The Annals of Statistics
2019, Vol. 47, No. 5, 2538–2566
https://doi.org/10.1214/18-AOS1756

SEMI-SUPERVISED INFERENCE: GENERAL THEORY AND ESTIMATION OF MEANS

BY ANRU ZHANG[*,1], LAWRENCE D. BROWN[†,2] AND T. TONY CAI[†,3]

University of Wisconsin—Madison[*] and University of Pennsylvania[†]

In memory of Lawrence D. Brown

We propose a general semi-supervised inference framework focused on the estimation of the population mean. As usual in semi-supervised settings, there exists an unlabeled sample of covariate vectors and a labeled sample consisting of covariate vectors along with real-valued responses ("labels"). Otherwise, the formulation is "assumption-lean" in that no major conditions are imposed on the statistical or functional form of the data. We consider both the ideal semi-supervised setting where infinitely many unlabeled samples are available, as well as the ordinary semi-supervised setting in which only a finite number of unlabeled samples is available.

Estimators are proposed along with corresponding confidence intervals for the population mean. Theoretical analysis on both the asymptotic distribution and ℓ_2-risk for the proposed procedures are given. Surprisingly, the proposed estimators, based on a simple form of the least squares method, outperform the ordinary sample mean. The simple, transparent form of the estimator lends confidence to the perception that its asymptotic improvement over the ordinary sample mean also nearly holds even for moderate size samples. The method is further extended to a nonparametric setting, in which the oracle rate can be achieved asymptotically. The proposed estimators are further illustrated by simulation studies and a real data example involving estimation of the homeless population.

1. Introduction. Semi-supervised learning arises naturally in statistics and machine learning when the labels are more difficult or more expensive to acquire than the unlabeled data. While numerous algorithms have been proposed for semi-supervised learning, they are mostly focused on classification, where the labels are discrete values representing the classes to which the samples belong [see, e.g., Ando and Zhang (2005, 2007), Blum and Mitchell (1998), Vapnik (2013), Wang and Shen (2007), Wang, Shen and Liu (2008), Wang, Shen and Pan (2009),

Received August 2017; revised August 2018.

[1]Supported in part by NSF Grant DMS-1811868 and NIH Grant R01-GM131399-01.

[2]Supported in part by NSF Grant DMS-10-07657.

[3]Supported in part NSF Grants DMS-1208982 and DMS-1403708, and NIH Grant R01 CA127334.

MSC2010 subject classifications. Primary 62F10, 62J05; secondary 62F12, 62G08.

Key words and phrases. Confidence interval, efficiency, estimation of mean, limiting distribution, semi-supervised inference.

Zhu (2008), Zhu and Goldberg (2009)]. The setting with continuous valued y has also been discussed in the literature; see, for example, Johnson and Zhang (2008), Lafferty and Wasserman (2008) and Chakrabortty and Cai (2018). For a survey of recent development in semi-supervised learning, readers are referred to Zhu and Goldberg (2009) and the references therein.

The general semi-supervised model can be formulated as follows. Let $(Y, X_1, X_2, \ldots, X_p)$ be a $(p+1)$-dimensional random vector following an unknown joint distribution $P = P(dy, dx_1, \ldots, dx_p)$. Denote by P_X the marginal distribution of $X = (X_1, X_2, \ldots, X_p)$. Suppose one observes n "labeled" samples from P,

$$[\mathbf{Y}, \mathbf{X}] = \{Y_k, X_{k1}, X_{k2}, \ldots, X_{kp}\}_{k=1}^{n}, \tag{1.1}$$

and, in addition, m "unlabeled" samples from the marginal distribution P_X

$$\mathbf{X}_{\text{add}} = \{X_{k1}, X_{k2}, \ldots, X_{kp}\}_{k=n+1}^{n+m}. \tag{1.2}$$

In this paper, we focus on estimation and statistical inference for one of the simplest features, namely the population mean $\theta = \mathbb{E}Y$. No specific distributional or marginal assumptions relating X and Y are made.

This inference of population mean under a general semi-supervised learning framework has a variety of applications. We discuss the estimation of treatment effect (ATE) in Section 5.1 and a prototypical example involving survey data in Section 5.2. It is noteworthy that for some other problems that do not at first look like mean estimation, one can recast them as mean estimation, possibly after an appropriate transformation. Examples include estimation of the variance of Y or covariance between Y and a given X_i. In work that builds on a portion of the present paper, Azriel et al. (2016) considers construction of linear predictors in semi-supervised learning settings.

To estimate $\theta = \mathbb{E}Y$, the most straightforward estimator is the sample average $\bar{\mathbf{Y}} := \frac{1}{n}\sum_{k=1}^{n} Y_k$. Surprisingly, as we show later, in the semi-supervised setting, a simple adjusted-least-squares estimator, which exploits the unknown association of Y and X, outperforms $\bar{\mathbf{Y}}$. We first consider an ideal setting where there are infinitely many unlabeled samples, that is, $m = \infty$. This is equivalent to the case of known marginal distribution P_X. We refer to this case as *ideal semi-supervised inference*. In this case, our proposed estimator is

$$\hat{\theta} = \bar{\mathbf{Y}} - \hat{\beta}_{(2)}^{\top}(\bar{\mathbf{X}} - \mu), \tag{1.3}$$

where $\bar{\mathbf{X}} \in \mathbb{R}^p$ such that $\bar{\mathbf{X}}_i = \frac{1}{n}\sum_{k=1}^{n} X_{ki}$, $\hat{\beta}_{(2)}$ is the p-dimensional least squares estimator for the regression slopes, and $\mu = \mathbb{E}X$ is the population mean of X. We emphasize again that although the estimator (1.3) has a linear structure we are not assuming that $\mathbb{E}(Y|X)$ is linearly related to X. This estimator is analyzed in detail in Section 2.2.

We then consider the more realistic setting where there are a finite number of unlabeled samples, that is, $m < \infty$. Here, one has only partial information about

P_X. We call this case *ordinary semi-supervised inference*. In this setting, we propose to estimate θ by

$$\hat{\theta} = \bar{\mathbf{Y}} - \hat{\beta}_{(2)}^{\top}(\bar{\mathbf{X}} - \hat{\mu}), \tag{1.4}$$

where $\hat{\mu}$ denotes the sample average of both the labeled and unlabeled X's. The detailed analysis of this estimator is given in Section 2.3.

We will investigate the properties of these estimators and in particular establish their asymptotic distributions and the ℓ_2 risk bounds. The limiting distribution results allow us to construct an asymptotically valid confidence interval based on the proposed estimators that is shorter than the traditional sample-mean-based confidence interval. Both the case of a fixed number of covariates and the case of a growing number of covariates are considered. The basic asymptotic theory in Section 2 begins with a setting in which the dimension, p, of X, is fixed and $n \to \infty$ (see Theorem 1). For ordinary semi-supervised learning, the asymptotic results are of nontrivial interest whenever $\liminf_{n\to\infty}(m_n/n) > 0$ [see Theorem 3(i)]. We then formulate and prove asymptotic results in the setting where p also grows with n. In general, these results require the assumption that $p = o(\sqrt{n})$ [see Theorems 2 and 3(ii)].

In Section 3, we propose a methodology for improving the results of Section 2 by introducing additional covariates as functions of those given in the original problem. We show the proposed estimator achieves an oracle rate asymptotically. This can be viewed as a nonparametric regression estimation procedure.

There are results in the sample-survey literature that are qualitatively related to what we propose. The earliest citation we are aware of is Cochran (1953), Chapter 7, for sample survey. See also Deng and Wu (1987) and more recently Lohr (2009), Chapter 3.2. In these references, one collects a finite sample, without replacement, from a (large) finite population. There is a response Y and a single, real covariate, X. The distribution of X within the finite population is known. The sample-survey target of estimation is the mean of Y within the full population. In the case in which the size of this population is infinitely large, sampling without replacement and sampling with replacement are indistinguishable. In that case, the results from this sampling theory literature coincide with our results for the ideal semi-supervised scenario with $p = 1$, both in terms of the proposed estimator and its asymptotic variance. Our work also relates to the control variates in Monte Carlo simulation [Bratley, Fox and Schrage (1987), Fishman (1996), Hickernell, Lemieux and Owen (2005)]. Suppose one is interested in evaluating the integral $\int_\Omega f(x)\,dx$, where f is a integrable function and Ω is a subset in the Euclidean space. The regular Monte Carlo estimator is $\frac{1}{n}\sum_{k=1}^{n} f(X_i)$, if $X_1, \ldots, X_n$ are i.i.d. uniform samples from Ω. One can further sharpen the estimator if one or more control variates $\{h_1(x), \ldots, h_p(x)\}$ and their integrals $\{\int_\Omega h_1(x)\,dx, \ldots, \int_\Omega h_p(x)\,dx\}$ are available a priori. From this perspective, the results from control variates Monte Carlo can be viewed as a special case in the ideal semi-supervised and noiseless

response setting, that is, $\mathbb{E}X$ is known and $\mathrm{Var}(Y|X) = 0$. Otherwise, the sample-survey and Monte Carlo theory results differ from those within our formulation, although there is a conceptual relationship. In particular, the theoretical population mean that is our target is different from the finite population mean that is the target of the sample-survey methods. In addition, we allow both the noisy response and $p > 1$, and as noted above, we also have asymptotic results for p growing with n. Most notably, our formulation includes the possibility of semi-supervised learning. We believe it should be possible, and sometimes of practical interest, to include semi-supervised sampling within a sampling survey and Monte Carlo simulation framework, but we do not do so in the present treatment.

Remarks at the end of Section 3 discuss in some detail the relation of our proposal to results in the semiparametric efficiency literature. In brief, it is known that $\bar{\mathbf{Y}}$ is not asymptotically semiparametric efficient; see Hasminskii and Ibragimov (1983) and Bickel, Ritov and Wellner (1991) for an asymptotically efficient estimator in the case of ideal semi-supervision. Chakrabortty and Cai (2018) deal with ideal semi-supervision and situations that are asymptotically equivalent to the ideal situation. They propose an estimator that is asymptotically efficient in this setting under mild regularity conditions. For situations in which there are many covariates, their estimator may not perform well in practice, and they propose a number of alternative estimators.

Our current primary objective is rather different. We describe simply expressed, easily implemented, effective improvements on $\bar{\mathbf{Y}}$. Our basic estimator asymptotically improves on $\bar{\mathbf{Y}}$, but is not asymptotically efficient. Virtually no regularity conditions are imposed for the asymptotic improvement in distribution. (Asymptotic improvement in quadratic risk requires a little more care.) Because of their simple form as well as the nature of our proofs it is heuristically clear that with finite samples our estimators usually improve on $\bar{\mathbf{Y}}$ even for quite moderate sample sizes. This is seen in the simulations reported in Tables 1 and 2. The series estimator we propose in Section 3 is semiparametric efficient under regularity conditions. (See Remarks 5 and 6.) But this is not a primary focus of our paper, so we do not concentrate on stating that asymptotic efficiency under the weakest possible conditions.

The rest of the paper is organized as follows. We introduce the fixed covariate procedures in Section 2. Specifically, ideal semi-supervised learning and ordinary semi-supervised learning are considered respectively in Sections 2.2 and 2.3, where we analyze the asymptotic properties for both estimators. We further give the ℓ_2-risk upper bounds for the two proposed estimators in Section 2.4. We extend the analysis in Section 3 to the regression model, where we show the proposed procedure achieves an oracle rate asymptotically. Simulation results are reported in Section 4. Applications to the estimation of average treatment effect is discussed in Section 5.1, and Section 5.2 describes a real data illustration involving estimation of the homeless population in a geographical region. The proofs of the main theorems are given in Section 6 and additional technical results are proved in the Supplementary Material [Zhang, Brown and Cai (2019)].

2. Procedures. We propose in this section a least squares estimator for the population mean in the semi-supervised inference framework. To better characterize the problem, we begin with a brief introduction of the random design regression model. More details of the model can be found in, for example, Buja et al. (2014, 2016).

2.1. *A random design regression model.* Let $(Y, X) \sim P$ represent the population response and predictors. Assume all second moments are finite. Denote $\vec{X} = (1, X^\top)^\top \in \mathbb{R}^{p+1}$ as the predictor with intercept. The following is a linear analysis, even though no corresponding linearity assumption is made about the true distribution P of (X, Y).

Some notation and definitions are needed. Let

$$\beta = \arg\min_{\gamma \in \mathbb{R}^{p+1}} \mathbb{E}(Y - \vec{X}^\top \gamma)^2$$

be the *population slopes*, and $\delta = Y - \beta^\top \vec{X}$ is called the *total deviation*. We also denote

$$\begin{aligned} &\tau^2 := \mathbb{E}\delta^2, \qquad \mu := \mathbb{E}X \in \mathbb{R}^p, \qquad \vec{\mu} := \mathbb{E}\vec{X} = \big(1, \mu^\top\big)^\top, \\ &\vec{\Xi} := \mathbb{E}\vec{X}\vec{X}^\top, \qquad \Sigma := \mathbb{E}(X - \mu)(X - \mu)^\top. \end{aligned} \tag{2.1}$$

It should be noted that under our general model, there is no independence assumption between X and δ, and $\mathbb{E}(\delta|X)$ is not necessarily zero. This is different from classical regression literature.

