Electronic Journal of Statistics
Vol. 18 (2024) 4640–4666
ISSN: 1935-7524
https://doi.org/10.1214/24-EJS2311

Robust improvement of efficiency using information on covariate distribution

Lu Mao

Department of Biostatistics and Medical Informatics, 207A WARF Office Building, 610 Walnut St., University of Wisconsin-Madison,
e-mail: lmao@biostat.wisc.edu

Abstract: The marginal inference of an outcome variable can be improved by closely related covariates with a structured distribution. This differs from standard covariate adjustment in randomized trials, which exploits covariate-treatment independence rather than knowledge on the covariate distribution. Yet it can also be done robustly against misspecification of the outcome-covariate relationship. Starting with a standard estimating function involving only the outcome, we first use a working regression model to compute its conditional expectation given the covariates, and then remove the uninformative part under the covariate distribution model. This effectively projects the initial function onto the joint tangent space of the full data, thereby achieving local efficiency when the regression model is correct. Importantly, even with a faulty working model, the estimator remains unbiased as the subtracted term is always asymptotically centered. Further improvement is possible if the outcome distribution also has its own structure. To demonstrate the process, we consider three examples: one with fully parametric covariates, one with a covariate following a partial parametric model against others, and another with mutually independent covariates.

MSC2020 subject classifications: Primary 62G35; secondary 62E20.
Keywords and phrases: Augmented estimator, efficient influence, projection, semiparametric theory, tangent space.

Received September 2023.

1. Introduction

Can the inference of an outcome Y benefit from knowledge on a closely related predictor X? Let Y be an indicator of mortality in COVID-19 patients, for example, and let X denote patient age. To estimate the overall mortality rate $\theta = E(Y)$ from a random n-sample (Y_i, X_i) $(i = 1, \ldots, n)$ of (Y, X), an immediate estimator is just the sample proportion $\mathbb{P}_n Y$, where $\mathbb{P}_n$ denotes the empirical measure, i.e., $\mathbb{P}_n f(Y, X) = n^{-1} \sum_{i=1}^{n} f(Y_i, X_i)$. However, because mortality rate differs widely across age groups (Koh, Geller and VanderWeele, 2021), it may be advantageous to start with the age-specific rates, and then average them by the age distribution in the population. This leads to $\hat{\theta}_{\text{adj}} = \sum_x p_x \{\mathbb{P}_n Y I(X = x)\}/\{\mathbb{P}_n I(X = x)\}$, where p_x is the population proportion of age group x, which may be obtained from census data, and $I(\cdot)$ is the indicator function. The hope is that the knowledge of p_x for the distribution of X helps to improve the

estimate of $E(Y)$. Indeed, standard analysis finds that $n\text{var}(\hat{\theta}_{\text{adj}}) \to E\{\text{var}(Y \mid X)\} = \text{var}(Y) - \text{var}\{E(Y \mid X)\}$. In this case, knowing the covariate distribution does increase efficiency of inference on the outcome, by eliminating outcome variations between different levels of the covariate.

Generalization of the above example faces two challenges. First, with multiple, continuous covariates, the curse of dimensionality makes it infeasible to estimate the covariate-specific mean $E(Y \mid X = x)$ nonparametrically. Instead, we need to rely on a lower-dimensional regression model. As a result, we can only hope for local efficiency when the regression model is correctly specified. This brings us to the more serious issue of robustness: a wrong model for the outcome-covariate relationship could not only fail to improve efficiency but also introduce bias, making information-borrowing counterproductive.

The ideal approach is to develop a robust system that requires only a working model for the outcome-covariate relationship. Such a system should improve efficiency when the model is correct, while still yielding valid results if the model is misspecified. In current semiparametric literature (Bickel et al., 1993), this type of robustness is typically achieved by combining auxiliary data for efficiency gain with a simple, standard estimator to guard against bias. For example, in a randomized trial, the inference of $E(\tilde{Y} \mid Z)$ — with $\tilde{Y}$ representing the response and Z the randomized group — can be made more efficient by adjusting for baseline covariates X through a working model. Any bias introduced by misspecification of the working model can be corrected by the unadjusted estimator based on $\tilde{Y}$ and Z alone (Tsiatis et al., 2008; Benkeser et al., 2021). In this case, with $Y = (\tilde{Y}, Z)$, the additional information comes not from the distribution of X but from the fact that $Z \perp\!\!\!\perp X$ (an aspect of Y-X relationship) due to randomization. Similar independence conditions underlie most doubly robust estimators in missing data problems (Robins, Rotnitzky and Zhao, 1994, 1995).

Our problem is fundamentally different. We are agnostic about the Y-X relationship but rather focus on any structure in the distribution of the latter. Such a structure can come in the form of exact knowledge (e.g., age distribution from census data), a parametric model with unknown parameters, or even qualitative features like independence among components, such as genetic versus environmental factors (Chatterjee, Kalaylioglu and Carroll, 2005).

To address this new problem, we return to the first principles of semiparametric inference (Bickel et al., 1993). Given a target estimand for the outcome, we first derive its theoretical efficient score by projecting a standard estimating function onto the joint tangent space of the outcome and covariates. This efficient score is then approximated under a working regression model and used as an improved estimating function. To ensure robustness, we modify the estimating function so that it remains centered even when the working model is misspecified.

2. General theory

2.1. Set-up

Consider estimating a d-dimensional parameter $\theta \equiv \theta(\mathbb{P}_Y) \in \mathbb{R}^d$, where $\mathbb{P}_Y$ is the probability measure of Y belonging to some model $\mathcal{M}_Y$. In the previous example, $\theta = \mathbb{P}_Y Y$ and $\mathcal{M}_Y$ is the nonparametric model $\mathcal{M}_Y^0$. From here on, we use $\mathbb{Q}f$ to denote the expectation of function f under a generic probability measure $\mathbb{Q}$. In addition, let $L_2(\mathbb{Q})$ and $L_2^0(\mathbb{Q})$ denote the space of $\mathbb{Q}$-square integrable functions and its subspace of mean-zero ones, respectively.

Without covariates, suppose that we can estimate θ using the Y-sample alone through a d-dimensional estimating function $\psi(Y;\theta)$ satisfying $\mathbb{P}_Y\psi(\cdot;\theta_0) = 0$, where θ_0 is the true value of θ. We can thus obtain a consistent initial estimator $\hat{\theta}_{\text{init}}$ by solving $\mathbb{P}_n\psi(Y;\hat{\theta}_{\text{init}}) \approx 0$. Mild regularity conditions guarantee that $\hat{\theta}_{\text{init}}$ is regular and asymptotically linear with expansion $n^{1/2}(\hat{\theta}_{\text{init}} - \theta_0) = -V(\theta_0)^{-1}n^{1/2}\mathbb{P}_n\psi(Y;\theta_0) + o_p(1)$, where $V(\theta) = \partial\mathbb{P}_Y\psi(Y;\theta)/\partial\theta$ (van der Vaart, 1998, Ch. 5). This means that the influence function of $\hat{\theta}_{\text{init}}$ is $-V(\theta_0)^{-1}\psi(Y;\theta_0)$, whose second moment corresponds to the asymptotic variance of the estimator. By standard theory, the most efficient estimator is the one with its influence function lying in the tangent space $\mathcal{H}_Y$, the closed linear span in $L_2^0(\mathbb{P}_Y)$ of all score functions generated by perturbing $\mathbb{P}_Y$ within model $\mathcal{M}_Y$ (see, e.g., Bickel et al., 1993, §3.2). As the influence function is a linear transform of the estimating function, it suffices that $\psi(Y;\theta_0) \in \mathcal{H}_Y$ component-wise. In the opening example, the initial estimator $\mathbb{P}_n Y$ is derived from $\psi(Y;\theta) = Y - \theta$, which is efficient since $\psi(Y;\theta_0) \in \mathcal{H}_Y = L_2^0(\mathbb{P}_Y)$. Any inefficient estimating function can be made efficient by projecting it orthogonally onto the tangent space $\mathcal{H}_Y$ (Bickel et al., 1993, §3.3). The projection gains efficiency by removing the noise in the score that is orthogonal to, and thus uncorrelated with, meaningful variations in $\mathcal{M}_Y$.

With concomitant X, the tangent space is no longer $\mathcal{H}_Y$ but needs to be re-derived. Let $\mathbb{P}_X$ denote the marginal probability measure of X constrained in a model $\mathcal{M}_X$, with corresponding tangent space $\mathcal{H}_X$. Because $\mathcal{M}_X$ represents our knowledge about $\mathbb{P}_X$, it should be smaller than the nonparametric model $\mathcal{M}_X^0$, so that $\mathcal{H}_X \neq L_2^0(\mathbb{P}_X)$. With $\mathbb{P}$ denoting the joint measure of (Y, X), it is then clear that $\mathbb{P} \in \mathcal{M}_Y \cap \mathcal{M}_X$.

2.2. Joint tangent space and efficient score

Let $L_2^0(\mathbb{P}_{Y|X}) = \{g \in L_2^0(\mathbb{P}) : E\{g(Y,X) \mid X\} = 0\}$ and $L_2^0(\mathbb{P}_{X|Y}) = \{g \in L_2^0(\mathbb{P}) : E\{g(Y,X) \mid Y\} = 0\}$. The next lemma, proved in Section 5, obtains the overall tangent space in a form amenable to projection operations. Let $\oplus$ denote the orthogonal sum.

Lemma 1 (Tangent space)**.** *The tangent space for* (Y, X) *under* $\mathcal{M}_Y \cap \mathcal{M}_X$ *is*

$$\mathcal{H} = \left\{L_2^0(\mathbb{P}_{Y|X}) \oplus \mathcal{H}_X\right\} \cap \left\{L_2^0(\mathbb{P}_{X|Y}) \oplus \mathcal{H}_Y\right\}. \tag{1}$$

In particular, if $\mathcal{M}_Y = \mathcal{M}_Y^0$ *and thus* $\mathcal{H}_Y = L_2^0(\mathbb{P}_Y)$*, then* $\mathcal{H} = L_2^0(\mathbb{P}_{Y|X}) \oplus \mathcal{H}_X$*.*

Use $\mathcal{H}_{X,\perp}$ and $\mathcal{H}_{Y,\perp}$ to denote the orthocomplements of $\mathcal{H}_X$ in $L_2^0(\mathbb{P}_X)$ and $\mathcal{H}_Y$ in $L_2^0(\mathbb{P}_Y)$, respectively. Because $L_2^0(\mathbb{P}) = L_2^0(\mathbb{P}_{Y|X}) \oplus L_2^0(\mathbb{P}_X) = L_2^0(\mathbb{P}_{X|Y}) \oplus L_2^0(\mathbb{P}_Y)$, the two intersecting spaces on the right hand side of (1) are just the orthocomplements of $\mathcal{H}_{X,\perp}$ and $\mathcal{H}_{Y,\perp}$ in $L_2^0(\mathbb{P})$, respectively. As a result, projection onto each space can be viewed as removing the noise under the corresponding marginal model. To carry this out, we will need the following operations.

Definition 1. *Define operator* $A : L_2(\mathbb{P}_Y) \to L_2(\mathbb{P}_X)$ *by* $Aa(X) = E\{a(Y) \mid X\}$ *and thus its adjoint* $A^{\mathrm{T}} : L_2(\mathbb{P}_X) \to L_2(\mathbb{P}_Y)$ *by* $A^{\mathrm{T}}h(X) = E\{h(X) \mid Y\}$*. Moreover, let* $\Pi_{\mathcal{H}_X}$ *and* $\Pi_{\mathcal{H}_Y}$ *denote the projection operators onto* $\mathcal{H}_X$ *and* $\mathcal{H}_Y$ *in* $L_2(\mathbb{P}_X)$ *and* $L_2(\mathbb{P}_Y)$*, respectively.*

Let $\Pi(\cdot \mid \mathcal{B})$ denote projection onto a subspace $\mathcal{B}$ in $L_2^0(\mathbb{P})$.

Lemma 2. *Given* $a(Y) \in L_2^0(\mathbb{P}_Y)$ *and* $h(X) \in L_2^0(\mathbb{P}_X)$*, we have that*

$$\Pi\left\{a(Y) \mid L_2^0(\mathbb{P}_{Y|X}) \oplus \mathcal{H}_X\right\} = a(Y) - (\mathcal{I}_X - \Pi_{\mathcal{H}_X})\, Aa(X),$$
$$\Pi\left\{h(X) \mid L_2^0(\mathbb{P}_{X|Y}) \oplus \mathcal{H}_Y\right\} = h(X) - (\mathcal{I}_Y - \Pi_{\mathcal{H}_Y})\, A^{\mathrm{T}}h(Y),$$

where $\mathcal{I}_X$ *and* $\mathcal{I}_Y$ *are the identity operators in* $L_2(\mathbb{P}_X)$ *and* $L_2(\mathbb{P}_Y)$*, respectively.*

Proof. For the first result,

$$\begin{aligned}
&\Pi\left\{a(Y) \mid L_2^0(\mathbb{P}_{Y|X}) \oplus \mathcal{H}_X\right\}\\
=&\Pi\left\{a(Y) \mid L_2^0(\mathbb{P}_{Y|X})\right\} + \Pi\left\{a(Y) \mid \mathcal{H}_X\right\}\\
=&a(Y) - \Pi\left\{a(Y) \mid L_2^0(\mathbb{P}_X)\right\} + \Pi\left[\Pi\left\{a(Y) \mid L_2^0(\mathbb{P}_X)\right\} \mid \mathcal{H}_X\right]\\
=&a(Y) - Aa(X) + \Pi_{\mathcal{H}_X} Aa(X)\\
=&a(Y) - (\mathcal{I}_X - \Pi_{\mathcal{H}_X})\, Aa(X),
\end{aligned}$$

where we have used $\Pi\left\{a(Y) \mid L_2^0(\mathbb{P}_X)\right\} = Aa(X)$. The other result follows similarly. □

Assume without loss of generality that $\psi(Y;\theta_0) \in \mathcal{H}_Y$ (otherwise we can always project it onto $\mathcal{H}_Y$ first). We can improve it in stages by first computing $\psi^{(1)}(Y,X;\theta_0) = \Pi\left\{\psi(Y;\theta_0) \mid L_2^0(\mathbb{P}_{Y|X}) \oplus \mathcal{H}_X\right\}$ component-wise. By Lemma 2,

$$\psi^{(1)}(Y,X;\theta_0) = \psi(Y;\theta_0) - (\mathcal{I}_X - \Pi_{\mathcal{H}_X})\, A\psi(\cdot;\theta_0)(X). \tag{2}$$

This will be our efficient score if $\mathcal{M}_Y = \mathcal{M}_Y^0$ (see remark below (1)). Otherwise, we can further project $\psi^{(1)}(Y;\theta_0)$ onto $L_2^0(\mathbb{P}_{X|Y}) \oplus \mathcal{H}_Y$, then back to $L_2^0(\mathbb{P}_{Y|X}) \oplus \mathcal{H}_X$, and so forth. That is, for $j = 1, 2, \ldots$, compute

$$\begin{aligned}
\psi^{(2j)}(Y,X;\theta_0) &= \Pi\left\{\psi^{(2j-1)}(Y,X;\theta_0) \mid L_2^0(\mathbb{P}_{X|Y}) \oplus \mathcal{H}_Y\right\}\\
\text{and } \psi^{(2j+1)}(Y,X;\theta_0) &= \Pi\left\{\psi^{(2j)}(Y,X;\theta_0) \mid L_2^0(\mathbb{P}_{Y|X}) \oplus \mathcal{H}_X\right\}.
\end{aligned} \tag{3}$$

The iteration will converge to the efficient score $\Pi\{\psi(Y;\theta_0) \mid \mathcal{H}\}$. However, if there is no structure to $\mathcal{M}_X$, we have that $\mathcal{H}_X = L_2^0(\mathbb{P}_X)$ so that $\Pi_{\mathcal{H}_X} = \mathcal{I}_X$. Consequently, $\psi(Y;\theta_0)$ will simply stay the same through (3) and no efficiency will be gained. The next theorem shows how to carry out the iteration in (3) in general.

