Biometrika (2022), **109**, 2, *pp.* 387–403 doi: 10.1093/biomet/asab042
Advance Access publication 17 March 2022

High-dimensional semi-supervised learning: in search of optimal inference of the mean

BY YUQIAN ZHANG

Institute of Statistics and Big Data, Renmin University of China,
59 Zhongguancun Avenue, Haidian District, Beijing 100872, P. R. China

AND JELENA BRADIC

Department of Mathematics, University of California San Diego,
9500 Gilman Drive, La Jolla, California 92093 0112, U.S.A.
jbradic@ucsd.edu

SUMMARY

A fundamental challenge in semi-supervised learning lies in the observed data's disproportional size when compared with the size of the data collected with missing outcomes. An implicit understanding is that the dataset with missing outcomes, being significantly larger, ought to improve estimation and inference. However, it is unclear to what extent this is correct. We illustrate one clear benefit: root-n inference of the outcome's mean is possible while only requiring a consistent estimation of the outcome, possibly at a rate slower than root n. This is achieved by a novel k-fold, cross-fitted, double robust estimator. We discuss both linear and nonlinear outcomes. Such an estimator is particularly suited for models that naturally do not admit root-n consistency, such as high-dimensional, nonparametric or semiparametric models. We apply our methods to estimating heterogeneous treatment effects.

Some key words: Coefficient of determination; Double robustness; Missing data; Model-lean inference.

1. INTRODUCTION

We consider a semi-supervised setting with n independent and identically distributed pairs $(X_i, Y_i)_{i=1}^n \sim P_{(X,Y)}$ of observations, with covariates $X_i \in \mathbb{R}^{p-1}$ and the outcome $Y_i \in \mathbb{R}$. We presuppose the existence of an additional set of m observations, $(X_i)_{i=n+1}^{n+m}$. With $\tau = \lim_{m,n\to\infty} n/(m+n) \in [0,1]$ denoting the ratio of the fully observed data and data with the missing outcomes, we are particularly focused on the case of $\tau = 0$, i.e., $m \gg n$. The semi-supervised learning setting can be viewed as a particular missing data setting, where the outcome is missing completely at random. Although the missing data literature, in general, addresses a more general setting of outcomes missing at random (Scharfstein et al., 1999), semi-supervised learning has a particular caveat that the missing data's size is enormous, $m \gg n$. With $m \gg n$, typical missing-at-random approaches (Bang & Robins, 2005) no longer apply. The positivity/overlap condition (see, e.g., Rotnitzky et al., 2012), is no longer satisfied; with $\tau = 0$, the probability of observing the outcome converges to zero, therefore implying that the semi-supervised setting is not a simple subset of the missing-at-random setting. Instead, we treat the missingness size, an impediment from the missing-at-random perspective, as a semi-supervised strength. In the case of

infinite missingness of the response, we are left with infinite additional information regarding the covariates' distribution, P_X. Mimicking the known P_X setting, we remove the bias in estimating the outcome model and show that semi-supervised double-robust inference is achievable.

Our main contribution is in constructing new semi-supervised estimates of $\theta = E(Y)$ and in providing root-n inferential guarantees while allowing for misspecification of the distribution of $Y \mid X$. An impediment to providing optimal inferences about θ lies in the inability to estimate $E(Y \mid X)$ with root-n guarantees. Sparse regularizers, random forests, nonparametric smoothing estimators or neural networks do not admit root-n consistency. While there is vast literature on semi-supervised learning, comparatively little is known about making inferences about θ; see Zhu (2005). Recent results of Wasserman & Lafferty (2008), El Alaoui et al. (2016) and Mai & Couillet (2018) consider the class of low-dimensional graph-oriented semi-supervised algorithms. Semi-supervised learning in the context of classification has had a long tradition; see Grandvalet & Bengio (2005) and Chapelle et al. (2009). A small but growing literature has considered the development of semi-supervised inferential procedures. The recent work of Zhang et al. (2019) is a special case of our construction. The authors utilize the least-squares approach in linear models and assume $p = o(n^{1/2})$. Our results are based on $n^{-1}\log(p) = o(1)$ together with many possible estimators, e.g., random forests and neural networks. Chakrabortty & Cai (2018) developed the semi-supervised regression method with improved efficiency when the linear model is misspecified. Gronsbell & Cai (2018) considered semi-supervised prediction, while Cai & Guo (2020) proposed semi-supervised explained variance estimates. We, therefore, view our contribution as complementary to this growing literature.

We believe that our new estimating tools will be useful beyond the specific class of environments studied here. We illustrate this point by applying our findings to heterogeneous treatment effects. The existing approaches of Chernozhukov et al. (2017, 2018) and Künzel et al. (2019) build learners that can conform to many machine learning methods (Athey et al., 2018; Wager & Athey, 2018). However, they do not consider the semi-supervised setting with the outcome and the treatment missing. We discover that the asymptotic variance size is reduced regardless of whether additional information on the treatment is available. Moreover, treatment assignment can potentially depend on all covariates with no explicit sparsity requirement. The method also shares the low-dimensional asymptotic efficiency of Cheng et al. (2020).

2. Efficient estimation of the mean

2.1. *From debiasing to double robustness*

Let $\beta^* \in \mathbb{R}^p$, the population slope, be an l_2 projection defined as $\beta^* = \arg\min_{\beta\in\mathbb{R}^p} E\left(Y - \beta_1 - X^{\mathrm{T}}\beta_{-1}\right)^2$. Here, β_{-j} denotes β with the jth coordinate removed. For $\varepsilon = Y - \beta_1^* - X^{\mathrm{T}}\beta_{-1}^*$ and $\sigma_\varepsilon^2 = \mathrm{var}(\varepsilon)$ with $E(\varepsilon \mid X) \neq 0$ we do not necessarily assume that the regression model is linear. With μ and C, denoting the mean and the covariance of X_i, respectively, we use $V_i = X_i - \mu$ and $Z_i = C^{-1/2}(X_i - \mu)$. With $\tilde{X}_i = (1, X_i^{\mathrm{T}})^{\mathrm{T}}$ and $\tilde{V}_i = (1, V_i^{\mathrm{T}})^{\mathrm{T}}$, let $\tilde{\mu} = (1, \mu^{\mathrm{T}})^{\mathrm{T}}$ and $\tilde{C} = \mathrm{cov}(\tilde{X})$ denote the mean and covariance of $\tilde{X} = (1, X^{\mathrm{T}})^{\mathrm{T}}$. The mean of the response, $\theta = E(Y)$, can be seen as a linear contrast of β^*: $\theta = \tilde{\mu}^{\mathrm{T}}\beta^*$.

When $p \gg n$, a good candidate estimate of β^* is a regularized estimator, $\hat{\beta}$, e.g., lasso (Tibshirani, 1997) or square-root lasso (Belloni et al., 2011). However, such estimators suffer from slower than root-n consistency: when the outcome model is linear, $\|\hat{\beta} - \beta^*\|_2^2 = o_P\{s\log(p)/n\}$ with $s = |\{j : \beta_j^* \neq 0\}|$. Hence, a plug-in estimate will not achieve root-n inference regarding θ, even if the outcome model is correct, unless s is a constant. The existing literature provides

easy solutions with many possible ways to remove the bias of regularization. Each of these could potentially achieve root-n inference of θ, but would, however, require strong assumptions on the models: the outcome must be well specified as well as sparse enough. For example, let $\hat{\beta}_{\mathrm{db}} = \hat{\beta} + n^{-1}\sum_{i=1}^{n} \hat{\Theta}\tilde{X}_i(Y_i - \tilde{X}_i^{\mathrm{T}}\hat{\beta})$ denote the debiased lasso (Van de Geer et al., 2014). Here, $\hat{\Theta}$ is a candidate estimate of $\tilde{\Sigma}^{-1}$, $\tilde{\Sigma} = E\tilde{X}\tilde{X}^{\mathrm{T}} \in \mathbb{R}^{p\times p}$. Root-$n$ inference of θ would then require outcome sparsity $s = o\{n^{1/2}/\log(p)\}$ as well as $|\{k \neq j : (\tilde{\Sigma}^{-1})_{j,k} \neq 0\}| = o\{n/\log(p)\}$ (Van de Geer et al., 2014).

However, $\hat{\beta}_{\mathrm{db}}$ does not directly use the additional covariate information available in the semi-supervised setting. Let us consider a particular case where P_X, and with it $\tilde{\Sigma}^{-1}$ and $\tilde{\mu}$, are known. In this case we could use an improved debiased semi-supervised estimator $\tilde{\beta} = \hat{\beta} + n^{-1}\sum_{i=1}^{n} \tilde{\Sigma}^{-1}\tilde{X}_i(Y_i - \tilde{X}_i^{\mathrm{T}}\hat{\beta})$, which then leads to $\tilde{\mu}^{\mathrm{T}}\tilde{\beta} = \tilde{\mu}^{\mathrm{T}}\hat{\beta} + n^{-1}\sum_{i=1}^{n} e_1^{\mathrm{T}}\tilde{X}_i^{\mathrm{T}}(Y_i - \tilde{X}_i\hat{\beta}) = \tilde{\mu}^{\mathrm{T}}\hat{\beta} + n^{-1}\sum_{i=1}^{n}(Y_i - \tilde{X}_i^{\mathrm{T}}\hat{\beta})$, where $e_1 = (1, 0, 0, \ldots, 0)^{\mathrm{T}}$. Interestingly, by algebraic manipulation it is not difficult to see that the right-hand side of the latter equation becomes $\bar{Y} + (\tilde{\mu} - \bar{X})^{\mathrm{T}}\hat{\beta}$, where $\bar{Y} = n^{-1}\sum_{i=1}^{n} Y_i$ and $\bar{X} = n^{-1}\sum_{i=1}^{n} \tilde{X}_i$, therefore matching the low-dimensional estimator of Zhang et al. (2019). There seems to be an intricate connection between the above estimator and the double-robust, missing-at-random estimators of Bang & Robins (2005). However, there is an important difference. If $T_j = 1$ for $j = 1, \ldots, n$ and zero otherwise, i.e., T is the indicator of the observed data, then missing-at-random estimators treat T as a random variable whereas semi-supervised learning treats T as fixed and nonrandom. Semi-supervised learning can be viewed as missing-at-random estimation conditional on $(T_i)_{i=1}^{m+n}$ being fixed. Then, the missing-at-random average treatment effect of the treated matches the above estimator,

$$\tilde{\mu}^{\mathrm{T}}\hat{\beta} + (n+m)^{-1}\sum_{i=1}^{n+m} T_i(Y_i - X_i\hat{\beta})/\mathrm{pr}(T_i = 1 \mid X_i),$$

where $\mathrm{pr}(T_i = 1 \mid X_i) = \mathrm{pr}(T_i = 1) = n/(n+m)$. However, missing-at-random double-robust estimates require $\mathrm{pr}(T_i = 1 \mid X_i) > 0$, whereas in the semi-supervised setting we have $\mathrm{pr}(T_i = 1 \mid X_i) \to 0$ with $m \gg n$.