For sample of observations $(Y_k, X_{k1}, X_{k2}, \ldots, X_{kp}) \overset{\text{i.i.d.}}{\sim} P$, $k = 1, \ldots, n$, let $\vec{X}_i = (1, \vec{X}_i^\top)^\top$ and denote the design matrix $\vec{\mathbf{X}} \in \mathbb{R}^{n \times (p+1)}$ as follows:

$$\vec{\mathbf{X}} := \begin{bmatrix} \vec{X}_1^\top \\ \cdots \\ \cdots \\ \vec{X}_n^\top \end{bmatrix} := \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}.$$

In our notation, $\vec{\cdot}$ means that the vector/matrix contains the intercept term; boldface indicates that the symbol is related to a multiple sample of observations. Meanwhile, denote the sample response and deviation as $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ and $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)^\top$. Now $\mathbf{Y}$ and $\mathbf{X}$ are connected by a regression model:

$$\mathbf{Y} = \vec{\mathbf{X}}\beta + \boldsymbol{\delta}, \quad \text{and} \quad Y_k = \vec{X}_k^\top \beta + \delta_k, \qquad k = 1, \ldots, n. \tag{2.2}$$

Let $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_{p+1})^\top$ be the usual least squares estimator, that is, $\hat{\beta} = (\vec{\mathbf{X}}^\top \vec{\mathbf{X}})^{-1} \vec{\mathbf{X}}^\top \mathbf{Y}$. β and $\hat{\beta}$ can be further split into two parts,

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_{(2)} \end{bmatrix}, \qquad \hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_{(2)} \end{bmatrix}, \qquad \beta_1, \hat{\beta}_1 \in \mathbb{R}, \qquad \beta_{(2)}, \hat{\beta}_{(2)} \in \mathbb{R}^p. \tag{2.3}$$

β_1, $\hat{\beta}_1$ and $\beta_{(2)}$, $\hat{\beta}_{(2)}$ play different roles in the analysis as we will see later. The ℓ_2 risk of the sample average $\bar{\mathbf{Y}}$ about the population mean $\theta = \mathbb{E}Y$ has the following decomposition.

PROPOSITION 1. *$\bar{\mathbf{Y}}$ is an unbiased estimator of θ and*

$$n\mathbb{E}(\bar{\mathbf{Y}} - \theta)^2 = n\,\mathrm{Var}(\bar{\mathbf{Y}}) = \tau^2 + \beta_{(2)}^\top \Sigma \beta_{(2)}. \tag{2.4}$$

From (2.4), we can see that as long as $\beta_{(2)} \neq 0$, that is, there is a significant linear relationship between Y and X, then the risk of $\bar{\mathbf{Y}}$ will be significantly greater than τ^2.

In the next two subsections, we discuss separately under the ideal semi-supervised setting and the ordinary semi-supervised setting.

2.2. *Improved estimator under the ideal semi-supervised setting.* We first consider the ideal setting where there are infinitely many unlabeled samples, or equivalently P_X is known. To improve $\bar{\mathbf{Y}}$, we propose the *least squares estimator*,

$$\hat{\theta}_{\mathrm{LS}} := \vec{\mu}^\top \hat{\beta} = \hat{\beta}_1 + \mu^\top \hat{\beta}_{(2)} = \bar{\mathbf{Y}} - \hat{\beta}_{(2)}^\top (\bar{\mathbf{X}} - \mu), \tag{2.5}$$

where $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_{(2)}^\top)^\top$ is the usual least square estimator.

When $(Y_i, X_i) \overset{\text{i.i.d.}}{\sim} P$ with no specific assumptions imposed on the relationship between Y_i and X_i, the following theorem provides the asymptotic distribution of the least squares estimator under the minimal conditions that $[Y, X]$ have finite second moments, $\vec{\Xi} = \mathbb{E}\vec{X}\vec{X}^\top$ be nonsingular and $\tau^2 = \mathbb{E}\delta^2 > 0$. In addition, a Berry–Esseen bound is given under the finite fourth moment condition.

THEOREM 1 (Asymptotic distribution of $\hat{\theta}_{\mathrm{LS}}$, fixed p). *Let $(Y_1, X_1), \ldots, (Y_n, X_n)$ be i.i.d. copies from P, and assume that $[Y, X]$ has finite second moments, $\vec{\Xi}$ is nonsingular and $\tau^2 > 0$. Then, under the setting that P is fixed and $n \to \infty$,*

$$\frac{\hat{\theta}_{\mathrm{LS}} - \theta}{\tau/\sqrt{n}} \xrightarrow{d} N(0, 1), \tag{2.6}$$

and

$$MSE/\tau^2 \xrightarrow{d} 1 \qquad \text{where } MSE := \frac{\sum_{i=1}^n (Y_i - \vec{X}_i^\top \hat{\beta})^2}{n - p - 1}. \tag{2.7}$$

Denote the cumulative distribution functions of $\frac{\hat{\theta}_{\mathrm{LS}} - \theta}{\tau/\sqrt{n}}$ and the standard normal variable by F_n and Φ, respectively. If P has finite fourth moment, then we further have

$$\left|F_n(x) - \Phi(x)\right| \leq Cn^{-1/4},$$

where C is a constant not depending on n.

In the more general setting where $P = P_{n,p}$ varies and $p = p_n$ grows, we need stronger conditions to analyze the asymptotic behavior of $\hat{\theta}_{\mathrm{LS}}$. Recall $\mathbb{E}X = \mu$, $\mathbb{E}(X-\mu)(X-\mu)^\top = \Sigma$, we consider the standardization of X as

$$Z \in \mathbb{R}^p, \qquad Z = \Sigma^{-1/2}(X-\mu). \tag{2.8}$$

Clearly, $\mathbb{E}Z = 0$, $\mathbb{E}ZZ^\top = I_p$. For this setting, we assume that Z, δ satisfy the following moment conditions for constants M_1, M_2, M_3:

$$\text{for some } \kappa > 0, \qquad \frac{\mathbb{E}\delta^{2+2\kappa}}{(\mathbb{E}\delta^2)^{1+\kappa}} \le M_1; \tag{2.9}$$

$$\forall v \in \mathbb{R}^p, \qquad \mathbb{E}\big|\langle v, Z\rangle\big|^{2+\kappa} \le M_2; \tag{2.10}$$

$$\frac{\mathbb{E}(\|Z\|_2^2\delta^2)}{(\mathbb{E}\|Z\|_2^2)\cdot(\mathbb{E}\delta^2)} \le M_3. \tag{2.11}$$

THEOREM 2 (Asymptotic result, growing p). *Let $(Y_1, X_1), \ldots, (Y_n, X_n)$ be i.i.d. copies from $P = P_{n,p}$, $p = p_n = o(\sqrt{n})$. Assume that the matrix of the second moments of X exists and is nonsingular and the standardized random variable Z given in* (2.8) *satisfies* (2.9), (2.10) *and* (2.11), *then the asymptotic behavior results* (2.6) *and* (2.7) *still hold.*

Based on Theorems 1 and 2, we can construct the asymptotic $(1-\alpha)$-level confidence interval for θ as

$$\left[\hat{\theta}_{\mathrm{LS}} - z_{1-\alpha/2}\sqrt{\frac{MSE}{n}}, \hat{\theta}_{\mathrm{LS}} + z_{1-\alpha/2}\sqrt{\frac{MSE}{n}}\right]. \tag{2.12}$$

REMARK 1. It is not difficult to see that, under the setting in Theorem 2,

$$MSE \xrightarrow{d} \tau^2, \qquad \hat{\sigma}_Y^2 \xrightarrow{d} \mathrm{Var}(Y) = \tau^2 + \beta_{(2)}^\top \Sigma \beta_{(2)}.$$

Then the traditional z-interval for the mean of Y,

$$\left[\bar{\mathbf{Y}} - z_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}_Y^2}{n}}, \bar{\mathbf{Y}} + z_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}_Y^2}{n}}\right], \tag{2.13}$$

is asymptotically longer than (2.12), which implies that the proposed least squares estimator is asymptotically more accurate than the sample mean.

2.3. *Improved estimator under the ordinary semi-supervised inference setting.* In the last section, we discussed the estimation of θ based on n full observations Y_k, X_k, $k = 1, \ldots, n$ with infinitely many unlabeled samples $\{X_k, k = n+1, \ldots\}$ (or equivalently with known marginal distribution P_X). However, having P_X known is rare in practice. A more realistic practical setting would assume

that distribution P_X is unknown and we only have finitely many i.i.d. samples $(X_{i+1}, X_{i+2}, \ldots, X_{i+m})$ without corresponding Y. This problem relates to the one in previous section since we are able to obtain partial information of P_X from the additional unlabeled samples.

When μ or $\vec{\mu}$ is unknown, we estimate by

$$\hat{\mu} = \frac{1}{n+m}\sum_{k=1}^{n+m} X_k, \qquad \hat{\vec{\mu}} = (1, \hat{\mu}^\top)^\top. \tag{2.14}$$

Recall that $\hat{\beta} = (\hat{\beta}_1, \beta_{(2)}^\top)^\top$ is the ordinary least squares estimator. Now, we propose the *semi-supervised least squares estimator* $\hat{\theta}_{\mathrm{SSLS}}$,

$$\hat{\theta}_{\mathrm{SSLS}} = \hat{\vec{\mu}}^\top \hat{\beta} = \bar{\mathbf{Y}} - \hat{\beta}_{(2)}^\top \left(\frac{\sum_{i=1}^{n} X_i}{n} - \frac{\sum_{i=1}^{n+m} X_i}{n+m} \right). \tag{2.15}$$

$\hat{\theta}_{\mathrm{SSLS}}$ has the following properties:

- when $m = \infty$, $\hat{\vec{\mu}} = \vec{\mu}$. Then $\hat{\theta}_{\mathrm{SSLS}}$ exactly equals $\hat{\theta}_{\mathrm{LS}}$ in (2.5);
- when $m = 0$, $\hat{\theta}_{\mathrm{SSLS}}$ exactly equals $\bar{\mathbf{Y}}$. As there are no additional samples of X so that no extra information for P_X is available, it is natural to use $\bar{\mathbf{Y}}$ to estimate θ.
- In the last term of (2.15), it is important to use $\frac{1}{n+m}\sum_{i=1}^{n+m} X_i$ rather than $\frac{1}{m}\sum_{i=1}^{m} X_i$, in spite of the fact that the latter might seem more natural because it is independent of the term $\frac{\sum_{i=1}^{n} X_i}{n}$ that precedes it.

Under the same conditions as Theorems 1, 2, we can show the following asymptotic results for $\hat{\theta}_{\mathrm{SSLS}}$, which relates to the ordinary semi-supervised setting described in the Introduction. The labeled sample size $n \to \infty$, the unlabeled sample size is $m = m_n \geq 0$ and the distribution P is fixed (but unknown) which, in particular, implies that p is a fixed dimension, not dependent on n. Let

$$\nu^2 = \sqrt{\tau^2 + \frac{n}{n+m}\beta_{(2)}^\top \Sigma \beta_{(2)}}.$$

THEOREM 3 (Asymptotic distribution of $\hat{\theta}_{\mathrm{SSLS}}$, fixed p). *Let $(Y_1, X_1), \ldots, (Y_n, X_n)$ be i.i.d. labeled samples from P, and let $X_{n+1}, \ldots, X_{n+m}$ be m additional unlabeled independent samples from P_X. Suppose $\vec{\Xi}$ is nonsingular and $\tau^2 > 0$. If P is fixed and $n \to \infty$, then*

$$\frac{\sqrt{n}(\hat{\theta}_{\mathrm{SSLS}} - \theta)}{\nu} \xrightarrow{d} N(0, 1), \tag{2.16}$$

and

$$\frac{\hat{\nu}^2}{\nu^2} \xrightarrow{d} 1, \tag{2.17}$$

where $\hat{\nu}^2 = \frac{m}{m+n} MSE + \frac{n}{m+n}\hat{\sigma}_Y^2$ with $MSE = \frac{1}{n-p-1}\sum_{k=1}^{n}(Y_i - \vec{X}_k^\top \hat{\beta})^2$ and $\hat{\sigma}_Y^2 = \frac{1}{n-1}\sum_{k=1}^{n}(Y_i - \bar{\mathbf{Y}})^2$.

The following statement refers to a setting in which $P = P_n$ and $p = p_n$ may depend on n as $n \to \infty$. Consequently, $\vec{\Xi} = \vec{\Xi}_n$, $\Sigma = \Sigma_n$ and $Z = Z_n$ [defined at (2.8)] may also depend on n.

THEOREM 4 (Asymptotic distribution of $\hat{\theta}_{\mathrm{SSLS}}$, growing p). *Let $n \to \infty$, $P = P_n$, and $p = p_n = o(\sqrt{n})$. Suppose $\vec{\Xi}_n$ is nonsingular, $\tau_n^2 > 0$ and the standardized random variable Z satisfies (2.9), (2.10) and (2.11). Then (2.16) and (2.17) hold.*

We can obtain asymptotic confidence interval for θ based on Theorems 3 or 4.

