SUPPLEMENTARY MATERIAL

A general framework for inference on algorithm-agnostic variable importance

B.D. Williamson, P.B. Gilbert, N.R. Simon & M. Carone

November 1, 2021

1 Proof of theorems

1.1 Proof of Theorem 1

Writing $r_n := \{V(f_n, P_n) - V(f_n, P_0)\} - \{V(f_0, P_n) - V(f_0, P_0)\}$, we first decompose

$$v_n - v_0 = \{V(f_0, P_n) - V(f_0, P_0)\} + \{V(f_n, P_0) - V(f_0, P_0)\} + r_n \ .$$

In view of condition (A2), the functional delta method is applicable and yields that

$$\begin{aligned} V(f_0, P_n) - V(f_0, P_0) &= \dot{V}(f_0, P_0; P_n - P_0) + o_P(n^{-1/2}) \\ &= \frac{1}{n}\sum_{i=1}^{n} \dot{V}(f_0, P_0; \delta_{Z_i} - P_0) + o_P(n^{-1/2}) \ , \end{aligned}$$

where $\dot{V}(f_0, P_0; h)$ is the Gâteaux derivative of the mapping $P \mapsto V(f_0, P)$ at P_0 in the direction h and δ_z is the degenerate distribution on z. Under condition (A1), we have that $|V(f_n, P_0) - V(f_0, P_0)| \leq C\|f_n - f_0\|_{\mathcal{F}}^2 = o_P(n^{-1/2})$ under condition (B1). It remains to show that $r_n = o_P(n^{-1/2})$ as well. For any given $\epsilon > 0$, $h \in \mathcal{Q}$ and $f \in \mathcal{F}$, we define

$$R_0(f, \epsilon, h) := \frac{V(f, P_0 + \epsilon h) - V(f, P_0)}{\epsilon} - \dot{V}(f, P_0; h) \ .$$

Setting $\epsilon_n := n^{-1/2}$ and $h_n := n^{1/2}(P_n - P_0)$, we have that

$$\begin{aligned} n^{1/2} r_n &= \frac{[\{V(f_n, P_n) - V(f_n, P_0)\} - \{V(f_0, P_n) - V(f_0, P_0)\}]}{\epsilon_n} \\ &= \{\dot{V}(f_n, P_0; h_n) + R_0(f_n, \epsilon_n, h_n)\} - \{\dot{V}(f_0, P_0; h_n) + R_0(f_0, \epsilon_n; h_n))\} = A_n + B_n \ , \end{aligned}$$

where $A_n := \dot{V}(f_n, P_0; h_n) - \dot{V}(f_0, P_0; h_n)$ and $B_n := R_0(f_n, \epsilon_n, h_n) - R_0(f_0, \epsilon_n, h_n)$, and so, we can write that $P_0\left(n^{1/2}|r_n| > \epsilon\right) \leq P_0\left(|A_n| > \epsilon/2\right) + P_0\left(|B_n| > \epsilon/2\right)$. On one hand, since we can rewrite $A_n = \dot{V}(f_n, P_0; h_n) - \dot{V}(f_0, P_0; h_n) = n^{1/2} \int g_n(z) d(P_n - P_0)(z)$, under conditions (B2) and (B3), an application of Lemma 19.24 of van der Vaart (2000) yields that $A_n = o_P(1)$ under P_0, and so, $P_0(|A_n| > \epsilon/2) \longrightarrow 0$. On the other hand, we can write

$$\begin{aligned} P_0\left(|B_n| > \epsilon/2\right) &= P_0\left(|B_n| > \epsilon/2, \|f_n - f_0\|_{\mathcal{F}} < \delta\right) + P_0\left(|B_n| > \epsilon/2, \|f_n - f_0\|_{\mathcal{F}} \geq \delta\right) \\ &\leq P_0\left(\sup_{f \in \mathcal{F}: \|f - f_0\|_{\mathcal{F}} < \delta} R_0(f, \epsilon_n, h_n) > \epsilon/4, \|f_n - f_0\|_{\mathcal{F}} < \delta\right) + P_0\left(\|f_n - f_0\|_{\mathcal{F}} \geq \delta\right) \\ &\leq P_0\left(\sup_{f \in \mathcal{F}: \|f - f_0\|_{\mathcal{F}} < \delta} R_0(f, \epsilon_n, h_n) > \epsilon/4\right) + P_0\left(\|f_n - f_0\|_{\mathcal{F}} \geq \delta\right). \end{aligned}$$

Since the first and second summands tend to zero by conditions (A2) and (B1), respectively, it follows that $P_0\left(|B_n| > \epsilon/2\right) \longrightarrow 0$. In summary, under conditions (A1)–(A2) and (B1)–(B3), we find that

$$v_n - v_0 = \frac{1}{n} \sum_{i=1}^{n} \dot{V}(f_0, P_0; \delta_{Z_i} - P_0) + o_P(n^{-1/2})$$

under sampling from P_0, as claimed.

Now, we verify the claim of asymptotic efficiency. Let s be any bounded element of $L_2^0(P_0)$. We construct the parametric submodel $\{P_{0,\epsilon}\}$ with univariate index ϵ defined in a neighborhood of zero and with corresponding distribution function defined pointwise as $F_{0,\epsilon}(z) := F_0(z) + \epsilon \int_{(-\infty, z]} s(u) F_0(du)$, where F_0 denotes the distribution function of P_0 and $(-\infty, z]$ is interpreted as an orthant in the dimension of $\mathcal{Z}$. We note that $z \mapsto \int_{(-\infty, z]} s(u) F_0(du)$ induces a finite signed measure $h(s) \in \mathcal{R}$ since it is cadlag and has finite total variation norm. We then write that