In the semi-supervised setting we aim to show that the above estimator's sample equivalent will suffice for root-n inference on θ. Let $\tilde{\theta} = \hat{\mu}^{\mathrm{T}}\hat{\beta} + n^{-1}\sum_{i=1}^{n}(Y_i - \tilde{X}_i^{\mathrm{T}}\hat{\beta})$ and $\hat{\mu} = (n+m)^{-1}\sum_{i=1}^{n}\tilde{X}_i$. Our estimator will use cross-fitting, which plays a crucial role in establishing the double-robust property of the proposed estimator, i.e., in controlling the term t_2 in the decomposition $\tilde{\theta} - \theta = t_1 + t_2 + t_3$, where $t_1 = \theta - n^{-1}\sum_{i=1}^{n} Y_i$, $t_2 = (n^{-1}\sum_{i=1}^{n}\tilde{X}_i - \hat{\mu})^{\mathrm{T}}(\hat{\beta} - \beta^*)$ and $t_3 = (n^{-1}\sum_{i=1}^{n}\tilde{X}_i - \hat{\mu})^{\mathrm{T}}\beta^*$. The cross-fitting technique helps in removing the bias arising from t_2. With the use of cross-fitting, the influence of $\hat{\beta}$ and X_i in t_2 are separated and tight control of t_2 is achieved under minimal conditions. Without cross-fitting, $|\tilde{\theta} - \theta| \leqslant \|n^{-1}\sum_{i=1}^{n}\tilde{X}_i - \hat{\mu}\|_\infty\|\hat{\beta} - \beta^*\|_1$, where the right-hand side is $O_P(n^{-1/2})$ as long as $s \leqslant n^{1/2}/\log(p)$. Instead, with the use of cross-fitting, we can guarantee root-n consistency as long as $s \leqslant n/\log(p)$. Cross-fitting can be traced back to the natural ideas of cross-validation. Historical background is provided by Stone (1974) and Geisser (1975), for example. More recently, Rinaldo et al. (2019) showed that sample splitting increases the accuracy and robustness of inference. Chernozhukov et al. (2017) used cross-fitting to define double-robust missing-at-random estimates.

We start by splitting the labelled observations into K sets, I_k, each of size N, and split the unlabelled observations into sets I'_k. Let $J_k = I_k \cup I'_k$, with $|J_k| = M$. Let $\hat{\beta}^{(-k)}$ denote an estimate of β^* computed on all but the kth labelled observations, $\hat{\beta}^{(-k)} = \hat{\beta}[\{(\tilde{X}_i, Y_i) : i \in$

$\{1, 2, \ldots, n\} \setminus I_k\}] \in \mathbb{R}^p$. Then, we propose

$$\hat{\theta}^{(k)} = \hat{\mu}^{(k)^{\mathrm{T}}} \hat{\beta}^{(-k)} + N^{-1} \sum_{i \in I_k} \left(Y_i - \tilde{X}_i^{\mathrm{T}} \hat{\beta}^{(-k)} \right), \qquad \hat{\mu}^{(k)} = M^{-1} \sum_{i \in J_k} \tilde{X}_i. \tag{1}$$

Finally, we propose the following semi-supervised estimator, which aggregates the above estimates:

$$\hat{\theta} = K^{-1} \sum_{k=1}^{K} \hat{\theta}^{(k)}.$$

We will show that this estimator becomes an unbiased estimator of θ, even in finite samples.

2.2. *From the mean to the coefficient of determination*

A crucial statistical problem is the estimation of the proportion of variance explained, PVE $= \mathrm{var}(\tilde{X}^{\mathrm{T}}\beta^*)/\sigma_Y^2$. Estimation of PVE with $p \gg n$ is difficult due to the numerous overfitting issues. In this subsection we propose a semi-supervised coefficient of determination, R^2, an estimator of PVE. The estimation of the explained variance, $b^2 = \mathrm{var}(\tilde{X}^{\mathrm{T}}\beta^*)$ (Cai & Guo, 2020), can be performed with the cross-fitted residuals

$$\hat{b}^{2(k)} = \hat{\beta}^{(-k)^{\mathrm{T}}} \hat{C}^{(k)} \hat{\beta}^{(-k)} + 2N^{-1} \sum_{i \in I_k} \hat{\beta}^{(-k)^{\mathrm{T}}} \hat{V}_i \hat{\varepsilon}_i, \qquad \hat{\varepsilon}_i = Y_i - \hat{\theta} - \hat{\beta}^{(-k)^{\mathrm{T}}} \hat{V}_i, \tag{2}$$

and $\hat{b} = K^{-1} \sum_{k=1}^{K} \hat{b}^{2(k)}$, where the estimates of $\tilde{V}_i$ are $\hat{V}_i = \tilde{X}_i - \hat{\mu}^{(k)}$ and their covariance $\hat{C}^{(k)} = M^{-1} \sum_{i \in J_k} \hat{V}_i \hat{V}_i^{\mathrm{T}}$. The motivation behind this careful construction is governed by bias propagation in the high-dimensional setting; as we will show, the residuals as defined above are, however, root-n consistent. This, in turn, provides a more stable estimate and enables theoretically weak conditions. To see that the naive estimate $Y_i - \hat{\beta}^{\mathrm{T}}\tilde{X}_i$ may not guarantee root-n consistency, we only need to observe that in such a case, $Y_i - \hat{\beta}^{\mathrm{T}}\tilde{X}_i = \varepsilon_i + (\theta - \hat{\beta}^{\mathrm{T}}\tilde{\mu}) - (\hat{\beta} - \beta^*)^{\mathrm{T}}(\tilde{X}_i - \tilde{\mu})$, while the term $\theta - \hat{\beta}^{\mathrm{T}}\tilde{\mu}$ is not necessarily root-n consistent whenever $p \gg n$. Our cross-fitted construction can be seen as a bias-corrected estimate of the residuals. We propose a new estimator of the variance of the response, $\sigma_Y^2 = \mathrm{var}(Y)$,

$$\hat{\sigma}_Y^{2(k)} = N^{-1} \sum_{i \in I_k} (Y_i - \hat{\theta})^2 + N^{-1} \sum_{i \in I_k} \hat{\beta}^{(-k)^{\mathrm{T}}} \left(\hat{C}^{(k)} - \hat{V}_i \hat{V}_i^{\mathrm{T}} \right) \hat{\beta}^{(-k)}, \tag{3}$$

and with it $\hat{\sigma}_Y^2 = K^{-1} \sum_{k=1}^{K} \hat{\sigma}_Y^{2(k)}$. Our results also hold for the truncated version $\hat{\sigma}_{Y,\mathrm{trunc}}^2 = \max(\hat{\sigma}_Y^2, 0)$. A classical estimate, the simple sample variance, $S_Y^2 = n^{-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$, does not utilize any additional knowledge of the covariates. Alternatively, one may consider $n^{-1} \sum_{i=1}^{n} (Y_i - \hat{\theta})^2$. However, both of these estimates can be improved. Our theoretical results demonstrate a persistent variance magnification, $n^{-1} \sum_{i=1}^{n} (Y_i - \hat{\theta})^2 = \sigma_Y^2 + n^{-1} \sum_{i=1}^{n} \{\beta^{*\mathrm{T}}(\tilde{V}_i \tilde{V}_i^{\mathrm{T}} - \tilde{C})\beta^*\} + T + O_P(n^{-1})$, where $E(T) = 0$ and $T = n^{-1} \sum_{i=1}^{n} (2\beta^{*\mathrm{T}} \tilde{V}_i \varepsilon_i + \varepsilon_i^2 - \sigma_\varepsilon^2)$; details are presented in the Supplementary Material. Hence, our estimator adds a correction term so that the contribution of the middle term disappears. Therefore, R^2 can be obtained by plugging in the estimators of b^2,

(2), and the variance of the response σ_Y^2, (3),

$$R^2 = K^{-1} \sum_{k=1}^{K} \hat{b}^{2^{(k)}} / \hat{\sigma}_Y^{2^{(k)}}. \tag{4}$$

2.3. *Root-n consistency*

We establish the root-n consistency of the proposed semi-supervised estimators. The constants in what follows, possibly changing from line to line, are independent of the sample size.

Condition 1. Let the covariance matrix C be such that $\lambda_{\min}(C) > 0$, $\lambda_{\max}(C) \leqslant c_1$ and $\sup_{\|a\|_2=1} E|a^{\mathrm{T}}Z|^{2+c} < c_1$, as well as $E|Y|^{2+c} < c_1$, for positive constants $c, c_1 > 0$.

Condition 2. The responses are such that $E|Y|^{4+c} < c_1$, whereas the covariance matrix C satisfies $\lambda_{\min}(C) > 0$, $\lambda_{\max}(C) \leqslant c_1$ and $\sup_{\|a\|_2=1} E|a^{\mathrm{T}}Z|^{4+c} < c_1$ for positive constants $c, c_1 > 0$.

Condition 3. Let $\hat{\beta}$ be an estimator for β^* that satisfies $\|\hat{\beta} - \beta^*\|_2 = O_P(1)$ as $n, p \to \infty$.

Conditions 1 or 2, used one at a time, provide a well-defined linear approximation model β^*. A bounded variance of Y simplifies exposition; all of the results still hold even if this condition is removed. However, the results would be less interpretable. Condition 3 allows for a wide variety of estimates of β^*: lasso, Dantzig, square-root lasso, elastic-net (Zou & Hastie, 2005) or slope (Bogdan et al., 2015) are plausible. Similarly, different structural forms of β^* are permissible; a considerably weaker form of sparsity, l_r sparsity with $r \in (0, 1)$, would be effective as long as $\|\beta^*\|_r^r = o[\{n/\log(p)\}^{1-r/2}]$ (Ye & Zhang, 2010), for example. As per Conditions 1 and 2, bounded $2 + c$ and $4 + c$ moments allow heavy-tailed distributions for the covariates as well as the noise; see, e.g., the Huber estimate of Sun et al. (2020).

