$$\rho(\cdot;\sigma_{w_a},\tau_{w_a}) = 1\{\sigma_{w_a}(\cdot) > \tau_{w_a}\} + \frac{1}{\tau_{w_a}}\sigma_{w_a}(\cdot)1\{\sigma_{w_a}(\cdot) \le \tau_{w_a}\},$$

where $w_a = (w_{a,1},\ldots,w_{a,d})^{\mathrm{T}}$ is the solution of the problem

$$\min_{w\in\mathcal{W}_a^\dagger}\left\{\sum_{j=1}^{d} w_j b_j^{-1}\xi_j - E\left[\sigma_w(V)\max\{\sigma_w(V),\tau_w\}\right]\right\},$$

$\sigma_w(\cdot) = \sqrt{\sum_{j=1}^{d} w_j b_j^{-1}\sigma_j^2(\cdot)}$, τ_w is the unique solution of $E[\rho(V;\sigma_w,\tau)] = \varpi$ with respect to τ, and $\mathcal{W}_a^\dagger = \{w = (w_1,\ldots,w_d) : \sum_{j=1}^{d} a_j w_j = 1,\ 0 \le w_j \le a_j^{-1},\ \text{for } j = 1,\ldots,d\}$. See Section S4.3 in Supplementary Material for numerical demonstrations on the effect of $\{a_j\}_{j=1}^d$ in prioritizing the efficiency of some parameter component.

4. Estimation of the optimal sampling rule

The optimal sampling rules proposed in Section 3.1 depend on the conditional variance of the full data efficient influence function, which is unknown. In this section, we consider the estimation of the optimal sampling rule, while the parameter of interest can be either scalar or multi-dimensional. We draw a simple random subsample at the beginning of the second phase, and use the pilot sample to estimate the optimal sampling rule. And then we draw the remaining subjects according to the estimated rule. Specifically, let $(V_1,U_1),\ldots,(V_n,U_n)$ be independent and identically distributed (i.i.d.) copies of (V,U). For $i = 1,\ldots,n$, V_i is observed in the first phase. Let R_{1i}, $i = 1,\ldots,n$, be independent Bernoulli random variables which equal to one with probability κ_n, where κ_n is the predetermined expected proportion of the pilot sample that may change with n. The ith subject is included in the pilot sample,

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0359

and U_i is measured if $R_{1i} = 1$, and not otherwise. Then the optimal sampling rules can be estimated based on the pilot sample. Subsequently, for each subject with $R_{1i} = 0$, we generate an independent Bernoulli random variables R_{2i} which equals one with probability $\tilde{\rho}(V_i)$, where $\tilde{\rho}(\cdot)$ is the estimated sampling rule. Afterwards, U_i is measured for subjects with $R_{2i} = 1$. Based on the data collected in the two phases and the pilot sample, θ_0 can be estimated using standard semiparametric methods, such as the targeted maximum likelihood (van der Laan and Rubin, 2006), the estimating equation method (Tsiatis, 2007), and the one-step estimation method (Bickel, 1982). For illustration, we state the one-step estimation procedure under two-phase designs in Section S3.3 in Supplementary Material. Next, we get down to the estimation of the sampling rule. The sampling rule for U in the above procedure is a mixture of the uniform rule and the estimated optimal rule due to the presence of the pilot sample. This sampling rule is nearly optimal provided the $\tilde{\rho}(\cdot)$ is consistent for the optimal sampling rule and the proportion of pilot sample is small in the sense that $\kappa_n/\varpi \to 0$.

We first consider the estimation of the optimal sampling rule $\rho_j(\cdot)$ for each component. The optimal sampling rule $\rho_s(\cdot)$ for a scalar parameter given in Theorem 1 can be regarded as a special case of $\rho_j(\cdot)$ with $j = d = 1$. Recall that $\rho_j(\cdot) = \rho(\cdot; \sigma_j, \tau_j)$ where τ_j satisfies $E[\rho(V; \sigma_j, \tau_j)] = \varpi$. The unknown quantity $\sigma_j(\cdot)$ can be estimated by well-established statistical methods based on the observations in the first phase and the pilot sample for $j = 1, \ldots, d$. For more details, see Section S3.2 in Supplementary Material. Denote the estimate for $\Pi_j(\cdot)$ and $\sigma_j(\cdot)$ by $\widetilde{\Pi}_j(\cdot)$ and $\tilde{\sigma}_j(\cdot)$, respectively, for $j = 1, \ldots, d$. The threshold τ_j is another quantity requiring estimating. Note that we aim to take ϖ as the expected proportion of subjects included in the second phase, and the pilot sample occupies an expected