$$\begin{aligned} &\left| \frac{V(f_{0,\epsilon}, P_{0,\epsilon}) - V(f_0, P_0)}{\epsilon} - \dot{V}(f_0, P_0; h(s)) \right| \\ &\leq \left| \frac{V(f_{0,\epsilon}, P_{0,\epsilon}) - V(f_{0,\epsilon}, P_0) + V(f_{0,\epsilon}, P_0) - V(f_0, P_0)}{\epsilon} - \dot{V}(f_{0,\epsilon}, P_0; h(s)) + \dot{V}(f_{0,\epsilon}, P_0; h(s)) - \dot{V}(f_0, P_0; h(s)) \right| \end{aligned}$$

$$\leq \; U_1(\epsilon) + U_2(\epsilon) + U_3(\epsilon) \;,$$

where we have defined the summands

$$U_1(\epsilon) := \left| \frac{V(f_{0,\epsilon}, P_{0,\epsilon}) - V(f_{0,\epsilon}, P_0)}{\epsilon} - \dot{V}(f_{0,\epsilon}, P_0; h(s)) \right|, \;\; U_2(\epsilon) := \left| \frac{V(f_{0,\epsilon}, P_0) - V(f_0, P_0)}{\epsilon} \right|$$

and $U_3(\epsilon) := |\dot{V}(f_{0,\epsilon}, P_0; h(s)) - \dot{V}(f_0, P_0; h(s))|$. By conditions (A1) and (A3), we can bound $U_2(\epsilon)$ above by $C\|f_{0,\epsilon} - f_0\|_{\mathcal{F}}^2/\epsilon = O(\epsilon)$. By condition (A4), we have that $U_3(\epsilon) = O(\epsilon)$. Since $\|f_{0,\epsilon} - f_0\|_{\mathcal{F}} = O(\epsilon)$ by condition (A3), then for small enough ϵ we have that

$$U_1(\epsilon) \;\leq \sup_{f \in \mathcal{F}: \|f - f_0\|_{\mathcal{F}} \leq \delta} \left| \frac{V(f, P_{0,\epsilon}) - V(f, P_0)}{\epsilon} - \dot{V}(f, P_0; h(s)) \right|,$$

where the right-hand side of the inequality itself tends to zero as $\epsilon \to 0$ in view of condition (A2). In other words, we find that $U_1(\epsilon) = O(\epsilon)$. Thus, we find that

$$\left| \frac{V(f_{0,\epsilon}, P_{0,\epsilon}) - V(f_0, P_0)}{\epsilon} - \dot{V}(f_0, P_0; h(s)) \right| = O(\epsilon) \;,$$

which implies that the derivative $\epsilon \mapsto V(f_{0,\epsilon}, P_{0,\epsilon})$ at $\epsilon = 0$ equals $\dot{V}(f_0, P_0; h(s))$. In view of Frangakis et al. (2015) and Luedtke et al. (2015), the evaluation of the nonparametric efficient influence function at observation value z is obtained by choosing s so that $h(s) = \delta_z - P_0$, establishing that v_n is indeed asymptotically efficient relative to a nonparametric model.

1.2 Proof of Theorem 2

As before, we denote by $B_n \in \{1, \ldots, K\}^n$ a random vector generated by sampling uniformly from $\{1, \ldots, K\}$ with replacement, and by D_k the subset of observations with index in $\{i : B_{n,i} = k\}$ for $k = 1, \ldots, K$. Additionally, we denote by $f_{k,n}$ an estimator of f_0 constructed using the data in $\cup_{j \neq k} D_j$, and we write $P_{k,n}$ for the empirical distribution estimator of P_0 based on the data in D_k. Recalling that $v_n^* = \frac{1}{K} \sum_{k=1}^K V(f_{k,n}, P_{k,n})$, we note that $v_n^* - v_0 = A_{1,K,n} + A_{2,K,n} + A_{3,K,n}$, where $A_{1,K,n} := \frac{1}{K} \sum_{k=1}^K \{V(f_0, P_{k,n}) - V(f_0, P_0)\}$, $A_{2,K,n} := \frac{1}{K} \sum_{k=1}^K \{V(f_{k,n}, P_0) - V(f_0, P_0)\}$ and $A_{3,K,n} := \frac{1}{K} \sum_{k=1}^K r_{k,n}$ with $r_{k,n} := \{V(f_{k,n}, P_{k,n}) - V(f_{k,n}, P_0)\} - \{V(f_0, P_{k,n}) - V(f_0, P_0)\}$. We will study separately each of these three summands.

Under condition (A2), the functional delta method can be used to establish the representation $V(f_0, P_{k,n}) - V(f_0, P_0) = \dot{V}(f_0, P_0; P_{k,n} - P_0) + o_P(n_k^{-1/2}) = \frac{1}{n_k} \sum_{i \in D_k} \dot{V}(f_0, P_0; \delta_{Z_i} - P_0) + o_P(n_k^{-1/2})$ for each $k \in \{1, \ldots, K\}$, from which it follows that

$$\begin{aligned} \left| A_{1,K,n} - \frac{1}{n} \sum_{i=1}^{n} \dot{V}(f_0, P_0; \delta_{Z_i} - P_0) \right| &\leq \max_k \left| \frac{n}{K n_k} - 1 \right| \cdot \frac{1}{n} \sum_{i=1}^{n} \dot{V}(f_0, P_0; \delta_{Z_i} - P_0) + \frac{1}{K} \sum_{k=1}^{K} o_P(n_k^{-1/2}) \\ &= O_P(n^{-1}) + o_P(n^{-1/2}) \;=\; o_P(n^{-1/2}) \; . \end{aligned}$$

Under conditions (A1) and (B1'), we have that

$$|A_{2,K,n}| \;\leq\; \max_k |V(f_{k,n}, P_0) - V(f_0, P_0)| \;\leq\; C \max_k \|f_{k,n} - f_0\|_{\mathcal{F}}^2 \;=\; o_P(n^{-1/2}) \; .$$