Theorem 1 (Efficient score)**.** *If $\mathcal{M}_Y = \mathcal{M}_Y^0$, then $\psi^{(1)}(Y,X;\theta_0)$ is the efficient score for θ. Otherwise, let $\kappa^{(1)}(X;\theta_0) = (\mathcal{I}_X - \Pi_{\mathcal{H}_X})\, A\psi(\cdot;\theta_0)(X)$. Then the projections in (3) are given by $\psi^{(2j)}(Y,X;\theta_0) = \psi^{(2j-1)}(Y,X;\theta_0) + \kappa^{(2j)}(Y;\theta_0)$ and $\psi^{(2j+1)}(Y,X;\theta_0) = \psi^{(2j)}(Y,X;\theta_0) - \kappa^{(2j+1)}(X;\theta_0)$, where*

$$\kappa^{(2j)}(Y;\theta_0) = (\mathcal{I}_Y - \Pi_{\mathcal{H}_Y})\, A^{\mathrm{T}}\kappa^{(2j-1)}(\cdot;\theta_0)(Y),$$
$$\text{and} \qquad \kappa^{(2j+1)}(X;\theta_0) = (\mathcal{I}_X - \Pi_{\mathcal{H}_X})\, A\kappa^{(2j)}(\cdot;\theta_0)(X). \tag{4}$$

Hence for $k = 1, 2, \ldots$, we have that

$$\mathbb{P}\left\{\psi^{(k)}(\cdot,\cdot;\theta_0)^{\otimes 2}\right\} = \mathbb{P}\left\{\psi^{(k-1)}(\cdot,\cdot;\theta_0)^{\otimes 2}\right\} - \mathbb{P}\left\{\kappa^{(k)}(\cdot;\theta_0)^{\otimes 2}\right\}, \tag{5}$$

where $\psi^{(0)}(Y,X;\theta_0) = \psi(Y;\theta_0)$ and $v^{\otimes 2} = vv^{\mathrm{T}}$ for any vector v. Moreover, $\mathbb{P}\{\kappa^{(k+1)}(\cdot;\theta_0)^{\otimes 2}\} \le \mathbb{P}\{\kappa^{(k)}(\cdot;\theta_0)^{\otimes 2}\}$, where two symmetric matrices $M_1 \le M_2$ means that $M_2 - M_1$ is nonnegative definite. Finally, $\|\psi^{(k)}(\cdot,\cdot;\theta_0) - \Pi\{\psi(\cdot;\theta_0) \mid \mathcal{H}\}\|_{L_2(\mathbb{P})} \to 0$ as $k \to \infty$, where $\|\cdot\|$ denotes the $L_2(\mathbb{P})$ norm.

Proof. To establish the iteration formulas, observe that, inductively,

$$\psi^{(2j-2)}(Y,X;\theta_0) \in L_2^0(\mathbb{P}_{X|Y}) \oplus \mathcal{H}_Y \quad \text{and } \psi^{(2j-1)}(Y,X;\theta_0) \in L_2^0(\mathbb{P}_{Y|X}) \oplus \mathcal{H}_X.$$

As a result, only the newly added $\kappa^{(2j-1)}(X;\theta_0)$ and $\kappa^{(2j)}(Y;\theta_0)$ require projections through Lemma 2, hence (4). The details are explained in Section 5. Next, (5) follows by the Pythagorean rule. The inequality $\mathbb{P}\{\kappa^{(k+1)}(\cdot;\theta_0)^{\otimes 2}\} \le \mathbb{P}\{\kappa^{(k)}(\cdot;\theta_0)^{\otimes 2}\}$ is a result of the projection operations in (4). The $L_2(\mathbb{P})$-convergence of $\psi^{(k)}$ follows by Theorem 1 of Halperin (1962). □

Remark 1. By Theorem 1, each iteration leads to a more efficient score, culminating in the semiparametric efficient one. Moreover, because the iterations amount to recursive projections between two subspaces, spatial intuition suggests that the convergence rate is geometrically fast, with each projection shrinking the distance by the cosine of the angle between the projectee and the target space. More formally, let $\mathcal{H}_1^{\perp}$ and $\mathcal{H}_2^{\perp}$ denote the orthocomplements of $\mathcal{H}$ in $L_2^0(\mathbb{P}_{X|Y}) \oplus \mathcal{H}_Y$ and $L_2^0(\mathbb{P}_{Y|X}) \oplus \mathcal{H}_X$, respectively. We show in the Supplementary Material (Mao, 2024) that

$$\|\psi^{(k)}(\cdot,\cdot;\theta_0) - \Pi\{\psi(\cdot;\theta_0) \mid \mathcal{H}\}\|_{L_2(\mathbb{P})} \le c\exp(-Mk) \text{ for some } c > 0, \tag{6}$$

where $M = \inf\{-\log\{|E(h_1h_2)|\} : \|h_j\|_{L_2(\mathbb{P})} \le 1, h_j \in \mathcal{H}_j^{\perp}, j = 1,2\} \in \mathbb{R}^+$ and plays the role of the negative log of the "cosine of the angle" between the two subspaces. In the case with $\mathcal{M}_Y = \mathcal{M}_Y^0$, we have that $\mathcal{H}_1^{\perp} = \{0\}$ so that $M = \infty$. Then (6) confirms that the first-step estimator ($k = 1$) achieves the efficient score.

In the opening example, the efficient score for θ is $\psi^{(1)}(Y,X;\theta_0) = Y - \theta_0 - \{\mu(X) - \theta_0 - \Pi_{\mathcal{H}_X}\mu(X)\}$, where $\mu(X) = E(Y \mid X)$ is the true age-specific mortality rate. For generality, suppose the age group distribution is not necessarily known exactly but only up to a model $\mathcal{M}_X$ indexed by parameter γ, that is, $\mathbb{P}(X = x) = p_x(\gamma)$. Then it can be shown easily that this efficient influence is achieved by $\hat{\theta}_{\text{adj}} = \sum_x p_x(\hat{\gamma})\{\mathbb{P}_n Y I(X = x)\}/\{\mathbb{P}_n I(X = x)\}$, where $\hat{\gamma}$ is any efficient estimator of γ. Straightforwardly, $E\{\psi^{(1)}(Y,X;\theta_0)^2\} = E\{\text{var}(Y \mid X)\} + \text{var}\{\Pi_{\mathcal{H}_X}\mu(X)\}$. Compared with the variance of the $\hat{\theta}_{\text{adj}}$ in Section 1, the extra term $\text{var}\{\Pi_{\mathcal{H}_X}\mu(X)\}$ reflects the cost of not knowing $\mathbb{P}_X$ exactly but only up to $\mathcal{M}_X$ (equivalently the cost of estimating γ). The term vanishes, of course, if we do know $\mathbb{P}_X$ exactly so that $\mathcal{H}_X = \{0\}$.

2.3. Robust augmented estimation

To apply Theorem 1, we need to know, or at least be able to approximate, operators A and A^{T}, which involve the yet unspecified relationship between Y and X. Let $\mathcal{M}^*_{Y|X}$ be a working model for $\mathbb{P}_{Y|X}$, the conditional measure of Y given X, indexed by a parameter ξ. We can then calculate $A\psi(\cdot;\theta)(X)$ in (2) by $\mu(X;\theta,\xi) = E^*\{\psi(Y;\theta) \mid X;\xi\}$, where $E^*(\cdot \mid X;\xi)$ denotes the conditional mean under $\mathcal{M}^*_{Y|X}$. To compute $\psi^{(1)}$, it is tempting to follow its form by taking $\psi^{(1)*}(Y;\theta,\xi) = \psi(Y;\theta) - \{\mu(X;\theta,\xi) - \hat{\Pi}_{\mathcal{H}_X}\mu(X;\theta,\xi)\}$, where $\hat{\Pi}_{\mathcal{H}_X}$ is some empirical approximation to $\Pi_{\mathcal{H}_X}$. This, however, leaves open the possibility that $\mathbb{P}\psi^{(1)*}(Y;\theta_0,\xi) \neq 0$ for any ξ due to $\mathbb{P}\mu(X;\theta_0,\xi) \neq 0$ under a wrong $\mathcal{M}^*_{Y|X}$. For robustness, we need to ensure that the subtracted term in the brackets is always centered. The following lemma provides a solution.

Lemma 3. *Given $\mu(X;\theta,\xi) \in L_2(\mathbb{P}_X)$, let $\hat{\mathbb{E}}_{\mathcal{M}_X}\mu(X;\theta,\xi)$ be an $\mathcal{M}_X$-efficient estimator of $\mathbb{P}\mu(X;\theta,\xi)$. Then*

$$\begin{aligned} &n^{1/2}(\mathbb{P}_n - \hat{\mathbb{E}}_{\mathcal{M}_X})\mu(X;\theta,\xi) \\ =&n^{1/2}\mathbb{P}_n\{\mu(X;\theta,\xi) - \mathbb{P}\mu(X;\theta,\xi) - \Pi_{\mathcal{H}_X}\mu(X;\theta,\xi)\} + o_p(1). \end{aligned} \tag{7}$$

Proof. As a regular estimator of $\mathbb{P}_X\mu(X;\theta,\xi)$, $\mathbb{P}_n\mu(X;\theta,\xi)$ has influence function $\mu(X;\theta,\xi) - \mathbb{P}\mu(X;\theta,\xi)$. Then any efficient estimator must have influence function $\Pi_{\mathcal{H}_X}\{\mu(X;\theta,\xi) - \mathbb{P}\mu(X;\theta,\xi)\} = \Pi_{\mathcal{H}_X}\mu(X;\theta,\xi)$ (Bickel et al., 1993, §3.3). Combine the two results to obtain (7). □

Therefore, $\hat{\mathbb{E}}_{\mathcal{M}_X}\mu(X;\theta,\xi)$ plays the dual role of centering $\mathbb{P}_n\mu(X;\theta,\xi)$ and achieving the desired projection in asymptotic expansion. To interpret the left hand side of (7), note that it is a difference between the empirical and $\mathcal{M}_X$-based estimators of $\mathbb{P}_X\mu(X;\theta,\xi)$, which can be considered as "noise" in light of the model.

Using Lemma 3, we can now construct a one-step estimator that is always consistent and asymptotically normal and that achieves the score $\psi^{(1)}$ when $\mathbb{P}_{Y|X} \in \mathcal{M}^*_{Y|X}$. This is done essentially by replacing the subtracted term on the right hand side of (2) with the left hand side of (7). For a rigorous exposition,

we need the following regularity conditions, most of which are standard in Z-estimation (van der Vaart and Wellner, 1996, Ch. 3.3). Assume that ξ belongs to some metric space (Ξ, ρ).

(C1) (Consistency): $\hat{\theta}_{\text{init}} \to_p \theta_0$ and $\rho(\hat{\xi}, \xi^*) \to_p 0$ for some $\xi^* \in \Xi$. The limit ξ^* is equal to the true parameter ξ_0 if $\mathbb{P}_{Y|X} \in \mathcal{M}^*_{Y|X}$.

(C2) (Donskerness): There exists $\delta > 0$ such that both $\{\psi(\cdot;\theta) : ||\theta - \theta_0|| < \delta\}$ and $\{\mu(\cdot;\theta;\xi) : ||\theta - \theta_0|| + \rho(\xi, \xi^*) < 2\delta\}$ are $\mathbb{P}$-Donsker, where $||\cdot||$ denotes the Euclidean norm.

(C3) (L_2-Continuity): As $\theta \to \theta_0$ and $\rho(\xi, \xi^*) \to 0$, we have that

$$\|\psi(\cdot;\theta) - \psi(\cdot;\theta_0)\|_{L_2(\mathbb{P})} \to 0 \text{ and } \|\mu(\cdot;\theta,\xi) - \mu(\cdot;\theta_0,\xi^*)\|_{L_2(\mathbb{P})} \to 0.$$

(C4) (Uniformity): The expansion (7) holds uniformly on $\{(\theta, \xi) : ||\theta - \theta_0|| + \rho(\xi, \xi^*) < 2\delta\}$.

(C5) (Nonsingularity): $V(\theta_0) = \partial \mathbb{P}\psi(Y;\theta)/\partial\theta|_{\theta=\theta_0}$ is nonsingular.

Theorem 2 (One-step estimator)**.** *Define the one-step augmented estimating function*

$$\varphi_n(\theta) = \mathbb{P}_n\psi(Y;\theta) - (\mathbb{P}_n - \hat{\mathbb{E}}_{\mathcal{M}_X})\mu(X;\hat{\theta}_{\text{init}},\hat{\xi})$$

and let $\hat{\theta}$ be such that $\varphi_n(\hat{\theta}) = o_p(n^{-1/2})$. Then under $\mathcal{M}_Y \cap \mathcal{M}_X$ and (C1)–(C5),

$$n^{1/2}(\hat{\theta} - \theta_0) = -n^{1/2}V(\theta_0)^{-1}\mathbb{P}_n\Big[\psi(Y;\theta_0) - \{\mu(X;\theta_0,\xi^*) - \mathbb{P}\mu(\cdot;\theta_0,\xi^*) - \Pi_{\mathcal{H}_X}\mu(X;\theta_0,\xi^*)\}\Big] + o_p(1). \quad (8)$$

In $\mathcal{M}_Y \cap \mathcal{M}_X \cap \mathcal{M}^*_{Y|X}$, the influence function on the right hand side of (8) reduces to

$$-V(\theta_0)^{-1}\psi^{(1)}(Y, X;\theta_0),$$

which is the efficient influence if $\mathcal{M}_Y = \mathcal{M}^0_Y$.