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0359

proportion of κ_n. Thus, the expected proportion of subjects included in the second phase with $R_{1i} = 0$ should be $\varpi - \kappa_n$. Hence we estimate the threshold τ_j by the solution $\tilde{\tau}_j$ of the equation

$$\frac{1}{n}\sum_{i=1}^{n}(1 - R_{1i})\left(1\{\tilde{\sigma}_j(V_i) > \tau\} + \frac{\tilde{\sigma}_j(V_i)}{\tau}1\{\tilde{\sigma}_j(V_i) \le \tau\}\right) = \varpi - \kappa_n. \tag{4.1}$$

Then $\rho_j(\cdot)$ can be estimated by

$$\tilde{\rho}_j(\cdot) = 1\{\tilde{\sigma}_j(\cdot) > \tilde{\tau}_j\} + \frac{\tilde{\sigma}_j(\cdot)}{\tilde{\tau}_j}1\{\tilde{\sigma}_j(\cdot) \le \tilde{\tau}_j\},$$

for $j = 1, \ldots, d$. The sampling rule $\rho_c(\cdot)$ for a multi-dimensional parameter can be estimated by $\tilde{\rho}_c(\cdot) = \rho_0(\cdot) + \sum_{j=1}^{d} \widetilde{w}_{c,j}(\tilde{\rho}_j(\cdot) - \rho_0(\cdot))$, where

$$\begin{aligned}\widetilde{w}_c &= (\widetilde{w}_{c,0}, \ldots, \widetilde{w}_{c,d}) \\ &= \arg\max_{w\in\mathcal{W}} \min_{j=1,\ldots,d} \left\{\tilde{b}_j^{-1}\left[\tilde{\xi}_j - \frac{1}{n}\sum_{i=1}^{n}\frac{\tilde{\sigma}_j^2(V_i)}{\rho_0(V_i) + \sum_{j=1}^{d} w_j\{\tilde{\rho}_j(V_i) - \rho_0(V_i)\}}\right]\right\},\end{aligned}$$

$\tilde{\xi}_j = n^{-1}\sum_{i=1}^{n}\tilde{\sigma}_j^2(V_i)/\rho_0(V_i)$ and $\tilde{b}_j = n^{-1}\sum_{i=1}^{n}\tilde{\sigma}_j^2(V_i)/\rho_0(V_i) + n^{-1}\sum_{i=1}^{n}\{\widetilde{\Pi}_j(V_i) - n^{-1}\sum_{i=1}^{n}\widetilde{\Pi}_j(V_i)\}^2$ are the estimates of ξ_j and b_j, for $j = 1, \ldots, d$. According to Theorem 2, to estimate $\rho_\varrho(\cdot)$, we first estimate w_ϱ by

$$\widetilde{w}_\varrho = \arg\min_{w\in\mathcal{W}^\dagger}\left\{\sum_{j=1}^{d} w_j\tilde{b}_j^{-1}\tilde{\xi}_j - \frac{1}{n}\sum_{i=1}^{n}\tilde{\sigma}_w(V_i)\max\{\tilde{\sigma}_w(V_i), \tilde{\tau}_w\}\right\},$$

where $\tilde{\sigma}_w(\cdot) = \sqrt{\sum_{j=1}^{d} w_j\tilde{b}_j^{-1}\tilde{\sigma}_j^2(\cdot)}$ and $\tilde{\tau}_w$ is the solution of (4.1) with $\tilde{\sigma}_j(\cdot)$ replaced by $\tilde{\sigma}_w(\cdot)$. Then the estimator for $\rho_\varrho(\cdot)$ is given by

$$\tilde{\rho}_\varrho(\cdot) = 1\{\tilde{\sigma}_{\widetilde{w}_\varrho}(\cdot) > \tilde{\tau}_{\widetilde{w}_\varrho}\} + \frac{\tilde{\sigma}_{\widetilde{w}_\varrho}(\cdot)}{\tilde{\tau}_{\widetilde{w}_\varrho}}1\{\tilde{\sigma}_{\widetilde{w}_\varrho}(\cdot) \le \tilde{\tau}_{\widetilde{w}_\varrho}\},$$

where $\tilde{\sigma}_{\widetilde{w}_\varrho}(\cdot)$ is $\tilde{\sigma}_w(\cdot)$ with w equaling to $\widetilde{w}_\varrho$. The following theorem shows that the above sampling procedure satisfies the exact budget constraint, i.e., the expected proportion of subjects included in the second phase is exactly ϖ.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0359