THEOREM 1. *Let Conditions 1 and 3 hold. Then, as $m, n, p \to \infty$, $\hat{\theta} - \theta = O_P(n^{-1/2})$. Moreover, if Condition 2 holds as well, $\hat{\sigma}_Y^2 - \sigma_Y^2 = O_P(n^{-1/2})$.*

Regarding $\hat{\theta}$, Condition 1 can be relaxed to bounded $1 + c$ moments. Importantly, we do not rely on a strong signal-to-noise ratio to achieve root-n consistency. If $s = p$, one can show that the lasso estimate equals zero with high probability, in which case the proposed estimate will be the same as the naive $\bar{Y}$. Hence, there is no loss in efficiency, and it seems that the semi-supervised mean estimate is advantageous in almost all cases. We discuss some aspects of the variance in the Supplementary Material.

2.4. *Asymptotic normality*

In this section we proceed to prove that semi-supervised estimates are asymptotically normal and that they improve the efficiency of estimation by borrowing strength from the additional dataset.

Condition 4. Let $\hat{\beta}$ be an estimator of β^* that satisfies $\|\hat{\beta} - \beta^*\|_2 = o_P(1)$ as $n, p \to \infty$.

THEOREM 2. *Let Conditions 1 and 4 hold. Then, as $m, n, p \to \infty$,*

$$n^{1/2}(\hat{\theta} - \theta) \to N\left(0, \sigma_\varepsilon^2 + \tau b^2\right) \tag{5}$$

in distribution, provided that $\sigma_\varepsilon^2 + \tau b^2 > c$ *for some constant* $c > 0$.

Compared with requirements for inference in high-dimensional linear models, Conditions 1 and 4 are milder. Where we require only moderately sparse regimes $s = o(n/\log p)$, high-dimensional and even doubly-robust methods require more strict settings; see, e.g., Bradic et al. (2019), Smucler et al. (2019) and Tan (2020a, 2020b). In particular, we do not require any sparsity structure on Σ^{-1}, a condition that has been typically assumed throughout the literature, if the variance is unknown. Lastly, we do not require homogeneity of the errors, ε.

Regarding efficiency, observe that $\text{var}(n^{1/2}\bar{Y}) = \sigma_Y^2 = \sigma_\varepsilon^2 + b^2 \geqslant \sigma_\varepsilon^2 + \tau b^2$, where $\sigma_\varepsilon^2 + \tau b^2$ is the asymptotic variance of $\hat{\theta}$ as in (5). Hence, the semi-supervised estimator $\hat{\theta}$ is asymptotically at least as accurate as $\bar{Y}$ and is often more accurate. Namely, the additional unlabelled data reduce the asymptotic variance by $(1-\tau)b^2$. The more unlabelled data we observe, the more accurate the proposed estimator $\hat{\theta}$ becomes. When $\tau = 0$, the asymptotic variance is equivalent to the case of known P_X.

Throughout the paper, we mainly focus on the case of the signal-to-noise ratio, $\text{SNR} = b^2/\sigma_\varepsilon^2$, being bounded away from 0 and ∞. However, observe that the two extremes are not particularly informative. Namely, the case of $\text{SNR} = 0$ illustrates that no estimator can improve the naive $\bar{Y}$. Conversely, the case of $\text{SNR} = \infty$ and $\tau = 0$ illustrates that the semi-supervised estimator can potentially lead to a better than $n^{1/2}$ convergence rate. Set $\rho_j = \text{Corr}(Z_j, Y)$ for each $j \in \{1, 2, \ldots, p-1\}$. Then, $b^2 = \beta_{-1}^{*\mathrm{T}} C \beta_{-1}^* = \{C^{-1}E(VY)\}^{\mathrm{T}} C C^{-1} E(VY) = \sigma_Y^2 \sum_{j=1}^{p-1} \rho_j^2$. If $\tau < 1$ and $\sigma_Y^2 \sum_{j=1}^{p-1} \rho_j^2 > c$ for some $c > 0$, i.e., when at least one of the covariates has positive marginal correlation with the response, $\hat{\theta}$ is asymptotically more accurate than $\bar{Y}$.

Our estimator is also optimal in the following sense. The asymptotic variance in Theorem 2 is the same as that of Zhang et al. (2019), proved under a low-dimensional setting; see their Theorem 2.4. Moreover, it also achieves the oracle lower bound presented in their Proposition 3.1. The following result presents theoretically valid root-n confidence intervals of θ, while only requiring consistency of $\hat{\beta}$ at an arbitrarily slow rate.

THEOREM 3. *Let Conditions 1 and 4 hold. With* $\hat{\varepsilon}_i$ *defined in* (2)*, we define* $\hat{\sigma}_\varepsilon^2 = n^{-1}\sum_{i=1}^n \hat{\varepsilon}_i^2$. *Then, whenever* $m, n, p \to \infty$, $\hat{\sigma}_\varepsilon^2 = \sigma_\varepsilon^2 + o_P(1)$, $\hat{b}^2 = b^2 + o_P(1)$ *and a valid confidence interval about* θ*, at significance level* α*, is defined as*

$$\text{CI}(\theta) = \big[\hat{\theta} - z_{1-\alpha/2}\{\hat{\sigma}_\varepsilon^2/n + \hat{b}^2/(m+n)\}^{1/2},\ \hat{\theta} + z_{1-\alpha/2}\{\hat{\sigma}_\varepsilon^2/n + \hat{b}^2/(m+n)\}^{1/2}\big],$$

with $z_{1-\alpha/2}$ *being the* $(1-\alpha/2)$*-quantile of a standard normal distribution.*

A few comments are in order. If we are willing to assume Condition 2, we show that $\hat{b}^2 - b^2 = O_P(\|\hat{\beta} - \beta^*\|_2^2 + n^{-1/2})$. In contrast, a naive plug-in estimate of b^2, $\hat{\beta}^{(-k)\mathrm{T}} \hat{C} \hat{\beta}^{(-k)}$, would only guarantee $O_P(\|\hat{\beta} - \beta^*\|_2)$. Therefore, our result on $\hat{b}^2$ can be seen as complementary to Cai & Guo (2020). We provide the same convergence rate whenever $b^2 > c$, $c > 0$, but with weaker assumptions: we allow heavy-tailed X and ε and a misspecified linear model. An asymptotically normal result holds once $\|\hat{\beta} - \beta^*\|_2 = o_P(n^{-1/4})$; the details of the asymptotic theory regarding $\hat{b}$ are contained in Theorem S2 of the Supplementary Material under a more general setting.

Next, we discuss the high-dimensional R^2 semi-supervised estimate. We begin by highlighting the asymptotic results on the variance estimate, followed by a simple corollary regarding the asymptotics of R^2.

THEOREM 4. *Let Conditions 2 and 4 hold. Then, as* $m, n, p \to \infty$,

$$n^{1/2}(\hat{\sigma}_Y^2 - \sigma_Y^2) \to N\{0, \mathrm{var}(\varepsilon^2 + 2\beta^{*\mathrm{T}}\tilde{V}\varepsilon) + \tau\,\mathrm{var}(\beta^{*\mathrm{T}}\tilde{V})^2\} \tag{6}$$

in distribution, provided that $\mathrm{var}(\varepsilon^2 + 2\beta^{*\mathrm{T}}\tilde{V}\varepsilon) + \tau\mathrm{var}(\beta^{*\mathrm{T}}\tilde{V})^2 > c$ *for some constant* $c > 0$. *Moreover, for* $\hat{\sigma}_v^2$ *and* $\hat{\sigma}_\xi^2$ *defined in the Supplementary Material, we have*

$$\hat{\sigma}_v^2 + n(m+n)^{-1}\hat{\sigma}_\xi^2 = \mathrm{var}(\varepsilon^2 + 2\beta^{*\mathrm{T}}\tilde{V}\varepsilon) + \tau\,\mathrm{var}(\beta^{*\mathrm{T}}\tilde{V})^2 + o_P(1).$$

A sufficient condition regarding Theorem 4 includes $\mathrm{var}(\varepsilon^2 + 2\beta^{*\mathrm{T}}\tilde{V}\varepsilon) > 0$: whenever $\sigma_\varepsilon^2 > c_1$ and $\mathrm{corr}(\varepsilon^2, \beta^{*\mathrm{T}}\tilde{V}\varepsilon) > -1 + c_2$ for some $c_1, c_2 > 0$, the asymptotic variance in (6) is positive. Now we are ready to state the asymptotic normality of R^2 as a simple corollary of a more general result; see Theorem S2 in the Supplementary Material.

COROLLARY 1. *Let Conditions 1 and 4 hold. Then, for* R^2 *defined in* (4), *we have* $R^2 = \mathrm{PVE} + o_P(1)$ *whenever* $m, n, p \to \infty$. *Moreover, if Condition 2 holds with* $\|\hat{\beta} - \beta^*\|_2 = o_P(n^{-1/4})$, *then, as* $m, n, p \to \infty$, $n^{1/2}V^{-1/2}(R^2)(R^2 - \mathrm{PVE}) \to N(0,1)$ *in distribution provided* $V(R^2) > 0$, *where* $V(R^2) = \mathrm{var}[\sigma_Y^{-4}b^2\varepsilon^2 + \sigma_Y^{-4}\sigma_\varepsilon^2\{2\varepsilon\beta^{*\mathrm{T}}\tilde{V} + \tau(\beta^{*\mathrm{T}}\tilde{V})^2\}] + \tau\sigma_Y^{-8}\sigma_\varepsilon^4\mathrm{var}\{(\beta^{*\mathrm{T}}\tilde{V})^2\}$.

3. BEYOND LINEAR OUTCOME MODELS

Recall that our estimation towards the mean depends on the linear projection of $g^0(x) = E(Y \mid X = x)$. A question arises naturally: can we use general machine learning algorithms to estimate $g^0(x)$ and design nonlinear projection for optimal estimation of θ? Are we able to construct confidence intervals, and will the asymptotic variances of the estimators be improved? We provide positive answers to both questions.