Remarkably, the asymptotic distribution of $\hat{\theta}$ is unaffected by $\hat{\theta}_{\text{init}}$ and $\hat{\xi}$ apart from their respective consistency as required in (C1). This is yet expected of an augmented estimator whose validity does not depend on the augmentation term (see, e.g., Tsiatis, 2006, §10.2). We can thus use (8) to construct a robust sandwich-type variance estimator $\hat{\text{var}}(\hat{\theta}) = n^{-1}\hat{V}(\hat{\theta})^{-1}\mathbb{P}_n[\{\psi(Y;\hat{\theta}) - \hat{\kappa}^{(1)}(X;\hat{\theta},\hat{\xi})\}^{\otimes 2}]\hat{V}(\hat{\theta})^{\intercal -1}$, where $\hat{V}(\theta) = \partial\mathbb{P}_n\psi(Y;\theta)/\partial\theta$ and $\hat{\kappa}^{(1)}(X;\theta,\xi) = \mu(X;\theta,\xi) - \mathbb{P}_n\mu(\cdot;\theta,\xi) - \hat{\Pi}_{\mathcal{H}_X}\mu(X;\theta,\xi)$, with $\hat{\Pi}_{\mathcal{H}_X}$ denoting some empirical approximation to $\Pi_{\mathcal{H}_X}$.

By the same principle, we can construct augmented estimators based on the $\psi^{(k)}$ $(k = 2, 3, \ldots)$ inductively using (4) (needed only when $\mathcal{M}_Y \neq \mathcal{M}^0_Y$). The workflow is detailed in Section 5. Under $\mathcal{M}_Y \cap \mathcal{M}_X$, each step-$k$ estimator $\hat{\theta}^{(k)}$ is consistent and asymptotically linear with robustly estimable variance. When $\mathcal{M}^*_{Y|X}$ is true, the sequence of estimators gain increasing efficiency in the sense of Theorem 1, achieving local efficiency as $k \to \infty$.

3. Applications

3.1. Covariates following a parametric model

Suppose that $\mathcal{M}_X$ consists of a parametric family of distributions, i.e.

$$\mathcal{M}_X = \{\mathbb{P}_X : \mathrm{d}\mathbb{P}_X/\mathrm{d}\eta = p(\cdot;\gamma), \gamma \in \Gamma \subset \mathbb{R}^m\} \tag{9}$$

with some dominating measure η. To construct the one-step estimator, let $\hat{\gamma}$ denote the maximum likelihood estimator (MLE) of parameter γ. It is then clear that the efficient estimator for $\mathbb{P}\mu(X;\theta,\xi)$ is

$$\hat{\mathbb{E}}_{\mathcal{M}_X}\mu(X;\theta,\xi) = \int \mu(x;\theta,\xi)p(x;\hat{\gamma})\mathrm{d}\eta(x), \tag{10}$$

to be evaluated by, e.g., numerical or Monte-Carlo integration. To see the efficiency, recall that by standard theory, the influence function of $\hat{\gamma}$ is $I_\gamma^{-1}\dot{l}_\gamma(X)$, where $\dot{l}_\gamma(X)$ and I_γ are the score function and Fisher information for γ, respectively. Proceeding heuristically, we have that

$$\begin{aligned}
&n^{1/2}\{\hat{\mathbb{E}}_{\mathcal{M}_X}\mu(X;\theta,\xi) - \mathbb{P}\mu(\cdot;\theta,\xi)\}\\
=&\int \mu(x;\theta,\xi)n^{1/2}\{p(x;\hat{\gamma}) - p(x;\gamma)\}\mathrm{d}\eta(x)\\
=&\int \mu(x;\theta,\xi)\dot{l}_\gamma(x)p(x;\gamma)\mathrm{d}\eta(x)n^{1/2}(\hat{\gamma}-\gamma) + o_p(1)\\
=&\mathbb{P}\{\mu(\cdot;\theta,\xi)\dot{l}_\gamma(\cdot)^{\mathrm{T}}\}I_\gamma^{-1}n^{1/2}\mathbb{G}_n\dot{l}_\gamma(X) + o_p(1),
\end{aligned}$$

where $\mathbb{G}_n = n^{1/2}(\mathbb{P}_n - \mathbb{P})$. Hence the influence function is indeed $\Pi_{\mathcal{H}_X}\mu(X;\theta,\xi)$, given that $\mathcal{H}_X$ is the linear span of $\dot{l}_\gamma(X)$ (see Example 2 of Tsiatis, 2006, Ch. 2). In this case, the uniformity condition (C4) holds so long as the model $p(\cdot;\gamma)$ is sufficiently smooth. It is also straightforward to estimate the projection by

$$\hat{\Pi}_{\mathcal{H}_X}\mu(X;\theta,\xi) = \mathbb{P}_n\left\{\mu(\cdot;\theta,\xi)\dot{l}_{\hat{\gamma}}(\cdot)^{\mathrm{T}}\right\}\left[\mathbb{P}_n\left\{\dot{l}_{\hat{\gamma}}(\cdot)^{\otimes 2}\right\}\right]^{-1}\dot{l}_{\hat{\gamma}}(X). \tag{11}$$

We can now use (10) and (11) to construct $\varphi_n(\theta)$ and estimate the variance of the resulting $\hat{\theta}$. Write $V = V(\theta_0)$ and $\mu(X) = E\{\psi(Y;\theta_0) \mid X\}$. The next result follows easily from (5).

Proposition 1 (Efficiency gain by parametric covariates)**.** *With $\mathcal{M}_X$ defined in (9), if $\mathcal{M}^*_{Y|X}$ is correct, then*

$$\begin{aligned}
&\lim_{n\to\infty}\left[nV\{\mathrm{var}(\hat{\theta}_{\mathrm{init}}) - \mathrm{var}(\hat{\theta})\}V^{\mathrm{T}}\right]\\
=&\mathbb{P}\left\{\mu(X)^{\otimes 2}\right\} - \mathbb{P}\left\{\mu(X)\dot{l}_\gamma(X)^{\mathrm{T}}\right\}I_\gamma^{-1}\mathbb{P}\left\{\dot{l}_\gamma(X)\mu(X)^{\mathrm{T}}\right\}.
\end{aligned}$$

The limit attains maximum $\mathbb{P}\{\mu(X)^{\otimes 2}\}$ if and only if $\mathbb{P}\{\mu(X)\dot{l}_\gamma(X)^{\mathrm{T}}\} = 0$ and minimum 0 if and only if $\mu(X) = B\dot{l}_\gamma(X)$ for some matrix $B \in \mathbb{R}^{d\times m}$.

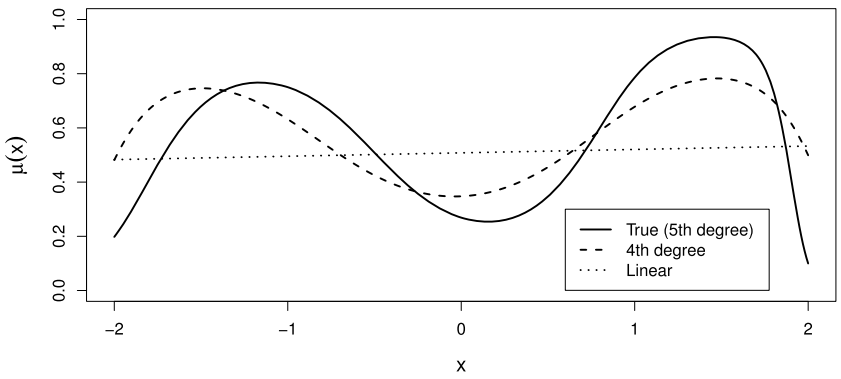

FIG 1. *True quintic curve of* $\mu(x) = \Pr(Y = 1 \mid X = x)$ *versus quartic and linear approximations.*

Example 1 (One Gaussian covariate)**.** In the opening example, suppose that, instead of a discrete X with known distribution, we have a continuous one following $N(\tau, \sigma^2)$ with unknown parameters. To gain efficiency over $\hat{\theta}_{\text{init}} = \mathbb{P}_n Y$, with $\psi(Y;\theta) = Y - \theta$, we use a working model $\mathcal{M}^*_{Y|X}$ to compute $\mu(X;\xi) = E^*(Y \mid X;\xi)$, where we drop a constant θ as is obviously allowed in Lemma 3. With $\gamma = (\tau, \sigma^2)$, compute $\hat{\mathbb{E}}_{\mathcal{M}_X}\mu(X;\xi) = \int \mu(x;\xi)\hat{\sigma}^{-1}\phi\{(x-\hat{\tau})/\hat{\sigma}\}\mathrm{d}x$, where $\phi(\cdot)$ is the density function of the standard normal distribution. Then the one-step, and also the locally efficient, estimator is just $\hat{\theta} = \mathbb{P}_n Y - \{\mathbb{P}_n \mu(X;\hat{\xi}) - \hat{\mathbb{E}}_{\mathcal{M}_X}\mu(X;\hat{\xi})\}$. Moreover, using the score vector $\dot{l}_\gamma(X) = \{X - \tau, (X-\tau)^2 - \sigma^2\}^{\mathrm{T}}$ we can calculate (11) explicitly for robust variance estimation. See Section 5 for details.

A simulation study is described below, in which the true $\mu(X;\xi_0)$ is a logistic function of a 5-degree polynomial of X. With correct or near-correct working models, $\hat{\theta}$ gains as much as 25% extra efficiency over $\mathbb{P}_n Y$. Meanwhile, gross misspecifications, such as fitting a straight line through the quintic curve, incur no apparent loss in accuracy or precision.

Specifically, we generated $X \sim N(\tau, \sigma^2)$ with $\tau = 0$ and $\sigma^2 = 1$, and Y from the quintic logistic regression model

$$\Pr(Y = 1 \mid X;\xi_0) = \frac{\exp(\xi_{01} + \xi_{02}X + \xi_{03}X^2 + \xi_{04}X^3 + \xi_{05}X^4 + \xi_{06}X^5)}{1 + \exp(\xi_{01} + \xi_{02}X + \xi_{03}X^2 + \xi_{04}X^3 + \xi_{05}X^4 + \xi_{06}X^5)}, \tag{12}$$

where $\xi_0 = (-1, -1, 3, 1.4, -0.8, -0.3)^{\mathrm{T}}$. Under this set-up, we have that $\theta_0 = 0.508$. We considered three working models for $\mathcal{M}^*_{Y|X}$: (a) a correct quintic logistic model; (b) a wrongly specified quartic, i.e., 4-degree polynomial, logistic model; and (c) a wrongly specified linear logistic model. Fig 1 shows approximations of the true $\mu(x) = \Pr(Y = 1 \mid X = x)$ by the latter two models. The quartic function is visibly inadequate, though it does capture some rough patterns of true curve. In comparison, the linear function completely misses important variations in the curve.

TABLE 1
Simulation results for efficient estimation of $\Pr(Y = 1)$ *with a Gaussian covariate.*

		Naive				Locally efficient				
n	Model	Bias	SE	SEE	CP	Bias	SE	SEE	CP	RE
200	+	0.000	3.52	3.53	0.953	−0.001	3.16	3.11	0.943	1.24
	−	0.001	3.55	3.53	0.950	0.001	3.26	3.21	0.943	1.19
	×	0.000	3.53	3.53	0.954	0.000	3.53	3.53	0.953	1.00
500	+	0.000	2.23	2.23	0.947	−0.001	1.99	1.97	0.949	1.26
	−	0.000	2.25	2.23	0.943	0.000	2.08	2.06	0.945	1.18
	×	0.000	2.24	2.23	0.946	0.000	2.24	2.23	0.947	1.00
1000	+	0.000	1.59	1.58	0.947	0.000	1.42	1.40	0.946	1.25
	−	0.000	1.58	1.58	0.950	0.000	1.46	1.44	0.949	1.17
	×	0.000	1.58	1.58	0.952	0.000	1.58	1.58	0.951	1.00
2000	+	0.000	1.11	1.12	0.952	0.000	1.00	0.98	0.947	1.25
	−	0.000	1.12	1.12	0.950	0.000	1.03	1.02	0.948	1.16
	×	0.000	1.11	1.12	0.950	0.000	1.11	1.12	0.949	1.00

+, correct quintic model; −, misspecified quartic model; ×, misspecified linear model. SE, empirical standard error of the estimator; SEE, empirical average of the standard error estimator; CP, empirical coverage rate of the 95% confidence interval. RE, relative efficiency, i.e., inverse ratio of the empirical variance, comparing the locally-efficient to the naive estimator. Each entry is based on 10,000 replicates.

Under each working model, we estimated its parameters ξ by standard logistic regression, and then used the procedures described above to construct the corresponding $\hat{\theta}$ and its robust variance. This was repeated on simulated samples of size $n = 200, 500, 1000$, and 2000. The results for both $\hat{\theta}$ and the naive $\hat{\theta}_{\text{init}} = \mathbb{P}_n Y$ are summarized side by side in Table 1. First, it is worth noting that all estimators are virtually unbiased, including the linear logistic model which grossly misspecifies the outcome-covariate relationship. In addition, the robust standard error estimators and associated confidence intervals also appear to be accurate, especially for the larger sample sizes. By successfully drawing on information about X, a correct working model allows $\hat{\theta}$ to gain as much as 25% extra efficiency over the naive $\hat{\theta}_{\text{init}}$. Interestingly, even the wrong quartic model results in substantial improvement, though to a lesser extent. Since the linear model is completely wrong, it fails to gain any efficiency over $\hat{\theta}_{\text{init}}$, but it does not lose any either.