COROLLARY 1. *The $(1-\alpha)$-level asymptotic confidence interval for θ can be written as*

$$\left[\hat{\theta}_{\mathrm{SSLS}} - z_{1-\alpha/2}\frac{\hat{\nu}}{\sqrt{n}}, \hat{\theta}_{\mathrm{SSLS}} + z_{1-\alpha/2}\frac{\hat{\nu}}{\sqrt{n}}\right]. \tag{2.18}$$

Since MSE $\le \hat{\sigma}_Y^2$ asymptotically (with equality only when $\beta_{(2)} = 0$), it follows that when $\beta_{(2)} \neq 0$ the asymptotic CI in (2.18) is shorter than the traditional sample-mean-based CI (2.13).

2.4. *ℓ_2 risk for the proposed estimators.* In this subsection, we analyze the ℓ_2 risk for both $\hat{\theta}_{\mathrm{LS}}$ and $\hat{\theta}_{\mathrm{SSLS}}$. Since the calculation of the proposed estimators involves the unstable process of inverting the Gram matrix $\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}$, for the merely theoretical purpose of obtaining the ℓ_2 risks we again consider the refinement

$$\hat{\theta}_{\mathrm{LS}}^1 := \mathrm{Trun}_{\mathbf{Y}}(\hat{\theta}_{\mathrm{LS}}), \quad \text{and} \quad \hat{\theta}_{\mathrm{SSLS}}^1 := \mathrm{Trun}_{\mathbf{Y}}(\hat{\theta}_{\mathrm{SSLS}}), \tag{2.19}$$

where

$$\mathrm{Trun}_{\mathbf{Y}}(x) = \begin{cases} (n+1)y_{\max} - ny_{\min} & \text{if } x > (n+1)y_{\max} - ny_{\min}, \\ x & \text{if } \left|x - \dfrac{y_{\max}+y_{\min}}{2}\right| \\ & \qquad \le \left(n+\dfrac{1}{2}\right)(y_{\max}-y_{\min}), \\ (n+1)y_{\min} - ny_{\max} & \text{if } x < (n+1)y_{\min} - ny_{\max}, \end{cases} \tag{2.20}$$

$y_{\max} = \max_{1\le k\le n} Y_k$, $y_{\min} = \min_{1\le k\le n} Y_k$. We emphasize that this refinement is mainly for theoretical reasons and is often not necessary in practice.

The regularization assumptions we need for analyzing the ℓ_2 risk are formally stated as below.

1. (*Moment conditions on δ.*) There exists constant $M_4 > 0$ such that

$$\mathbb{E}\delta^4 = \mathbb{E}\delta_n^4 \le M_4. \tag{2.21}$$

2. (*Sub-Gaussian condition.*) Let $Z = Z_n$ be the standardization of $X = X_n$,

$$Z_n \in \mathbb{R}^p, \qquad Z_n = \Sigma_n^{-1/2}(X_n - \mu_n), \qquad \Sigma_n = \mathbb{E}(X_n - \mu_n)(X_n - \mu_n)^\top.$$

Assume Z_n satisfies

$$\forall u \in \{u \in \mathbb{R}^{p+1} : \|u\|_2 = 1\}, \qquad \|u^\top Z_n\|_{\psi_2} \le M_5 \tag{2.22}$$

for constant $M_t > 0$. Here $\|\cdot\|_{\psi_2}$ is defined as $\|x\|_{\psi_2} = \sup_{q \ge 1} q^{-1/2}(\mathbb{E}|x|^q)^{1/q}$ for any random variable x.

$2'$. (*Bounded condition.*) The standardization Z_n satisfies

$$\|Z_n\|_\infty \le M_5 \qquad \text{almost surely.} \tag{2.23}$$

[If the dimension p remains bounded as $n \to \infty$, then (2.23) implies (2.22). However, if p increases without bound, as in Section 3, then there are rather unusual examples in which (2.23) holds but (2.22) does not.]

We also note $\Sigma_{\delta 1} = \mathbb{E}(X - \mu)\delta(X - \mu)^\top$, $\Sigma_{\delta 2} = \mathbb{E}(X - \mu)\delta^2(X - \mu)^\top$. Under the regularization assumptions above, we provide the ℓ_2 risks for $\hat\theta^1_{\mathrm{LS}}$ and $\hat\theta^1_{\mathrm{SSLS}}$, respectively, in the next two theorems.

THEOREM 5 (ℓ_2 risk of $\hat\theta^1_{\mathrm{LS}}$). *Let $(Y_1, X_1), \dots, (Y_n, X_n)$ be i.i.d. copies from P_n. Assume Assumption 1 holds. In addition, either Assumptions 2 or $2'$ hold, $p = p_n = o(\sqrt{n})$. Recall $\tau^2 = \tau_n^2 = \mathbb{E}(Y - \vec{X}\beta)^2$ depends on n. Then we have the following estimate for the risk of $\hat\theta^1_{\mathrm{LS}}$:*

$$n\mathbb{E}\big(\hat\theta^1_{\mathrm{LS}} - \theta\big)^2 = \tau_n^2 + s_n, \tag{2.24}$$

where

$$s_n = \frac{p^2}{n} A_{n,p} + \frac{p^2}{n^{5/4}} B_{n,p}, \qquad \max\big(|A_{n,p}|, |B_{n,p}|\big) \le C \tag{2.25}$$

for a constant C that depends on M_0, M_1 and M_2. The formula for $A_{n,p}$ is

$$\begin{aligned} A_{n,p} = \frac{1}{p^2}\Big(&\big[\mathrm{tr}\big(\Sigma^{-1}\Sigma_{\delta 1}\big)\big]^2 + 3\big\|\Sigma^{-1}\Sigma_{\delta 1}\big\|_F^2 - \mathrm{tr}\big(\Sigma^{-1}\Sigma_{\delta 2}\big) \\ &+ 2\mathbb{E}\big(\delta^2(X - \mu)^\top\big) \\ &\times \mathbb{E}\big(\Sigma^{-1}(X - \mu)(X - \mu)^\top\Sigma^{-1}(X - \mu)\big) + 2p\tau^2\Big). \end{aligned} \tag{2.26}$$

THEOREM 6 (ℓ_2 risk of $\hat\theta^1_{\mathrm{SSLS}}$). *Let $(Y_1, X_1), \dots, (Y_n, X_n)$ be i.i.d. labeled samples from P, and let $X_{n+1}, \dots, X_{n+m}$ be additional m unlabeled independent samples from P_X. Assume Assumption 1 holds. In addition, either Assumptions 2 or $2'$ hold, $p = o(\sqrt{n})$. We have the following estimate of the risk for $\hat\theta^1_{\mathrm{SSLS}}$:*

$$n\mathbb{E}\big(\hat\theta^1_{\mathrm{SSLS}} - \theta\big)^2 = \tau_n^2 + \frac{n}{n+m}\beta_{(2),n}^\top \Sigma_n \beta_{(2),n} + s_{n,m}, \tag{2.27}$$

where

$$|s_{n,m}| \le \frac{Cp^2}{n}, \tag{2.28}$$

for constant C only depends on M_0, M_1 and M_2 in Assumptions (2.21)–(2.23).

REMARK 2. Comparing Proposition 1, Theorems 5 and 6, we can see as long as

$$\beta_{(2),n}^\top \Sigma_n \beta_{(2),n} > 0,$$

that is, $\mathbb{E}(Y|X)$ has nonzero correlation with X, $\hat{\theta}_{\mathrm{LS}}^1$ and $\hat{\theta}_{\mathrm{SSLS}}^1$ outperform $\bar{\mathbf{Y}}$ asymptotically in ℓ_2-risk. We can also see the risk of $\hat{\theta}_{\mathrm{SSLS}}$ is approximately a linear combination of $\bar{\mathbf{Y}}$ and $\hat{\theta}_{\mathrm{LS}}$ with weight based on m and n,

$$\mathbb{E}\big(\hat{\theta}_{\mathrm{SSLS}}^1 - \theta\big)^2 \approx \frac{n}{n+m}\mathbb{E}(\bar{\mathbf{Y}} - \theta)^2 + \frac{m}{m+n}\mathbb{E}\big(\hat{\theta}_{\mathrm{LS}}^1 - \theta\big)^2.$$

REMARK 3. The proposed $\hat{\theta}_{\mathrm{SSLS}}$ is a direct and simple estimator that achieves good finite sample performance for both estimation and confidence interval. An improved semi-supervised least square estimator that achieves semiparametric efficiency will be further introduced and discussed later in Section 3.2.

REMARK 4 (Gaussian design). Theorems 5 and 6 only provide upper bound of the ℓ_2 risks since only moment conditions on the distribution of Y, X are assumed. In fact, under Gaussian design of Y, X, we can obtain an exact expression for the ℓ_2-risk of both $\hat{\theta}_{\mathrm{LS}}$ and $\hat{\theta}_{\mathrm{SSLS}}$. It is noteworthy that the truncation refinement is not necessary for both estimators under Gaussian design. All results are nonasymptotic.

PROPOSITION 2. *Assume $X \sim N_p(\mu, \Sigma)$ and $Y|X \sim N_p(X\beta, \tau^2 I)$, where Σ is nonsingular. If $\{Y_k, X_k\}_{k=1}^n$ are n i.i.d. copies, then*

$$n\mathbb{E}(\hat{\theta}_{\mathrm{LS}} - \theta)^2 = \tau^2 + \frac{p\tau^2}{(n-p-2)}. \tag{2.29}$$

If we further have m additional unlabeled samples $\{X_k\}_{k=n+1}^{n+m}$, we also have

$$n\mathbb{E}(\hat{\theta}_{\mathrm{SSLS}} - \theta)^2 = \tau^2 + \frac{m}{n+m}\frac{p\tau^2}{n-p-2} + \frac{n}{n+m}\beta_{(2)}^\top \Sigma \beta_{(2)}. \tag{2.30}$$

The result in Proposition 2 matches with the general expression of (2.24) and (2.26) as $\frac{p\tau^2}{(n-p-2)} = \frac{p\tau^2}{n} + O(\frac{p^2}{n^2})$ if $p = o(\sqrt{n})$. By comparing (2.29), (2.30), we can also see

$$n\mathbb{E}(\hat{\theta}_{\mathrm{SSLS}} - \theta)^2 = \frac{n}{n+m} n\mathbb{E}(\bar{\mathbf{Y}} - \theta)^2 + \frac{m}{n+m} n\mathbb{E}(\hat{\theta}_{\mathrm{LS}} - \theta)^2.$$

3. Further improvements—oracle optimality. In the previous sections, we proposed and analyzed $\hat{\theta}_{\mathrm{LS}}$ and $\hat{\theta}_{\mathrm{SSLS}}$ under the semi-supervised learning settings. These estimators are based on linear regression and best linear approximation of Y by X. We consider further improvement in this section. Before we illustrate how the improved estimator works, it is helpful to take a look at the oracle risk for estimating the mean $\theta = \mathbb{E}Y$, which can serve as a benchmark for the performance of the improved estimator.

3.1. *Oracle estimator and risk.* Define $\xi(X) = \mathbb{E}_P(Y|X)$ as the response surface and suppose

$$\xi(x) = \xi_0(x) + c + o(1/\sqrt{n})$$

for some unknown constant c. Here, the $o(1/\sqrt{n})$ term is uniform in X and $\xi_0(x)$ represents any approximately "location-free shape" of $\xi_0(x)$ in the sense that $\xi(x) - \xi_0(x)$ is nearly a constant: $|\xi(x) - \xi_0(x) - c| \le o(1/\sqrt{n})$. Given samples $\{(Y_k, X_k)\}_{k=1}^n$, our goal is to estimate $\mathbb{E}Y = \theta$. Now assume an oracle has knowledge of $\xi_0(x)$, but not of $\theta = \mathbb{E}(Y)$, c, nor the distribution of $Y - \xi_0(X)$. In this case, the model can be written as

$$\begin{aligned} Y_k - \xi_0(X_k) &= c + \varepsilon_k, \qquad k = 1, \dots, n \qquad \text{where } \mathbb{E}\varepsilon_k = o(1/\sqrt{n}); \\ \theta &= \mathbb{E}\xi_0(X) + c + o(1/\sqrt{n}). \end{aligned} \tag{3.1}$$

Under the ideal semi-supervised setting, P_X, ξ_0 and $\mathbb{E}\xi_0(X)$ are known. To estimate θ, the natural idea is to use the following estimator:

$$\hat{\theta}^* = \bar{\mathbf{Y}} - \bar{\xi}_0 + \mathbb{E}\xi_0(X) = \frac{1}{n}\sum_{k=1}^n \big(Y_k - \xi_0(X_k)\big) + \mathbb{E}\xi_0(X). \tag{3.2}$$

Consider a sample $\{Y_i : i = 1, \dots, n\}$ with no covariates. It is known that $\bar{\mathbf{Y}}$ is an asymptotically efficient estimator of $\mathbb{E}(Y)$, locally on a neighborhood of the true distribution of Y. In much the same way, $\bar{Y} - \overline{\xi_0(X)}$ is an asymptotically efficient estimator of $\mathbb{E}(Y - \xi_0(X))$, even when the ancillary statistics, $\{X_i\}$, are also observed. For details, see the proof of Proposition 3 in the Supplementary Material [Zhang, Brown and Cai (2019)]. Thus, $\hat{\theta}^*$ is an asymptotically efficient estimator of $\mathbb{E}(Y) = \mathbb{E}(Y - \xi_0(X)) + E(\xi_0(X))$, and

$$\begin{aligned} n\mathbb{E}\big(\hat{\theta}^* - \theta\big)^2 &= n\,\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n \big(Y_i - \xi_0(X_i)\big)\right) = \mathrm{Var}\big(Y_i - \xi(X_i)\big) \\ &= \mathbb{E}_X\big(\mathbb{E}_Y\big(Y - \xi(X)|X\big)^2\big) := \sigma^2. \end{aligned} \tag{3.3}$$

This defines the oracle risk for population mean estimation under the ideal semi-supervised setting as $\sigma^2 = \mathbb{E}_X(\mathbb{E}_Y(Y - \mathbb{E}(Y|X))^2)$.