Finally, we show that $|A_{3,K,n}| = o_P(n^{-1/2})$ by showing that $|r_{k,n}| = o_P(n^{-1/2})$ for each k. Similarly as in the proof of Theorem 1, setting $\epsilon_{k,n} := n_k^{-1/2}$ and $h_{k,n} := n_k^{1/2}(P_{k,n} - P_0)$, we can write that $n_k^{1/2} r_{k,n} = A_{k,n} + B_{k,n}$, where we have defined the terms $A_{k,n} := \dot{V}(f_{k,n}, P_0; h_{k,n}) - \dot{V}(f_0, P_0; h_{k,n})$ and $B_{k,n} := R_0(f_{k,n}, \epsilon_{k,n}, h_{k,n}) - R_0(f_0, \epsilon_{k,n}, h_{k,n})$. Following the same argument made for B_n in the proof of Theorem 1, we can show that $B_{k,n} = o_P(1)$. We then note that $A_{k,n} = n_k^{1/2} \int g_{k,n}(z) d(P_{k,n} - P_0)(z)$. For any $\varepsilon > 0$, by Chebyshev's inequality, we have that

$$0 \;\leq\; P_0 \left(|A_{k,n}| > \varepsilon \,|\, \cup_{j \neq k} D_j \right) \;\leq\; \frac{var_0 \left[g_{k,n}(Z) \,|\, \cup_{j \neq k} D_j \right]}{\varepsilon^2} \;\leq\; \frac{P_0 g_{k,n}^2}{\varepsilon^2} \; .$$

Thus, by condition (B2'), we have that $P_0 \left(|A_{k,n}| > \varepsilon \,|\, \cup_{j \neq k} D_j \right) = o_P(1)$. Since $P_0 \left(|A_{k,n}| > \varepsilon \,|\, \cup_{j \neq k} D_j \right)$ is uniformly bounded by virtue of being a probability, this implies that $E_0[P_0 \left(|A_{k,n}| > \varepsilon \,|\, \cup_{j \neq k} D_j \right)] = o(1)$, and so, $P_0 \left(|A_{k,n}| > \varepsilon \right) = o(1)$. Thus, we find that $A_{k,n} = o_P(1)$. As such, we have found that $|r_{k,n}| = o_P(n_k^{-1/2})$, and since $n/n_k \xrightarrow{P} K$, this implies that $|r_{k,n}| = o_P(n^{-1/2})$.

The proof of nonparametric asymptotic efficiency is identical to that provided for Theorem 1.

1.3 Proof of Theorem 3

Fix an arbitrary $h \in \mathcal{H}$, and let $\{P_{0,\epsilon}\} \subset \mathcal{M}$ be an arbitrary regular univariate parametric submodel through P_0 at $\epsilon = 0$ and with score h for ϵ at $\epsilon = 0$. Write $f_{0,\epsilon} := f_{P_{0,\epsilon}}$ for brevity. We note that

$$V(f_{0,\epsilon}, P_{0,\epsilon}) - V(f_0, P_0) \;=\; V(f_{0,\epsilon}, P_{0,\epsilon}) - V(f_0, P_{0,\epsilon}) + V(f_0, P_{0,\epsilon}) - V(f_0, P_0)$$

$$= V(f_{0,\epsilon}, P_0) - V(f_0, P_0) + V(f_0, P_{0,\epsilon}) - V(f_0, P_0) + o(\epsilon) \ , \qquad \text{(S1)}$$

where the second line follows from the first in view of condition (A5a). By the nonparametric pathwise differentiability of $P \mapsto V(f_0, P)$ at P_0, we have that $V(f_0, P_{0,\epsilon}) - V(f_0, P_0) = \epsilon \int d_0(z)h(z)dP_0(z) + O(\epsilon^2)$, where d_0 is the nonparametric EIF of $P \mapsto V(f_0, P)$ at P_0. Condition (A5b) and (A5c) together indicate that

$$\frac{d}{d\epsilon} V(f_{0,\epsilon}, P_0)\bigg|_{\epsilon=0} = 0 \ ,$$

and furthermore, that $V(f_{0,\epsilon}, P_0) - V(f_0, P_0) = o(\epsilon)$. So, in view of equation S1, we obtain the representation $V(f_{0,\epsilon}, P_{0,\epsilon}) - V(f_0, P_0) = \epsilon \int d_0(z)h(z)dP_0(z) + o(\epsilon)$, which implies that $P \mapsto V(f_P, P)$ is pathwise differentiable at P_0 relative to the nonparametric model $\mathcal{M}$ and has nonparametric EIF d_0.

2 Explicit description of estimation procedure for Examples 1–4

In this section, we provide the explicit form of our proposed estimator for Examples 1–4. For each example, we describe both the simple plug-in estimator and the cross-fitted estimator. When we discuss cross-fitting, recall that we generate a random partition assignment vector $B_n \in \{1, \ldots, K\}^n$ by sampling uniformly from $\{1, \ldots, K\}$ with replacement, and denote by D_k the subset of observations with index in $\{i : B_{n,i} = k\}$ for $k = 1, \ldots, K$. For each $k = 1, \ldots, K$, we denote by $f_{k,n}$ and $f_{k,n,s}$ estimators of f_0 and $f_{0,s}$, respectively, constructed on the data in $\bigcup_{j \neq k} D_j$, and we denote by $P_{k,n}$ the empirical distribution estimator of P_0 based on the data in D_k.