3.2. Covariates following a conditional model

Let $X = (X_{\cdot 1}, X_{\cdot 2})$. Suppose $X_{\cdot 1}$ given $X_{\cdot 2}$ follows a parametric model but the distribution of $X_{\cdot 2}$ is unspecified. This can be suitable setup when $X_{\cdot 1}$ is some exposure, $X_{\cdot 2}$ is a set of baseline predictors, and both affect the outcome Y. More formally, let

$$\mathcal{M}_X = \{\mathbb{P}_X = \mathbb{P}_{X_{\cdot 1}|X_{\cdot 2}} \times \mathbb{P}_{X_{\cdot 2}} : \mathrm{d}\mathbb{P}_{X_{\cdot 1}|X_{\cdot 2}}(x_1 \mid x_2) = p(x_1 \mid x_2; \gamma)\mathrm{d}\eta(x_1), \\ \gamma \in \Gamma \subset \mathbb{R}^m, \text{ and } \mathbb{P}_{X_{\cdot 2}} \in \mathcal{M}^0_{X_{\cdot 2}}\}, \tag{13}$$

with some dominating measure η, where $\mathbb{P}_{X_{\cdot 1}|X_{\cdot 2}}$ and $\mathbb{P}_{X_{\cdot 2}}$ are the conditional law of $X_{\cdot 1}$ given $X_{\cdot 2}$ and the marginal law of $X_{\cdot 2}$, respectively.

Let $\hat{\gamma}$ denote an efficient estimator (e.g., MLE) of γ. By similar intuition to (10), we can show that an efficient estimator for $\mu(X;\theta,\xi)$ is

$$\hat{\mathbb{E}}_{\mathcal{M}_X}\mu(X;\theta,\xi)=\mathbb{P}_n\int\mu(x_1,X_{\cdot 2};\theta,\xi)p(x_1\mid X_{\cdot 2};\hat{\gamma})\mathrm{d}\eta(x_1), \tag{14}$$

that is, $\mu(X;\theta,\xi)$ weighted by the model-based conditional distribution of $X_{\cdot 1}$ given $X_{\cdot 2}$, and then averaged by the empirical distribution of the latter.

To show that (14) is indeed efficient, note that (13) entails a tangent space

$$\mathcal{H}_X=[\dot{l}_\gamma(X_{\cdot 1}\mid X_{\cdot 2})]\oplus L_2^0(\mathbb{P}_{X_{\cdot 2}}),$$

where $[\dot{l}_\gamma(X_{\cdot 1}\mid X_{\cdot 2})]$ is the linear span of the conditional score $\dot{l}_\gamma(X_{\cdot 1}\mid X_{\cdot 2})=\partial\log p(X_{\cdot 1}\mid X_{\cdot 2};\gamma)/\partial\gamma$. As a result, the projection is given by

$$\begin{aligned}\Pi_{\mathcal{H}_X}\mu(X;\theta,\xi)&=\Pi\left\{\mu(X;\theta,\xi)\mid[\dot{l}_\gamma(X_{\cdot 1}\mid X_{\cdot 2})]\right\}+\Pi\left\{\mu(X;\theta,\xi)\mid L_2^0(\mathbb{P}_{X_{\cdot 2}})\right\}\\&=\{\mathbb{P}\mu(\cdot;\theta,\xi)\dot{l}_\gamma(\cdot\mid\cdot)^{\mathrm{T}}\}I_\gamma^{-1}\dot{l}_\gamma(X_{\cdot 1}\mid X_{\cdot 2})\\&\quad+\mu_2(X_{\cdot 2};\theta,\xi,\gamma)-\mathbb{P}\mu(\cdot;\theta,\xi),\end{aligned}$$

where $I_\gamma=\mathbb{P}\dot{l}_\gamma(X_{\cdot 1}\mid X_{\cdot 2})^{\otimes 2}$ and $\mu_2(X_{\cdot 2};\theta,\xi,\gamma)=E\{\mu(X;\theta,\xi)\mid X_{\cdot 2};\gamma\}$.

On the other hand, (14) gives

$$\begin{aligned}&n^{1/2}\{\hat{\mathbb{E}}_{\mathcal{M}_X}\mu(X;\theta,\xi)-\mathbb{P}\mu(X;\theta,\xi)\}\\=&\mathbb{P}_n\int\mu(x_1,X_{\cdot 2};\theta,\xi)n^{1/2}\{p(x_1\mid X_{\cdot 2};\hat{\gamma})-p(x_1\mid X_{\cdot 2};\gamma)\}\mathrm{d}\eta(x_1)\\&+n^{1/2}\mathbb{P}_n\left\{\int\mu(x_1,X_{\cdot 2};\theta,\xi)p(x_1\mid X_{\cdot 2};\gamma)\mathrm{d}\eta(x_1)-\mathbb{P}\mu(\cdot;\theta,\xi)\right\}\\=&\mathbb{P}\int\mu(x_1,\cdot;\theta,\xi)\dot{l}_\gamma(x_1\mid\cdot)^{\mathrm{T}}p(x_1\mid\cdot;\gamma)\mathrm{d}\eta(x_1)n^{1/2}(\hat{\gamma}-\gamma)\\&+\mathbb{G}_n\mu_2(X_{\cdot 2};\theta,\xi,\gamma)+o_p(1)\\=&\{\mathbb{P}\mu(\cdot;\theta,\xi)\dot{l}_\gamma(\cdot\mid\cdot)^{\mathrm{T}}\}I_\gamma^{-1}\mathbb{G}_n\dot{l}_\gamma(X_{\cdot 1}\mid X_{\cdot 2})+\mathbb{G}_n\mu_2(X_{\cdot 2};\theta,\xi,\gamma)+o_p(1)\\=&\mathbb{G}_n\Pi_{\mathcal{H}_X}\mu(X;\theta,\xi)+o_p(1),\end{aligned}$$

where we have used the fact that $n^{1/2}(\hat{\gamma}-\gamma)=I_\gamma^{-1}\mathbb{G}_n\dot{l}_\gamma(X_{\cdot 1}\mid X_{\cdot 2})$. For robust variance estimation, we can approximate the influence function by

$$\begin{aligned}\hat{\Pi}_{\mathcal{H}_X}\mu(X;\theta,\xi)=&\{\mathbb{P}_n\mu(\cdot;\theta,\xi)\dot{l}_{\hat{\gamma}}(\cdot\mid\cdot)^{\mathrm{T}}\}\left[\mathbb{P}_n\dot{l}_{\hat{\gamma}}(\cdot\mid\cdot)^{\otimes 2}\right]^{-1}\dot{l}_{\hat{\gamma}}(X_{\cdot 1}\mid X_{\cdot 2})\\&+\mu_2(X_{\cdot 2};\theta,\xi,\hat{\gamma})-\mathbb{P}_n\mu_2(\cdot;\theta,\xi,\hat{\gamma}).\end{aligned} \tag{15}$$

Combining (14) and (15) allows us to construct $\hat{\theta}$ and its robust variance estimator using Theorem 2.

Write $V=V(\theta_0)$, $\mu(X)=E\{\psi(Y;\theta_0)\mid X\}$, and $\mu_2(X_{\cdot 2})=E\{\mu(X)\mid X_{\cdot 2}\}$. The efficiency gain by $\hat{\theta}$ can be quantified similarly to Proposition 1.

Proposition 2 (Efficiency gain by a partial model on covariates)**.** *With* $\mathcal{M}_X$ *defined in* (13)*, if* $\mathcal{M}^*_{Y|X}$ *is correct, then*

$$\begin{aligned}&\lim_{n\to\infty}\left[nV\{\mathrm{var}(\hat{\theta}_{\mathrm{init}})-\mathrm{var}(\hat{\theta})\}V^{\mathrm{T}}\right]\\ =&\mathbb{P}\left\{\mu(X)^{\otimes 2}\right\}-\mathbb{P}\left\{\mu(X)\dot{l}_\gamma(X_{\cdot 1}\mid X_{\cdot 2})^{\mathrm{T}}\right\}I_\gamma^{-1}\mathbb{P}\left\{\dot{l}_\gamma(X_{\cdot 1}\mid X_{\cdot 2})\mu(X)^{\mathrm{T}}\right\}\\ &-\mathbb{P}\left\{\mu_2(X_{\cdot 2})^{\otimes 2}\right\}.\end{aligned}$$

The limit attains maximum $\mathbb{P}\{\mu(X)^{\otimes 2}\}$ *if and only if* $\mathbb{P}\left\{\mu(X)\dot{l}_\gamma(X_{\cdot 1}\mid X_{\cdot 2})^{\mathrm{T}}\right\}=0$ *and* $\mu_2(X_{\cdot 2})\equiv 0$*, and minimum 0 if and only if* $\mu(X)=B\dot{l}_\gamma(X_{\cdot 1}\mid X_{\cdot 2})+b(X_{\cdot 2})$ *for some matrix* $B\in\mathbb{R}^{d\times m}$ *and some* $b\in L_2^0(\mathbb{P}_{X_{\cdot 2}})^{\otimes d}$*.*

Example 2 (A logistic covariate model)**.** Let $X_{\cdot 1}=1,0$ be a binary exposure following a logistic regression model against a vector of predictors $X_{\cdot 2}$:

$$\Pr(X_{\cdot 1}=1\mid X_{\cdot 2};\gamma)=\frac{\exp(\gamma^{\mathrm{T}}X_{\cdot 2})}{1+\exp(\gamma^{\mathrm{T}}X_{\cdot 2})},\tag{16}$$

where we assume $X_{\cdot 2}$ includes 1 as a component to allow for an intercept. As in Example 1, consider a binary Y with an initial empirical estimator $\hat{\theta}_{\mathrm{init}}=\mathbb{P}_nY$, so that $\psi(Y;\theta)=Y-\theta$. Under a working model $\mathcal{M}^*_{Y|X}$, e.g., another binary regression model for Y against $X=(X_{\cdot 1},X_{\cdot 2})$, the projectee of interest is $\mu(X_{\cdot 1},X_{\cdot 2};\xi)=E^*(Y\mid X;\xi)$. Let $p(X_{\cdot 2};\hat{\gamma})=\Pr(X_{\cdot 1}=1\mid X_{\cdot 2};\hat{\gamma})$ in (16) and apply it to (14). We find that

$$\hat{\mathbb{E}}_{\mathcal{M}_X}\mu(X;\xi)=\mathbb{P}_n\left[p(X_{\cdot 2};\hat{\gamma})\mu(1,X_{\cdot 2};\xi)+\{1-p(X_{\cdot 2};\hat{\gamma})\}\mu(0,X_{\cdot 2};\xi)\right]$$

and that, after some algebraic manipulation, the locally efficient estimator has the nice form

$$\hat{\theta}=\mathbb{P}_n\left[Y-\{X_{\cdot 1}-p(X_{\cdot 2};\hat{\gamma})\}\left\{\mu(1,X_{\cdot 2};\hat{\xi})-\mu(0,X_{\cdot 2};\hat{\xi})\right\}\right].\tag{17}$$

We use this setup to explore how the results hold up for a possibly high-dimensional $X_{\cdot 2}$ in simulations. By (17), we only need two classification models to calculate $\hat{\theta}$: one for $p(X_{\cdot 2};\hat{\gamma})$, the class probability of $X_{\cdot 1}=1$ given $X_{\cdot 2}$; the other for $\mu(X_{\cdot 1},X_{\cdot 2};\hat{\xi})$, the class probability of $Y=1$ given X. In the Supplementary Material (Mao, 2024), we consider two logistic models for $X_{\cdot 1}$ and Y with a p-dimension of $X_{\cdot 2}$, where $p=10,50$, and 500. Under sample size $n=200,500$, and 1000, we use L_1-regularized logistic regression to perform the classifications. The resulting $\hat{\theta}$ has a clear efficiency gain over the naive estimator when $p<<n$. While its advantage diminishes as p becomes comparable to or larger than n, the augmented estimator remains unbiased, suggesting its robustness even in high-dimensional settings (see Section S2.1 of Supplementary Material (Mao, 2024)).

3.3. Mutually independent covariates

Suppose that $X=(X_{\cdot 1},\ldots,X_{\cdot p})$ with independent components, i.e.,

$$\mathcal{M}_X=\{\mathbb{P}_X=\mathbb{P}_{X_{\cdot 1}}\times\cdots\times\mathbb{P}_{X_{\cdot p}}:\mathbb{P}_{X_{\cdot j}}\in\mathcal{M}^0_{X_{\cdot j}},j=1,\ldots,p\},\tag{18}$$

where $\mathbb{P}_{X_{\cdot j}}$ is the marginal law of $X_{\cdot j}$ and $\mathcal{M}^0_{X_{\cdot j}}$ the corresponding nonparametric model (no restriction on the component-wise distributions). Due to mutual independence, a sample of $(X_{\cdot 1}, \ldots, X_{\cdot p})$ is equivalent to p independent samples of the $X_{\cdot j}$ $(j = 1, \ldots, p)$. Intuitively, the expectation of $\mu(X; \theta, \xi) = \mu(X_{\cdot 1}, \ldots, X_{\cdot p}; \theta, \xi)$ should be efficiently estimated by a p-sample U-statistic.

Lemma 4. *For $\mathcal{M}_X$ in (18), $\mathcal{H}_X = L_2^0(\mathbb{P}_{X_1}) \oplus \cdots L_2^0(\mathbb{P}_{X_p})$. Hence,*

$$\Pi_{\mathcal{H}_X} \mu(X; \theta, \xi) = \sum_{j=1}^{p} \mu_j(X_{\cdot j}; \theta, \xi),$$

where $\mu_j(x_j; \theta, \xi) = E\{\mu(X_{\cdot 1}, \ldots, X_{\cdot, j-1}, x_j, X_{\cdot, j+1} \ldots, X_{\cdot p}; \theta, \xi)\}$. In addition, an efficient estimator for $\mathbb{P}\mu(X; \theta, \xi)$ is

$$\hat{\mathbb{E}}_{\mathcal{M}_X} \mu(X; \theta, \xi) = (\mathbb{P}_{X_{\cdot 1}, n} \times \cdots \times \mathbb{P}_{X_{\cdot p}, n}) \mu(X; \theta, \xi), \tag{19}$$

where $\mathbb{P}_{X_{\cdot j}, n}$ denotes the empirical measure based on the n-sample of $X_{\cdot j}$.

Proof. The form of $\mathcal{H}$ is a simple consequence of the product measure. The projection formula follows immediately from Hájek (1968). That the right hand side of (10) yields the efficient influence $\sum_{j=1}^{p} \mu_j(X_{\cdot j}; \theta, \xi)$ can be proved directly using Theorem 11.2 of van der Vaart (1998). The details can be found in Section 5. □

With (19), the condition (C4) is implied by the Donskerness of $\mu(X; \theta, \xi)$ in (C2) by standard U-process theory (Arcones and Giné, 1993). We can estimate the projection by $\hat{\Pi}_{\mathcal{H}_X} \mu(X; \theta, \xi) = \sum_{j=1}^{p} \hat{\mu}_j(X_{\cdot j}; \theta, \xi)$, where $\hat{\mu}_j(x_j; \theta, \xi) = \mathbb{P}_n \mu(X_{\cdot 1}, \ldots, X_{\cdot, j-1},$
$x_j, X_{\cdot, j+1} \ldots, X_{\cdot p}; \theta, \xi)$. This along with (19) will allow us to construct $\hat{\theta}$ and estimate its variance.