Theorem 3. *If $\tilde{\rho}(\cdot) = \tilde{\rho}_c(\cdot)$, $\tilde{\rho}_\varrho(\cdot)$, or $\tilde{\rho}_j(\cdot)$ ($j \in \{1, \ldots, d\}$), we have*

$$E\left[\frac{1}{n}\sum_{i=1}^{n}(R_{1i} + R_{2i})\right] = \varpi.$$

The proof of Theorem 3 is provided in Section S2.5 in Supplementary Material. We next investigate the convergence rate of the proposed sampling rule estimators, starting with the component-wise optimal sampling rule $\tilde{\rho}_j(\cdot)$ for $j = 1, \ldots, d$. For any function $f(\cdot)$, let $\|f\|_\infty = \sup_v |f(v)|$ be the infinity norm of f. To establish the theoretical result, it is required that the size of the pilot sample and the conditional standard deviation estimator satisfy the following condition.

Condition 1. $\kappa_n \to 0$, $n\kappa_n \to \infty$, $\|\widetilde{\Pi}_j - \Pi_j\|_\infty = O_P\left\{(n\kappa_n)^{-\delta}\right\}$ and $\|\tilde{\sigma}_j - \sigma_j\|_\infty = O_P\left\{(n\kappa_n)^{-\delta}\right\}$ for $j = 1, \ldots, d$ and some $0 < \delta \leq 1/2$.

The first two rate conditions can be satisfied by properly chosen κ_n. In practice, we recommend to take κ_n as a sequence that converges slowly to zero, e.g., $\kappa_n = \varpi/\log(\varpi n)$, to ensure the stability of the estimates based on the pilot sample. Convergence rates required by Condition 1 are well established in the nonparametric estimation literature. Specifically, the convergence rate is readily available under certain regularity conditions if the conditional variance is estimated using the kernel method (Hardle et al., 1988) or the sieve method (Chen and Christensen, 2015). Under suitable regularity conditions, we have the following theorem. See Section S2.6 of Supplementary Material for the proof of the theorem.

Theorem 4. *Suppose the parameter θ_0 is scalar. Under Condition 1 and Conditions S1–S2 in Supplementary Material, we have*

$$\|\tilde{\rho}_s - \rho_s\|_\infty = O_P\left\{(n\kappa_n)^{-\delta} + \kappa_n\right\}.$$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0359

Theorem 4 can be directly extended to prove $\|\tilde{\rho}_j - \rho_j\|_\infty = O_P\left\{(n\kappa_n)^{-\delta} + \kappa_n\right\}$ for $j = 1, \ldots, d$ when the parameter is multi-dimensional. From the proof of Theorem 4, it can be seen that the term κ_n in the convergence rate can be removed if $(1 - \kappa_n)\varpi$ is used instead of $\varpi - \kappa_n$ on the right-hand side of (4.1). However, we use $\varpi - \kappa_n$ in (4.1) to ensure the exact budget constraint stated in Theorem 3. For the estimation of $\rho_c(\cdot)$ and $\rho_g(\cdot)$, we have the following convergence result.

Theorem 5. *Suppose $\rho_0(\cdot)$ is bounded away from zero. Then under Condition 1 and Conditions S1–S5 in Supplementary Material, we have*

$$\|\tilde{\rho}_c - \rho_c\|_\infty = O_P\left\{\sqrt{(n\kappa_n)^{-\delta} + \kappa_n}\right\} \text{ and } \|\tilde{\rho}_g - \rho_g\|_\infty = O_P\left\{\sqrt{(n\kappa_n)^{-\delta} + \kappa_n}\right\}.$$

The proof of Theorem 5 can be found in Section S2.7 of Supplementary Material.

5. Simulation study

In this section, we use "uni" to denote the uniform rule, "S-opt" to denote the optimal sampling rule for a scalar parameter $\tilde{\rho}_s$, "C-opt" to denote the sampling rule $\tilde{\rho}_c$, and "G-opt" to denote the sampling rule $\tilde{\rho}_g$. We consider the average treatment effect estimation problem introduced in Example S3. The problem is akin to the causal effect estimation problem in observational studies with missing confounders discussed in Lin and Chen (2014) and Yang and Ding (2020). Let Z be a q-dimensional covariate vector with independent $U[-2.5, 2.5]$ components, where $U[-2.5, 2.5]$ is the uniform distribution on $[-2.5, 2.5]$. Suppose the confounder X follows the model

$$X = \zeta_q^{\mathrm{T}} Z + \epsilon_x,$$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0359

where $\zeta_q = (0.5/\sqrt{q}, \ldots, 0.5/\sqrt{q})^{\mathrm{T}}$ and ϵ_x follows a normal distribution with mean zero and variance 0.25. The potential outcomes Y_0 and Y_1 follow the model