Example 1: R^2

The difference in R^2 VIM estimator is

$$\psi_{n,s} = \left[1 - \frac{\sum_{i=1}^n \{Y_i - f_n(X_i)\}^2}{\sum_{i=1}^n (Y_i - \overline{Y}_n)^2}\right] - \left[1 - \frac{\sum_{i=1}^n \{Y_i - f_{n,s}(X_i)\}^2}{\sum_{i=1}^n (Y_i - \overline{Y}_n)^2}\right],$$

where $\overline{Y}_n := \frac{1}{n}\sum_{i=1}^n Y_i$ is the marginal empirical mean of Y. In this example, $f_n = \mu_n$ and $f_{n,s} = \mu_{n,s}$, where μ_n and $\mu_{n,s}$ are estimators of μ_0 and $\mu_{0,s}$, respectively. For each $k = 1, \ldots, K$, the fold-specific difference in R^2 VIM estimator is

$$\psi_{k,n,s} = \left[1 - \frac{\frac{1}{n_k}\sum_{i \in D_k} \{Y_i - f_{k,n}(X_i)\}^2}{\frac{1}{n_k}\sum_{i \in D_k} (Y_i - \overline{Y}_{k,n})^2}\right] - \left[1 - \frac{\frac{1}{n_k}\sum_{i \in D_k} \{Y_i - f_{k,n,s}(X_i)\}^2}{\frac{1}{n_k}\sum_{i \in D_k} (Y_i - \overline{Y}_{k,n})^2}\right],$$

where $n_k := \sum_{i=1}^{n} I(i \in D_k)$ is the number of observations in fold k, and $\overline{Y}_{k,n} := \frac{1}{n_k} \sum_{i \in D_k}^{n} Y_i$ is the marginal empirical mean of Y in fold k. The cross-fitted estimator is then $\psi_{n,s}^* = \frac{1}{K} \sum_{k=1}^{K} \psi_{k,n,s}$.

Example 2: deviance

The difference in deviance VIM estimator is

$$\psi_{n,s} = \left[1 - \frac{\frac{1}{n}\sum_{i=1}^{n}\{Y_i \log f_n(X_i) + (1-Y_i)\log(1-f_n(X_i))\}}{\pi_n \log(\pi_n) + (1-\pi_n)\log(1-\pi_n)}\right] - \left[1 - \frac{\frac{1}{n}\sum_{i=1}^{n}\{Y_i \log f_{n,s}(X_i) + (1-Y_i)\log(1-f_{n,s}(X_i))\}}{\pi_n \log(\pi_n) + (1-\pi_n)\log(1-\pi_n)}\right],$$

where $\pi_n := \frac{1}{n}\sum_{i=1}^{n} Y_i$ is the empirical estimator of the marginal probability $P_0(Y=1)$. Again, in this example, $f_n = \mu_n$ and $f_{n,s} = \mu_{n,s}$. For each $k = 1, \ldots, K$, the fold-specific difference in deviance VIM estimator is

$$\psi_{k,n,s} = \left[1 - \frac{\frac{1}{n_k}\sum_{i \in D_k}\{Y_i \log f_{k,n}(X_i) + (1-Y_i)\log(1-f_{k,n}(X_i))\}}{\pi_{k,n} \log(\pi_{k,n}) + (1-\pi_{k,n})\log(1-\pi_{k,n})}\right] - \left[1 - \frac{\frac{1}{n_k}\sum_{i \in D_k}\{Y_i \log f_{k,n,s}(X_i) + (1-Y_i)\log(1-f_{k,n,s}(X_i))\}}{\pi_{k,n} \log(\pi_{k,n}) + (1-\pi_{k,n})\log(1-\pi_{k,n})}\right],$$

where $\pi_{k,n} := \frac{1}{n_k}\sum_{i \in D_k} Y_i$ is the marginal estimator of $P_0(Y=1)$ in fold k. The cross-fitted estimator is then $\psi_{n,s}^* = \frac{1}{K}\sum_{k=1}^{K} \psi_{k,n,s}$.

Example 3: classification accuracy

The difference in classification accuracy VIM estimator is $\psi_{n,s} = \frac{1}{n}\sum_{i=1}^{n} I\{Y_i = f_n(X_i)\} - \frac{1}{n}\sum_{i=1}^{n} I\{Y_i = f_{n,s}(X_i)\}$. Sensible estimators of f_0 and $f_{0,s}$ are given by

$$f_n : x \mapsto I\{\mu_n(x) > 0.5\} \quad \text{and} \quad f_{n,s} : x \mapsto I\{\mu_{n,s}(x) > 0.5\}.$$

The fold-specific difference in classification accuracy VIM estimator is

$$\psi_{k,n,s} = \frac{1}{n_k}\sum_{i \in D_k} I\{Y_i = f_{k,n}(X_i)\} - \frac{1}{n_k}\sum_{i \in D_k} I\{Y_i = f_{k,n,s}(X_i)\}.$$

The cross-fitted estimator is then $\psi_{n,s}^* = \frac{1}{K}\sum_{k=1}^{K} \psi_{k,n,s}$.

Example 4: area under the ROC curve

The difference in AUC VIM estimator is

$$\psi_{n,s} = \frac{1}{n_0 n_1} \sum_{i=1}^{n} \sum_{j=1}^{n} I\{f_n(X_i) < f_n(X_j)\}(1 - Y_i)Y_j - \frac{1}{n_0 n_1} \sum_{i=1}^{n} \sum_{j=1}^{n} I\{f_{n,s}(X_i) < f_{n,s}(X_j)\}(1 - Y_i)Y_j \ ,$$

where $n_1 := \sum_{i=1}^{n} Y_i$ is the number of observations with corresponding $Y = 1$ and $n_0 := n - n_1$. As above, in this example, we can take $f_n = \mu_n$ and $f_{n,s} = \mu_{n,s}$. The fold-specific difference in AUC VIM estimator is

$$\begin{aligned} \psi_{k,n,s} = & \frac{1}{n_{k,0} n_{k,1}} \sum_{i \in D_k} \sum_{j \in D_k} I\{f_{k,n}(X_i) < f_{k,n}(X_j)\}(1 - Y_i)Y_j \\ & - \frac{1}{n_{k,0} n_{k,1}} \sum_{i \in D_k} \sum_{j \in D_k} I\{f_{k,n,s}(X_i) < f_{k,n,s}(X_j)\}(1 - Y_i)Y_j, \end{aligned}$$

where $n_{k,1} := \sum_{i \in D_k} I(Y_i = 1)$ is the number of observations with corresponding $Y = 1$ in fold k and $n_{k,0} := n_k - n_{k,1}$. The cross-fitted estimator is then $\psi^*_{n,s} = \frac{1}{K} \sum_{k=1}^{K} \psi_{k,n,s}$.