Proposition 3 (Efficiency gain by independent covariates)**.** *With $\mathcal{M}_X$ defined in (18), if $\mathcal{M}^*_{Y|X}$ is correct, then*

$$\lim_{n \to \infty} \left[nV\{\mathrm{var}(\hat{\theta}_{\mathrm{init}}) - \mathrm{var}(\hat{\theta})\} V^{\mathrm{T}} \right] = \mathbb{P}\left\{\mu(X)^{\otimes 2}\right\} - \sum_{j=1}^{p} \mathbb{P}\left\{\mu_j(X_{\cdot j})^{\otimes 2}\right\},$$

where $\mu_j(x_j) = E\{\mu(X_{\cdot 1}, \ldots, X_{\cdot, j-1}, x_j, X_{\cdot, j+1} \ldots, X_{\cdot p})\}$. The limit attains maximum $\mathbb{P}\{\mu(X)^{\otimes 2}\}$ if and only if the $\mu_j(X) \equiv 0$, and minimum 0 if and only if $\mu(X) = \sum_{j=1}^{p} \mu_j(X_{\cdot j})$.

Example 3 (Two independent covariates)**.** Let Y be a trinomial random variable denoting a diploid genotype in $\{GG, Gg, gg\}$, with distribution following the Hardy–Weinberg law (basically assuming the two alleles are independent) (Edwards, 2008). Consider $\theta = \Pr(Y = gg) = q^2$, where q is the allele frequency of g. The maximum likelihood estimator of q is $\hat{q}_{\mathrm{init}} = 2^{-1}\{\mathbb{P}_n(Y_1 - Y_2) + 1\}$, where $Y_1 = I(Y = gg)$ and $Y_2 = I(Y = GG)$. Let $X_{\cdot 1}$ and $X_{\cdot 2}$ be two independent predictors of Y. We can then compute $\chi_k(X; \xi) = E^*(Y_k \mid X; \xi)$ $(k = 1, 2)$

by, e.g., a multinomial logistic regression. With details relegated to Section 5, we obtain the one-step estimator $\hat{q} = \hat{q}_{\text{init}} - 2^{-1}(\mathbb{P}_n - \mathbb{U}_n)\{\chi_1(X;\hat{\xi}) - \chi_2(X;\hat{\xi})\}$, where $\mathbb{U}_n h(X) = (\mathbb{P}_{X_{\cdot 1},n} \times \mathbb{P}_{X_{\cdot 2},n})h(X_{\cdot 1}, X_{\cdot 2})$. Then $\hat{\theta} = \hat{q}^2$.

Because the distribution of Y is constrained by a one-parameter model, there is potential gain in the subsequent step-k estimators, whose calculations are derived in Section 5. In a simulation study described below, however, we see efficiency gain plateau at $\hat{\theta}$ (somewhat expected given Remark 1). But if we start with an inefficient estimator, $\hat{\theta}^{(2)}$ can still improve upon $\hat{\theta}$, mainly by regaining the efficiency initially lost in the outcome space.

Specifically, we generated $X_{\cdot j} \sim N(0,1)$ and Y from the following model:

$$\Pr(Y_1 = 1 \mid X; \xi_0) = \frac{\Phi(\xi_{01}^{\mathrm{T}}\tilde{X})}{1 + \Phi(\xi_{01}^{\mathrm{T}}\tilde{X}) + \Phi(\xi_{02}^{\mathrm{T}}\tilde{X})}$$
$$\text{and} \quad \Pr(Y_2 = 1 \mid X; \xi_0) = \frac{\Phi(\xi_{02}^{\mathrm{T}}\tilde{X})}{1 + \Phi(\xi_{01}^{\mathrm{T}}\tilde{X}) + \Phi(\xi_{02}^{\mathrm{T}}\tilde{X})}, \tag{20}$$

where $\tilde{X} = (1, X_{\cdot 1}, X_{\cdot 2}, X_{\cdot 1}X_{\cdot 2})^{\mathrm{T}}$, $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution, $\xi_{01} = (-1.35, -1, -1, 2)^{\mathrm{T}}$, $\xi_{02} = (2.48, 1, 1, -1.5)^{\mathrm{T}}$. This set-up yields $\Pr(Y_1 = 1) = q_0^2 = 0.16$ and $\Pr(Y_2 = 1) = (1 - q_0)^2 = 0.36$ with $q_0 = 0.4$. So the marginal Hardy–Weinberg model holds.

Instead of (20), we used a trinomial logistic regression model for $\mathcal{M}^*_{Y|X}$. This means replacing the $\Phi(\cdot)$ in (20) with $\exp(\cdot)$ and thus gives rise to a rather different model (as $\exp(\cdot)$ amplifies rather than constrains large values of the linear predictor). Then we used the procedures described earlier as well as in Section 5.8 to compute $\hat{\theta}^{(k)}$ for $k = 1, 2, 3$. This was done with the $\mathcal{H}_Y$-efficient initial estimator $\hat{\theta}_{\text{init}} = \hat{q}^2_{\text{init}}$ and also the simple but inefficient estimator $\mathbb{P}_n Y_1$. The simulation results with sample size $n = 200, 500, 1000$, and 2000 are summarized in Table 2. The relative efficiencies of the different estimators with reference to $\hat{\theta}_{\text{init}}$ are plotted in Fig 2. The sequence starting with $\hat{\theta}_{\text{init}}$ gains about 20% extra efficiency at step 1, but plateaus thereafter. By contrast, the one that starts with the inefficient $\mathbb{P}_n Y_1$, which is almost only half as efficient as $\hat{\theta}_{\text{init}}$, keeps improving until step 2, where the initial loss of efficiency is regained. However, its eventual gain is slightly lower than that of the first sequence, likely an artifact due to the misspecified $\mathcal{M}^*_{Y|X}$. Simulations in the Supplementary Material (Mao, 2024) confirm that an $\mathcal{H}_Y$-inefficient initial estimator can gain full efficiency under a correctly specified $\mathcal{M}^*_{Y|X}$.

4. Discussions

Clearly, the extent of efficiency gain depends on the size of $\mathcal{M}_X$, which reflects how restrictive the covariate model is. If $\mathcal{M}_X$ is a singleton, meaning $\mathbb{P}_X$ is known exactly, the efficiency gain is maximized. At the other extreme, when $\mathcal{M}_X = \mathcal{M}^0_X$—where no information about $\mathbb{P}_X$ is available—no improvement can be made. Yet another key factor is the alignment of the $\mathcal{H}_Y$-efficient score

TABLE 2
Simulation results for iterative augmented estimation of $\Pr(Y_1 = 1)$ *under a Hardy-Weinberg model with two independent covariates.*

		Step 0: $\mathcal{H}_Y$-efficient					Step 0: $\mathcal{H}_Y$-inefficient				
n	S	Bias	SE	SEE	CP	RE	Bias	SE	SEE	CP	RE
200	0	0.001	1.97	1.95	0.945	1.00	0.000	2.60	2.58	0.945	0.57
	1	0.000	1.80	1.78	0.947	1.20	0.000	2.44	2.41	0.948	0.65
	2	0.000	1.80	1.78	0.946	1.20	0.001	1.86	1.84	0.943	1.12
	3	0.000	1.80	1.78	0.946	1.20	0.001	1.87	1.85	0.943	1.11
500	0	0.000	1.23	1.24	0.951	1.00	0.000	1.64	1.64	0.948	0.56
	1	0.000	1.12	1.13	0.954	1.21	0.000	1.52	1.53	0.951	0.65
	2	0.000	1.12	1.13	0.953	1.21	0.000	1.15	1.17	0.954	1.14
	3	0.000	1.12	1.13	0.953	1.21	0.000	1.16	1.17	0.953	1.13
1000	0	0.000	0.88	0.88	0.948	1.00	0.000	1.17	1.16	0.950	0.57
	1	0.000	0.81	0.80	0.947	1.20	0.000	1.09	1.09	0.951	0.65
	2	0.000	0.81	0.80	0.947	1.20	0.000	0.83	0.83	0.947	1.13
	3	0.000	0.81	0.80	0.947	1.20	0.000	0.84	0.83	0.947	1.12
2000	0	0.000	0.63	0.62	0.946	1.00	0.000	0.83	0.82	0.947	0.58
	1	0.000	0.57	0.57	0.949	1.20	0.000	0.77	0.77	0.947	0.66
	2	0.000	0.57	0.57	0.948	1.20	0.000	0.59	0.59	0.946	1.14
	3	0.000	0.57	0.57	0.948	1.20	0.000	0.59	0.59	0.946	1.13

See note to Table 1. S, step; RE, relative efficiency with respect to the $\mathcal{H}_Y$-efficient initial estimator.

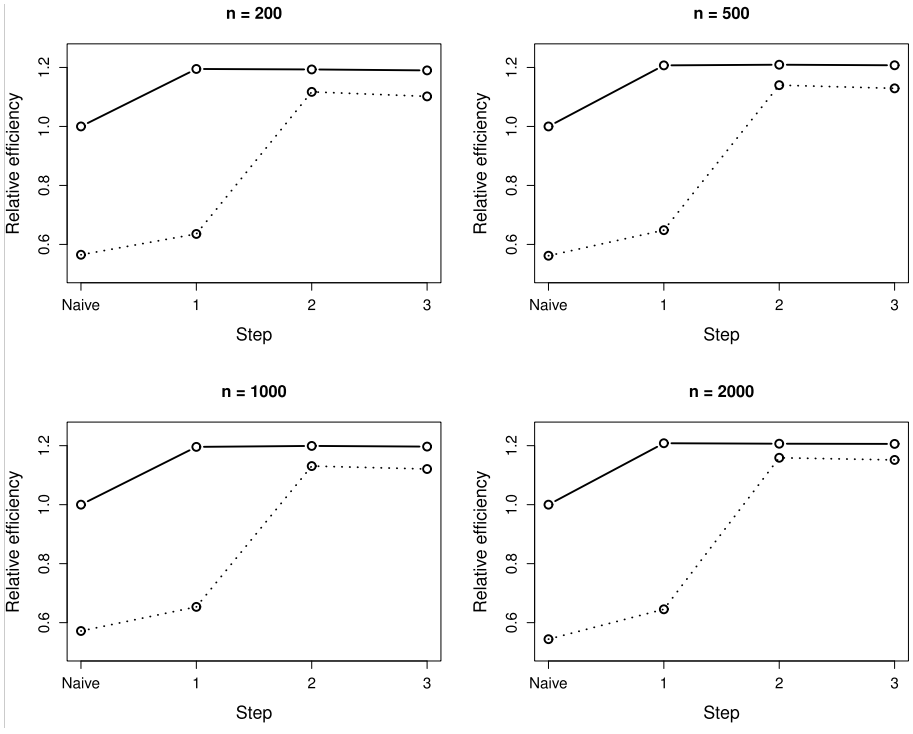

FIG 2. *Empirical relative efficiency of step-k estimator* $\hat{\theta}^{(k)}$ *with respect to the* $\mathcal{H}_Y$*-efficient initial estimator based on 10,000 replicates. Solid line, starting with an* $\mathcal{H}_Y$*-efficient initial estimator; dotted line, starting with an* $\mathcal{H}_Y$*-inefficient initial estimator.*

with the structure of $\mathcal{H}_X$. If the conditional mean of the former lies entirely within the latter, meaning the initial estimator is already fully compatible with $\mathcal{M}_X$, then there would be no residual noise to reduce, regardless of how restrictive $\mathcal{M}_X$ might be. Lastly, the efficiency gain also depends on the strength of the association between Y and X, and how well this relationship is captured by $\mathcal{M}^*_{Y|X}$. The stronger the association and the more accurately $\mathcal{M}^*_{Y|X}$ approximates it, the more information can be drawn from the covariates to improve the outcome estimation.

The marginal inference of Y should be broadly understood as any inference that is not conditioned on X, including inferences on the relationship between its components. For example, with $Y = (\tilde{Y}, Z)$, the current framework can be applied to the estimation in a restricted mean model $E(\tilde{Y} \mid Z) = \theta^{\mathrm{T}} Z$, where θ contains the regression coefficients of Z. More details can be found in Section 5.

The augmented estimator in Theorem 2 ensures robustness against misspecification of $\mathcal{M}^*_{Y|X}$, but not against misspecification of $\mathcal{M}_X$. In fact, if the true distribution $\mathbb{P}_X$ does not lie in $\mathcal{M}_X$, the bias in the estimating function $\varphi_n(\theta)$ is given by the likely nonzero $(\mathbb{P}^*_X - \mathbb{P}_X)\mu(X; \theta_0, \xi^*)$ for some $\mathbb{P}^*_X \in \mathcal{M}_X$. Moreover, this bias persists even when $\mathcal{M}^*_{Y|X}$ is correctly specified, because in general $\mathbb{P}^*_X \mu(X; \theta_0, \xi^*) \neq 0$ even though $\mathbb{P}_X \mu(X; \theta_0, \xi^*) = 0$. This means that the augmented estimator is *not* doubly robust. On the other hand, if $\mathbb{P}^*_X$ is estimated using (some version of) the MLE, as in all three examples in Section 3, it should be the closest element in $\mathcal{M}_X$ to $\mathbb{P}_X$ in terms of the Kullback–Liebler (KL) divergence. This allows us to bound the bias using this minimized KL divergence via its relationship with the total variation distance between $\mathbb{P}^*_X$ and $\mathbb{P}_X$. Such bounds are particularly useful if the estimating function is uniformly bounded (as in all three examples). Some details are provided in Section 5.10.

Since the validity of the new estimator depends on the correctness of $\mathcal{M}_X$, it is crucial that the covariate model be justifiable based on substantive knowledge. In this paper we have assumed the model as given, and have provided only simple examples to illustrate the technical derivation. Real applications may involve more complex scenarios. One application we have in mind, for example, is to use multiple quantitative imaging markers (e.g., those measuring fat or iron content in the liver) to improve disease diagnosis. The fact that these markers are obtained under strictly controlled conditions and by well-defined algorithms often makes certain models (e.g., linear mixed effects models) a natural fit (see, e.g., Hernando et al., 2023). While these more elaborate models present no new theoretical challenges, their derivations and computations would be far more involved than any of the examples in Section 3. A future study is needed to explore them in greater depth.