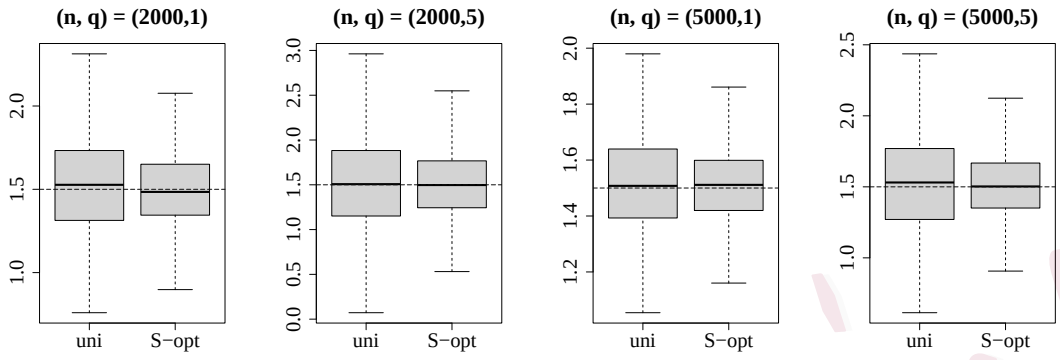
$$Y_0 = 0.5\zeta_q^{\mathrm{T}} Z + X + \exp(2\zeta_q^{\mathrm{T}} Z)\epsilon_y,$$

$$Y_1 = \theta_0 + 0.5\zeta_q^{\mathrm{T}} Z - X + \exp(2\zeta_q^{\mathrm{T}} Z)\epsilon_y.$$

where ϵ_y is a standard normal error. Under the above model, $\theta_0 = E[Y_1 - Y_0]$ is the average treatment effect. We take $\theta_0 = 1.5$ in the simulation. The treatment indicator T is a Bernoulli variable that is equal to one with probability $1/\{1+\exp(0.1\zeta_q^{\mathrm{T}} Z - 0.5X)\}$. Then the outcome variable is $Y = TY_1 + (1-T)Y_0$. Let $V = (Y, T, Z^{\mathrm{T}})^{\mathrm{T}}$ be the vector of first-phase variables and $U = X$ the second-phase variable in this example. In the simulation, we generate $n = 2000$ or 5000 i.i.d. copies of (V, U), draw a pilot sample and estimate the sampling rules based on the nonparametric method suggested in Section S3.2 in Supplementary Material. We apply different sampling rules to draw the sample in the second phase. Then, the one-step estimation method is adopted to estimate θ_0. The one-step estimation method (Bickel, 1982) is a classic semiparametric method to construct an efficient estimator based on the efficient influence function. Please refer to Section S3.3 in Supplementary Material for more details. We consider $q = 1$ and 5 and take the proportion of pilot sample $\kappa_n = \varpi/[1 + \log\{\varpi n/(q+1)\}]$. The quantity $\varpi/[1 + \log\{\varpi n/(q+1)\}]$ increases as q grows and decreases to zero slowly as $n \to \infty$. Thus, the choice of κ_n ensures Condition 1 while taking the dimension of the first-phase variable vector into consideration. Figure 1 presents boxplots of the average treatment effect estimator over 500 simulations with $n = 2000, 5000$ and $q = 1, 5$.

Figure 1 shows that the efficiency of the average treatment effect estimator is improved under $\tilde{\rho}_s$ compared to the uniform rule. The improvement is consistently observed across

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0359



(a) (b) (c) (d)

Figure 1: Boxplots for the average treatment effect estimation with different combinations of n and q; dashed lines are the true values.

different combinations of n and q. We compute the relative efficiency (RE) under $\tilde{\rho}_s$ with respect to the uniform rule. The RE of a sampling rule is the variance ratio of the estimator under the uniform rule to that under the considered sampling rule. Higher RE indicates greater improvement than the uniform rule. The REs are 1.6995, 2.1835, 1.8907 and 2.3239 for $(n, q) = (2000, 1), (2000, 5), (5000, 1)$, and $(5000, 5)$, respectively, which indicates significant efficiency improvements across different scenarios.