A natural extension of $\hat{\theta}$ can be defined as

$$\hat{\theta}_{\mathrm{gen}} = K^{-1}\sum_{k=1}^{K}\hat{\theta}_{\mathrm{gen}}^{(k)}, \qquad \text{where } \hat{\theta}_{\mathrm{gen}}^{(k)} = M^{-1}\sum_{i\in J_k}\hat{g}^{(-k)}(X_i) + N^{-1}\sum_{i\in I_k}\{Y_i - \hat{g}^{(-k)}(X_i)\} \tag{7}$$

and $\hat{g}^{(-k)}$ is the estimate of g^0 computed on all but the kth labelled observations. We suppose the existence of some $g^* = g_d^*\colon \mathbb{R}^p \to \mathbb{R}$ such that $\mu_{2,X}\{\hat{g}^{(-k)}(x) - g^*(X)\} = o_P(1)$ as $n \to \infty$ and possibly $p, q \to \infty$, and where $\mu_r(f) = E\{f - E(f)\}^r$ is the rth central moment and $\mu_{r,X}(f) = E_X\{f - E_X(f)\}^r$, with E_X denoting the conditional expectation on the marginal distribution P_X. Here, d denotes the degree of freedom of the working model. Note that $g^*(x) = g^0(x)$ is unnecessary. Here, $g^* = g_d^*$ can be chosen as the projection of the underlying curve $g^0(x)$ to a functional class $\mathcal{G}_d$, i.e.,

$$g^* = \arg\min_{g\in\mathcal{G}_d} E\{g^*(X) - g^0(X)\}^2. \tag{8}$$

With a small abuse of notation, let $\varepsilon = Y - g^*(X)$ denote the unexplained error of the model. To better interpret our results, we assume that $E(\varepsilon) = 0$ and $E\{\varepsilon g^*(X)\} = 0$, which is satisfied once $b + ag \in \mathcal{G}_d$ for all $a, b \in \mathbb{R}$ and $g \in \mathcal{G}_d$. We demonstrate in Theorem 5 that $\hat{\theta}_{\mathrm{gen}}$ of (7) is asymptotically normal with asymptotic variance $V_{\mathrm{gen}}(\theta) = \sigma_{\varepsilon,\mathrm{gen}}^2 + \tau b_{\mathrm{gen}}^2$, where

$b^2_{\text{gen}} = \text{var}\{g^*(X)\}$ denotes the explained variance of the model g, and $\sigma^2_{\varepsilon,\text{gen}} = E\{Y - g^*(X)\}^2 = \text{var}(Y) - b^2_{\text{gen}}$ denotes the unexplained variance. When g^* is defined as in (8), b^2_{gen} and $\sigma^2_{\varepsilon,\text{gen}}$ are the largest explained variance and smallest unexplained variance among the functional class $\mathcal{G}_d$, respectively. The unexplained variance can be estimated using a cross-fitting scheme,

$$\hat{\sigma}^2_{\varepsilon,\text{gen}} = n^{-1} \sum_{k=1}^{K} \sum_{i \in I_k} \left\{ Y_i - \hat{\theta}_{\text{gen}} - \hat{h}^{(-k)}(X_i) \right\}^2,$$

with $\hat{h}^{(-k)}(X_i) = \hat{g}^{(-k)}(X_i) - M^{-1} \sum_{i \in J_k} \hat{g}^{(-k)}(X_i)$. As for the explained variance, (2) can be generalized through a bias-corrected cross-fitting estimator

$$\hat{b}^2_{\text{gen}} = (m+n)^{-1} \sum_{k=1}^{K} \sum_{i \in J_k} \{\hat{h}^{(-k)}(X_i)\}^2 + 2n^{-1} \sum_{k=1}^{K} \sum_{i \in I_k} \hat{h}^{(-k)}(X_i) \{Y_i - \hat{\theta}_{\text{gen}} - \hat{h}^{(-k)}(X_i)\}.$$

Now, $\hat{V}_{\text{gen}}(\theta) = \hat{\sigma}^2_{\varepsilon,\text{gen}} + n\hat{b}^2_{\text{gen}}/(m+n)$, and an α-level confidence interval can be constructed as

$$\text{CI}_{\text{gen}}(\theta) = \left[\hat{\theta}_{\text{gen}} - z_{1-\alpha/2} \left\{ \hat{V}_{\text{gen}}(\theta)/n \right\}^{1/2}, \hat{\theta}_{\text{gen}} + z_{1-\alpha/2} \left\{ \hat{V}_{\text{gen}}(\theta)/n \right\}^{1/2} \right]. \tag{9}$$

Due to space constraints, we relegate the asymptotic normality of nonlinear R^2 to the Supplementary Material.

THEOREM 5. *Suppose that* $E|Y|^{2+c} < C$ *and* $E|g^*(X)|^{2+c} < C$ *for some* $C < \infty$. *Then, as long as* $\mu_{2,X}\{\hat{g}^{(-k)}(x) - g^*(X)\} = o_P(1)$ *for each* k, *as* $n, p \to \infty$ *or* $n, p, d \to \infty$, $\hat{\theta}_{\text{gen}}$ *satisfies* $n^{1/2} V^{-1/2}_{\text{gen}}(\theta)(\hat{\theta}_{\text{gen}} - \theta) \to N(0,1)$, $\hat{V}_{\text{gen}}(\theta) = V_{\text{gen}}(\theta) + o_P(1)$ *provided that* $V_{\text{gen}}(\theta) > 0$.

The asymptotic variance above depends on the explained variance b^2_{gen}: the larger the explained variance, the more efficient the estimation of θ. In particular, a worst case of $b^2_{\text{gen}} = 0$ corresponds to the sample mean estimator. When $g^*(x) = g^0(x)$, the asymptotic variance is optimal; it matches the oracle lower bound of Proposition 3.1 in Zhang et al. (2019), and one can see a clear efficiency gain through $b^2_{\text{gen}}(g^*) \leqslant b^2_{\text{gen}}(g^0)$.

4. HETEROGENEOUS TREATMENT EFFECTS

Suppose that in addition to the previous settings we have access to a treatment indicator $D_i \in \{0,1\}$, $i = 1, \ldots, m+n$. Following the potential outcomes framework (Rubin, 1974; Holland, 1988; Splawa-Neyman et al., 1990) we then hypothesize the presence of potential outcomes $Y_i(0)$ and $Y_i(1)$ corresponding to, respectively, the response the ith subject would have experienced with and without the treatment. We then observe that the average treatment effect $\delta = E\{E(Y \mid X, D=1) - E(Y \mid X, D=0)\} = \tau_1 - \tau_0$.

Similarly to § 2, we hypothesize the existence of the l_2 slopes $\beta^*_w = \min_{\beta \in \mathbb{R}^p} E\{(Y - \tilde{X}^{\mathrm{T}}\beta)^2 \mid D = w\}$, defined at the population level for $w \in \{0,1\}$. A standard way of constructing the average treatment effects estimates is to posit a model on the treatment assignment and then adjust for possible confounding. Treatments are assigned to subjects according to an underlying scheme that depends on the subjects' features. Their dependence can be captured by $D_i = e(X_i) + \zeta_i$, where

$e(X_i)$ is an unknown propensity score function (Rosenbaum & Rubin, 1983). In the following, we assume two primitive conditions: a widely regarded overlap condition regarding the treatment missingness, and an identifiability condition.

Condition 5. Let $\text{pr}\{c \leqslant e(X) \leqslant 1-c\} = 1$ and $\text{pr}\{c \leqslant \hat{e}(X) \leqslant 1-c\} = 1$ for some constant $c \in (0,1)$. For $\varepsilon_i = Y_i(D_i) - \{D_i\beta_1^* + (1-D_i)\beta_0^*\}^{\mathrm{T}}\tilde{X}_i$, let $E(\zeta \mid X) = 0$, as well as $\text{pr}\{E(\varepsilon^2 \mid X) < C\} = 1$ with some constant $C > 0$.

Let $\hat{\beta}_1$, $\hat{\beta}_0$ and $\hat{e}$ denote estimators for β_1^*, β_0^* and e, respectively, satisfying $E_{P_X}\{(\hat{\beta}_w^{(-k)} - \beta_w^{*})^{\mathrm{T}}\tilde{X}\}^2 = O_P(a_{n,p}^2)$, $E_{P_X}\{\hat{e}^{(-k)}(X) - e(X)\}^2 = O_P(b_{m+n,p}^2)$ and $E[\{E(Y \mid X) - \beta_w^{*\mathrm{T}}\tilde{X}\}^2 \mid D = w] = O_P(c_p^2)$. Here, $a_{n,p}$, $b_{m+n,p}$ and c_p are nonnegative sequences of numbers, with c_p describing how close the linear model is to the true underlying curve. The semi-supervised estimator (1) needs to be adjusted for the confounding effects. To that end, we introduce

$$\hat{\tau}_\omega^{(k)} = \hat{\mu}^{(k)\mathrm{T}}\hat{\beta}_\omega^{(-k)} + N^{-1}\sum_{i\in I_k} w_i^{(-k)}(\omega)\big(Y_i - \tilde{X}_i^{\mathrm{T}}\hat{\beta}_\omega^{(-k)}\big), \qquad \hat{\mu}^{(k)} = M^{-1}\sum_{i\in J_k}\tilde{X}_i.$$

In the above, the weights $w_i^{(-k)}(\omega)$ correspond to the ratio of the observed treatment proportion; then, the framework from § 2.1 will still lead to root-n consistent estimates. We denote these weights as $w_i^{(-k)}(\omega) = \omega D_i/\hat{e}^{(-k)}(X_i) + (1-\omega)(1-D_i)/\{1-\hat{e}^{(-k)}(X_i)\}$. Then, the estimate of the average treatment effect can be defined as the difference of $\hat{\delta}^{(k)} = \hat{\tau}_1^{(k)} - \hat{\tau}_0^{(k)}$ and $\hat{\delta} = K^{-1}\sum_{k=1}^K \hat{\delta}^{(k)}$.