In practice, covariates may not only exhibit complex relationships but also be high-dimensional. Such scenarios are not fully addressed by the current theory, which relies on first-order asymptotics under a fixed dimension. Our experimentation in Example 2, however, suggests that an extension to high-dimensional settings may be feasible. Given recent advances in double machine learning and cross-fitting (see, e.g., Díaz et al., 2021), an appealing approach would be to

build a nonparametric model for $Y \mid X$ and solve it using methods like cross-fitted targeted maximum likelihood estimation (TMLE) (van der Laan and Rose, 2018). If validated, these techniques could help automate variable and model selection, thereby enabling efficiency gain in more complex, high-dimensional contexts.

5. Proofs and technical details

5.1. Variance of $\hat{\theta}_{\text{adj}}$ in opening example

We want to show that

$$n^{1/2}(\hat{\theta}_{\text{adj}} - \theta) \to_d N\left[0, E\{\text{var}(Y \mid X)\}\right]. \tag{21}$$

Proof. Write $\hat{p}_x = \mathbb{P}_n I(X = x)$ and $\mu(x) = E(Y \mid X = x)$. Then $\hat{\theta}_{\text{adj}} = \sum_x p_x \hat{p}_x^{-1} \mathbb{P}_n Y I(X = x)$ and $\theta = \sum_x p_x \mu(x)$. The left hand side of (21) is thus

$$\begin{aligned}
&n^{1/2} \sum_x p_x \left\{\hat{p}_x^{-1} \mathbb{P}_n Y I(X = x) - \mu(x)\right\} \\
=&n^{1/2} \sum_x p_x \hat{p}_x^{-1} \left\{\mathbb{P}_n Y I(X = x) - \mu(x) \mathbb{P}_n I(X = x)\right\} \\
=&n^{1/2} \sum_x p_x \hat{p}_x^{-1} \mathbb{P}_n I(X = x) \left\{Y - \mu(x)\right\} \\
=&n^{1/2} \sum_x \mathbb{P}_n I(X = x) \left\{Y - \mu(x)\right\} + \underbrace{n^{1/2} \sum_x (p_x \hat{p}_x^{-1} - 1) \mathbb{P}_n I(X = x) \left\{Y - \mu(x)\right\}}_{O_p(n^{-1/2})} \\
=&n^{1/2} \mathbb{P}_n \left\{Y - \mu(X)\right\} + o_p(1),
\end{aligned}$$

with asymptotic variance

$$E\left[\left\{Y - \mu(X)\right\}^2\right] = E\left(E\left[\left\{Y - \mu(X)\right\}^2 \mid X\right]\right) = E\{\text{var}(Y \mid X)\}.$$

This completes the proof. □

5.2. Proof of Lemma 1

Let $p_0(y \mid x)$ denote the true conditional density of Y given X with respect to some dominating measure ν and $p_0(x)$ the true marginal density of X with respect to some dominating measure η. Then the joint density of Y, X with respect to $\nu \times \eta$ is $p_0(y, x) = p_0(y \mid x) p_0(x)$. For simplicity, we assume that all scores are generated by taking pointwise derivatives of the log-density functions with respect to smoothly indexed parameters. For a fully general version based on differentiation in the quadratic mean, see §3.3 of Bickel et al. (1993).

Given a joint score $g \in \mathcal{H}$ generated by $g(y, x) = \partial \log\{p_\varepsilon(y, x)\}/\partial\varepsilon|_{\varepsilon=0}$, where $\{p_\varepsilon(y, x)\} \subset \mathcal{M}_Y \cap \mathcal{M}_X$ is a one-dimensional submodel parametrized

by ε and passing through $p_0(y,x)$ at $\varepsilon = 0$. Let $p_\varepsilon(x) = \int p_\varepsilon(y,x)\mathrm{d}\nu(y)$ and $p_\varepsilon(y \mid x) = p_\varepsilon(y,x)/p_\varepsilon(x)$ denote the corresponding marginal density of X and conditional density of $Y \mid X$, respectively. Since $p_\varepsilon(y,x) = p_\varepsilon(y \mid x)p_\varepsilon(x)$, we have that

$$g(y,x) = \underbrace{\partial \log\{p_\varepsilon(y \mid x)\}/\partial\varepsilon|_{\varepsilon=0}}_{\in L_2^0(\mathbb{P}_{Y|X})} + \underbrace{\partial \log\{p_\varepsilon(x)\}/\partial\varepsilon|_{\varepsilon=0}}_{\in \mathcal{H}_X},$$

where the first containment can be seen by $\int p_\varepsilon(y \mid x)\mathrm{d}\nu(y) \equiv 1$ and the second follows as the implied marginal model $\{p_\varepsilon(x)\} \subset \mathcal{M}_X$. This shows that $\mathcal{H} \subset L_2^0(\mathbb{P}_{Y|X}) \oplus \mathcal{H}_X$. By symmetry between Y and X, we thus have that

$$\mathcal{H} \subset \left\{L_2^0(\mathbb{P}_{Y|X}) \oplus \mathcal{H}_X\right\} \cap \left\{L_2^0(\mathbb{P}_{X|Y}) \oplus \mathcal{H}_Y\right\}.$$

To show the reverse containment, let $a \in L_2^0(\mathbb{P}_{Y|X})$ and $h \in \mathcal{H}_X$. Suppose that the latter is generated by $h(x) = \partial \log\{p_\varepsilon(x)\}/\partial\varepsilon|_{\varepsilon=0}$ for some submodel $\{p_\varepsilon(x)\} \subset \mathcal{M}_X$. If $a(y,x)$ is uniformly bounded, then we can construct the joint density

$$p_\varepsilon(y,x) = \underbrace{p_0(y \mid x)\{1 + \varepsilon a(y,x)\}}_{\text{conditional density for small } \varepsilon} p_\varepsilon(x),$$

which generates the score $\partial \log\{p_\varepsilon(y,x)\}/\partial\varepsilon|_{\varepsilon=0} = a(y,x) + h(x)$. The uniform boundedness of $a(y,x)$ can be relaxed since such functions are dense in $L_2^0(\mathbb{P}_{Y|X})$. Therefore,

$$\mathcal{H} \supset L_2^0(\mathbb{P}_{Y|X}) \oplus \mathcal{H}_X \supset \left\{L_2^0(\mathbb{P}_{Y|X}) \oplus \mathcal{H}_X\right\} \cap \left\{L_2^0(\mathbb{P}_{X|Y}) \oplus \mathcal{H}_Y\right\}.$$

This completes the proof.

5.3. Details for proof of Theorem 1

We prove inductively that

$$\begin{aligned}
\psi^{(2j)}(Y,X;\theta_0) &= \Pi\left\{\psi^{(2j-1)}(Y,X;\theta_0) \mid L_2^0(\mathbb{P}_{X|Y}) \oplus \mathcal{H}_Y\right\} \\
&= \psi^{(2j-1)}(Y,X;\theta_0) + \kappa^{(2j)}(Y;\theta_0) \\
\text{and } \psi^{(2j+1)}(Y,X;\theta_0) &= \Pi\left\{\psi^{(2j)}(Y,X;\theta_0) \mid L_2^0(\mathbb{P}_{Y|X}) \oplus \mathcal{H}_X\right\} \\
&= \psi^{(2j)}(Y,X;\theta_0) - \kappa^{(2j+1)}(X;\theta_0). \qquad (22)
\end{aligned}$$

To start, we show it holds for $j = 1$. Indeed,

$$\begin{aligned}
&\psi^{(2)}(Y,X;\theta_0) \\
=&\Pi\left\{\psi^{(1)}(Y,X;\theta_0) \mid L_2^0(\mathbb{P}_{X|Y}) \oplus \mathcal{H}_Y\right\} \\
=&\Pi\left\{\psi(Y;\theta_0) \mid L_2^0(\mathbb{P}_{X|Y}) \oplus \mathcal{H}_Y\right\} - \Pi\left\{\kappa^{(1)}(X;\theta_0) \mid L_2^0(\mathbb{P}_{X|Y}) \oplus \mathcal{H}_Y\right\}
\end{aligned}$$

$$
\begin{aligned}
&=\psi(Y;\theta_0)-\left\{\kappa^{(1)}(X;\theta_0)-(\mathcal{I}_Y-\Pi_{\mathcal{H}_Y})A^{\mathrm{T}}\kappa^{(1)}(\cdot;\theta_0)(Y)\right\}\\
&=\left\{\psi(Y;\theta_0)-\kappa^{(1)}(X;\theta_0)\right\}+(\mathcal{I}_Y-\Pi_{\mathcal{H}_Y})A^{\mathrm{T}}\kappa^{(1)}(\cdot;\theta_0)(Y)\\
&=\psi^{(1)}(Y,X;\theta_0)+\kappa^{(2)}(Y;\theta_0),
\end{aligned}
$$

where the third equality follows by $\psi(Y;\theta_0)\in\mathcal{H}_Y$ and Lemma 2. Likewise,

$$
\begin{aligned}
&\psi^{(3)}(Y,X;\theta_0)\\
=&\Pi\left\{\psi^{(2)}(Y,X;\theta_0)\mid L_2^0(\mathbb{P}_{Y|X})\oplus\mathcal{H}_X\right\}\\
=&\Pi\left\{\psi^{(1)}(Y,X;\theta_0)\mid L_2^0(\mathbb{P}_{Y|X})\oplus\mathcal{H}_X\right\}+\Pi\left\{\kappa^{(2)}(Y;\theta_0)\mid L_2^0(\mathbb{P}_{Y|X})\oplus\mathcal{H}_X\right\}\\
=&\psi^{(1)}(Y,X;\theta_0)+\left\{\kappa^{(2)}(Y;\theta_0)-(\mathcal{I}_X-\Pi_{\mathcal{H}_X})A\kappa^{(2)}(\cdot;\theta_0)(X)\right\}\\
=&\left\{\psi^{(1)}(Y,X;\theta_0)+\kappa^{(2)}(Y;\theta_0)\right\}-(\mathcal{I}_X-\Pi_{\mathcal{H}_X})A\kappa^{(2)}(\cdot;\theta_0)(X)\\
=&\psi^{(2)}(Y,X;\theta_0)-\kappa^{(3)}(X;\theta_0),
\end{aligned}
$$

where the third equality follows by $\psi^{(1)}(Y,X;\theta_0)=\Pi\{\psi(Y;\theta_0)\mid L_2^0(\mathbb{P}_{Y|X})\oplus\mathcal{H}_X\}\in L_2^0(\mathbb{P}_{Y|X})\oplus\mathcal{H}_X$ and Lemma 2. Now, suppose that (22) holds for any particular j, we can use similar arguments to show that it also holds for $(j+1)$. This proves the result.

5.4. *Proof of Theorem 2*

With established regularity conditions on $\psi(Y;\theta)$, such as those in Theorem 5.9 of van der Vaart (1998), it is easy to see that $\hat{\theta}$ is consistent because $||(\mathbb{P}_n-\hat{\mathbb{E}}_{\mathcal{M}_X})\mu(X;\hat{\theta}_{\text{init}},\hat{\xi})||\to_p 0$ by (C2) and (C4). Therefore, we focus on the asymptotic expansion in (8). Use $\mathbb{G}_n=n^{1/2}(\mathbb{P}_n-\mathbb{P})$ to denote the standardized empirical process. By (C1)–(C3) and $\hat{\theta}\to_p\theta_0$, the following stochastic continuity properties hold (see, e.g., van der Vaart and Wellner, 1996, Ch. 3.3):

$$
\begin{aligned}
\mathbb{G}_n\left\{\mu(X;\hat{\theta}_{\text{init}},\hat{\xi})-\Pi_{\mathcal{H}_X}\mu(X;\hat{\theta}_{\text{init}},\hat{\xi})\right\}&=\mathbb{G}_n\left\{\mu(X;\theta_0,\xi^*)-\Pi_{\mathcal{H}_X}\mu(X;\theta_0,\xi^*)\right\}\\
&\quad+o_p(1)\\
\text{and}\qquad \mathbb{G}_n\psi(Y;\hat{\theta})&=\mathbb{G}_n\psi(Y;\theta_0)+o_p(1),
\end{aligned}
\tag{23}
$$

where the continuity of $\mathbb{G}_n$ on the projected class $\{\Pi_{\mathcal{H}_X}\mu(X;\theta,\xi):||\theta-\theta_0||+\rho(\xi,\xi^*)<2\delta\}$ can be verified directly by tightness arguments.

By assumption,

$$
\begin{aligned}
o_p(1)&=n^{1/2}\varphi_n(\hat{\theta})\\
&=n^{1/2}\mathbb{P}_n\psi(Y;\hat{\theta})-(\mathbb{P}_n-\hat{\mathbb{E}}_{\mathcal{M}_X})\mu(X;\hat{\theta}_{\text{init}},\hat{\xi})\\
&=n^{1/2}\mathbb{P}_n\psi(Y;\hat{\theta})\\
&\quad-n^{1/2}\mathbb{P}_n\left\{\mu(X;\hat{\theta}_{\text{init}},\hat{\xi})-\mathbb{P}\mu(X;\hat{\theta}_{\text{init}},\hat{\xi})-\Pi_{\mathcal{H}_X}\mu(X;\hat{\theta}_{\text{init}},\hat{\xi})\right\}+o_p(1)
\end{aligned}
$$

$$
\begin{aligned}
&= \mathbb{G}_n \psi(Y;\hat{\theta}) - \mathbb{G}_n \left\{ \mu(X;\hat{\theta}_{\text{init}},\hat{\xi}) - \Pi_{\mathcal{H}_X} \mu(X;\hat{\theta}_{\text{init}},\hat{\xi}) \right\} \\
&\qquad + n^{1/2} \mathbb{P} \psi(Y;\hat{\theta}) + o_p(1) \\
&= \mathbb{G}_n \psi(Y;\theta_0) - \mathbb{G}_n \left\{ \mu(X;\theta_0,\xi^*) - \Pi_{\mathcal{H}_X} \mu(X;\theta_0,\xi^*) \right\} \\
&\qquad + n^{1/2} \mathbb{P} \psi(Y;\hat{\theta}) + o_p(1) \\
&= \mathbb{G}_n \psi(Y;\theta_0) - \mathbb{G}_n \left\{ \mu(X;\theta_0,\xi^*) - \Pi_{\mathcal{H}_X} \mu(X;\theta_0,\xi^*) \right\} \\
&\qquad + V(\theta_0) n^{1/2} (\hat{\theta} - \theta_0) + o_p\left(1 + n^{1/2} ||\hat{\theta} - \theta_0||\right),
\end{aligned}
$$

where the third equality follows by (C4), the fourth by $\mathbb{P}\Pi_{\mathcal{H}_X} \mu(X;\hat{\theta}_{\text{init}},\hat{\xi}) = 0$ as a result of $\Pi_{\mathcal{H}_X} \mu(X;\theta,\xi) \in L_2^0(\mathbb{P}_X)$ for all θ and ξ, and the fifth by (23). By (C5), the above display implies that $n^{1/2}||\hat{\theta} - \theta_0|| = O_p(1)$ and that

$$
n^{1/2}(\hat{\theta} - \theta_0) = -V(\theta_0)^{-1} \mathbb{G}_n \left[\psi(Y;\theta_0) - \left\{ \mu(X;\theta_0,\xi^*) - \Pi_{\mathcal{H}_X} \mu(X;\theta_0,\xi^*) \right\} \right] + o_p(1).
$$

This proves (8). When $\mathbb{P}_{Y|X} \in \mathcal{M}^*_{Y|X}$, we have that $\mu(X;\theta_0,\xi^*) = \mu(X;\theta_0,\xi_0) = A\psi(\cdot;\theta_0)(X) \in L_2^0(\mathbb{P}_X)$. Thus the influence function becomes

$$
\begin{aligned}
&- V(\theta_0)^{-1} \left[\psi(Y;\theta_0) - \left\{ A\psi(\cdot;\theta_0)(X) - \mathbb{P} A\psi(\cdot;\theta_0)(X) - \Pi_{\mathcal{H}_X} A\psi(\cdot;\theta_0)(X) \right\} \right] \\
&= - V(\theta_0)^{-1} \left[\psi(Y;\theta_0) - (\mathcal{I}_X - \Pi_{\mathcal{H}_X}) A\psi(\cdot;\theta_0)(X) \right] \\
&= - V(\theta_0)^{-1} \psi^{(1)}(Y,X;\theta_0).
\end{aligned}
$$

This completes the proof.