In practice, there may be multiple treatments. In this case, the average treatment effects of different treatments compared to the control are of interest. Next, we consider the average treatment effect estimation with two different treatments (denoted by "1", and "2") and a control group (denoted by "0"). Define Z in the same way as in the scalar case. Suppose X follows the model

$$X = \zeta_q^{\mathrm{T}} Z + \epsilon_x,$$

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0359

where ϵ_x follows a normal distribution with mean zero and variance 0.25. The potential outcomes are generated from

$$Y_0 = 0.5\zeta_q^{\mathrm{T}} Z + X + \{0.5 + \nu_2(Z)\}\epsilon_y,$$

$$Y_1 = \theta_{01} + 0.5\zeta_q^{\mathrm{T}} Z - X + \{\nu_1(Z) + \nu_2(Z)\}\epsilon_y,$$

$$Y_2 = \theta_{02} - 0.5\zeta_q^{\mathrm{T}} Z - 0.5X + \nu_2(Z)\epsilon_y,$$

where $\theta_{01} = 1$, $\theta_{02} = 0.5$, ϵ_y follows a standard normal distribution, $\nu_1(z) = \sqrt{0.1 + (2\zeta_q^{\mathrm{T}} z)^4}$, and $\nu_2(z) = \exp(2\zeta_q^{\mathrm{T}} z)$. It is easy to verify that the average treatment effects $E[Y_1 - Y_0]$ and $E[Y_2 - Y_0]$ equal to θ_{01} and θ_{02}, respectively. Suppose the treatment indicator T equals to 0, 1, and 2 with probabilities $1 - p_1(X, Z) - p_2(X, Z)$, $p_1(X, Z)$ and $p_2(X, Z)$, where

$$p_1(X, Z) = \frac{\exp(-0.1\zeta_q^{\mathrm{T}} Z + 0.25X)}{1 + \exp(-0.1\zeta_q^{\mathrm{T}} Z + 0.25X) + \exp(0.1\zeta_q^{\mathrm{T}} Z - 0.25X)},$$

$$p_2(X, Z) = \frac{\exp(0.1\zeta_q^{\mathrm{T}} Z - 0.25X)}{1 + \exp(-0.1\zeta_q^{\mathrm{T}} Z + 0.25X) + \exp(0.1\zeta_q^{\mathrm{T}} Z - 0.25X)}.$$

The observed outcome Y equals to Y_t if $T = t$ for $t = 0, 1$, and 2. Table 1 presents the bias, standard error (SE) of the estimator, and the RE compared to the uniform rule based on 500 simulations.

Table 1 shows that the SEs are smaller under the sampling rules $\tilde{\rho}_c$ and $\tilde{\rho}_g$ than under the uniform rule. Moreover, the improvement is consistent and significant across different parameter components and values of (n, q). In some cases, the RE is close to or even larger than two. The improvement of $\tilde{\rho}_g$ tends to be greater than $\tilde{\rho}_c$ in most cases, probably because $\tilde{\rho}_g$ estimates the optimal solution over the full space, while $\tilde{\rho}_c$ estimates the optimal solution over only a subspace. In addition, the simulation results in this section indicate that the proposed method's performance remains stable even as the dimension of the first-phase

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0359

Table 1: Bias, SE, and RE in two-dimensional average treatment effect estimation

(n, q)	Rule	Estimate of θ_{01}			Estimate of θ_{02}		
		Bias	SE	RE	Bias	SE	RE
$(2000, 1)$	uni	-0.0055	0.5630	1.0000	0.0150	0.4755	1.0000
	C-opt	0.0022	0.4558	1.5257	0.0220	0.3620	1.7254
	G-opt	0.0157	0.4597	1.4999	0.0237	0.3596	1.7485
$(2000, 5)$	uni	0.0220	0.9095	1.0000	-0.0417	0.8307	1.0000
	C-opt	-0.0171	0.7734	1.3829	-0.0545	0.6452	1.6577
	G-opt	-0.0206	0.7384	1.5171	-0.0354	0.6199	1.7957
$(5000, 1)$	uni	0.0200	0.3349	1.0000	-0.0043	0.2765	1.0000
	C-opt	0.0289	0.2564	1.7061	-0.0010	0.2179	1.6102
	G-opt	0.0189	0.2459	1.8549	-0.0043	0.2113	1.7123
$(5000, 5)$	uni	-0.0353	0.5363	1.0000	-0.0119	0.4667	1.0000
	C-opt	-0.0161	0.3970	1.8249	-0.0181	0.3332	1.9618
	G-opt	-0.0006	0.3836	1.9546	-0.0110	0.3115	2.2447

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0359

variable increases. This suggests that the proposed method is not highly sensitive to the dimensionality of the first-phase variable despite the involvement of nonparametric methods in the estimation procedure.