3 Additional technical details

3.1 Bayes classifier maximizes classification accuracy

Suppose that $Y \in \{0, 1\}$ is a binary random variable. Define the Bayes classifier $b_0 : x \mapsto I\{\mu_0(x) > 1/2\}$ with $\mu_0(x) = E_0(Y \mid X = x)$. For any fixed $x \in \mathcal{X}$, we have that

$$\begin{aligned} P_0\{f(X) = Y \mid X = x\} &= P_0\{Y = 1, f(X) = 1 \mid X = x\} + P_0\{Y = 0, f(X) = 0 \mid X = x\} \\ &= f(x) P_0(Y = 1 \mid X = x) + \{1 - f(x)\} P_0(Y = 0 \mid X = x) \\ &= f(x)\mu_0(x) + \{1 - f(x)\}\{1 - \mu_0(x)\} \ , \end{aligned}$$

which allows us to write that

$$\begin{aligned} & P_0\{f(X) = Y \mid X = x\} - P_0\{b_0(X) = Y \mid X = x\} \\ & \quad = \mu_0(x)\{f(x) - b_0(x)\} + \{1 - \mu_0(x)\}[\{1 - f(x)\} - \{1 - b_0(x)\}] \\ & \quad = \{2\mu_0(x) - 1\}\{f(x) - b_0(x)\} \ \leq \ 0 \end{aligned}$$

by definition of b_0. It follows then that

$$\begin{aligned} P_0\{f(X) = Y\} - P_0\{b_0(X) = Y\} &= E_0\left[P_0\{f(X) = Y \mid X\}\right] - E_0\left[P_0\{b_0(X) = Y \mid X\}\right] \\ &= E_0\left[P_0\{f(X) = Y \mid X\} - P_0\{b_0(X) = Y \mid X\}\right] \leq 0 , \end{aligned}$$

so that b_0 is the maximizer of the classification accuracy $P_0\{Y = f(X)\}$.

3.2 Conditional mean maximizes the area under the ROC curve

Suppose that $Y \in \{0, 1\}$ is a binary random variable. For a given function $f \in \mathcal{F}$, we define the conditional distribution functions

$$F_1(P_0, f)(c) := P_0\{f(X) \leq c \mid Y = 1\} \quad \text{and} \quad F_0(P_0, f)(c) := P_0\{f(X) \leq c \mid Y = 0\} .$$

If Y denotes the presence of a disease, then $1 - F_1(P_0, f)(c)$ and $F_0(P_0, f)(c)$ denote the sensitivity and specificity of a medical test that flags the presence of disease if and only if $f(X) > c$. The AUC value corresponding to f and P_0 can be written as

$$\begin{aligned} P_0\{f(X_1) < f(X_2) \mid Y_1 = 0, Y_2 = 1\} &= \int_0^\infty \{1 - F_1(P_0, f)(c)\} F_0(P_0, f)(dc) \\ &= \int_0^1 \left\{1 - F_1(P_0, f)(F_0^{-1}(P_0, f)(w))\right\} dw . \end{aligned}$$

For a fixed w, the integrand $1 - F_1(P_0, f)(F_0^{-1}(P_0, f)(w))$ is the sensitivity of a test based on f and a cutoff that results in specificity w. By an application of the Neyman-Pearson Lemma, it is known that, for any fixed specificity level, any strictly increasing transformation of the likelihood ratio mapping $x \mapsto P_0(Y = 1 \mid X = x)/P_0(Y = 0 \mid X = x) = \mu_0(x)/\{1 - \mu_0(x)\}$ gives an optimal choice of f. in particular, the function $f : x \mapsto \mu_0(x)$ is optimal. Since this is true irrespective of the fixed specificity level, it holds uniformly across specificity levels and hence also maximizes the AUC value, as claimed.

3.3 Verification of conditions (A1) and (A2) for Examples 1–4

Example 1: R^2

We have that $|V(f, P_0) - V(f_0, P_0)| = E_0\{f(X) - f_0(X)\}^2/\sigma^2(P_0)$ so that $|V(f, P_0) - V(f_0, P_0)| =$

$O(\|f-f_0\|_{\mathcal{F}}^2)$ and condition (A1) holds. We can verify that $\dot{V}(f, P_0; h) = -\int\{y-f(x)\}^2 h(dz)/\sigma^2(P_0)$. Since $P \mapsto E_P\{Y-f(X)\}^2$ is linear and thus Hadamard differentiable uniformly in f, condition (A2) can be shown to hold for any $\delta > 0$ provided the marginal distribution of Y under P_0 has bounded support.

Example 2: deviance

Using that $f_0 = \mu_0$ and setting $a_0 := -2/\{\log P_0(Y=0) + \log P_0(Y=1)\}$, a standard argument based on Taylor approximations allows to write that

$$\begin{aligned} |V(f, P_0) - V(f_0, P_0)| &= a_0 \left| E_0 \left[f_0(X) \log\left\{\frac{f(x)}{f_0(x)}\right\} + \{1-f_0(x)\} \log\left\{\frac{1-f(x)}{1-f_0(x)}\right\}\right]\right| \\ &\leq \frac{a_0}{2} E_0 \left[\{f(x)-f_0(x)\}^2 \left\{\frac{f_0(x)}{\xi_0(x)} + \frac{1-f_0(x)}{1-\xi_1(x)}\right\}\right] \end{aligned}$$

for some $\xi_0, \xi_1 : \mathcal{X} \to \mathcal{Y}$ lying pointwise between f and f_0. If $f(X), f_0(X) \in (\delta, 1-\delta)$ almost surely under P_0, then we find that $|V(f, P_0) - V(f_0, P_0)| \leq a_0 \left(\frac{1-\delta}{\delta}\right) \|f-f_0\|_{\mathcal{F}}^2$. Thus, condition (A1) then holds with $\alpha = 2$. Since $P \mapsto E_P\left[Y \log f(X) + (1-Y)\log\{1-f(X)\}\right]$ is linear and thus Hadamard differentiable uniformly in f, condition (A2) can again be shown to hold for any $\delta > 0$.