For the ordinary semi-supervised setting, where P_X is unknown but m additional unlabeled samples $\{X_k\}_{k=n+1}^{n+m}$ are available, we propose the semi-supervised oracle estimator as

$$\hat{\theta}_{\mathrm{ss}}^{*} = \bar{\mathbf{Y}} - \frac{1}{n}\sum_{k=1}^{n}\xi_0(X_k) + \frac{1}{n+m}\sum_{k=1}^{n+m}\xi_0(X_k).$$

Then one can calculate that

$$n\mathbb{E}\big(\hat{\theta}_{ss}^{*} - \theta\big)^2 = \sigma^2 + \frac{n}{n+m}\operatorname{Var}_{P_X}\big(\xi(X)\big). \tag{3.4}$$

The detailed calculation of (3.4) is provided in the Supplementary Material [Zhang, Brown and Cai (2019)].

The preceding motivation for σ^2 and $\sigma^2 + \frac{n}{n+m}\operatorname{Var}_{P_X}(\xi(X))$ as the oracle risks are partly heuristic, but it corresponds to formal minimax statements, as in the following Propositions 3 and 4. Particularly, Proposition 3 proposes the general lower bounds for both ideal and semiparametric settings. Proposition 4 develops the asymptotic lower bound on a more restrictive set, that is, the least favorable one-dimensional family of conditional means of any specific distribution P, under ideal semi-supervision. Proposition 4 further yields an asymptotic semiparametric efficiency result as we will illustrate later in Remark 6.

PROPOSITION 3 (Oracle lower bound). *Let $\sigma^2 > 0$, $\xi_0 : \mathbb{R}^p \to \mathbb{R}$ be a measurable function, and P_X be a p-dimensional distribution of X. Suppose*

$$\mathcal{P}_{\xi_0(\cdot),P_X,\sigma^2} = \big\{P : P_X \text{ is the marginal distribution of } P,\ \mathbb{E}_P(Y|X=x) = \xi_0(x) + c, \sigma^2 = \mathbb{E}_X\big(\mathbb{E}_Y\big(Y - \mathbb{E}(Y|X)\big)^2\big)\big\}.$$

Then based on observations $\{Y_i, X_i\}_{i=1}^{n}$ and known marginal distribution P_X,

$$\inf_{\tilde{\theta}}\sup_{P\in\mathcal{P}_{P_X,\xi_0,\sigma^2}}\big[\mathbb{E}_P\big(n(\tilde{\theta}-\theta)^2\big)\big] \geq \sigma^2. \tag{3.5}$$

Let $\sigma^2, \sigma_\xi^2 > 0$, $\xi_0(X) : \mathbb{R}^p \to \mathbb{R}$ be a linear function,

$$\mathcal{P}^{\mathrm{ss}}_{\xi_0,\sigma_\xi^2,\sigma^2} = \big\{P : \xi_0(x) = \mathbb{E}(Y|X=x) - c, \sigma_\xi^2 = \operatorname{Var}\big(\xi(X)\big), \sigma^2 = \mathbb{E}_X\big(\mathbb{E}_Y\big(Y - \mathbb{E}(Y|X)\big)^2\big)\big\},$$

based on observations $\{Y_i, X_i\}_{i=1}^{n}$ and $\{X_i\}_{i=n+1}^{n+m}$,

$$\inf_{\tilde{\theta}}\sup_{P\in\mathcal{P}^{\mathrm{ss}}_{\xi_0,\sigma_\xi^2,\sigma^2}}\big[\mathbb{E}_P\big(n(\tilde{\theta}-\theta)^2\big)\big] \geq \sigma^2 + \frac{n}{n+m}\sigma_\xi^2. \tag{3.6}$$

PROPOSITION 4 (Asymptotic oracle lower bound for ideal semi-supervised setting). *Let* $\sigma^2 = \mathbb{E}_P[(Y - \xi(X))^2] > 0$ *and*

$$\mathcal{P}_K = \{P : P_X = P_X^0, |\mathbb{E}_P(Y|X) - \xi^0(X)| \le K\sigma^2(X)/\sigma^2 + 1/(K\sqrt{n}),$$
$$|\mathbb{E}_P[(Y - \xi(X))^2] - \sigma^2| < 1/K, |c| < K\}.$$

Then

$$\lim_{K\to\infty} \liminf_{n\to\infty} \inf_{\hat\theta} \sup_{P\in\mathcal{P}_k} n\mathbb{E}(\hat\theta - \mathbb{E}_P(Y))^2 \ge \sigma^2. \tag{3.7}$$

3.2. *Improved procedure.* In order to approach oracle optimality, we propose to augment the set of covariates $X_1, \dots, X_p$ with additional covariates $g_1(X), \dots, g_q(X)$. (Of course, these additional covariates need to be chosen without knowledge of ξ_0. We will discuss their choice later in this section.) In all, there are now $p^\bullet = p + q$ covariates, say

$$X^\bullet = (X_1^\bullet, \dots, X_p^\bullet, X_{p+1}^\bullet, \dots, X_{p+q}^\bullet) = (X_1, \dots, X_p, g_1(X), \dots, g_q(X)).$$

For both ideal and ordinary semi-supervision we propose to let $q = q_n$ as $n \to \infty$, and to use the estimator $\hat\theta_{\mathrm{LS}}^\bullet$ and $\hat\theta_{\mathrm{SSLS}}^\bullet$. For merely theoretical purpose of ℓ_2 risks, we consider the refinement again

$$\hat\theta_{\mathrm{LS}}^{\bullet 1} = \mathrm{Trun}_{\mathbf{Y}}(\hat\theta_{\mathrm{LS}}^\bullet) \quad \text{and} \quad \hat\theta_{\mathrm{SSLS}}^{\bullet 1} = \mathrm{Trun}_{\mathbf{Y}}(\hat\theta_{\mathrm{SSLS}}^\bullet),$$

where $\mathrm{Trun}_{\mathbf{Y}}(\cdot)$ is defined as (2.20). Apply previous theorems for asymptotic distributions and moments. For convenience of statement and proof, we assume that the support of X is compact, $\xi(X)$ is bounded and Y is sub-Gaussian. These assumptions can each be somewhat relaxed at the cost of additional technical assumptions and complications. Here is a formal statement of the result.

THEOREM 7. *Assume the support of* X *is compact,* $\xi(X) = \mathbb{E}(Y|X)$ *is bounded, and* Y *is sub-Gaussian. Consider asymptotics as* $n \to \infty$ *for the case of both ideal and ordinary semi-supervision. Assume also that either* (i) $\xi(X)$ *is continuous or* (ii) *that* P_X *is absolutely continuous with respect to Lebesgue measure on* $\{X\}$. *Let* $\{g_k(x) : k = 1, \dots\}$ *be a bounded basis for the continuous functions on* $\{X\}$ *in case* (i) *and be a bounded basis for the ordinary* ℓ_2 *Hilbert space on* $\{X\}$ *in case* (ii). *Suppose* $q_n \to \infty$ *satisfying* $q_n = o(\sqrt{n})$, *Assumption* 1 *holds, and either Assumptions* 2 *or* $2'$ *are satisfied, then*:

- *The estimator* $\hat\theta_{LS}^{\bullet 1}$ *for the problem with observations* $\{Y_i, X_{p+q_n}^\bullet : i = 1, \dots, n\}$ *asymptotically achieves the ideal oracle risk, that is,*

$$\lim_{n\to\infty} n\mathbb{E}(\hat\theta_{\mathrm{LS}}^{\bullet 1} - \theta)^2 = \sigma^2. \tag{3.8}$$

- *Now we suppose* $\lim_{n\to\infty} \frac{n}{n+m_n} = \rho$ *for some fixed value* $0 \le \rho \le 1$. *Applying the estimator* $\hat{\theta}^{\bullet}_{\mathrm{SSLS}}$ *for the problem with observations* $\{Y_i, X^{\bullet}_{p+q_n} : i = 1, \dots, n\}$ *and* $\{X^{\bullet}_i\}_{i=n+1}^{n+m_n}$. *Then*

$$\lim_{n\to\infty} n\mathbb{E}\big(\hat{\theta}^{\bullet 1}_{\mathrm{SSLS}} - \theta\big)^2 = \sigma^2 + \rho \operatorname{Var}_{P_X}\big(\xi(X)\big). \tag{3.9}$$

Finally, $\hat{\theta}^{\bullet}_{\mathrm{LS}}$ *and* $\hat{\theta}^{\bullet}_{\mathrm{SSLS}}$ *are asymptotically unbiased and normal with the corresponding variances.*

Equations (3.8) and (3.9) show that the proposed estimators asymptotically achieve the oracle values in (3.5) and (3.6). On the other hand, one could use the simpler ordinary estimators $\hat{\theta}^{\bullet}_{\mathrm{LS}}$ and $\hat{\theta}^{\bullet}_{\mathrm{SSLS}}$ in place of $\hat{\theta}^{\bullet 1}_{\mathrm{LS}}$ and $\hat{\theta}^{\bullet 1}_{\mathrm{SSLS}}$ in practice, since $\hat{\theta}^{\bullet}_{\mathrm{LS}}$ and $\hat{\theta}^{\bullet}_{\mathrm{SSLS}}$ converge in distribution with asymptotic variance as in (3.5) and (3.6).

REMARK 5. There are several results in the semiparametric regression literature [Bickel, Ritov and Wellner (1991), Bickel et al. (1998), Chakrabortty and Cai (2018), Hansen (2017), Hasminskii and Ibragimov (1983), Peng and Schick (2002), van der Vaart (2002)] that show similar aspects to our results. For example, Bickel, Ritov and Wellner (1991) discusses semiparametric inference for the joint distribution of bivariate $(Y, X) \in \mathbb{R}^2$ given known marginal distributions P_X and/or P_Y. With P_X known and P_Y unknown, this corresponds to our ideal semi-supervised setting. Their estimator is built from a suitable, binned nonparametric regression estimator of $\xi(x)$. It can be shown using comments in Section 4 of their article that their procedure will yield a semiparametric efficient estimator of $\mathbb{E}Y$ for our ideal semi-supervised problem when X is real. (Generalization to multivariate X is relatively straightforward.) Chakrabortty and Cai (2018) develop several different semiparametric efficient estimators for the population regression slopes in ideal semi-supervised semiparametric regression, or when $m/n \to \infty$. It can be shown with a little extra work that these also yield semiparametric efficient estimators of the mean of Y for such a setting. Though it shares some common features with each of these approaches our series estimator also shows some fundamental differences to any of these proposals. We remark below that our series estimator is also semiparametric efficient under suitable regularity conditions. But our emphasis remains on its directness, simplicity and the implications of this for good finite sample performance (including confidence intervals) relative to $\bar{\mathbf{Y}}$.

REMARK 6. The oracle optimality in Proposition 4 involves an asymptotically least favorable one-dimensional family of conditional means under ideal semi-supervised setting. It also places no special restriction on the conditional distribution of $Y - \xi_0(X)$. Consequently, the conditional sample mean (if a large conditional sample were available) would be the asymptotically efficient estimator of

$\xi_0(X)$. It follows that the oracle optimal rates in (3.3) is equal to the asymptotic semiparametric efficiency bound. Hence the series estimator $\hat{\theta}^{\bullet}_{\mathrm{LS}}$ of Section 3.2 is asymptotically efficient under the regularity conditions of Theorem 7. We believe Proposition 4 can be further extended to a version applying to ordinary semi-supervision and yields the corresponding semiparametric efficiency bound.

Although the preceding argument is informal, it can be made precise. Bickel, Ritov and Wellner (1991) and Chakrabortty and Cai (2018) contain more detailed, conventional arguments for estimating regression slopes in the ideal semi-supervised case, and the result for estimating $\mathbb{E}Y$ can be drawn from there via standard reasoning. Some remarks in the latter paper about MAR data can be used to extend the treatment to ordinary semi-supervision, as can a specialization of the MAR results in Graham (2011). A detailed argument for all cases can be found in Kuchibhotla (2017).

REMARK 7. Theorem 7 suggests that the number of terms in the series should be $q_n = o(\sqrt{n})$. As a crude rule of thumb, we could suggest choosing $q_n \approx n^{1/3}$. Hence, if $n = 100$, one could choose $q_n = 5$. Our estimator in a problem having such n and q is not optimal in any sense, but one can be fairly confident that it will at least be noticeably better than $\bar{\mathbf{Y}}$.

4. Simulation results. In this section, we investigate the numerical performance of the proposed estimators in various settings in terms of estimation errors and coverage probability as well as length of confidence intervals. All the simulations are repeated for 1000 times.