6. Real data analysis

In this section, we illustrate the proposed approach using AIDS Clinical Trials Group (ACTG) 175 data (Hammer et al., 1996), which can be found in the R package *speff2trial*. The ACTG175 clinical trial evaluates four therapies for treating human immunodeficiency virus infection. The four treatments are: 1, zidovudine; 2, zidovudine and didanosine; 3, zidovudine and zalcitabine; and 4, didanosine. In the trial, 2139 subjects were assigned randomly to the four treatment groups. Covariates such as age, gender, weight, medication history, and symptom indicator were recorded before the treatment, and CD4 and CD8 cell counts were measured after 20 weeks of the treatment. The CD4/CD8 ratio describes the overall immune function of an individual. A low value of CD4/CD8 ratio indicates a severe immune dysfunction. We use the CD4/CD8 ratio as the outcome variable and consider the average treatment effects of the three zidovudine-treated groups (groups 1, 2, and 3) compared to the non-zidovudine-treated group (group 4). Let θ_{0j} be the average treatment effect of the treatment j compared to the treatment 4 for $j = 1, \ldots, 3$.

The original experiment measured CD4 and CD8 cell counts for all 2139 subjects after 20 weeks of the treatment. However, it can be costly to measure the CD4 and CD8 cell counts for more than two thousand people. The cost of the trials would be drastically reduced if a two-phase design were adopted and the CD4 and CD8 cell counts were measured for only

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0359

a subset of patients. We applied the proposed method to design the second-phase sampling rule and obtained results as if the experiment was conducted under the two-phase design. We use six variables as the first-phase variables V, including age, gender, weight, medication history, symptom indicator, and treatment assignment. The outcome is considered as the second-phase variable U. In the implementation, we take $\varpi = 0.3$ and $\kappa_n = \varpi/[1 + \log(0.1n\varpi)]$. A pilot sample is randomly drawn with inclusion probability κ_n. The sampling rules $\tilde{\rho}_c$ and $\tilde{\rho}_g$ are constructed based on the pilot sample and the full data efficient influence function in Example S3 in Section S1.1 in Supplementary Material. Then, a sample is drawn according to the resulting sampling rule. The full data efficient influence function involves the propensity score and the outcome regression function of each group. The propensity score for each group is $1/4$ because subjects were randomly assigned to each group with equal probability. We adopt a linear model for the outcome regression function.

We employ the one-step estimation method to estimate the average treatment effects. In line with the simulation, "uni", "C-opt", and "G-opt" represent the uniform rule, the sampling rule $\tilde{\rho}_c$, and the sampling rule $\tilde{\rho}_g$, respectively. Table 2 provides the estimates for θ_{01}, θ_{02}, and θ_{03} under different sampling rules. Additionally, it includes the estimated SEs based on the asymptotic variance of the estimator and p-values associated with the Wald tests for testing whether each average treatment effect equals to zero.

Table 2 shows that the estimators for θ_{01}, θ_{02}, and θ_{03} all have smaller estimated SEs under $\tilde{\rho}_c$ and $\tilde{\rho}_g$ compared to the uniform rule. The p-values under $\tilde{\rho}_c$ and $\tilde{\rho}_g$ suggest that, at the 0.01 significance level, treatment 1 performs differently than treatment 4 in increasing the CD4/CD8 ratio, while there is no significant difference between treatments 2, 3, and

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0359

Table 2: Point estimates (Est), estimated SEs, and p-values associated with the estimation of θ_{01}, θ_{02}, and θ_{03} in the ACTG175 trial under different sampling rules

Rule	Estimation of θ_{01}			Estimation of θ_{02}			Estimation of θ_{03}		
	Est	SE	p-value	Est	SE	p-value	Est	SE	p-value
uni	-0.0489	0.0223	0.0283	0.0166	0.0247	0.5013	-0.0034	0.0228	0.8805
C-opt	-0.0606	0.0211	0.0041	0.0182	0.0235	0.4398	-0.0012	0.0216	0.9574
G-opt	-0.0630	0.0211	0.0029	0.0147	0.0231	0.5227	-0.0006	0.0216	0.9794

treatment 4. In contrast, the results under the uniform rule discover no significant difference among treatments 1, 2, 3, and treatment 4 under the significance level 0.01. For comparison, we calculate the full-cohort augmented inverse probability weighted estimator for θ_{01}, θ_{02}, and θ_{03} based on the first-phase variables and the outcome of all subjects. It is worth noting that the full-cohort estimator requires measuring the outcome of all 2139 subjects, while the estimates in Table 2 only needs the outcome of approximately 30% of the subjects. The all data-based estimates are -0.0500, 0.0166, and 0.0054 for θ_{01}, θ_{02} and θ_{03}, respectively, with p-values 0.0005, 0.2851, and 0.7140. These p-values support the same conclusions as those suggested by the estimators under $\tilde{\rho}_c$ and $\tilde{\rho}_g$.