An asymptotic $(1-\alpha)$-level confidence interval for the average treatment effect could then be defined as

$$\big(\hat{\delta} - z_{1-\alpha/2}\hat{V}_\delta^{1/2}n^{-1/2},\ \hat{\delta} + z_{1-\alpha/2}\hat{V}_\delta^{1/2}n^{-1/2}\big). \tag{10}$$

The estimator of $V_\delta = V_1 + \tau V_2$, (12), is defined as $\hat{V}_\delta = K^{-1}\sum_{k=1}^K\big\{\hat{V}_1^{(k)} + n(m+n)^{-1}\hat{V}_2^{(k)}\big\}$. Observe that $V_1 = \text{var}\{r(Y - \beta_1^{*\mathrm{T}}\tilde{X}) - \rho(Y - \beta_0^{*\mathrm{T}}\tilde{X})\}$. Then, a natural plug-in estimator can be defined as $\hat{V}_1^{(k)} = N^{-1}\sum_{i\in I_k}\nu_{\delta,i}^2$, where $\nu_{\delta,i} = r_i^{(-k)}(Y_i - \hat{\beta}_1^{(-k)\mathrm{T}}\tilde{X}_i) - \rho_i^{(-k)}(Y_i - \hat{\beta}_0^{(-k)\mathrm{T}}\tilde{X}_i) - \{\hat{\delta} - (\hat{\beta}_1^{(-k)} - \hat{\beta}_0^{(-k)})^{\mathrm{T}}\hat{\mu}^{(k)}\}$, recall that $\hat{\mu}^{(k)}$ is defined in (1). The second component, $V_2 = E\{(\beta_1^* - \beta_0^*)^{\mathrm{T}}(\tilde{X} - \tilde{\mu})\}^2$, is estimated as $\hat{V}_2^{(k)} = N^{-1}\sum_{i\in I_k}\xi_{\delta,i}^2$ for $\xi_{\delta,i} = (\hat{\beta}_1^{(-k)} - \hat{\beta}_0^{(-k)})^{\mathrm{T}}(\tilde{X}_i - \hat{\mu})$. The next theorem is the main result of this section.

Theorem 6. *Let Conditions 1 and 5 hold. Then, under the setting of this section,* $\hat{\delta} - \delta = O_P(n^{-1/2} + a_{n,p}b_{m+n,p} + b_{m+n,p}c_p)$ *whenever* $a_{n,p} = O(1)$. *Therefore, whenever* $a_{n,p}b_{m+n,p} = o(1)$ *and* $b_{m+n,p}c_p = o(1)$, $\hat{\delta}$ *is consistent. If, however,* $a_{n,p}b_{m+n,p} = O(n^{-1/2})$ *and* $b_{m+n,p}c_p = O(n^{-1/2})$, $\hat{\delta}$ *is an* $n^{1/2}$*-consistent estimate of* δ. *Additionally, the asymptotic normality follows*

$$n^{1/2}(\hat{\delta} - \delta) \to N(0, V_\delta) \tag{11}$$

in distribution, whenever $a_{n,p} = o(1)$, $b_{m+n,p} = o(1)$, $a_{n,p}b_{m+n,p} = o(n^{-1/2})$ *and* $b_{m+n,p}c_p = o(n^{-1/2})$, *with an asymptotic variance*

$$V_\delta = \text{var}\,(\varepsilon\zeta/[e(X)\{1-e(X)\}]) + \tau(\beta_1^* - \beta_0^*)^{\mathrm{T}}\tilde{C}(\beta_1^* - \beta_0^*), \tag{12}$$

provided that $V_\delta > c$ *for some* $c > 0$, *and* $\tau = \lim_{m,n\to\infty} n/(m+n)$. *Moreover,* $\hat{V}_\delta = V_\delta + o_P(1)$.

Suppose the sparsity of the outcome and the treatment model are s_Y and s_D, respectively. For illustration purposes suppose that both models are parametric and linear. Then, $c_p = 0$ and the rates $a_{n,p}$ and $b_{m+n,p}$ for a lasso estimate become $a_{n,p} = O_P[\{s_Y \log(p)/n\}^{1/2}]$, $b_{m+n,p} = O_P[\{s_D \log(p)/(m+n)\}^{1/2}]$. Therefore, $s_Y = o\{n/\log(p)\}$, $s_D = o\{(m+n)/\log(p)\}$ and $s_Y s_D = o[(m+n)/\{\log(p)\}^2]$ are required to achieve asymptotic normality. Then, when m is large enough, in that $s_D n \log p/m \to 0$, we require $s_Y = o\{n/\log(p)\}$, which is extremely mild, i.e., consistency in estimation of the propensity model at any arbitrary rate. If both D and Y were unavailable in the unlabelled data, the estimation error on the propensity score would depend on n rather than $m + n$ with the same sparsity assumptions as in Chernozhukov et al. (2017), Smucler et al. (2019) and others. At the same time, we achieve a more efficient estimator regardless of whether D is available in the unlabelled data or not, i.e., reducing the size of the asymptotic variance. When the outcome model is misspecified, even if $c_p = O(1)$, such that the linear model does not reach the underlying curve as p grows, we can still obtain the asymptotic normality (11) provided m is large enough that $b_{m+n,p} = o(n^{-1/2})$. Supervised settings have more stringent conditions; see, e.g., Smucler et al. (2019) and Tan (2020b). If one is only interested in obtaining root-n consistency, the outcome model can be completely misspecified, including completely dense high-dimensional models. They can be estimated using machine learning methods, such as random forests, Bayesian classification, regression tree or deep neural networks; one just needs to replace the linear projection $\tilde{X}_i^{\mathrm{T}} \hat{\beta}^{(-k)}$ by any $\hat{g}^{(-k)}(w, X_i)$,

$$\hat{\delta}_{\mathrm{gen}} = (m+n)^{-1} \sum_{k=1}^{K} \sum_{i \in J_k} \left\{ \hat{g}^{(-k)}(1, X_i) - \hat{g}^{(-k)}(0, X_i) \right\} + n^{-1} \sum_{k=1}^{K} \sum_{i \in I_k} \left[\frac{D_i \{ Y_i - \hat{g}^{(-k)}(1, X_i) \}}{\hat{e}^{(-k)}(X_i)} - \frac{(1 - D_i) \{ Y_i - \hat{g}^{(-k)}(0, X_i) \}}{1 - \hat{e}^{(-k)}(X_i)} \right],$$

where $\hat{g}^{(-k)}(w, X_i)$ is an estimate of $E(Y \mid X, D = w)$ trained on $(D_i, Y_i, X_i)_{i \in \{1,2,\ldots,n\} \setminus I_k}$ for $w \in \{0, 1\}$. Moreover, an asymptotic confidence interval can be extended from (10) by replacing the linear outcome model with a general nonlinear model.

5. FINITE-SAMPLE EXPERIMENTS

5.1. *Numerical experiments*

In this section we illustrate the finite-sample properties of $\hat{\theta}$. The estimation of the variance can be found in the Supplementary Material. We consider semi-supervised estimators based on ordinary least squares (SSL-OLS), the 10-fold cross-validated lasso (SSL-Lasso), the additive model (SSL-Additive), XGBoost (SSL-XGBoost), multilayer perceptron (SSL-MLP) and random forest (SSL-RF) for which vanilla, preset tuning parameters are used. We compare with the sample mean $\bar{Y} = n^{-1} \sum_{i=1}^{n} Y_i$ and with the semi-supervised least squares estimator proposed in Zhang et al. (2019) whenever $p < n$. We consider confidence intervals (9), where the significance level is $\alpha = 0.05$ throughout. Each set of results is based on 200 repetitions with $K = 5$. The black solid line in all the plots denotes the optimal ratio $\{\sigma_Y^2 - m b_{\mathrm{gen}}^2(g^0)/(m+n)\}/\sigma_Y^2$. We will see that, as long as the sample size n is large enough, our proposed semi-supervised estimator $\hat{\theta}$ is better than the sample mean $\bar{Y}$ in the sense of mean squared error.

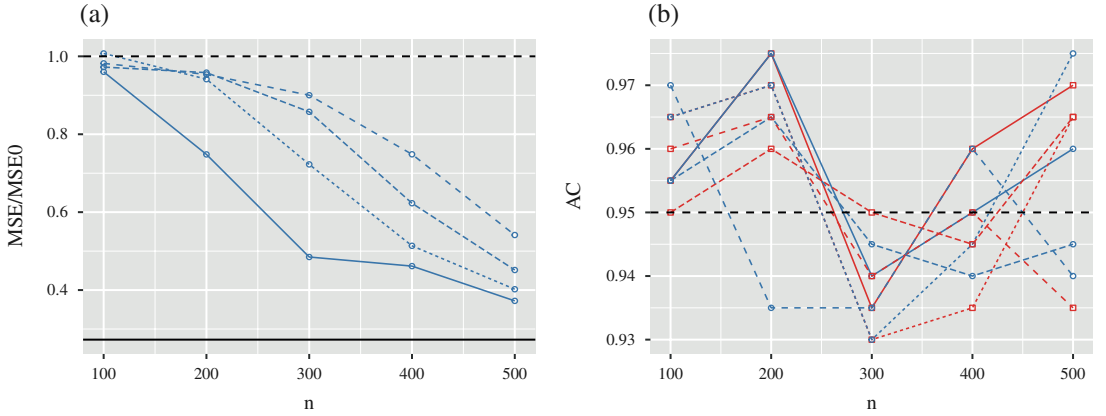

Fig. 1. Model 1: Comparison of SSL-lasso and the sample mean. (a) The ratio of mean squared errors. (b) The average coverage of $\bar{Y}$ and $\hat{\theta}$. The mean squared error of the sample mean is denoted as MSE0. The plot includes sample mean (red squares) and SSL-lasso (blue circles) estimates. The sparsity level of the linear coefficients, s, is denoted with long dashed, dashed, dotted and solid lines for $s = 90$, $s = 70$, $s = 50$ and $s = 30$, respectively.