5.5. Details for step-k estimators

For projection back on the outcome space, we will need a working model to implement A^{T}. If $\mathcal{M}^*_{Y|X}$ specifies a class of densities $p(y \mid x;\xi)$, then we can use the Bayes rule to estimate $A^{\mathrm{T}} h(y)$ by $\{\mathbb{P}_n p(y \mid X;\xi) h(X)\} / \mathbb{P}_n p(y \mid X;\xi)$ for any $h(X) \in L_2(\mathbb{P}_X)$. Likewise, we will need $\hat{\mathbb{E}}_{\mathcal{M}_Y} a(Y)$ as an $\mathcal{M}_Y$-efficient estimator of $\mathbb{P}a(Y)$ for any $a(Y) \in L_2(\mathbb{P}_Y)$ and an empirical $\hat{\Pi}_{\mathcal{H}_Y}$ to approximate $\Pi_{\mathcal{H}_Y}$. To initialize, let $\varphi_n^{(1)}(\theta) = \varphi_n(\theta)$, $\hat{\mu}^{(1)}(X;\theta,\xi) = \mu(X;\theta,\xi)$, and $\hat{\Psi}^{(1)}(Y,X) = \psi(Y;\hat{\theta}) - \hat{\kappa}^{(1)}(X;\hat{\theta},\hat{\xi})$.

Given $\varphi_n^{(2j-1)}(\theta)$, $\hat{\mu}^{(2j-1)}(X;\theta,\xi)$, $\hat{\kappa}^{(2j-1)}(X;\theta,\xi)$, and $\hat{\Psi}^{(2j-1)}(Y,X)$, compute

$$
\begin{aligned}
\varphi_n^{(2j)}(\theta) &= \varphi_n^{(2j-1)}(\theta) + (\mathbb{P}_n - \hat{\mathbb{E}}_{\mathcal{M}_Y}) \hat{\mu}^{(2j)}(Y;\hat{\theta}_n,\hat{\xi}), \\
\text{and} \quad \varphi_n^{(2j+1)}(\theta) &= \varphi_n^{(2j)}(\theta) - (\mathbb{P}_n - \hat{\mathbb{E}}_{\mathcal{M}_X}) \hat{\mu}^{(2j+1)}(X;\hat{\theta}_n,\hat{\xi}),
\end{aligned} \tag{24}
$$

where $\hat{\mu}^{(2j)}(y;\theta,\xi) = \{\mathbb{P}_n \hat{\kappa}^{(2j-1)}(\cdot;\theta,\xi) p(y \mid \cdot;\xi)\} / \mathbb{P}_n p(y \mid \cdot;\xi)$, $\hat{\kappa}^{(2j)}(Y;\theta,\xi) = \hat{\mu}^{(2j)}(Y;\theta,\xi) - \mathbb{P}_n \hat{\mu}^{(2j)}(\cdot;\theta,\xi) - \hat{\Pi}_{\mathcal{H}_Y} \hat{\mu}^{(2j)}(Y;\theta,\xi)$, and $\hat{\mu}^{(2j+1)}(X;\theta,\xi) = E^*\{\hat{\kappa}^{(2j)}(Y;\theta,\xi) \mid X;\xi\}$. Then compute $\hat{\theta}^{(2j)}$ and $\hat{\theta}^{(2j+1)}$ by solving

$$
\varphi_n^{(2j)}\left\{\hat{\theta}^{(2j)}\right\} = o_p(n^{-1/2}) \text{ and } \varphi_n^{(2j+1)}\left\{\hat{\theta}^{(2j+1)}\right\} = o_p(n^{-1/2}),
$$

respectively. Also compute

$$\hat{\Psi}^{(2j)}(Y,X) = \hat{\Psi}^{(2j-1)}(Y,X) + \hat{\kappa}^{(2j)}\left\{Y;\hat{\theta}^{(2j)},\hat{\xi}\right\}$$
$$\text{and} \qquad \hat{\Psi}^{(2j+1)}(Y,X) = \hat{\Psi}^{(2j)}(Y,X) - \hat{\kappa}^{(2j+1)}\left\{X;\hat{\theta}^{(2j+1)},\hat{\xi}\right\},$$

where $\hat{\kappa}^{(2j+1)}(X;\theta,\xi) = \hat{\mu}^{(2j+1)}(X;\theta,\xi) - \mathbb{P}_n\hat{\mu}^{(2j+1)}(\cdot;\theta,\xi) - \hat{\Pi}_{\mathcal{H}_X}\hat{\mu}^{(2j+1)}(X;\theta,\xi)$. Now that we have the needed $\varphi_n^{(2j+1)}(\theta)$, $\hat{\mu}^{(2j+1)}(X;\theta,\xi)$, $\hat{\kappa}^{(2j+1)}(X;\theta,\xi)$, and $\hat{\Psi}^{(2j+1)}(Y,X)$, we can move on to the next cycle.

The step-k estimators $\hat{\theta}^{(k)}$ $(k = 1,2,3,\ldots)$ satisfy the stated robustness and efficient properties. To estimate their variances, use the sandwich-type estimator

$$\hat{\text{var}}\left\{\hat{\theta}^{(k)}\right\} = n^{-1}\hat{V}\left\{\hat{\theta}^{(k)}\right\}^{-1}\mathbb{P}_n\left\{\hat{\Psi}^{(k)}(Y,X)^{\otimes 2}\right\}\hat{V}\left\{\hat{\theta}^{(k)}\right\}^{\mathrm{T}-1}$$

5.6. Details for Example 1

Clearly, $\dot{l}_\gamma(X) = \{X-\tau, (X-\tau)^2-\sigma^2\}^{\mathrm{T}}$ is the influence function of $(\hat{\tau},\hat{\sigma}^2)$ and is thus the efficient influence. Since the efficient influence is a linear transform of the score function, it can be used in place of the latter. We then have that

$$I_\gamma = E\{\dot{l}_\gamma(X)^{\otimes 2}\} = \begin{pmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{pmatrix},$$

where the second diagonal element is calculated from $E[\{(X-\tau)^2-\sigma^2\}^2] = E\{(X-\tau)^4\} - \sigma^4 = 3\sigma^4 - \sigma^4 = 2\sigma^4$ (see, e.g., Papoulis and Pillai, 2002). Then for any $h(X) \in L_2(\mathbb{P}_X)$, the projection formula is given by

$$\begin{aligned}\Pi_{\mathcal{H}_X}h(X) &= \mathbb{P}\{h(\cdot)l_\gamma(\cdot)^{\mathrm{T}}\}I_\gamma^{-1}\dot{l}_\gamma(X)\\ &= b_1(\gamma)(X-\tau) + b_2(\gamma)\{(X-\tau)^2-\sigma^2\},\end{aligned}$$

where $b_1(\gamma) = \sigma^{-2}\text{cov}\{h(X),X\}$ and $b_2(\gamma) = 2^{-1}\sigma^{-4}\text{cov}\{h(X),(X-\tau)^2\}$. We can thus estimate $\Pi_{\mathcal{H}_X}h(X)$ by plugging in the estimators $(\hat{\tau},\hat{\sigma}^2)$ and replacing the covariances in $b_1(\gamma)$ and $b_2(\gamma)$ with their empirical analogs.

5.7. Details for proof of Lemma 4

We want to show that

$$n^{1/2}(\mathbb{P}_{X_{\cdot 1},n}\times\cdots\times\mathbb{P}_{X_{\cdot p},n} - \mathbb{P})\mu(X;\theta,\xi) = \mathbb{G}_n\left\{\sum_{j=1}^p \mu_j(X_{\cdot j};\theta,\xi)\right\} + o_p(1) \quad (25)$$

Write $X_i = (X_{i1},\ldots,X_{ip})$ $(i = 1,\ldots,n)$, where the X_{ij} are independent copies of $X_{\cdot j}$. Let $\mathbb{P}_{X_{ij}}$ denote the underlying measure of X_{ij}. We show that the projection of the left hand side of (25) onto $\dot{\mathcal{P}} = \oplus_{i=1}^n \oplus_{j=1}^p L_2^0(\mathbb{P}_{X_{ij}})$, where $\mathbb{P}_{X_{ij}}$ is the measure for X_{ij}, is the first term on the right hand side. Indeed,

$$\Pi\left\{n^{1/2}(\mathbb{P}_{X_{\cdot 1},n}\times\cdots\times\mathbb{P}_{X_p,n} - \mathbb{P}_n)\mu(X;\theta,\xi) \mid \dot{\mathcal{P}}\right\}$$

$$
\begin{aligned}
&= n^{1/2} n^{-p} \sum_{i_1=1}^{n} \cdots \sum_{i_p=1}^{n} \Pi \left\{ \mu(X_{i_1 1}, \ldots, X_{i_p p}; \theta, \xi) \mid \dot{\mathcal{P}} \right\} \\
&= n^{1/2} n^{-p} \sum_{i_1=1}^{n} \cdots \sum_{i_p=1}^{n} \sum_{j=1}^{p} \left\{ \mu_j(X_{i_j j}; \theta, \xi) - \mathbb{P}\mu(X; \theta, \xi) \right\} \qquad \text{(by Hájek (1968))} \\
&= n^{1/2} \sum_{j=1}^{p} n^{-1} \sum_{i_j=1}^{n} \left\{ \mu_j(X_{i_j j}; \theta, \xi) - \mathbb{P}\mu(X; \theta, \xi) \right\} \\
&= n^{1/2} \sum_{j=1}^{p} \mathbb{P}_n \left\{ \mu_j(X_{\cdot j}; \theta, \xi) - \mathbb{P}\mu(\cdot; \theta, \xi) \right\} \\
&= \mathbb{G}_n \left\{ \sum_{j=1}^{p} \mu_j(X_j; \theta, \xi) \right\}.
\end{aligned}
$$

Now by Theorem 11.2 of van der Vaart (1998), it suffices to show that the variance of the left hand side of (25) tends to $\sum_{j=1}^{p} \operatorname{var}\{\mu_j(X_{\cdot j}; \theta, \xi)\}$, the asymptotic variance of the right hand side. We can do so by extending the methods used in Theorems 12.3 and 12.6 of van der Vaart (1998) for one- and two-sample U-statistics. Specifically,

$$
\begin{aligned}
&\operatorname{var} \left\{ n^{1/2} (\mathbb{P}_{X_{\cdot 1}, n} \times \cdots \times \mathbb{P}_{X_{\cdot p}, n} - \mathbb{P}) \mu(X; \theta, \xi) \right\} \\
&= n^{1-2p} \operatorname{var} \left[\sum_{i_1=1}^{n} \cdots \sum_{i_p=1}^{n} \mu(X_{i_1 1}, \ldots, X_{i_p p}; \theta, \xi) \right] \\
&= n^{1-2p} \Bigg[n (n^2)^{p-1} \\
&\quad \times \sum_{j=1}^{p} \operatorname{cov} \{ \underbrace{\mu(X_{i_1 1}, \ldots, X_{i_j j}, \ldots, X_{i_p p}; \theta, \xi), \mu(X_{i_1^* 1}, \ldots, X_{i_j j}, \ldots, X_{i_p^* p}; \theta, \xi)}_{i_k \neq i_k^* (k \neq j)} \} \\
&\qquad + \sum_{k=2}^{p} O \left\{ n^k (n^2)^{p-k} \right\} \Bigg] \\
&= \sum_{j=1}^{p} \operatorname{var} \{ \mu_j(X_{\cdot j}; \theta, \xi) \} + O(n^{-1}),
\end{aligned}
$$

where the second equality follows by first laying out the covariances between terms that share a single $X_{i_j j}$, and then those that share two and more. This completes the proof.

5.8. Details for Example 3

Write

$$
\mathcal{D}_n(c_1, c_2; \xi) = (\mathbb{P}_n - \mathbb{U}_n) \{ c_1 \chi_1(X; \xi) + c_2 \chi_2(X; \xi) \}.
$$

It is then clear that the one-step estimator for q is $\hat{q} = \hat{q}_{\text{init}} - \mathcal{D}_n(2^{-1}, -2^{-1}; \hat{\xi})$. Moreover, define

$$\hat{\kappa}(c_1, c_2)(X;\xi) = \sum_{s=1}^{2} c_s \left[\chi_s(X;\xi) - \mathbb{P}_n \chi_s(\cdot;\xi) - \sum_{j=1}^{2} \{ \hat{\chi}_{sj}(X_{\cdot j};\xi) - \mathbb{U}_n \chi_s(\cdot;\xi) \} \right],$$

where $\hat{\chi}_{s1}(x_1;\xi) = \mathbb{P}_n \chi_s(x_1, X_{\cdot 2};\xi)$ and $\hat{\chi}_{s2}(x_2;\xi) = \mathbb{P}_n \chi_s(X_{\cdot 1}, x_2;\xi)$. Then we have that

$$\hat{\kappa}^{(1)}(X;\xi) = \hat{\kappa}(2^{-1}, -2^{-1})(X;\xi), \quad \hat{\Psi}^{(1)}(Y,X) = 2^{-1}(Y_1 - Y_2 + 1) - \hat{q} - \hat{\kappa}^{(1)}(X;\hat{\xi}), \tag{26}$$

and thus $\hat{\text{var}}(\hat{q}) = n^{-1}\mathbb{P}_n\{\hat{\Psi}^{(1)}(Y,X)^2\}$. Use the delta method to compute the variance of $\hat{\theta} = \hat{q}^2$.