7. Discussion

This paper proposes a unified framework for designing sampling rules in two-phase studies. The proposed method offers promising guarantees in estimation efficiency and budget control. It exhibits broad applicability across diverse research domains, such as the electronic

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0359

health record data analysis (Lotspeich et al., 2022; Zhang et al., 2024), genetic association analysis (Lin et al., 2013), causal inference (Lin and Chen, 2014; Yang and Ding, 2020), and measurement error problems. Our framework accommodates multi-dimensional parameters, making it well-suited for association studies involving multiple traits and the estimation of causal effects with multiple treatments. The current framework relies on the semiparametric efficiency bound for a fixed-dimensional parameter. An important topic for future research is the development of suitable criteria for high-dimensional problems and the corresponding optimal designs.

The implementation of the proposed design method requires the full-data efficient influence function. The efficient influence functions are available from the literature and easy to calculate in many important examples such as the examples presented in Section S1.1 in Supplementary Material, as well as in a lot of other examples (Newey, 1993; Tsiatis, 2007). However, the derivation and calculation of the full-data efficient influence function may be challenging under some complicated semiparametric models such as those with missing covariates or complex censoring. Additional theoretical or computational efforts are required to implement the proposed method in these scenarios.

Supplementary material

Supplementary Material contains proofs of the theoretical results; a nonparametric procedure to estimate the conditional mean and variance simultaneously; a description of the one-step estimation method based on the data collected in the two phases and the pilot sample; and additional simulation results.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0359

Acknowledgments

We would like to thank the editor, associate editor, and the anonymous reviewer for their very insightful and helpful comments, which led to a significant improvement of our paper. Wang's research was supported by the National Natural Science Foundation of China (Grant No. 12271510), and a grant from the Key Lab of Random Complex Structure and Data Science, CAS. Miao's research was supported by the National Key R&D Program (2022YFA1008100) and the National Natural Science Foundation of China (Grant No. 12071015).

References

Bickel, P. (1982). On adaptive estimation. *The Annals of Statistics 10*(3), 647–671.

Chatterjee, N., Y.-H. Chen, and N. E. Breslow (2003). A pseudoscore estimator for regression problems with two-phase sampling. *Journal of the American Statistical Association 98*(461), 158–168.

Chen, T. and T. Lumley (2022). Optimal sampling for design-based estimators of regression models. *Statistics in Medicine 41*(8), 1482–1497.

Chen, X. and T. M. Christensen (2015). Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions. *Journal of Econometrics 188*(2), 447–465.

Cochran, W. G. (2007). *Sampling Techniques* (3rd ed.). John Wiley & Sons.

Cox, D. R. and D. V. Hinkley (1979). *Theoretical statistics*. CRC Press.

Dette, H. (1997). Designing experiments with respect to 'standardized'optimality criteria. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 59*(1), 97–110.

Fattorini, L., M. Marcheselli, C. Pisani, and L. Pratelli (2017). Design-based asymptotics for two-phase sampling strategies in environmental surveys. *Biometrika 104*(1), 195–205.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0359

Gilbert, P. B., X. Yu, and A. Rotnitzky (2014). Optimal auxiliary-covariate-based two-phase sampling design for semiparametric efficient estimation of a mean or mean difference, with application to clinical trials. *Statistics in Medicine 33*(6), 901–917.

Green, D. M., N. E. Breslow, J. B. Beckwith, J. Z. Finklestein, P. E. Grundy, P. R. Thomas, T. Kim, S. J. Shochat, G. M. Haase, M. L. Ritchey, P. P. Kelalis, and G. J. D'Angio (1998). Comparison between single-dose and divided-dose administration of dactinomycin and doxorubicin for patients with wilms' tumor: a report from the national wilms' tumor study group. *Journal of Clinical Oncology 16*(1), 237–245.

Hammer, S. M., D. A. Katzenstein, M. D. Hughes, H. Gundacker, R. T. Schooley, R. H. Haubrich, W. K. Henry, M. M. Lederman, J. P. Phair, M. Niu, M. S. Hirsch, and T. C. Merigan (1996). A trial comparing nucleoside monotherapy with combination therapy in hiv-infected adults with cd4 cell counts from 200 to 500 per cubic millimeter. *The New England Journal of Medicine 335*, 1081–1090.

Hardle, W., P. Janssen, and R. Serfling (1988). Strong uniform consistency rates for estimators of conditional functionals. *The Annals of Statistics*, 1428–1449.

Kiefer, J. (1959). Optimum experimental designs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 21*(2), 272–304.