Example 3: classification accuracy

Using that $f_0 : x \mapsto I\{\mu_0(x) > 1/2\}$ is an optimizer of accuracy, and writing any candidate prediction function $f : \mathcal{X} \to \{0,1\}$ as $f(x) = I\{\mu(x) > 1/2\}$ for some function $\mu : \mathcal{X} \to [0,1]$, we can write

$$\begin{aligned} 0 \leq\ & P_0\{Y = f_0(X)\} - P_0\{Y = f(X)\} = E_0\left[I\{Y = f_0(X)\} - I\{Y = f(X)\}\right] \\ =\ & P_0\{Y = f_0(X), Y \neq f(X)\} - P_0\{Y \neq f_0(X), Y = f(X)\} \\ =\ & P_0\{f_0(X) = 1, f(X) = 0, Y = 1\} + P_0\{f_0(X) = 0, f(X) = 1, Y = 0\} \\ & - P_0\{f_0(X) = 0, f(X) = 1, Y = 1\} - P_0\{f_0(X) = 1, f(X) = 0, Y = 0\} \\ =\ & [P_0\{Y = 1 \mid \mu_0(X) \geq \tfrac{1}{2} > \mu(X)\} - P_0\{Y = 0 \mid \mu_0(X) \geq \tfrac{1}{2} > \mu(X)\}]\, P_0\{\mu_0(X) \geq \tfrac{1}{2} > \mu(X)\} \\ & + [P_0\{Y = 0 \mid \mu(X) \geq \tfrac{1}{2} > \mu_0(X)\} - P_0\{Y = 1 \mid \mu(X) \geq \tfrac{1}{2} > \mu_0(X)\}]\, P_0\{\mu(X) \geq \tfrac{1}{2} > \mu_0(X)\} \\ =\ & [2P_0\{Y = 1 \mid \mu_0(X) \geq \tfrac{1}{2} > \mu(X)\} - 1]\, P\{\mu_0(X) \geq \tfrac{1}{2} > \mu(X)\} \\ & + [2P_0\{Y = 0 \mid \mu(X) \geq \tfrac{1}{2} > \mu_0(X)\} - 1]\, P\{\mu(X) \geq \tfrac{1}{2} > \mu_0(X)\}\ . \end{aligned}$$

Now, on one hand, we note that

$$\begin{aligned} P_0\{Y = 1 \mid \mu_0(X) \geq \tfrac{1}{2} > \mu(X)\} - \tfrac{1}{2} &= E_0\{Y \mid \mu_0(X) \geq \tfrac{1}{2} > \mu(X)\} - \tfrac{1}{2} \\ &= E_0\{\mu_0(X) \mid \mu_0(X) \geq \tfrac{1}{2} > \mu(X)\} - \tfrac{1}{2} \;=\; E_0\{\mu_0(X) - \tfrac{1}{2} \mid \mu_0(X) \geq \tfrac{1}{2} > \mu(X)\} \;, \end{aligned}$$

and so it follows that $|P_0\{Y = 1 \mid \mu_0(X) \geq \frac{1}{2} > \mu(X)\} - \frac{1}{2}| \leq \|\mu - \mu_0\|_\infty$. We can similarly show that $|P_0\{Y = 0 \mid \mu(X) \geq \frac{1}{2} > \mu_0(X)\} - \frac{1}{2}| \leq \|\mu - \mu_0\|_\infty$. On the other hand, in view of the margin condition we impose, we have that

$$P_0\{\mu_0(X) \geq \tfrac{1}{2} > \mu(X)\} \;\leq\; P_0\{|\mu_0(X) - \tfrac{1}{2}| < |\mu(X) - \mu_0(X)|\} \;\leq\; \kappa\|\mu - \mu_0\|_\infty$$

and similarly, $P_0\{\mu(X) \geq \frac{1}{2} > \mu_0(X)\} \leq \kappa\|\mu - \mu_0\|_\infty$. Combining the inequalities we have derived, we conclude that $0 \leq P_0\{Y = f_0(X)\} - P_0\{Y = f(X)\} \leq 4\kappa\|\mu - \mu_0\|_\infty$.

Example 4: Area under the ROC curve

We begin by writing

$$\begin{aligned} 0 \;&\leq\; P_0\left\{f_0(X_1) < f_0(X_2), Y_1 = 0, Y_2 = 1\right\} - P_0\left\{f(X_1) < f(X_2), Y_1 = 0, Y_2 = 1\right\} \\ &=\; E_0\left[I\left\{f_0(X_1) < f_0(X_2), Y_1 = 0, Y_2 = 1\right\} - I\left\{f(X_1) < f(X_2), Y_1 = 0, Y_2 = 1\right\}\right] \\ &=\; \tfrac{1}{2}\, E_0\left[I\left\{f_0(X_1) < f_0(X_2), Y_1 = 0, Y_2 = 1\right\} + I\left\{f_0(X_1) \geq f_0(X_2), Y_1 = 1, Y_2 = 0\right\}\right] \\ &\qquad - \tfrac{1}{2}\, E_0\left[I\left\{f(X_1) < f(X_2), Y_1 = 0, Y_2 = 1\right\} + I\left\{f(X_1) \geq f(X_2), Y_1 = 1, Y_2 = 0\right\}\right] \\ &=\; \tfrac{1}{2}\, E_0\left[(Y_2 - Y_1) I\left\{f_0(X_1) < f_0(X_2), f(X_1) \geq f(X_2)\right\}\right] \\ &\qquad + \tfrac{1}{2}\, E_0\left[(Y_1 - Y_2) I\left\{f_0(X_1) \geq f_0(X_2), f(X_1) < f(X_2)\right\}\right] \\ &=\; \tfrac{1}{2}\, E_0\left[\{f_0(X_2) - f_0(X_1)\} I\left\{f_0(X_1) < f_0(X_2), f(X_1) \geq f(X_2)\right\}\right] \\ &\qquad + \tfrac{1}{2}\, E_0\left[\{f_0(X_1) - f_0(X_2)\} I\left\{f_0(X_1) \geq f_0(X_2), f(X_1) < f(X_2)\right\}\right] \\ &\leq\; \tfrac{1}{2}\, E_0\left[|f_0(X_1) - f_0(X_2)| I\left\{[f_0(X_1) - f_0(X_2)][f(X_1) - f(X_2)] < 0\right\}\right]. \end{aligned}$$