We analyze the linear least squares estimators $\hat{\theta}_{\mathrm{LS}}$ and $\hat{\theta}_{\mathrm{SSLS}}$ proposed in Section 2 in the following three settings:

1. (Gaussian X and quadratic ξ.) We generate the design and parameters as follows, $\mu \sim N(0, I_p)$, $\Sigma \in \mathbb{R}^{p\times p}$, $\Sigma_{ij} = I\{i = j\} + \frac{1}{2p} I\{i \neq j\}$, $\beta \sim N(0, I_{p+1})$. Then we draw i.i.d. samples $\mathbf{Y}$, $\mathbf{X}$ as

$$X_k \sim N(\mu, \Sigma), \qquad Y_k = \xi(X_k) + \varepsilon_k,$$

where

$$\xi(X_k) = \big(\|X_k\|_2^2 - p\big) + \vec{\mathbf{X}}^{\top}\beta, \qquad \varepsilon_k \sim N\big(0, 2\|X_k\|_2^2/p\big).$$

It is easy to calculate that $\theta = \mathbb{E}Y = \beta_1$ in this setting.

2. (Heavy tailed X and Y.) We randomly generate

$$\{X_{ki}\}_{1\le k\le n, 1\le i\le p} \overset{\text{i.i.d.}}{\sim} P_3, \qquad Y_k = \sum_{i=1}^{p} \big(\sin(X_{ki}) + X_{ki}\big) + 0.5\cdot\varepsilon_k, \qquad \varepsilon_k \overset{\text{i.i.d.}}{\sim} P_3,$$

where P_3 has density $f_{P_3}(x) = \frac{1}{1+|x|^3}$, $-\infty < x < \infty$. Here, the distribution P_3 has no third or higher moments. In this case, $\mu = \mathbb{E}X = 0$, $\theta = \mathbb{E}Y = 0$.

3. (Poisson X and Y.) Then we also consider a setting where

$$\{X_{ki}\}_{1\le k\le n, 1\le i\le p} \overset{\text{i.i.d.}}{\sim} \text{Poisson}(10), \qquad Y_k|X_k \overset{\text{i.i.d.}}{\sim} \text{Poisson}(10X_{k1}).$$

In this case, $\mu = \mathbb{E}X = (10, \ldots, 10)^\top \in \mathbb{R}^p$, $\theta = \mathbb{E}\mathbf{Y} = 100$.

We compare the average ℓ_2-loss of $\bar{\mathbf{Y}}$, $\hat{\theta}_{\text{LS}}$ and $\hat{\theta}_{\text{SSLS}}$ for various choices of n, p and m. The results are summarized in Table 1. The primary message to notice is that in every case, our estimator is preferable to Y-bar. An interesting aspect is even when p grows faster than $n^{1/2}$, $\hat{\theta}_{\text{LS}}$ and $\hat{\theta}_{\text{SSLS}}$ are still preferable estimators to $\bar{\mathbf{Y}}$. It is also noteworthy that although our theoretical analysis for the ℓ_2-risk focused on the refined estimators $\hat{\theta}^1_{\text{LS}}$ and $\hat{\theta}^1_{\text{SSLS}}$ with bounded or sub-Gaussian designs, the refinement and assumptions are for technical asymptotic needs, which might not be necessary in practice as we can see from these examples.

We also compute the 95% confidence interval for each setting above and list the average length and coverage probability in Table 2. It can be seen that under the

TABLE 1

Average squared loss of sample mean estimator $\bar{\mathbf{Y}}$, the least squares estimator $\hat{\theta}_{\text{LS}}$ and the semi-supervised least squares estimators $\hat{\theta}_{\text{SSLS}}$ under different values of (p, n) and various settings

		$(\hat{\theta}_{\text{SSLS}} - \theta)^2$			
(p, n)	$(\bar{\mathbf{Y}} - \theta)^2$	$m = 100$	$m = 1000$	$m = 10{,}000$	$(\hat{\theta}_{\text{LS}} - \theta)^2$
		Setting 1: Gaussian X and Quadratic ξ			
(1, 100)	0.304	0.184	0.075	0.063	0.056
(10, 100)	2.73	1.529	0.518	0.313	0.296
(50, 100)	13.397	7.961	3.967	2.988	2.868
(10, 500)	0.526	0.464	0.211	0.067	0.045
(50, 500)	2.668	2.278	1.089	0.373	0.273
(200, 500)	10.743	9.135	4.615	2.345	1.949
		Setting 2: Heavy tailed X and Y			
(1, 100)	0.732	0.410	0.244	0.196	0.188
(10, 100)	7.791	5.428	2.505	1.959	1.831
(50, 100)	107.363	47.036	17.754	14.201	13.435
(10, 500)	2.575	2.097	0.988	0.354	0.261
(50, 500)	12.569	10.481	5.619	2.342	1.780
(200, 500)	43.997	36.123	30.856	13.175	9.642
		Setting 3: Poisson X and Y			
(1, 100)	97.912	50.510	10.168	2.036	1.015
(10, 100)	98.337	50.772	10.535	2.085	1.061
(50, 100)	94.475	52.166	10.951	3.146	2.100
(10, 500)	20.062	16.765	6.890	1.104	0.186
(50, 500)	19.915	15.793	6.541	1.165	0.225
(200, 500)	20.933	17.639	7.159	1.300	0.333

TABLE 2
Average length and coverage probability (in the parenthesis) 95%-CI based on $\bar{\mathbf{Y}}$, $\hat{\theta}_{\mathrm{LS}}$ *and* $\hat{\theta}_{\mathrm{SSLS}}$ *under different values of* (p, n) *and various settings*

		$\hat{\boldsymbol{\theta}}_{\mathbf{SSLS}}$			
$(\boldsymbol{p}, \boldsymbol{n})$	$\bar{\mathbf{Y}}$	$\boldsymbol{m} = \mathbf{100}$	$\boldsymbol{m} = \mathbf{1000}$	$\boldsymbol{m} = \mathbf{10{,}000}$	$\hat{\boldsymbol{\theta}}_{\mathbf{LS}}$
		Setting 1: Gaussian X and Quadratic ξ			
(1, 100)	1.902 (0.945)	1.521 (0.954)	1.074 (0.951)	0.940 (0.939)	0.921 (0.936)
(5, 100)	4.430 (0.942)	3.301 (0.930)	1.911 (0.945)	1.467 (0.941)	1.400 (0.931)
(10, 100)	6.318 (0.952)	4.678 (0.942)	2.655 (0.937)	2.010 (0.924)	1.913 (0.916)
(1, 500)	0.845 (0.959)	0.793 (0.958)	0.608 (0.959)	0.451 (0.958)	0.413 (0.954)
(10, 500)	2.818 (0.955)	2.596 (0.959)	1.768 (0.952)	1.023 (0.949)	0.832 (0.936)
(25, 500)	4.558 (0.949)	4.194 (0.961)	2.837 (0.946)	1.606 (0.942)	1.288 (0.922)
		Setting 2: Heavy tailed X and Y			
(1, 100)	3.349 (0.961)	2.069 (0.941)	1.596 (0.939)	1.446 (0.956)	1.420 (0.962)
(5, 100)	7.332 (0.950)	4.885 (0.918)	3.384 (0.933)	2.920 (0.937)	2.847 (0.952)
(10, 100)	11.292 (0.956)	7.436 (0.921)	5.073 (0.922)	4.343 (0.943)	4.225 (0.956)
(1, 500)	1.573 (0.954)	1.205 (0.945)	0.970 (0.923)	0.773 (0.937)	0.723 (0.942)
(10, 500)	5.947 (0.957)	4.427 (0.939)	3.217 (0.916)	2.180 (0.931)	1.904 (0.953)
(25, 500)	8.582 (0.960)	7.079 (0.945)	5.197 (0.928)	3.617 (0.931)	3.229 (0.953)
		Setting 3: Poisson X and Y			
(1, 100)	39.164 (0.937)	27.831 (0.939)	12.386 (0.944)	5.506 (0.953)	3.895 (0.925)
(5, 100)	39.396 (0.947)	28.003 (0.957)	12.485 (0.933)	5.600 (0.938)	4.004 (0.930)
(10, 100)	39.143 (0.935)	27.832 (0.946)	12.443 (0.936)	5.655 (0.942)	4.105 (0.937)
(1, 500)	17.548 (0.946)	16.035 (0.946)	10.232 (0.950)	4.195 (0.957)	1.753 (0.946)
(10, 500)	17.621 (0.947)	16.102 (0.938)	10.276 (0.952)	4.216 (0.950)	1.768 (0.957)
(25, 500)	17.632 (0.947)	16.113 (0.948)	10.285 (0.949)	4.229 (0.955)	1.795 (0.939)

condition $p = o(n^{1/2})$, the proposed confidence intervals based on $\hat{\theta}_{\mathrm{LS}}$ and $\hat{\theta}_{\mathrm{SSLS}}$ are close to valid and shorter on average than the traditional z-confidence interval centered at $\bar{\mathbf{Y}}$.

5. Applications. In this section, we apply the proposed procedures to the average treatment effect estimation and a real data example on homeless population.

5.1. *Application to average treatment effect estimation.* We first discuss an application of the proposed least squares estimator to Average Treatment Effect (ATE) estimation. Suppose Y_T and Y_C are the responses for the treatment population and control population, respectively, then ATE is then defined as

$$d = \mathbb{E}Y_T - \mathbb{E}Y_C. \tag{5.1}$$

Under Neyman's paradigm [Rubin (1990), Spłlawa-Neyman (1990)], a total number of $(n_t + n_c)$ subjects are randomly assigned to the treatment group and

control group. Suppose $Y_{t,1}, \ldots, Y_{t,n_t}$ are the responses under treatment, while $Y_{t,1}, \ldots, Y_{t,n_c}$ are the responses of the control group. The straight forward idea for estimating ATE is the sample average treatment effect (SATE), which simply takes the difference of average effects between the two groups. In addition, the covariates associated with the responses are often available and helpful to improve the estimation of ATE.

In the estimation of ATE, we follow the model setting of Pitkin et al. (2013). Suppose n_t, n_c people are from treatment group and control group respectively, where their response and predictor satisfies

$$(Y_t, X_t) \overset{\text{i.i.d.}}{\sim} P^t, \qquad (Y_c, X_c) \overset{\text{i.i.d.}}{\sim} P^c.$$

Here, due to the randomization setting, it is reasonable to assume P^t and P^c share the same marginal distribution of X: $P_X^t = P_X^c = P_X$. There are also m additional samples possibly coming from drop-outs or any other subjects that also represent the population P_X. In summary, the available samples include

$$\{(Y_{t,k}, X_{t,k})\}_{k=1}^{n_t}, \qquad \{(Y_{c,k}, X_{c,k})\}_{k=1}^{n_c}, \qquad \{(X_{a,k})\}_{k=1}^{m}. \tag{5.2}$$

We again introduce the population slope for both treatment and control group to measure the relationship between Y_t, X_t and Y_c, X_c, respectively,

$$\beta_t = \operatorname*{arg\,min}_{\gamma \in \mathbb{R}^{p+1}} \mathbb{E}\big(Y_t - \vec{X}_t^\top \gamma\big)^2, \qquad \beta_c = \operatorname*{arg\,min}_{\gamma \in \mathbb{R}^{p+1}} \mathbb{E}\big(Y_c - \vec{X}_c^\top \gamma\big)^2. \tag{5.3}$$

Based on Lemma 1 in the Supplementary Material [Zhang, Brown and Cai (2019)], β_t, β_c has the following closed form when P_t, P_c have nondegenerate second moment:

$$\beta_t = \big(\mathbb{E}\vec{X}_t\vec{X}_t^\top\big)^{-1}\big(\mathbb{E}\vec{X}_t Y_t\big), \qquad \beta_c = \big(\mathbb{E}\vec{X}_c\vec{X}_c^\top\big)^{-1}\big(\mathbb{E}\vec{X}_c Y_c\big). \tag{5.4}$$

Our target, the population ATE, is defined as $d = \mathbb{E}Y_c - \mathbb{E}Y_t$. We propose the corresponding semi-supervised least squares estimator

$$\hat{d}_{\text{SSLS}} = \hat{\mu}^\top(\hat{\beta}_t - \hat{\beta}_c). \tag{5.5}$$

Here, $\hat{\beta}_t, \hat{\beta}_c \in \mathbb{R}^{p+1}$ are the least squares estimators for treatment and control group, respectively; $\hat{\vec{\mu}}$ is the mean of all available predictors,

$$\hat{\beta}_t = \big(\vec{\mathbf{X}}_t^\top\vec{\mathbf{X}}_t\big)^{-1}\vec{\mathbf{X}}_t^\top\mathbf{Y}_t, \qquad \hat{\beta}_c = \big(\vec{\mathbf{X}}_c^\top\vec{\mathbf{X}}_c\big)^{-1}\vec{\mathbf{X}}_c^\top\mathbf{Y}_c, \tag{5.6}$$

where

$$\hat{\vec{\mu}} = \begin{pmatrix} 1 \\ \hat{\mu} \end{pmatrix}, \qquad \hat{\mu} = \frac{1}{n_t + n_c + m}\left(\sum_{k=1}^{n_t} X_{t,k} + \sum_{k=1}^{n_c} X_{c,k} + \sum_{k=1}^{m} X_{a,k}\right). \tag{5.7}$$

Based on the analysis, we have in the previous section, the proposed $\hat{d}_{\text{SSLS}}$ has the following asymptotic distribution with a fixed p, P^t and P^c.