Lin, D.-Y., D. Zeng, and Z.-Z. Tang (2013). Quantitative trait analysis in sequencing studies under trait-dependent sampling. *Proceedings of the National Academy of Sciences 110*(30), 12247–12252.

Lin, H.-W. and Y.-H. Chen (2014). Adjustment for missing confounders in studies based on observational databases: 2-stage calibration combining propensity scores from primary and validation data. *American Journal of Epidemiology 180*(3), 308–317.

Liu, Y., Z. Li, and X. Lin (2022). A minimax optimal ridge-type set test for global hypothesis with applications in whole genome sequencing association studies. *Journal of the American Statistical Association 117*(538), 897–908.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0359

Lotspeich, S. C., B. E. Shepherd, G. G. Amorim, P. A. Shaw, and R. Tao (2022). Efficient odds ratio estimation under two-phase sampling using error-prone data from a multi-national hiv research cohort. *Biometrics 78*(4), 1674–1685.

McGowan, C. C., P. Cahn, E. Gotuzzo, D. Padgett, J. W. Pape, M. Wolff, M. Schechter, and D. R. Masys (2007). Cohort profile: Caribbean, central and south america network for hiv research (ccasanet) collaboration within the international epidemiologic databases to evaluate aids (iedea) programme. *International Journal of Epidemiology 36*(5), 969–976.

McIsaac, M. A. and R. J. Cook (2014). Response-dependent two-phase sampling designs for biomarker studies. *Canadian Journal of Statistics 42*(2), 268–284.

McIsaac, M. A. and R. J. Cook (2015). Adaptive sampling in two-phase designs: A biomarker study for progression in arthritis. *Statistics in Medicine 34*(21), 2899–2912.

McNamee, R. (2002). Optimal designs of two-stage studies for estimation of sensitivity, specificity and positive predictive value. *Statistics in Medicine 21*(23), 3609–3625.

Meinshausen, N. and P. Bühlmann (2015). Maximin effects in inhomogeneous large-scale data. *The Annals of Statistics 43*(4), 1801–1830.

Nab, L., M. van Smeden, R. de Mutsert, F. R. Rosendaal, and R. H. Groenwold (2021). Sampling strategies for internal validation samples for exposure measurement–error correction: A study of visceral adipose tissue measures replaced by waist circumference measures. *American Journal of Epidemiology 190*(9), 1935–1947.

Newey, W. K. (1993). 16 efficient estimation of models with conditional moment restrictions. In *Econometrics*, Volume 11 of *Handbook of Statistics*, pp. 419–454. Elsevier.

Reilly, M. and M. S. Pepe (1995). A mean score method for missing and auxiliary covariate data in regression models. *Biometrika 82*(2), 299–314.

Tao, R., D. Zeng, and D.-Y. Lin (2017). Efficient semiparametric inference under two-phase sampling, with applications to genetic association studies. *Journal of the American Statistical Association 112*(520), 1468–1476.

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0359

Tao, R., D. Zeng, and D.-Y. Lin (2020). Optimal designs of two-phase studies. *Journal of the American Statistical Association 115*(532), 1946–1959.

Tsiatis, A. (2007). *Semiparametric theory and missing data*. Springer Science & Business Media.

van der Laan, M. J. and J. M. Robins (2012). *Unified Methods for Censored Longitudinal Data and Causality*. Springer Science & Business Media.

van der Laan, M. J. and D. Rubin (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics 2*(1), 1–38.

Yang, S. and P. Ding (2020). Combining multiple observational data sources to estimate causal effects. *Journal of the American Statistical Association 115*(531), 1540–1554.

Zhang, G., L. J. Beesley, B. Mukherjee, and X. Shi (2024). Patient recruitment using electronic health records under selection bias: A two-phase sampling framework. *The Annals of Applied Statistics 18*(3), 1858–1878.

Zhou, H., W. Xu, D. Zeng, and J. Cai (2014). Semiparametric inference for data with a continuous outcome from a two-phase probability-dependent sampling scheme. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 1*(76), 197–215.

Department of Biostatistics, Harvard T.H. Chan School of Public Health, 677 Huntington Avenue, Boston, MA, 02115, USA

E-mail: ruoyuwang@hsph.harvard.edu

Academy of Mathematics and Systems Science, Chinese Academy of Sciences, 55 Zhongguancun East Road, Haidian District, Beijing 100190, China

E-mail: qhwang@amss.ac.cn

School of Mathematical Sciences, Peking University, 5 Summer Palace Road, Haidian District, Beijing 100871, China

Statistica Sinica: Preprint
doi:10.5705/ss.202024.0359

E-mail: mwfy@pku.edu.cn

Statistica Sinica