Defining $A := \{f(X_1) - f_0(X_1)\} + \{f_0(X_2) - f(X_2)\}$, $B := f_0(X_1) - f_0(X_2)$ and $t : x \mapsto |f(x) - f_0(x)|$, we note that

$$\{[f_0(X_1) - f_0(X_2)][f(X_1) - f(X_2)] < 0\} \;=\; \{B(A + B) < 0\} \;=\; \{(\tfrac{1}{2}A + B)^2 - \tfrac{1}{4}A^2 < 0\}$$

$$= \{|A| > |B|, AB < 0\} \subseteq \{|A| > |B|\} \subseteq \{|f_0(X_1) - f_0(X_2)| < t(X_1) + t(X_2)\} .$$

Using this result and the inequality derived above, and defining $\alpha_0 := \{P_0(Y = 1)P_0(Y = 0)\}^{-1}$, we have that

$$\begin{aligned}
0 &\leq AUC(f_0, P_0) - AUC(f, P_0) \\
&= \alpha_0 \left[P_0 \{f_0(X_1) < f_0(X_2), Y_1 = 0, Y_2 = 1\} - P_0 \{f(X_1) < f(X_2), Y_1 = 0, Y_2 = 1\}\right] \\
&\leq \tfrac{1}{2}\alpha_0 E_0 \left[|f_0(X_1) - f_0(X_2)| I \{[f_0(X_1) - f_0(X_2)][f(X_1) - f(X_2)] < 0\}\right] \\
&\leq \tfrac{1}{2}\alpha_0 E_0 \left[|f_0(X_1) - f_0(X_2)| I \{|f_0(X_1) - f_0(X_2)| < t(X_1) + t(X_2)\}\right] \\
&\leq \tfrac{1}{2}\alpha_0 E_0 \left[|f_0(X_1) - f_0(X_2)| I \{|f_0(X_1) - f_0(X_2)| < 2\|t\|_\infty\}\right] \\
&\leq \alpha_0 \|t\|_\infty P_0 \{|f_0(X_1) - f_0(X_2)| < 2\|t\|_\infty\} \leq 2\alpha_0 \kappa \|t\|_\infty^2 ,
\end{aligned}$$

where the last inequality follows from the margin condition we impose.

3.4 Derivation of the EIFs for Examples 5 and 6

Example 5: Mean outcome under a binary intervention rule

The nonparametric EIF for this example is derived in, for example, Sections 2 and 3 of Luedtke and van der Laan (2016) and in Section A.1 of its supplement.

Example 6: Classification accuracy under outcome missingness

Recall that, in this example, the ideal-data structure consists of $\mathbb{Z} := (X, Y) \sim \mathbb{P}$, and the observed data structure is $Z := (X, \Delta, U)$, where Δ is the indicator of having observed the outcome Y, and we have defined $U := \Delta Y$. The ideal-data nonparametric EIF at $\mathbb{P}$, following Appendix A, is given by

$$\phi_{\mathbb{P}}^F(x, y) = I\{y = f_{\mathbb{P}}(x)\} - V(f_{\mathbb{P}}, \mathbb{P}).$$

Based on results in Chapter 25.5.3 of van der Vaart (2000), the observed-data nonparametric EIF at P is given by

$$\phi_P(z) = \frac{\delta}{g_P(x)}\phi_{\mathbb{P}}^F(z) + \left\{1 - \frac{\delta}{g_P(x)}\right\} E_P\{\phi_{\mathbb{P}}^F(Z) \mid \Delta = 1, X = x\} . \tag{S2}$$

Table S1: Approximate values of $\psi_{0,s}$ in the numerical experiments.

Importance measure	Scenario	X_1	X_2	X_3	X_4	(X_1, X_3)	(X_2, X_4)
Accuracy	(1,2,3)	0.136	0.236	0	0	0.136	0.236
	4	0.081	0.228	0	0	0.136	0.236
Area under the ROC curve	(1,2,3)	0.105	0.221	0	0	0.105	0.221
	4	0.052	0.211	0	0	0.105	0.221

Defining the nuisance function $Q_P(x) := P\{Y = f_P(X) \mid \Delta = 1, X = x\}$, simple algebraic manipulations then yield that $E_P\{\phi^F_{\mathbb{P}}(Z) \mid \Delta = 1, X = x\} = Q_P(x) - V(f_P, P)$. Plugging this into (S2) yields the desired form of the EIF.

4 Additional numerical experiments

4.1 Replicating all numerical experiments

All numerical experiments presented here and in the main manuscript can be replicated using code available on GitHub. In all cases, we generate data by:

1 : drawing $X \sim MVN(0, \Sigma)$

2 : drawing $\epsilon \sim N(0, 1)$ independent of X, and setting $Y = I\{x\beta_0 + \epsilon > 0\}$ given $X = x$,

where Σ is the $p \times p$ identity matrix and $\beta_0 = (2.5, 3.5, 0, \ldots, 0)^\top$. The dimension p is determined by the scenario. The approximate true values of variable importance based on accuracy and AUC under all scenarios considered here are provided in Table S1. The specification of each individual algorithm for estimating f_0 and $f_{0,s}$ is provided in Table S2, while the specification of the candidate algorithms used in the Super Learner is provided in Table S3.