THEOREM 8 (Asymptotic behavior of $\hat{d}_{\mathrm{SSLS}}$). *Suppose P^t, P^c are fixed distribution with finite and nondegenerate second moments, then we have the following asymptotic distribution if the sample size n_t, t_c grow to infinity*:

$$\frac{\hat{d}_{\mathrm{SSLS}} - d}{V} \xrightarrow{d} N(0,1), \qquad \frac{\hat{V}^2}{V^2} \xrightarrow{d} 1. \tag{5.8}$$

Here,

$$\begin{aligned} V^2 &= \frac{\tau_t^2}{n_t} + \frac{\tau_c^2}{n_c} + \frac{1}{n_t + n_c + m}(\beta_{t,(2)} - \beta_{c,(2)})^\top \\ &\quad \times \mathbb{E}(X-\mu)(X-\mu)^\top(\beta_{t,(2)} - \beta_{c,(2)}), \end{aligned} \tag{5.9}$$

$$\hat{V}^2 = \frac{MSE_t}{n_t} + \frac{MSE_c}{n_c} + \frac{1}{n_t + n_c + m}(\hat{\beta}_t - \hat{\beta}_c)^\top \hat{\mathbf{\Sigma}}_X (\hat{\beta}_t - \hat{\beta}_c), \tag{5.10}$$

$$MSE_t = \frac{1}{n_t - p - 1}\sum_{k=1}^{n_t}\big(Y_{t,k} - \vec{X}_{t,k}^\top \hat{\beta}_t\big)^2,$$

$$MSE_c = \frac{1}{n_c - p - 1}\sum_{k=1}^{n_c}\big(Y_{c,k} - \vec{X}_{c,k}^\top \hat{\beta}_c\big)^2,$$

$$\begin{aligned} \hat{\mathbf{\Sigma}}_X = \frac{1}{n_t + n_c + m}\Bigg(&\sum_{k=1}^{n_t}(X_{t,k} - \hat{\mu})(X_{t,k} - \hat{\mu})^\top \\ &+ \sum_{k=1}^{n_c}(X_k - \hat{\mu})(X_k - \hat{\mu})^\top + \sum_{k=1}^{m}(X_k - \hat{\mu})(X_k - \hat{\mu})^\top\Bigg). \end{aligned}$$

REMARK 8. Similar to the procedure in Proposition 1, we can calculate that for the sample average treatment effect, that is,

$$\hat{d} = \sum_{k=1}^{n_t}\frac{Y_{t,k}}{n_t} - \sum_{k=1}^{n_c}\frac{Y_{t,c}}{n_c},$$

$$\begin{aligned} \mathrm{Var}(\hat{d}) = &\frac{\tau_t^2 + \beta_{t,(2)}^\top \mathbb{E}(X-\mu)(X-\mu)^\top \beta_{t,(2)}}{n_t} \\ &+ \frac{\tau_c^2 + \beta_{c,(2)}^\top \mathbb{E}(X-\mu)(X-\mu)^\top \beta_{t,(2)}}{n_c}. \end{aligned}$$

We can check that asymptotically $V^2 \le \mathrm{Var}(\hat{d})$, which also shows the merit of the proposed semi-supervised least squares estimator.

REMARK 9. The asymptotic behavior of $\hat{d}_{\mathrm{SSLS}}$ and the ℓ_2 risk for a refined $\hat{d}_{\mathrm{SSLS}}$ for growing p can be elaborated similar to the previous sections.

5.2. *Real data example*: *Estimating homeless in Los Angeles County.* We now consider an application to estimate the number of homeless people in Los Angeles County. Homelessness has been a significant public issue for the United States since nearly a century ago [Rossi (1991)]. A natural question for the demographers is to estimate the number of homeless in a certain region. Estimating the number of homeless in metropolitan area is an important but difficult task due to the following reasons. In a typical design of U.S. Census, demographers visit people through their place of residence. In this case, most of the homeless will not be contacted [Rossi (1991)] through this process. Visiting homeless shelters or homeless service centers may collect some information of the homeless, but a large number of homeless still cannot be found since they may use the service anonymously or simply not use the service.

Los Angeles County includes land of 2000 square miles, total population of 10 million and 2054 census tracts. In 2004–2005, the Los Angeles Homeless Services Authority (LAHSA) conducted a study of the homeless population. Due to the cost of performing street visits for all census tracts, LAHSA used a stratified sampling plan. First, 244 tracts that were believed to have large amount of homeless were preselected and visited. Next, for the rest of the tracts, 265 of them were randomly selected and visited. This design leaves 1545 tracts unvisited. Besides the number of homeless, some predictors were available for all 2054 tracts. In our analysis, 7 of them were included, `Perc.Industrial`, `Perc.Residential`, `Perc.Vacant`, `Perc.Commercial`, `Perc.OwnerOcc`, `Perc.Minority`, `MedianHouseholdIncome`. These predictors have been used and were known to have a useful correlation with the response [Kriegler and Berk (2010)].

Suppose T_{total} is the total number of homeless in Los Angeles, T_{pre} is the number of homeless in 244 pre-selected tracts, θ_{ran} is average number of homeless per tract in all 1810 nonpreselected tracts. Clearly,

$$T_{\text{total}} = T_{\text{pre}} + 1810 \cdot \theta_{\text{ran}}. \tag{5.11}$$

The proposed semi-supervised inference framework fit into the 1810 samples with 265 labeled and 1545 unlabeled samples. We can apply the proposed semi-supervised least squares estimator $\hat{\theta}^1_{\text{SSLS}}$ to estimate θ_{ran} and use (5.11) to calculate the estimate and 95% confidence interval for T_{total}. In contrast, the estimate via sample-mean estimator was also calculated. The results are shown in Table 3. It is easy to see that the estimate via $\hat{\theta}^1_{\text{SSLS}}$ is slightly larger than the one via $\bar{\mathbf{Y}}$.

TABLE 3
Estimated total number of homeless in Los Angeles County

via $\hat{\theta}_{\text{SSLS}}$	95%-CI	via $\bar{\mathbf{Y}}$	95%-CI
53,824	[47,120, 60,529]	52,527	[45,485, 59,570]

TABLE 4
Diagnostic table for Los Angeles data example

	$\hat{\beta}$	$\bar{\mathbf{X}}_{\text{full}} - \bar{\mathbf{X}}$	$\bar{\mathbf{X}}$	$\bar{\mathbf{X}}_{\text{full}}$
Intercept	21.963			
Perc.Industrial	0.027	0.143	61.293	61.149
Perc.Residential	−0.087	−0.075	4.066	4.141
Perc.Vacant	1.404	−0.075	4.066	4.141
Perc.Commercial	0.338	−0.542	15.130	15.672
Perc.OwnerOcc	−0.233	2.489	54.039	51.550
Perc.Minority	0.058	0.833	50.890	50.057
MedianInc (in $K)	0.074	0.638	48.805	48.167
		Adjustment: $\hat{\beta}_{(2)}^{\top}(\bar{\mathbf{X}}_{\text{full}} - \bar{\mathbf{X}}) = -0.768$		

To further investigate and diagnose, we calculated the least squares estimator $\hat{\beta}$, the average predictor values across all 1810 nonpreselected tracts $\bar{\mathbf{X}}_{\text{full}}$ and the average predictor values across 265 randomly selected tracts $\bar{\mathbf{X}}$. These values are listed in Table 4.

We can see from Table 4 that due to insufficiency of sampling, there is difference between $\bar{\mathbf{X}}$ and $\bar{\mathbf{X}}_{\text{full}}$, especially for the predictor `Perc.OwnerOcc`. When there is association between these predictors and response, it is more reasonable to adjust for this discrepancy from taking the mean. Recall the proposed estimator

$$\hat{\theta}_{\text{SSLS}} = \bar{\mathbf{Y}} + \hat{\beta}_{(2)}^{\top}(\bar{\mathbf{X}}_{\text{full}} - \bar{\mathbf{X}}) \qquad \text{where } \bar{\mathbf{X}}_{\text{full}} = \frac{1}{n+m}\sum_{k=1}^{n+m} X_k, \bar{\mathbf{X}} = \frac{1}{n}\sum_{k=1}^{n} X_k.$$

The difference between two estimates exactly originated from the adjustment term $\hat{\beta}_{(2)}^{\top}(\bar{\mathbf{X}}_{\text{full}} - \bar{\mathbf{X}})$, which has been justified in both theoretical analysis and simulation studies in the previous sections.

6. Proofs of the main results. We prove the main results in this section. The proofs of other technical results are provided in the Supplementary Material [Zhang, Brown and Cai (2019)].

6.1. *Proofs for ideal semi-supervised inference estimator* $\hat{\theta}_{\text{LS}}$. PROOF OF THEOREM 1. We first show that $\hat{\theta}_{\text{LS}}$ is invariant under simultaneous affine translation on both $\mathbf{X}$ and μ. Specifically, suppose $X_k = U \cdot Z_k + \alpha$, $(k = 1, \ldots, n)$ for any fixed invertible matrix $U \in \mathbb{R}^{p\times p}$ and vector $\alpha \in \mathbb{R}^p$. Then one has

$$\vec{X}_k = \begin{bmatrix} 1 & 0 \\ \alpha & U \end{bmatrix} \vec{Z}_k, \qquad \vec{\mathbf{X}} = \vec{\mathbf{Z}} \begin{bmatrix} 1 & \alpha^{\top} \\ 0 & U^{\top} \end{bmatrix},$$

$$\hat{\theta}_{\text{LS}} = \vec{\mu}^{\top}(\vec{\mathbf{X}}^{\top}\vec{\mathbf{X}})^{-1}\vec{\mathbf{X}}^{\top}\mathbf{Y}$$

$$
\begin{aligned}
&= (1, \mu^\top)\left(\begin{bmatrix}1 & 0\\ \alpha & U\end{bmatrix}\vec{\mathbf{Z}}^\top\vec{\mathbf{Z}}\begin{bmatrix}1 & \alpha^\top\\ 0 & U^\top\end{bmatrix}\right)^{-1}\begin{bmatrix}1 & 0\\ \alpha & U\end{bmatrix}\vec{\mathbf{Z}}^\top\mathbf{Y}\\
&= (1, \mu^\top)\begin{bmatrix}1 & \alpha^\top\\ 0 & U^\top\end{bmatrix}^{-1}(\vec{\mathbf{Z}}^\top\vec{\mathbf{Z}})^{-1}\vec{\mathbf{Z}}^\top\mathbf{Y}\\
&= (1, (U^{-1}(\mu-\alpha))^\top)(\vec{\mathbf{Z}}^\top\vec{\mathbf{Z}})^{-1}\vec{\mathbf{Z}}^\top\mathbf{Y}.
\end{aligned}
$$

Since $\mathbb{E}Z_k = U^{-1}(\mu-\alpha)$, we know $\hat{\theta}_{\mathrm{LS}}$ is invariant under simultaneous affine translation on $\mathbf{X}$ and μ.

Based on the affine transformation invariant property, we only need to consider the situation when $\mathbb{E}X = \mu = 0$, $\mathrm{Cov}(X) = I_p$, where I_p is the p-by-p identity matrix. Next, we discuss the asymptotic behavior for $\hat{\theta}_{\mathrm{LS}}$. For simplicity, we note $1_n = (\overbrace{1, \ldots, 1}^{n})^\top$, $\mathbb{P}_{\vec{\mathbf{X}}} = \vec{\mathbf{X}}(\vec{\mathbf{X}}^\top\vec{\mathbf{X}})^{-1}\vec{\mathbf{X}}^\top \in \mathbb{R}^{(p+1)\times(p+1)}$ as the projection matrix onto the column space of $\vec{\mathbf{X}}$. $\bar{\mathbf{X}} = \frac{1}{n}\sum_{k=1}^n X_k$. Clearly, 1_n lies in the column space of $\vec{\mathbf{X}}$, which means $\mathbb{P}_{\vec{\mathbf{X}}}1_n = 1_n$. Then

$$
\begin{aligned}
\hat{\theta}_{\mathrm{LS}} - \theta &= \vec{\mu}^\top\hat{\beta} - \theta = \vec{\mu}^\top(\vec{\mathbf{X}}^\top\vec{\mathbf{X}})^{-1}\vec{\mathbf{X}}^\top\mathbf{Y} - \theta\\
&= \vec{\mu}^\top(\vec{\mathbf{X}}^\top\vec{\mathbf{X}})^{-1}\vec{\mathbf{X}}^\top(\vec{\mathbf{X}}\beta + \boldsymbol{\delta}) - \theta = \vec{\mu}^\top(\vec{\mathbf{X}}^\top\vec{\mathbf{X}})^{-1}\vec{\mathbf{X}}^\top\boldsymbol{\delta}\\
&= \frac{1_n^\top}{n}\vec{\mathbf{X}}(\vec{\mathbf{X}}^\top\vec{\mathbf{X}})^{-1}\vec{\mathbf{X}}^\top\boldsymbol{\delta} - \frac{1_n^\top}{n}(\vec{\mathbf{X}} - 1_n\vec{\mu}^\top)(\vec{\mathbf{X}}^\top\vec{\mathbf{X}})^{-1}\vec{\mathbf{X}}^\top\boldsymbol{\delta}\\
&= \frac{1_n^\top\mathbb{P}_{\vec{\mathbf{X}}}}{n}\boldsymbol{\delta} - \Big(0, \frac{1_n^\top}{n}\mathbf{X}\Big)(\vec{\mathbf{X}}^\top\vec{\mathbf{X}})^{-1}\vec{\mathbf{X}}^\top\boldsymbol{\delta}\\
&= \frac{1_n^\top}{n}\boldsymbol{\delta} - \Big(0, \frac{1_n^\top}{n}\mathbf{X}\Big)\Big(\frac{1}{n}\vec{\mathbf{X}}^\top\vec{\mathbf{X}}\Big)^{-1}\Big(\frac{1}{n}\vec{\mathbf{X}}^\top\boldsymbol{\delta}\Big)\\
&= \bar{\boldsymbol{\delta}} - (0, \bar{\mathbf{X}}^\top)\Big(\frac{1}{n}\vec{\mathbf{X}}^\top\vec{\mathbf{X}}\Big)^{-1}\Big(\frac{1}{n}\vec{\mathbf{X}}^\top\boldsymbol{\delta}\Big),
\end{aligned}
\tag{6.1}
$$