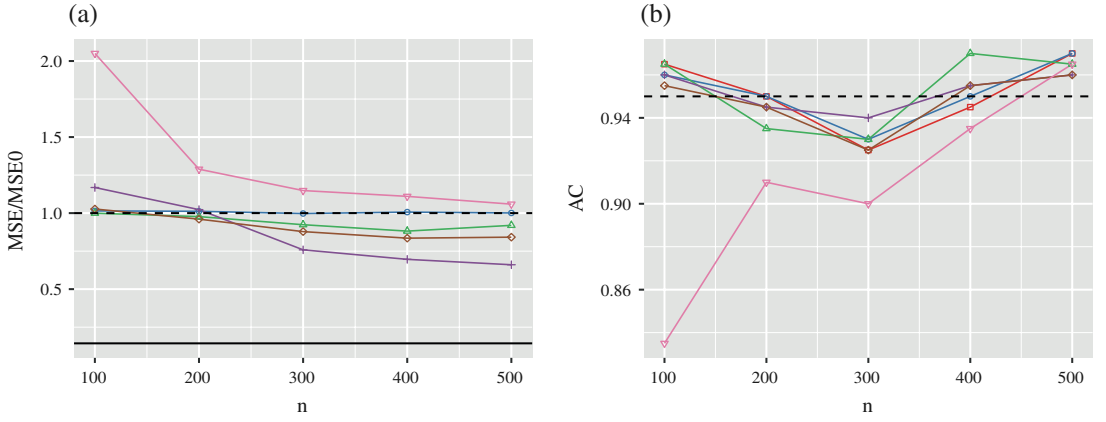

Fig. 2. Model 2: Comparison between Zhang et al. (2019) (SSLS) and our SSL estimators. (a) The ratio of mean squared errors. (b) The average coverage. The plot includes sample mean (red squares), SSL-lasso (blue circles), SSL-additive (green up triangles), SSL-XGBoost (purple pluses), SSL-RF (brown diamonds) and SSLS (pink down triangles) estimates.

Model 1. Let $X_i \overset{\text{iid}}{\sim} N_{p-1}(0, I_{p-1})$ with $p = 500$ and $m = 10n$, and $Y_i = s^{-1/2} \sum_{j=1}^{s} X_{ij} + \delta_i$, $s \in \{30, 50, 70, 90\}$, $\delta_i \overset{\text{iid}}{\sim} N(0, 0.25)$. The results are presented in Fig. 1, where we observe that our SSL-lasso estimator is more efficient than the sample mean, Fig. 1(a), regardless of the level of sparsity. Figure 1(b) illustrates robustness in terms of the average coverage probability of the SSL-lasso estimate.

Model 2. Let X_i and δ_i be as in Model 1 and consider a nonlinear model $Y_i = 3\cos(X_{i1} + X_{i2} + X_{i3}) + \delta_i$, with $p = 51$, $m = 10n$. We compare our SSL estimator with a variety of baseline procedures and the semi-supervised least squares estimator $\hat{\theta}_{\text{SSLS}}$ of Zhang et al. (2019). Figure 2(a) illustrates that SSLS is less efficient than the sample-mean estimator, that our SSL-lasso is equivalent to the sample-mean, and that all other SSL-methods are more efficient, with SSL-XGBoost outperforming the rest. Figure 2(b) demonstrates extremely poor finite-sample coverage of SSLS and nominal coverage of our proposal.

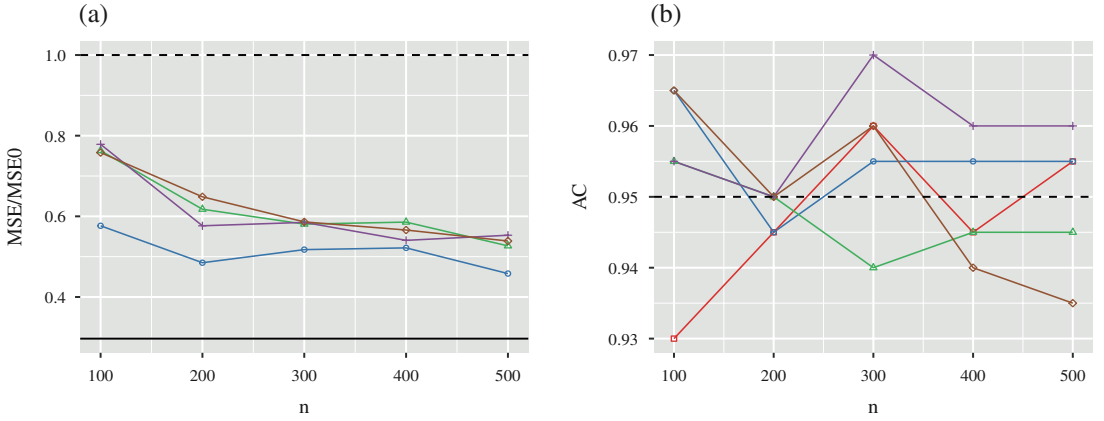

Fig. 3. Model 3: Comparison of SSL-method with the sample mean. (a) The ratio of mean squared errors. (b) The average coverage. The plot includes sample mean (red squares), SSL-lasso (blue circles), SSL-additive (green up triangles), SSL-XGBoost (purple pluses) and SSL-RF (brown diamonds) estimates.

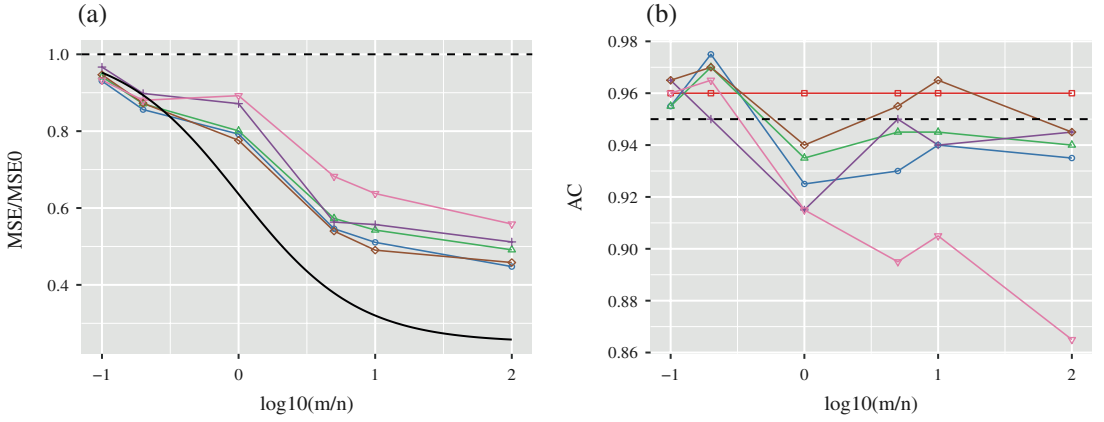

Fig. 4. Model 4: Impact of the size of additional data. (a) The ratio of mean squared errors. (b) The average coverage. The plot shows sample mean (red squares), SSL-lasso (blue circles), SSL-additive (green up triangles), SSL-XGBoost (purple pluses), SSL-RF (brown diamonds) and SSLS (pink down triangles) estimates.

Model 3. Let $X_i \overset{\text{iid}}{\sim} N_{p-1}(0, C)$ be equicorrelated with $C_{ij} = \{1 - 1/(2p)\}1_{\{i=j\}} + 1/(2p)1_{\{i\neq j\}}$, with $p = 1001$, $m = 10n$. We consider a nonlinear additive outcome model $Y_i = \sum_{j=1}^{p-1} 0.7^{j-1} \sin(X_{ij}) + \delta_i$, where $\delta_i \overset{\text{iid}}{\sim} N(0, 0.25)$. Figure 3(a) demonstrates significant gain in reduction of MSE for the proposed method, with the SSL-lasso in the lead. Figure 3(b) presents strong finite sample coverage.

Model 4. Here we observe behaviour with varying m. Let X_i and δ_i be as in Model 1 and consider the nonlinear outcome of Model 3. Set $p = 201$, $n = 500$ and let m vary from $0.1n$ to $10n$. We compare with $\bar{Y}$ and SSLS of Zhang et al. (2019). We see substantial gains in efficiency. SSL-RF dominates the other estimators, both in terms of MSE, Fig. 4(a), and coverage, Fig. 4(b). SSLS loses coverage with a larger m. When m is small, the ordinary least squares estimate's impact is not significant, and SSLS is similar to the sample mean $\bar{Y}$. As m grows, the instability of least squares and the unfitness of SSLS is exposed.

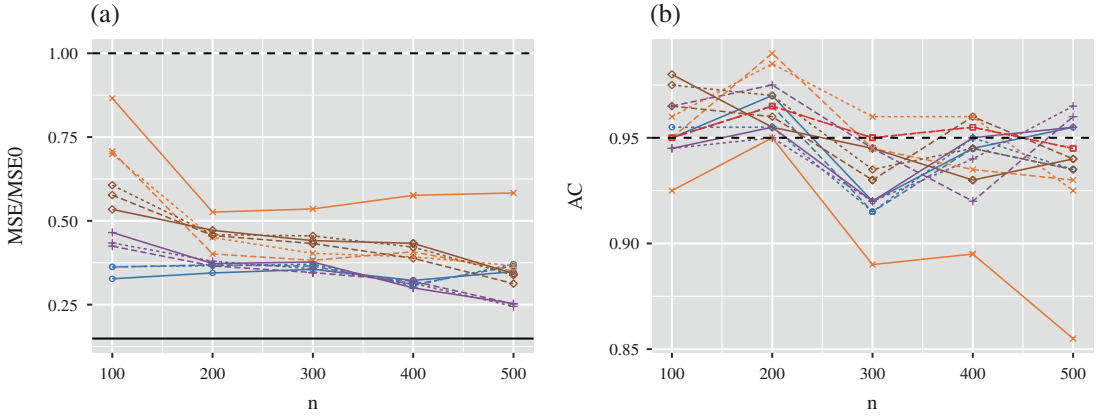

Fig. 5. Model 5: Is sample splitting needed? (a) The ratio of mean squared errors. (b) The average coverage. The plot includes sample mean (red squares), SSL-Lasso (blue circles), SSL-XGBoost (purple pluses), SSL-MLP (orange crosses) and SSL-RF (brown diamonds) estimates. The number of folds, K, is denoted with solid, dashed and long dashed lines for $K = 1$ (without cross-fitting), $K = 5$ and $K = 20$, respectively.