Now, we construct the step-k estimator $\hat{q}^{(k)}$, and thus $\hat{\theta}^{(k)} = \hat{q}^{(k)2}$, following the instructions in Section 5.5. Write $Y_3 = 1 - Y_1 - Y_2$, $\chi_3(X;\xi) = 1 - \chi_1(X;\xi) - \chi_2(X;\xi)$, $\delta_{1n} = \mathbb{P}_n Y_1 - \hat{q}_{\text{init}}^2$, and $\delta_{2n} = \mathbb{P}_n Y_2 - (1 - \hat{q}_{\text{init}})^2$. The latter two are the differences between the empirical and $\mathcal{M}_Y$-efficient estimators of $\mathbb{P}Y_1$ and $\mathbb{P}Y_2$, respectively.

Given $\hat{q}^{(2j-1)}$, $\hat{\kappa}^{(2j-1)}(X;\xi)$, and $\hat{\Psi}^{(2j-1)}(Y,X)$, compute

$$\hat{c}_s^{(2j)}(\xi) = \frac{\mathbb{P}_n \hat{\kappa}^{(2j-1)}(X;\xi)\chi_s(X;\xi)}{\mathbb{P}_n \chi_s(X;\xi)} - \frac{\mathbb{P}_n \hat{\kappa}^{(2j-1)}(X;\xi)\chi_3(X;\xi)}{\mathbb{P}_n \chi_3(X;\xi)} \quad (s = 1, 2).$$

Then the model-based conditional expectation of $\hat{\kappa}^{(2j-1)}(X;\xi)$ given Y is up to a constant

$$\hat{\mu}^{(2j)}(Y;\xi) = \hat{c}_1^{(2j)}(\xi) Y_1 + \hat{c}_2^{(2j)}(\xi) Y_2. \tag{27}$$

So the next step estimator is just

$$\hat{q}^{(2j)} = \hat{q}^{(2j-1)} + \sum_{s=1}^{2} \hat{c}_s^{(2j)}(\hat{\xi}) \delta_{sn}.$$

The approximated influence function of the added term is

$$\hat{\kappa}^{(2j)}(Y;\xi) = \hat{\mu}^{(2j)}(Y;\xi) - \mathbb{P}_n \hat{\mu}^{(2j)}(\cdot;\xi) - 2^{-1}\hat{\sigma}^{-2} \left\{ \mathbb{P}_n \hat{\mu}^{(2j)}(\cdot;\xi) \dot{l}(\cdot) \right\} (Y_1 - Y_2 + 1), \tag{28}$$

where $\dot{l}(Y) = 2^{-1}(Y_1 - Y_2 + 1) - \hat{q}_{\text{init}}$ and $\hat{\sigma}^2 = \mathbb{P}_n\{\dot{l}(Y)^2\}$, and so $\hat{\Psi}^{(2j)}(Y,X) = \hat{\Psi}^{(2j-1)}(Y,X) + \hat{\kappa}^{(2j)}(Y;\hat{\xi})$. As a result, we can estimate the variance of $\hat{q}^{(2j)}$ by $\hat{\text{var}}\{\hat{q}^{(2j)}\} = n^{-1}\mathbb{P}_n\{\hat{\Psi}^{(2j)}(Y,X)^2\}$.

Proceeding to step $(2j+1)$, we can see from (27) and (28) that $\hat{\kappa}^{(2j)}(Y;\xi)$ is up to a constant $\hat{c}_1^{(2j+1)}(\xi) Y_1 + \hat{c}_2^{(2j+1)}(\xi) Y_2$, where

$$\hat{c}_1^{(2j+1)}(\xi) = \hat{c}_1^{(2j)}(\xi) - 2^{-1}\hat{\sigma}^{-2} \left\{ \mathbb{P}_n \hat{\mu}^{(2j)}(\cdot;\xi) \dot{l}(\cdot) \right\},$$
$$\hat{c}_2^{(2j+1)}(\xi) = \hat{c}_2^{(2j)}(\xi) + 2^{-1}\hat{\sigma}^{-2} \left\{ \mathbb{P}_n \hat{\mu}^{(2j)}(\cdot;\xi) \dot{l}(\cdot) \right\},$$

As a result, the next-step estimator is just

$$\hat{q}^{(2j+1)} = \hat{q}^{(2j)} - \mathcal{D}_n \left\{ \hat{c}_1^{(2j+1)}(\hat{\xi}), \hat{c}_2^{(2j+1)}(\hat{\xi}); \hat{\xi} \right\}.$$

Then similarly to (26),

$$\hat{\kappa}^{(2j+1)}(X;\xi) = \hat{\kappa}\left\{ \hat{c}_1^{(2j+1)}(\xi), \hat{c}_2^{(2j+1)}(\xi) \right\}(X;\xi),$$
$$\hat{\Psi}^{(2j+1)}(Y,X) = \hat{\Psi}^{(2j+1)}(Y,X) - \hat{\kappa}^{(2j+1)}(X;\hat{\xi}).$$

This completes the cycle.

5.9. Details on restricted mean model

Assume the general restricted mean model

$$E(\tilde{Y} \mid Z) = g(Z;\theta)$$

for some known function g indexed by θ. By standard result (e.g., Tsiatis, 2006, §4.5), the $\mathcal{H}_Y$-efficient score for θ is

$$\psi(Y;\theta_0) = D(Z)\Sigma(Z)^{-1}\{\tilde{Y} - g(Z;\theta_0)\},$$

where $D(Z) = \partial g(Z;\theta)|_{\theta=\theta_0}$ and $\Sigma(Z) = \mathrm{var}(\tilde{Y} \mid Z)$. We then have that

$$A\psi(\cdot;\theta)(X) = E\left[D(Z)\Sigma(Z)^{-1}\{\Upsilon(Z,X) - g(Z;\theta)\} \mid X \right],$$

where $\Upsilon(Z,X) = E(\tilde{Y} \mid Z, X)$. Given this form, it is convenient to posit a two-step working model, first $\mathcal{M}^*_{\tilde{Y}|Z,X}$ for the conditional distribution of $\tilde{Y} \mid Z, X$ and then $\mathcal{M}^*_{Z|X}$ for the conditional distribution of $Z \mid X$, indexed by ξ_1 and ξ_2, respectively. Then we can compute the working model-based $A\psi(\cdot;\theta)(X)$ by

$$\mu(X;\theta,\xi) = E^*\left[D(Z)\Sigma(Z)^{-1}\{\Upsilon(Z,X;\xi_1) - g(Z;\theta)\} \mid X;\xi_2 \right],$$

where $\Upsilon(Z,X;\xi_1) = E^*(\tilde{Y} \mid Z, X;\xi_1)$ and $\xi = (\xi_1,\xi_2)$. Provided that $\mu(X;\theta,\xi)$ can be calculated from the above, depending on what $\mathcal{M}_X$ is, we can then derive the augmented estimators similarly to the three examples in Section 3.

5.10. Bias under a misspecified $\mathcal{M}_X$

Proposition 4. *Suppose $\mathbb{P}_X \notin \mathcal{M}_X$, and the "efficient" estimator $\hat{\mathbb{E}}_{\mathcal{M}_X}\mu(X;\theta,\xi)$ is constructed by $\hat{\mathbb{E}}_{\mathcal{M}_X}\mu(X;\theta,\xi) = \hat{\mathbb{P}}_X\mu(X;\theta,\xi)$, where*

$$\hat{\mathbb{P}}_X = \arg\max_{\tilde{\mathbb{P}}_X \in \mathcal{M}_X} \mathbb{P}_n \log \mathrm{d}\tilde{\mathbb{P}}_X(X).$$

As a measure of distance between $\mathcal{M}_X$ and $\mathbb{P}_X$, define

$$\mathrm{KL}(\mathcal{M}_X \mid \mathbb{P}_X) = \inf_{\mathbb{P}^*_X \in \mathcal{M}_X} \mathrm{KL}(\mathbb{P}^*_X \mid \mathbb{P}_X) := \inf_{\mathbb{P}^*_X \in \mathcal{M}_X} \mathbb{P}_X \left\{ \log \frac{\mathrm{d}\mathbb{P}_X}{\mathrm{d}\mathbb{P}^*_X}(X) \right\}.$$

Further suppose that $\|\psi(Y;\theta)\| \le M_0$ *for all* Y *and* θ*, where* $\|\cdot\|$ *denotes the Euclidean or matrix norm, whenever appropriate. Then as* $n \to \infty$*, we have that* $\hat{\theta} \to_p \theta^*$ *with*

$$\|\theta^* - \theta_0\| \le M_0 \|V(\theta_0)\|^{-1} \sqrt{2\mathrm{KL}(\mathcal{M}_X \mid \mathbb{P}_X)}. \tag{29}$$

Proof. By a standard result on the MLE, $\hat{\mathbb{P}}_X \to \mathbb{P}_X^*$ for some $\mathbb{P}_X^* \in \mathcal{M}_X$ that minimizes the KL divergence with $\mathbb{P}_X$ (Shao, 2003), that is, $\mathrm{KL}(\mathbb{P}_X^* \mid \mathbb{P}_X) = \mathrm{KL}(\mathcal{M}_X \mid \mathbb{P}_X)$. But by the form of $\hat{\theta}$ in Theorem 2 and the discussion in Section 4, we have that

$$\begin{aligned}
\|\theta^* - \theta_0\| &\le \|V(\theta_0)\|^{-1} \|(\mathbb{P}_X^* - \mathbb{P}_X)\mu(X)\| \\
&\le M_0 \|V(\theta_0)\|^{-1} \int |\mathrm{d}\mathbb{P}_X^* - \mathrm{d}\mathbb{P}_X| \\
&\le M_0 \|V(\theta_0)\|^{-1} \sqrt{2\mathrm{KL}(\mathbb{P}_X^* \mid \mathbb{P}_X)} \\
&= M_0 \|V(\theta_0)\|^{-1} \sqrt{2\mathrm{KL}(\mathcal{M}_X \mid \mathbb{P}_X)},
\end{aligned}$$

where the third inequality uses the Pinsker inequality that bounds the total variation by the KL divergence (see, e.g., Csiszár and Körner, 2011) □

In Example 1, for instance, we have that $M_0 = 1$ and $\|V(\theta_0)\| = 1$. Then we can use (29) to bound the bias when the true $\mathbb{P}_X$ is, say, exponential rather than Gaussian.

Acknowledgments

I thank the editor, associate editor, and three anonymous referees for helpful comments.

Funding

This research was supported by the U.S. National Science Foundation grant DMS2015526 and National Institutes of Health grant R01HL149875.

Supplementary Material

Supplement to "Robust improvement of efficiency using information on covariate distribution"
(doi: 10.1214/24-EJS2311SUPP; .pdf). Supplementary Material online includes technical details and additional numerical studies.

References

Arcones, M. A. and Giné, E. (1993). Limit theorems for U-processes. *Ann. Prob.* 1494–1542. MR1235426

Benkeser, D., Díaz, I., Luedtke, A., Segal, J., Scharfstein, D. and Rosenblum, M. (2021). Improving precision and power in randomized trials for COVID-19 treatments using covariate adjustment, for binary, ordinal, and time-to-event outcomes. *Biometrics* **77** 1467–1481. MR4357852

Bickel, P. J., Klaassen, C. A., Bickel, P. J., Ritov, Y., Klaassen, J., Wellner, J. A. and Ritov, Y. (1993). *Efficient and adaptive estimation for semiparametric models*. Baltimore: Johns Hopkins University Press. MR1245941

Chatterjee, N., Kalaylioglu, Z. and Carroll, R. J. (2005). Exploiting gene-environment independence in family-based case-control studies: increased power for detecting associations, interactions and joint effects. *Genet. Epidemiol.* **28** 138–156.

Csiszár, I. and Körner, J. (2011). *Information theory: coding theorems for discrete memoryless systems*. Cambridge: Cambridge University Press. MR2839250

Díaz, I., Hejazi, N. S., Rudolph, K. E. and van Der Laan, M. J. (2021). Nonparametric efficient causal mediation with intermediate confounders. *Biometrika* **108** 627–641. MR4298768

Edwards, A. (2008). Anecdotal, Historical and Critical Commentaries on Genetics: GH Hardy (1908) and Hardy–Weinberg Equilibrium. *Genetics* **179** 1143.

Hájek, J. (1968). Asymptotic normality of simple linear rank statistics under alternatives. *Ann. Math. Statist.* 325–346. MR0222988

Halperin, I. (1962). The product of projection operators. *Acta. Sci. Math.* **23** 96–99. MR0141978

Hernando, D., Zhao, R., Yuan, Q., Aliyari Ghasabeh, M., Ruschke, S., Miao, X., Karampinos, D. C., Mao, L., Harris, D. T., Mattison, R. J. et al. (2023). Multicenter reproducibility of liver iron quantification with 1.5-T and 3.0-T MRI. *Radiology* **306** e213256.

Koh, H. K., Geller, A. C. and VanderWeele, T. J. (2021). Deaths from COVID-19. *J. Am. Med. Assoc.* **325** 133–134.

Mao, L. (2024). Supplement to "Robust improvement of efficiency using information on covariate distribution". https://doi.org/10.1214/24-EJS2311SUPP.

Papoulis, A. and Pillai, S. U. (2002). *Probability, random variables, and stochastic processes*. New York: McGraw-Hill. MR0176501

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Am. Statist. Assoc.* **89** 846–866. MR1294730

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Am. Statist. Assoc.* **90** 106–121. MR1325118

Shao, J. (2003). *Mathematical statistics*. New York: Springer. MR2002723

Tsiatis, A. A. (2006). *Semiparametric theory and missing data*. New York: Springer. MR2233926

Tsiatis, A. A., Davidian, M., Zhang, M. and Lu, X. (2008). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: a principled yet flexible approach. *Statist. Med.* **27** 4658–4677. MR2528575

van der Laan, M. J. and Rose, S. (2018). *Targeted learning in data science*. Springer. MR3791826

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge: Cambridge University Press. MR1652247

van der Vaart, A. W. and Wellner, J. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. New York: Springer. MR1385671