$$
\begin{aligned}
&\frac{n-p-1}{n}MSE\\
&\quad= \frac{1}{n}\|\mathbf{Y} - \vec{\mathbf{X}}\hat{\beta}\|_2^2 = \frac{1}{n}\|\boldsymbol{\delta} + \vec{\mathbf{X}}\beta - \vec{\mathbf{X}}(\vec{\mathbf{X}}^\top\vec{\mathbf{X}})^{-1}\vec{\mathbf{X}}^\top(\mathbf{X}\beta + \boldsymbol{\delta})\|_2^2\\
&\quad= \frac{1}{n}\|\boldsymbol{\delta} - \vec{\mathbf{X}}(\vec{\mathbf{X}}^\top\vec{\mathbf{X}})^{-1}\vec{\mathbf{X}}^\top\boldsymbol{\delta}\|_2^2 = \frac{1}{n}(\boldsymbol{\delta}^\top\boldsymbol{\delta} - \boldsymbol{\delta}^\top\vec{\mathbf{X}}(\vec{\mathbf{X}}^\top\vec{\mathbf{X}})^{-1}\vec{\mathbf{X}}^\top\boldsymbol{\delta})\\
&\quad= \Big(\frac{1}{n}\boldsymbol{\delta}^\top\boldsymbol{\delta} - \Big(\frac{1}{n}\vec{\mathbf{X}}^\top\boldsymbol{\delta}\Big)^\top\Big(\frac{1}{n}\vec{\mathbf{X}}^\top\vec{\mathbf{X}}\Big)^{-1}\Big(\frac{1}{n}\vec{\mathbf{X}}^\top\boldsymbol{\delta}\Big)\Big).
\end{aligned}
\tag{6.2}
$$

Since P is fixed and has finite second moment, by the law of large numbers, one can show as $n \to \infty$,

$$\begin{aligned}
&\frac{1}{n}\boldsymbol{\delta}^\top \boldsymbol{\delta} = \frac{1}{n}\sum_{k=1}^{n} \delta_k^2 \xrightarrow{d} \mathbb{E}\delta^2 = \tau^2, \\
&\left\| \frac{\vec{\mathbf{X}}^\top \boldsymbol{\delta}}{n} \right\|_2^2 \xrightarrow{d} \|\mathbb{E}\vec{X}\delta\|_2^2 = 0, \\
&\frac{1}{n}\vec{\mathbf{X}}^\top \vec{\mathbf{X}} \xrightarrow{d} \mathbb{E}\vec{X}\vec{X}^\top = \begin{bmatrix} 1 & 0 \\ 0 & \operatorname{Cov}(X) \end{bmatrix}.
\end{aligned} \tag{6.3}$$

Since $\operatorname{Cov}(X) = I_p$ is invertible, we know

$$\left(\frac{1}{n}\vec{\mathbf{X}}^\top \vec{\mathbf{X}}\right)^{-1} \xrightarrow{d} \begin{bmatrix} 1 & 0 \\ 0 & \operatorname{Cov}(X)^{-1} \end{bmatrix}.$$

Additionally, since $\mathbb{E}X = 0$, and $X_1, \ldots, X_n$ are independent,

$$\begin{aligned}
\mathbb{E}\left\| \frac{1_n^\top \mathbf{X}}{\sqrt{n}} \right\|_2^2 &= \frac{1}{n}\mathbb{E}\left(\sum_{k=1}^{n} X_k\right)\left(\sum_{k=1}^{n} X_k\right)^\top \\
&= \frac{1}{n}\mathbb{E}\sum_{k=1}^{n} X_k^\top X_k = \mathbb{E}\operatorname{tr}\big(XX^\top\big) = \operatorname{tr}\big(\operatorname{Cov}(X)\big) < \infty.
\end{aligned} \tag{6.4}$$

Based on the asymptotic distributions above, for any $\varepsilon > 0$, we have

$$\begin{aligned}
&\mathbb{P}\left(\sqrt{n}(0, \bar{\mathbf{X}}^\top)\left(\frac{1}{n}\vec{\mathbf{X}}^\top \vec{\mathbf{X}}\right)^{-1}\left(\frac{1}{n}\vec{\mathbf{X}}^\top \boldsymbol{\delta}\right) \geq \varepsilon\right) \\
&\quad \leq \mathbb{P}\left(\left\| \frac{1_n^\top \mathbf{X}}{\sqrt{n}} \right\|_2 \left\| \frac{\vec{\mathbf{X}}^\top \boldsymbol{\delta}}{n} \right\|_2 \cdot \left\| \left(\frac{1}{n}\vec{\mathbf{X}}^\top \vec{\mathbf{X}}\right)^{-1} \right\| \geq \varepsilon\right) \\
&\quad \leq \mathbb{P}\left(\left\| \frac{1_n^\top \mathbf{X}}{\sqrt{n}} \right\|_2 \geq \varepsilon/\varepsilon_n\right) + \mathbb{P}\left(\left\| \frac{\vec{\mathbf{X}}^\top \boldsymbol{\delta}}{n} \right\|_2 \geq \varepsilon_n/\big(2(\|\operatorname{Cov}(X)^{-1}\| + 1)\big)\right) \\
&\qquad + \mathbb{P}\left(\left\| \left(\frac{1}{n}\vec{\mathbf{X}}^\top \vec{\mathbf{X}}\right)^{-1} \right\| \geq 2(\|\operatorname{Cov}(X)^{-1}\| + 1)\right),
\end{aligned}$$

where ε_n grows slowly with n to ensure that $\mathbb{P}(\|\frac{\vec{\mathbf{X}}^\top \boldsymbol{\delta}}{n}\|_2 \geq \varepsilon_n/(2(\|\operatorname{Cov}(X)^{-1}\| + 1))) \to 0$ as $n \to \infty$. By such an argument,

$$\forall \varepsilon > 0, \qquad \lim_{n\to\infty} \mathbb{P}\left(\sqrt{n}(0, \bar{\mathbf{X}}^\top)\left(\frac{1}{n}\vec{\mathbf{X}}^\top \vec{\mathbf{X}}\right)^{-1}\left(\frac{1}{n}\vec{\mathbf{X}}^\top \boldsymbol{\delta}\right) \geq \varepsilon\right) = 0,$$

which means $\sqrt{n}(0, \bar{\mathbf{X}}^\top)(\frac{1}{n}\vec{\mathbf{X}}^\top \vec{\mathbf{X}})^{-1}(\frac{1}{n}\vec{\mathbf{X}}^\top \boldsymbol{\delta}) \xrightarrow{d} 0$. Next, by the central limit theorem,

$$\sqrt{n}\bar{\boldsymbol{\delta}}/\tau \xrightarrow{d} N(0, 1).$$

Combining (6.1), (6.2) and the previous asymptotic arguments, we know

$$\sqrt{n}(\hat{\theta}_{\mathrm{LS}}-\theta)/\tau=\sqrt{n}\bar{\boldsymbol{\delta}}/\tau-\sqrt{n}\big(0,\bar{\mathbf{X}}^{\top}\big)\bigg(\frac{1}{n}\vec{\mathbf{X}}^{\top}\vec{\mathbf{X}}\bigg)^{-1}\bigg(\frac{1}{n}\vec{\mathbf{X}}^{\top}\boldsymbol{\delta}\bigg)/\tau\to N(0,1),$$

$$\frac{n-p-1}{n}MSE\xrightarrow{d}\tau^2,$$

in the case that P_X fixed and $n\to\infty$.

Next, we use C and c to denote generic constants which does not depend on n (but may depend on the distribution P). When P further has a finite fourth moment, by Berry–Esseen's CLT,

$$\big|\mathbb{P}(\sqrt{n}\bar{\boldsymbol{\delta}}/\tau\geq x)-\Phi(x)\big|\leq\frac{C}{\sqrt{n}}. \tag{6.5}$$

We also have a finer estimation for $\|(\tilde{\mathbf{X}}^{\top}\boldsymbol{\delta})/n\|_2^2$ than the one in (6.3). Note that

$$\begin{aligned}\mathbb{E}\bigg\|\frac{\vec{\mathbf{X}}^{\top}\boldsymbol{\delta}}{n}\bigg\|_2^2&=\frac{1}{n^2}\mathbb{E}\bigg(\sum_{k=1}^{n}\vec{X}_k\delta_k\bigg)^{\top}\bigg(\sum_{k=1}^{n}\vec{X}_k\delta_k\bigg)\\&=\frac{1}{n^2}\sum_{k=1}^{n}\mathbb{E}\vec{X}_k^{\top}\vec{X}_k\delta_k^2\leq\frac{C}{n},\end{aligned} \tag{6.6}$$

$$\begin{aligned}\mathbb{E}\bigg\|\frac{1}{n}\vec{\mathbf{X}}^{\top}\vec{\mathbf{X}}-I_{p+1}\bigg\|_F^2&=\mathbb{E}\,\mathrm{tr}\bigg(\frac{1}{n}\sum_{k=1}^{n}(\vec{X}_k\vec{X}_k^{\top}-\mathbb{E}\vec{X}\vec{X}^{\top})\bigg)^2\\&=\frac{1}{n^2}\sum_{k=1}^{n}\mathbb{E}\,\mathrm{tr}\big(\vec{X}_k\vec{X}_k^{\top}-\mathbb{E}\vec{X}_k\vec{X}_k^{\top}\big)^2\\&=\frac{1}{n}\mathbb{E}\,\mathrm{tr}\big(\vec{X}\vec{X}^{\top}-\mathbb{E}\vec{X}\vec{X}^{\top}\big)^2\leq\frac{C}{n}.\end{aligned} \tag{6.7}$$

By Markov's inequality,

$$\begin{aligned}&\mathbb{P}\bigg(\sqrt{n}\big(0,\bar{\mathbf{X}}^{\top}\big)\bigg(\frac{1}{n}\vec{\mathbf{X}}^{\top}\vec{\mathbf{X}}\bigg)^{-1}\bigg(\frac{1}{n}\vec{\mathbf{X}}^{\top}\boldsymbol{\delta}\bigg)\geq\frac{C}{n^{1/4}}\bigg)\\&\quad=\mathbb{P}\bigg(\bigg\|\frac{1_n^{\top}\mathbf{X}^{\top}}{\sqrt{n}}\bigg\|_2\cdot\bigg\|\bigg(\frac{1}{n}\vec{\mathbf{X}}^{\top}\vec{\mathbf{X}}\bigg)^{-1}\bigg\|\cdot\bigg\|\frac{\vec{\mathbf{X}}^{\top}\boldsymbol{\delta}}{n}\bigg\|_2\geq\frac{C}{n^{1/4}}\bigg)\\&\quad\leq\mathbb{P}\bigg(\bigg\|\frac{1_n^{\top}\mathbf{X}}{\sqrt{n}}\bigg\|_2\geq Cn^{1/8}\bigg)+\mathbb{P}\bigg(\bigg\|\frac{1}{n}\vec{\mathbf{X}}^{\top}\vec{\mathbf{X}}-I_{p+1}\bigg\|_F\leq1/2\bigg)\\&\qquad+\mathbb{P}\bigg(\bigg\|\frac{\vec{\mathbf{X}}^{\top}\boldsymbol{\delta}}{n}\bigg\|_2\geq\frac{C}{n^{3/8}}\bigg)\end{aligned} \tag{6.8}$$

$$\le \frac{\mathbb{E}\|1_n^\top \mathbf{X}/\sqrt{n}\|_2^2}{Cn^{1/4}} + \frac{\mathbb{E}\|\frac{1}{n}\vec{\mathbf{X}}^\top\vec{\mathbf{X}} - I_{p+1}\|_F^2}{1/2} + \frac{\mathbb{E}\|\vec{\mathbf{X}}^\top\boldsymbol{\delta}/n\|_2^2}{Cn^{-3/4}}$$

$$\overset{(6.4)(6.6)(6.7)}{\le} Cn^{-1/4}.$$

Finally, for any $x > 0$,

$$\mathbb{P}\left(\frac{\sqrt{n}(\hat{\theta}_{LS} - \theta)}{\tau} \le x\right)$$

$$\overset{(6.1)}{\le} \mathbb{P}\left(\frac{\sqrt{n}\bar{\boldsymbol{\delta}}}{\tau} \le x + \frac{C}{n^{1/4}}\right)$$

$$+ \mathbb{P}\left(-\sqrt{n}(0, \bar{\mathbf{X}}^\top)\left(-\frac{1}{n}\vec{\mathbf{X}}^\top\vec{\mathbf{X}}\right)^{-1}\left(\frac{1}{n}\vec{\mathbf{X}}^\top\boldsymbol{\delta}\right)/\tau \le -\frac{C}{n^{1/4}}\right)$$

$$\overset{(6.8)}{\le} \Phi\left(x + \frac{C}{n^{1/4}}\right) + Cn^{-1/4} \le \Phi(x) + Cn^{-1/4}.$$