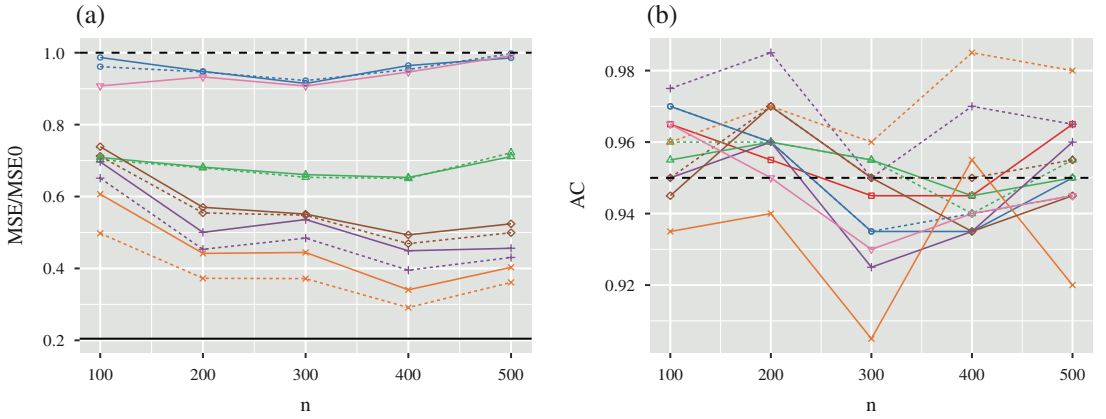

Fig. 6. Model 6: Does partitioning matter? (a) The ratio of mean squared errors. (b) The average coverage. The plot includes sample mean (red squares), SSL-OLS (blue circles), SSL-additive (green up triangles), SSL-XGBoost (purple pluses), SSL-MLP (orange crosses), SSL-RF (brown diamonds) and SSL-SSLS (pink down triangles) estimates. The number of cross-fitting repetitions, S, is denoted with solid and dashed lines for $S = 1$ and $S = 5$, respectively.

Model 5. Is sample splitting needed? Let $X_i \overset{\text{iid}}{\sim} \text{Lognormal}_{p-1}(0, C)$, with C as in Model 3 and $p = 101$, $m = 10n$. Let $Y_i = \sum_{j=1}^{3}\{\log(X_{ij}+1)^2 + 0.1\} + \delta_i$, where $\delta_i \overset{\text{iid}}{\sim} N(0, 0.25)$. We varied K from 1 to 5 and then to 20. We observe that some methods, like SSL-MLP, benefit significantly from sample splitting: without it, they under-cover, Fig. 5(b), and have the largest MSE, Fig. 5(a).

Model 6. In finite samples, the randomness from the K-partition creates an additional variance. We repeat the random K-partition S times, and for each time we obtain an estimate $\hat{\theta}^s$ and the corresponding estimated asymptotic variance $\hat{V}(\hat{\theta}^s)$. Here we compare $\hat{\theta}^1$ with the average $\tilde{\theta} = S^{-1}\sum_{s=1}^{S}\hat{\theta}^s$. An asymptotic confidence interval based on $\tilde{\theta}$ can be constructed using an estimated variance $\tilde{V}(\tilde{\theta}) = S^{-1}\sum_{s=1^S}\{\hat{V}(\hat{\theta}^s) + (\hat{\theta}^s - \tilde{\theta})^2\}$. The outcome model is nonlinear with one interaction term, $Y_i = X_{i1}X_{i2} + 0.5(X_{i3}+0.5)^2 + \delta_i$, and X_i and δ_i are as in Model 1 with $p = 4$, $m = 10n$. Figures 6(a) and 6(b) illustrate that partitions do not matter much for the least-squares procedure: SSL-lasso, SSL-additive and SSL-RF do not vary much. However, highly

Table 1. *Experiments for the average treatment effect*

Estimator	Bias	Emp SE	ASE	RMSE	AC
$n = 100, m = 200$			Linear outcome		
Zhang & Bradic (ridge+ridge)	0.0010	0.0881	0.0812	0.0879	0.935
Chernozhukov et al. (2017) (ridge+ridge)	0.0097	0.1295	0.1238	0.1295	0.930
Cheng et al. (2020)	−0.0147	0.0885	0.0801	0.0895	0.925
$n = 500, m = 1000$			Linear outcome		
Zhang & Bradic (ridge+ridge)	−0.0025	0.0333	0.0351	0.0333	0.945
Chernozhukov et al. (2017) (ridge+ridge)	−0.0052	0.0588	0.0546	0.0588	0.965
Cheng et al. (2020)	−0.0093	0.0329	0.0352	0.0341	0.940
$n = 200, m = 400$			Nonlinear outcome		
Zhang & Bradic (ridge+ridge)	0.0031	0.0660	0.0672	0.0659	0.965
Chernozhukov et al. (2017) (ridge+ridge)	0.0051	0.0714	0.0737	0.0714	0.955
Zhang & Bradic (additive+ridge)	0.0027	0.0622	0.0638	0.0621	0.960
Chernozhukov et al. (2017) (additive+ridge)	0.0054	0.0705	0.0731	0.0706	0.960
Zhang & Bradic (mlp+ridge)	−0.0027	0.0518	0.0497	0.0518	0.935
Chernozhukov et al. (2017) (mlp+ridge)	0.0015	0.0570	0.0596	0.0569	0.960
Cheng et al. (2020)	−0.0209	0.0637	0.0655	0.0669	0.970
$n = 500, m = 1000$			Nonlinear outcome		
Zhang & Bradic (ridge+ridge)	−0.0005	0.0384	0.0413	0.0383	0.970
Chernozhukov et al. (2017) (ridge+ridge)	−0.0014	0.0433	0.0457	0.0432	0.955
Zhang & Bradic (additive+ridge)	−0.0001	0.0385	0.0395	0.0383	0.975
Chernozhukov et al. (2017) (additive+ridge)	−0.0006	0.0436	0.0455	0.0435	0.960
Zhang & Bradic (mlp+ridge)	−0.0025	0.0256	0.0275	0.0255	0.975
Chernozhukov et al. (2017) (mlp+ridge)	−0.0017	0.0361	0.0354	0.0361	0.940
Cheng et al. (2020)	−0.0143	0.0377	0.0408	0.0402	0.945

Bias, average of the estimation biases; Emp SE, empirical standard error; ASE, average of estimated standard errors; RMSE, root-mean-square error; AC, average coverage of the 95% confidence intervals.

nonlinear methods, such as SSL-MLP and SSL-XGBoost, benefit significantly from repeating the partitioning process.

Model 7. (Average treatment effect). Consider $X_{ij} \overset{\text{iid}}{\sim} \text{Un}(-1, 1)$ with $p = 11$, and $D_i \sim \text{Ber}[1/\{1+\exp(5^{1/2}\sum_{j=1}^{5} X_{ij}/2)\}]$. In the linear setting, the outcome model is $Y_i = D_i(1+\beta_1^{\text{T}}X_i)+(1-D_i)\beta_0^{\text{T}}X_i + \delta_i$, where $\delta_i \overset{\text{iid}}{\sim} N(0, 0.2^2)$ and $\beta_0 = -(0.5^{1/2}, 0.5, 0.5^{3/2}, 0.5^2, 0.5^2, 0, 0, 0, 0, 0)$, $\beta_1 = -\beta_0$. In the nonlinear setting, the outcome model is $Y_i = D_i\{X_{i1}X_{i2} + 0.5(X_{i3} + 0.5)^2\} + (1 - D_i)\{X_{i1}X_{i2} - 0.5(X_{i3} + 0.5)^2\} + \delta_i$. For the linear setting our proposed estimator and the estimator of Chernozhukov et al. (2017) estimate the propensity and the outcome model by cross-validated generalized and linear ridge regression. For the nonlinear setting, the outcome models are estimated by ridge regression, additive model and multilayer perceptron. The parameters α and β of Cheng et al. (2020) are estimated by cross-validated adaptive lasso, where the initial weights are estimated by linear regression or generalized linear regression; the parameter γ is estimated by cross-validated lasso; the kernel is chosen to be sixth-order Gaussian, and the bandwidth is estimated by the plug-in method. Table 1 contains all the results. We found that the biases of our SSL and the supervised estimator of Chernozhukov et al. (2017) are not sensitive to the choice of the tuning parameters, while the bias of Cheng et al. (2020) is. Under the linear outcome models, the two SSL estimators have smaller mean squared errors than the supervised estimator; under nonlinear outcome models, our semi-supervised mlp+ridge estimator outperforms the others.

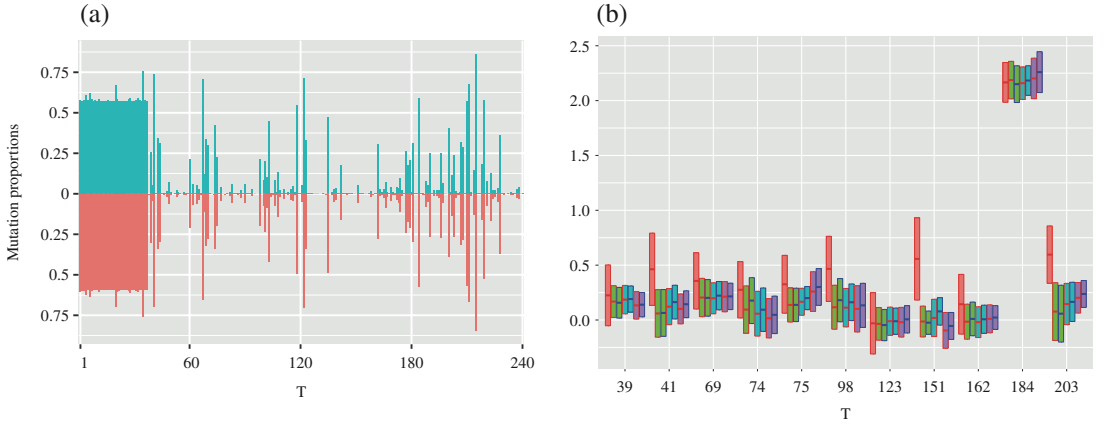

Fig. 7. Real data. (a) A back-to-back bar chart comparing the labelled and unlabelled groups' mutation proportions on reverse transcriptase positions between 1 and 240. The blue colour on the top denotes the unlabelled group, and the red colour on the bottom denotes the labelled group. (b) Confidence intervals of the average treatment effect. We compare the sample mean of the labelled samples (red border and red fill), supervised Chernozhukov et al. (2017) estimators (red border), and our SSL-method estimators (blue border). Estimators of the propensity score and the outcome model are: logistic + lasso (green fill), XGBoost + XGBoost (aqua fill), RF + RF (purple fill).