Here, the last inequality is due to the fact that the c.d.f. of the standard normal distribution $\Phi(\cdot)$ is a Lipschitz continuous function. Similarly,

$$\mathbb{P}\left(\frac{\sqrt{n}(\hat{\theta}_{LS} - \theta)}{\tau} \le x\right) \ge \Phi(x) - Cn^{-1/4}.$$

These together complete the proof of this theorem. □

PROOF OF THEOREM 2. First, based on the proof of Theorem 1, the affine transformation on $\mathbf{X}$ would not affect the property of $\hat{\theta}_{\mathrm{LS}}$. Without loss of generality, we assume that $\mathbb{E}X = 0$, $\mathrm{Var}(X) = I$. In other words, $\mathbf{Z} = \mathbf{X}$. Next, based on formulas (6.1) and (6.2), we have

$$\sqrt{n}(\hat{\theta}_{\mathrm{LS}} - \theta)/\tau = \frac{\sqrt{n}\bar{\boldsymbol{\delta}}}{\tau} - \sqrt{n}(0, \bar{\mathbf{X}}^\top)\left(\frac{1}{n}\vec{\mathbf{X}}^\top\vec{\mathbf{X}}\right)^{-1}\left(\frac{1}{n}\vec{\mathbf{X}}^\top\boldsymbol{\delta}\right),$$

$$\left|\frac{\sqrt{n}}{\tau}(0, \bar{\mathbf{X}}^\top)\left(\frac{1}{n}\vec{\mathbf{X}}^\top\vec{\mathbf{X}}\right)^{-1}\left(\frac{1}{n}\vec{\mathbf{X}}^\top\boldsymbol{\delta}\right)\right| \le \left\|\frac{1_n\mathbf{X}^\top}{n^{3/4}}\right\|_2 \cdot \lambda_{\min}^{-1}\left(\frac{1}{n}\vec{\mathbf{X}}^\top\vec{\mathbf{X}}\right) \cdot \left\|\frac{\vec{\mathbf{X}}^\top\boldsymbol{\delta}}{n^{3/4}\tau}\right\|_2,$$

then we only need to prove the following asymptotic properties in order to finish the proof of Theorem 2:

$$\frac{\sqrt{n}\bar{\boldsymbol{\delta}}}{\tau} \xrightarrow{d} N(0, 1), \tag{6.9}$$

$$\left\|\frac{1_n\mathbf{X}}{n^{3/4}}\right\|_2 \xrightarrow{d} 0, \qquad \left\|\frac{\vec{\mathbf{X}}^\top\boldsymbol{\delta}}{n^{3/4}}\right\|_2/\tau \xrightarrow{d} 0, \tag{6.10}$$

for some uniform $t_1 > t_2 > 0$,

$$P\left(t_1 \ge \lambda_{\max}\left(\frac{1}{n}\vec{\mathbf{X}}^\top\vec{\mathbf{X}}\right) \ge \lambda_{\min}\left(\frac{1}{n}\vec{\mathbf{X}}^\top\vec{\mathbf{X}}\right) \ge t_2\right) \to 1. \tag{6.11}$$

Here, $\lambda_{\max}, \lambda_{\min}(\cdot)$ represent the largest and least eigenvalues of the given matrix. Next, we will show (6.9), (6.10) and (6.11) separately.

- Based on the assumption of the theorem, $\frac{\delta_1}{\tau}, \ldots, \frac{\delta_n}{\tau}$ are i.i.d. samples with mean 0, variance 1 and bounded $(2+2\varepsilon)$th moment, (6.9) holds by Lyapunov's central limit theorem.
- Since $X_1, \ldots, X_k$ are i.i.d. samples with mean 0 and covariance I_p, we can calculate that

$$\mathbb{E}\left\| \frac{1_n \mathbf{X}^\top}{n^{3/4}} \right\|_2^2 = \frac{1}{n^{3/2}} \cdot n\mathbb{E}\|X\|_2^2 = \frac{p}{n^{1/2}} \to 0 \qquad \text{as } n \to 0.$$

  Since $X_1\delta_1, \ldots, X_n\delta_n$ are i.i.d. samples with mean 0 and satisfying (2.11), we have

$$\mathbb{E}\left\| \frac{\vec{\mathbf{X}}^\top \boldsymbol{\delta}}{n^{3/4}} \right\|_2^2 = \frac{1}{n^{3/2}} \cdot n\mathbb{E}\|\vec{X}\delta\|_2^2 \le \frac{M_3}{n^{1/2}}\mathbb{E}\|X\|_2^2 \cdot \mathbb{E}\delta^2 = \frac{p}{n^{1/2}} M_3 \tau^2.$$

  Thus, $\mathbb{E}\|\frac{\vec{\mathbf{X}}^\top \boldsymbol{\delta}}{n^{3/4}}\|_2^2/\tau^2 \to 0$ as $n \to \infty$. Thus, we have (6.10).
- For (6.11), since $\mathbb{E}X = 0$, $\mathrm{Cov}(X) = I_p$ and Assumption (2.10) holds, (6.11) is directly implied by Theorem 2 in Yaskov (2014). □

Acknowledgments. The authors thank Arun Kuchibhotla for many helpful discussions. The authors also thank the Editor, the Associate Editor and anonymous referees for many helpful comments, which greatly helped to improve the presentation of this paper.

SUPPLEMENTARY MATERIAL

Supplement to "Semi-supervised inference: General theory and estimation of means" (DOI: 10.1214/18-AOS1756SUPP; .pdf). The supplement contains additional proofs for the main results of the paper.

REFERENCES

ANDO, R. K. and ZHANG, T. (2005). A framework for learning predictive structures from multiple tasks and unlabeled data. *J. Mach. Learn. Res.* **6** 1817–1853. MR2249873

ANDO, R. K. and ZHANG, T. (2007). Two-view feature generation model for semi-supervised learning. In *Proceedings of the* 24*th International Conference on Machine Learning* 25–32. ACM, New York.

AZRIEL, D., BROWN, L. D., SKLAR, M., BERK, R., BUJA, A. and ZHAO, L. (2016). Semi-supervised linear regression. Preprint. Available at arXiv:1612.02391.

BICKEL, P. J., RITOV, Y. and WELLNER, J. A. (1991). Efficient estimation of linear functionals of a probability measure P with known marginal distributions. *Ann. Statist.* **19** 1316–1346. MR1126327

BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1998). *Efficient and Adaptive Estimation for Semiparametric Models*. Springer, New York. Reprint of the 1993 original. MR1623559

BLUM, A. and MITCHELL, T. (1998). Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory* (*Madison, WI*, 1998) 92–100. ACM, New York. MR1811574

BRATLEY, P., FOX, B. L. and SCHRAGE, L. E. (1987). *A Guide to Simulation*. Springer, New York.

BUJA, A., BERK, R., BROWN, L., GEORGE, E., PITKIN, E., TRASKIN, M., ZHAN, K. and ZHAO, L. (2014). Models as approximations, Part I: A conspiracy of nonlinearity and random regressors in linear regression. Preprint. Available at arXiv:1404.1578.

BUJA, A., BERK, R., BROWN, L., GEORGE, E., KUCHIBHOTLA, A. K. and ZHAO, L. (2016). Models as approximations—Part II: A general theory of model-robust regression. Preprint. Available at ArXiv:1612.03257.

CHAKRABORTTY, A. and CAI, T. (2018). Efficient and adaptive linear regression in semi-supervised settings. *Ann. Statist.* **46** 1541–1572. MR3819109

COCHRAN, W. G. (1953). *Sampling Techniques*. Wiley, New York. MR0054199

DENG, L.-Y. and WU, C.-F. J. (1987). Estimation of variance of the regression estimator. *J. Amer. Statist. Assoc.* **82** 568–576. MR0898360

FISHMAN, G. S. (1996). *Monte Carlo*: *Concepts*, *Algorithms*, *and Applications*. Springer, New York. MR1392474

GRAHAM, B. S. (2011). Efficiency bounds for missing data models with semiparametric restrictions. *Econometrica* **79** 437–452. MR2809376

HANSEN, B. E. (2017). Econometrics. Book draft. Available at http://www.ssc.wisc.edu/~bhansen/econometrics/.

HASMINSKII, R. Z. and IBRAGIMOV, I. A. (1983). On asymptotic efficiency in the presence of an infinite-dimensional nuisance parameter. In *Probability Theory and Mathematical Statistics* (*Tbilisi*, 1982). *Lecture Notes in Math.* **1021** 195–229. Springer, Berlin. MR0735986

HICKERNELL, F. J., LEMIEUX, C. and OWEN, A. B. (2005). Control variates for quasi-Monte Carlo. *Statist. Sci.* **20** 1–31. MR2182985

JOHNSON, R. and ZHANG, T. (2008). Graph-based semi-supervised learning and spectral kernel design. *IEEE Trans. Inform. Theory* **54** 275–288. MR2446753

KRIEGLER, B. and BERK, R. (2010). Small area estimation of the homeless in Los Angeles: An application of cost-sensitive stochastic gradient boosting. *Ann. Appl. Stat.* **4** 1234–1255. MR2751340

KUCHIBHOTLA, A. (2017). Research notes on efficiency in semi-supervised problems. Available from the author at arunku@wharton.upenn.edu.

LAFFERTY, J. D. and WASSERMAN, L. (2008). Statistical analysis of semi-supervised regression. In *Advances in Neural Information Processing Systems* 801–808.

LOHR, S. (2009). *Sampling*: *Design and Analysis*. Nelson Education.

PENG, H. and SCHICK, A. (2002). On efficient estimation of linear functionals of a bivariate distribution with known marginals. *Statist. Probab. Lett.* **59** 83–91. MR1925291

PITKIN, E., BERK, R., BROWN, L., BUJA, A., GEORGE, E., ZHANG, K. and ZHAO, L. (2013). Improved precision in estimating average treatment effects. Preprint. Available at arXiv:1311.0291.

ROSSI, P. H. (1991). Strategies for homeless research in the 1990s. *Hous. Policy Debate* **2** 1027–1055.

RUBIN, D. B. (1990). Comment on J. Neyman and causal inference in experiments and observational studies: "On the application of probability theory to agricultural experiments. Essay on principles. Section 9" [*Ann. Agric. Sci.* **10** (1923), 1–51]. *Statist. Sci.* **5** 472–480. MR1092987

SPŁAWA-NEYMAN, J. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statist. Sci.* **5** 465–472. Translated from the Polish and edited by D. M. Dabrowska and T. P. Speed. MR1092986

VAN DER VAART, A. (2002). Semiparametric statistics. In *Lectures on Probability Theory and Statistics* (*Saint-Flour*, 1999). *Lecture Notes in Math.* **1781** 331–457. Springer, Berlin. MR1915446

VAPNIK, V. N. (2013). *The Nature of Statistical Learning Theory*. Springer, Berlin.
WANG, J. and SHEN, X. (2007). Large margin semi-supervised learning. *J. Mach. Learn. Res.* **8** 1867–1891. MR2353822
WANG, J., SHEN, X. and LIU, Y. (2008). Probability estimation for large-margin classifiers. *Biometrika* **95** 149–167. MR2409720
WANG, J., SHEN, X. and PAN, W. (2009). On efficient large margin semisupervised learning: Method and theory. *J. Mach. Learn. Res.* **10** 719–742. MR2491755
YASKOV, P. (2014). Lower bounds on the smallest eigenvalue of a sample covariance matrix. *Electron. Commun. Probab.* **19** no. 83, 10. MR3291620
ZHANG, A., BROWN, L. D. and CAI, T. T. (2019). Supplement to "Semi-supervised inference: General theory and estimation of means." DOI:10.1214/18-AOS1756SUPP.
ZHU, X. (2008). Semi-supervised learning literature survey. Technical report.
ZHU, X. and GOLDBERG, A. B. (2009). Introduction to semi-supervised learning. *Synth. Lect. Artif. Intell. Mach. Learn.* **3** 1–130. DOI:10.2200/S00196ED1V01Y200906AIM006.

A. ZHANG
DEPARTMENT OF STATISTICS
UNIVERSITY OF WISCONSIN—MADISON
MADISON, WISCONSIN 53706
USA
E-MAIL: anruzhang@stat.wisc.edu

L. D. BROWN
T. T. CAI
DEPARTMENT OF STATISTICS
THE WHARTON SCHOOL
UNIVERSITY OF PENNSYLVANIA
PHILADELPHIA, PENNSYLVANIA 19104-6340
USA
E-MAIL: lbrown@wharton.upenn.edu
tcai@wharton.upenn.edu
URL: http://www-stat.wharton.upenn.edu/~lbrown/
http://www-stat.wharton.upenn.edu/~tcai/