5.2. *HIV drug resistance*

We consider the dataset of Baxter et al. (2006), available at the Stanford University HIV Drug Resistance Database (Rhee et al., 2003), `https://hivdb.stanford.edu`. It is known that mutations are common in HIV, and some of the mutations may affect HIV drug resistance. We provide estimation and inference for the average treatment effect of a specific mutation on the reverse transcriptase to the drug resistance. The outcome is lamivudine (3TC), a nucleoside reverse transcriptase inhibitor (NRTI), drug resistance. The treatment, D, denotes the existence of a mutation on the Tth position of the HIV's reverse transcriptase. Explanatory variables X_j, where $j \in \{1, 2, \ldots, 240\} \setminus \{T\}$, denote the existence of a mutation on the jth position. We consider the subtype B sequence. Redundant viruses obtained from the same individuals were excluded. We obtained $n = 423$ pairs of supervised data $(D_{i,T}, Y_i, \{X_{i,j}\}_{j \neq T})_{i=1}^{n}$ and $m = 2458$ pairs of additional unlabelled covariates $(D_{i,T}, \{X_{i,j}\}_{j \neq T})_{i=n+1}^{m+n}$. Fix $T \in \{1, 2, \ldots, 240\}$. Before we perform our semi-supervised methods, we first check whether there is a significant difference between the distribution of X in the two groups; see the back-to-back bar chart of the labelled and unlabelled groups' mutation proportions on different reverse transcriptase positions in Fig. 7(a). The p-value based on the Pearson statistic was obtained using a permutation distribution (Agresti & Klingenberg, 2005) and resulted in a value of 0.178. We do not have any significant evidence that the covariates' distributions differ between the supervised and unlabelled groups. Estimators of the propensity score and the outcome model are: (logistic) lasso + lasso, XGBoost + XGBoost, and random forest + random forest. In order to improve the stability of the estimator, we trim each $\hat{e}^{(-k)}(X_i)$ to $(0.01, 0.99)$. We compare with the sample estimator $(\sum_{i=1}^{n} D_i)^{-1} \sum_{i=1}^{n} D_i Y_i - \{\sum_{i=1}^{n}(1-D_i)\}^{-1} \sum_{i=1}^{n}(1-D_i) Y_i$, suitable only for homogeneous effects. Figure 7(b) shows the confidence intervals for δ on several positions based on different estimators. We can see that there is a large average treatment effect on position 184, a small average treatment effect on positions 39 and 69, and potentially a small average treatment effect on positions 41, 75 and 203. The sample estimator is most different from the rest on positions 41, 98, 151 and 203. The sample estimator is biased when the distribution of X on treated and control is different. It implies that the mutations on positions 41, 98, 151 and 203 are significantly dependent on the other positions' mutations. Moreover, our confidence intervals are shorter than those of Chernozhukov et al. (2017). This coincides with the fact that additional unlabelled data provide improved asymptotic efficiency.

ACKNOWLEDGEMENT

Zhang was previously at the Department of Mathematics, University of California San Diego during the initial preparation of this work. Bradic is also affiliated with the Halıcıoğlu Data Science Institute at the University of California San Diego.

SUPPLEMENTARY MATERIAL

Supplementary Material available at *Biometrika* online includes detailed proofs of all theoretical results and an additional section on the missing-at-random extension of our proposal.

REFERENCES

AGRESTI, A. & KLINGENBERG, B. (2005). Multivariate tests comparing binomial probabilities, with application to safety studies for drugs. *Appl. Statist.* **54**, 691–706.

ATHEY, S., IMBENS, G. W. & WAGER, S. (2018). Approximate residual balancing: Debiased inference of average treatment effects in high dimensions. *J. R. Statist. Soc.* B **80**, 597–623.

BANG, H. & ROBINS, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61**, 962–73.

BAXTER, J. D., SCHAPIRO, J. M., BOUCHER, C. A., KOHLBRENNER, V. M., HALL, D. B., SCHERER, J. R. & MAYERS, D. L. (2006). Genotypic changes in human immunodeficiency virus type 1 protease associated with reduced susceptibility and virologic response to the protease inhibitor tipranavir. *J. Virology* **80**, 10794–801.

BELLONI, A., CHERNOZHUKOV, V. & WANG, L. (2011). Square-root lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika* **98**, 791–806.

BOGDAN, M., VAN DEN BERG, E., SABATTI, C., SU, W. & CANDÈS, E. J. (2015). Slope-adaptive variable selection via convex optimization. *Ann. Appl. Statist.* **9**, 1103.

BRADIC, J., WAGER, S. & ZHU, Y. (2019). Sparsity double robust inference of average treatment effects. *arXiv:*1905.00744.

CAI, T. & GUO, Z. (2020). Semisupervised inference for explained variance in high-dimensional linear regression and its applications. *J. R. Statist. Soc.* B **82**, 391–419.

CHAKRABORTTY, A. & CAI, T. (2018). Efficient and adaptive linear regression in semi-supervised settings. *Ann. Statist.* **46**, 1541–72.

CHAPELLE, O., SCHÖLKOPF, B. & ZIEN, A. (2009). Semi-supervised learning. *IEEE Trans. Neural Networks* **20**, 542.

CHENG, D., ANANTHAKRISHNAN, A. & CAI, T. (2020). Robust and efficient semi-supervised estimation of average treatment effects with application to electronic health records data. *arXiv:*1804.00195v2.

CHERNOZHUKOV, V., CHETVERIKOV, D., DEMIRER, M., DUFLO, E., HANSEN, C. & NEWEY, W. (2017). Double/debiased/Neyman machine learning of treatment effects. *Am. Econ. Rev.* **107**, 261–5.

CHERNOZHUKOV, V., CHETVERIKOV, D., DEMIRER, M., DUFLO, E., HANSEN, C., NEWEY, W. & ROBINS, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Economet. J.* **21**, C1–C68.

EL ALAOUI, A., CHENG, X., RAMDAS, A., WAINWRIGHT, M. J. & JORDAN, M. I. (2016). Asymptotic behavior of ℓ_p-based Laplacian regularization in semi-supervised learning. *Proc. Mach. Learn. Res.* **49**, 879–906.

GEISSER, S. (1975). The predictive sample reuse method with applications. *J. Am. Statist. Assoc.* **70**, 320–28.

GRANDVALET, Y. & BENGIO, Y. (2005). Semi-supervised learning by entropy minimization. In *Proc. 17th Int. Conf. Neural Information Processing Systems*, eds. L. K. Saul, Y. Weiss & L. Bottou, Cambridge, MA: MIT Press, pp. 529–36.

GRONSBELL, J. L. & CAI, T. (2018). Semi-supervised approaches to efficient evaluation of model prediction performance. *J. R. Statist. Soc.* B **80**, 579–94.

HOLLAND, P. W. (1988). Causal inference, path analysis and recursive structural equations models. *Sociol. Methodol.* **18**, 449–84.

KÜNZEL, S. R., SEKHON, J. S., BICKEL, P. J. & YU, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proc. Nat. Acad. Sci.* **116**, 4156–65.

MAI, X. & COUILLET, R. (2018). A random matrix analysis and improvement of semi-supervised learning for large dimensional data. *J. Mach. Learn. Res.* **19**, 3074–100.

RHEE, S.-Y., GONZALES, M. J., KANTOR, R., BETTS, B. J., RAVELA, J. & SHAFER, R. W. (2003). Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Research* **31**, 298–303.

RINALDO, A., WASSERMAN, L. & G'SELL, M. (2019). Bootstrapping and sample splitting for high-dimensional, assumption-lean inference. *Ann. Statist.* **47**, 3438–69.

ROSENBAUM, P. R. & RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.
ROTNITZKY, A., LEI, Q., SUED, M. & ROBINS, J. M. (2012). Improved double-robust estimation in missing data and causal inference models. *Biometrika* **99**, 439–56.
RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66**, 688.
SCHARFSTEIN, D. O., ROTNITZKY, A. & ROBINS, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *J. Am. Statist. Assoc.* **94**, 1096–120.
SMUCLER, E., ROTNITZKY, A. & ROBINS, J. M. (2019). A unifying approach for doubly-robust ℓ_1 regularized estimation of causal contrasts. *arXiv:*1904.03737v3.
SPLAWA-NEYMAN, J., DABROWSKA, D. M. & SPEED, T. P. (1990). On the application of probability theory to agricultural experiments. *Statist. Sci.* **5**, 465–72.
STONE, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. R. Statist. Soc.* B **36**, 111–33.
SUN, Q., ZHOU, W.-X. & FAN, J. (2020). Adaptive huber regression. *J. Am. Statist. Assoc.* **115**, 254–65.
TAN, Z. (2020a). Model-assisted inference for treatment effects using regularized calibrated estimation with high-dimensional data. *Ann. Statist.* **48**, 811–37.
TAN, Z. (2020b). Regularized calibrated estimation of propensity scores with model misspecification and high-dimensional data. *Biometrika* **107**, 137–58.
TIBSHIRANI, R. (1997). The lasso method for variable selection in the Cox model. *Statist. Med.* **16**, 385–95.
VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. & DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42**, 1166–202.
WAGER, S. & ATHEY, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *J. Am. Statist. Assoc.* **113**, 1228–42.
WASSERMAN, L. & LAFFERTY, J. D. (2008). Statistical analysis of semi-supervised regression. In *Proc. 20th Int. Conf. Neural Information Processing Systems*, eds. J. C. Platt, D. Koller, Y. Singer & S. T. Roweis, Cambridge, MA: MIT Press, pp. 801–08.
YE, F. & ZHANG, C.-H. (2010). Rate minimaxity of the lasso and Dantzig selector for the ℓ_q loss in ℓ_r balls. *J. Mach. Learn. Res.* **11**, 3519–40.
ZHANG, A., BROWN, L. D. & CAI, T. T. (2019). Semi-supervised inference: General theory and estimation of means. *Ann. Statist.* **47**, 2538–66.
ZHU, X. (2005). Semi-supervised learning literature survey. *World* **10**, 10.
ZOU, H. & HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Statist. Soc.* B **67**, 301–20.

[*Received on* 29 *May* 2019. *Editorial decision on* 8 *April* 2021]