4.2 Properties of our proposal under the alternative hypothesis

In this section, we present additional results under Scenario 1. In this case, $p = 2$. For each scenario presented here, we generated 1000 random datasets of size $n \in \{100, 500, 1000, \ldots, 4000\}$, and considered the importance of both X_1 and X_2. We highlight results for both features using the AUC and for X_1 using accuracy, and we provide the coverage of nominal 95% confidence intervals. We assess performance in the same way as in the main manuscript.

Algorithm	R Implementation	Tuning Parameter(s) and possible values	Tuning parameter description
Generalized linear models	`glm`	–	–
Generalized additive models	`mgcv` (Wood, 2011)	method = `"GCV.Cp"`	Smoothing parameter estimation method
Random forests	`ranger` (Wright and Ziegler, 2017)	`ntree`‡	Number of variables to possibly split at in each node
		`max.depth`‡	Maximum tree depth
		`min.node.size`‡	Minimum node size

Table S2: Individual algorithms considered with their R implementation, tuning parameter values, and description of the tuning parameters. All tuning parameters besides those listed here are set to their default values. In particular, the random forests are grown with 500 trees, `mtry` = $\sqrt{p}$ †, and a subsampling fraction of 1; five-fold cross-validation over the grid defined by (`ntree`, `max.depth`, `min.node.size`) was used to select the tuning parameter combination that minimized log-likelihood loss.
†: p denotes the total number of predictors.
‡: For setting 1, `ntree` $\in \{100, 500, 1000\}$, `max.depth` $= 5$, `min.node.size` $= 1$; for all other settings, `ntree` $\in \{500, 1000, 1500, 2000, 5000\}$, `max.depth` $\in \{1, 3, 5\}$, `min.node.size` $= 10$.

Candidate Learner	R Implementation	Tuning Parameter and possible values	Tuning parameter description
Generalized linear models	`glm`	–	–
Generalized additive models	`gam` (Hastie, 2019)	degree = 2	Degree of smooth terms
Random forests	`ranger` (Wright and Ziegler, 2017)	`mtry` = $\sqrt{p}$ †	Number of variables to possibly split at in each node
Gradient boosted trees	`xgboost` (Chen et al., 2019)	`max.depth` = 1	Maximum tree depth
Elastic net‡	`glmnet` (Friedman et al., 2010)	mixing parameter α = 1	Trade-off between ℓ_1 and ℓ_2 regularization

Table S3: Candidate learners in the Super Learner ensemble along with their R implementation, tuning parameter values, and description of the tuning parameters. All tuning parameters besides those listed here are set to their default values. In particular, the random forests are grown with 500 trees, a minimum node size of 5 for continuous outcomes and 1 for binary outcomes, and a subsampling fraction of 1; the boosted trees are grown with a maximum of 1000 trees, shrinkage rate of 0.1, and a minimum of 10 observations per node; and the lasso ℓ_1 tuning parameter is chosen using 10-fold cross-validation.
†: p denotes the total number of predictors.
‡: lasso is only included in cases where $p \geq 4$.

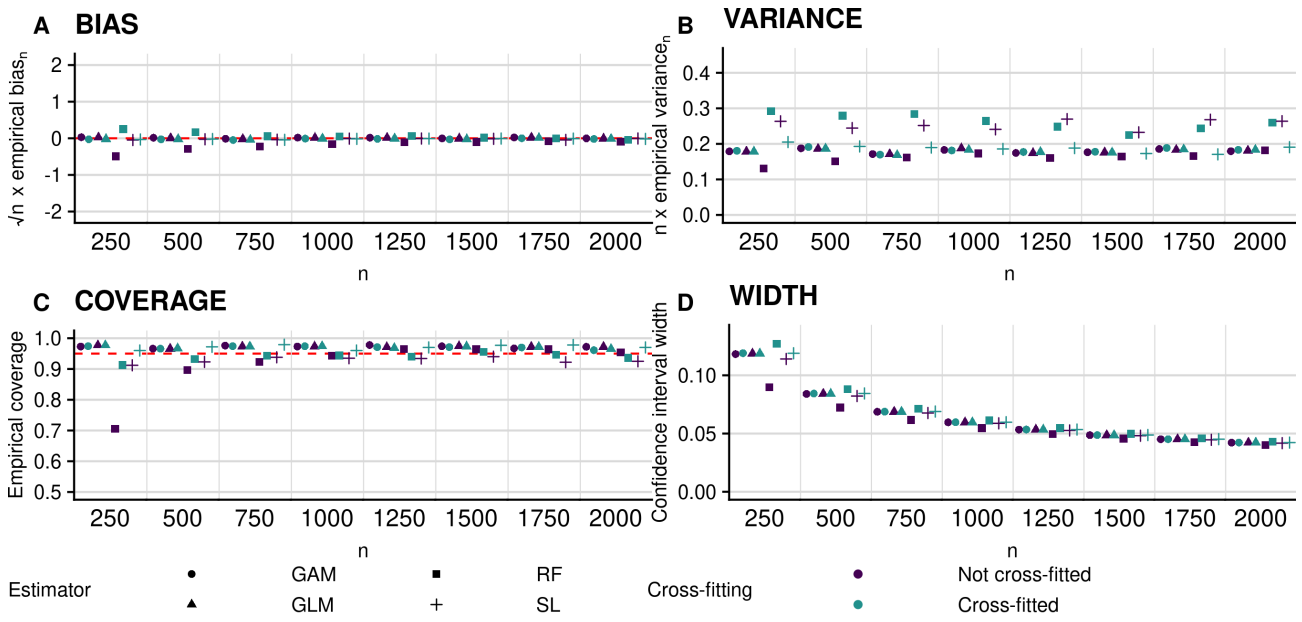

Figure S1: Performance of plug-in estimators for estimating (non-zero) importance of X_1 in terms of accuracy under Scenario 1 (all features have non-zero importance). Clockwise from top left: empirical bias of the proposed plug-in estimator scaled by $n^{1/2}$; empirical variance scaled by n; empirical coverage of nominal 95% confidence intervals; and width of these intervals. Circles, triangles, squares, and plus symbols denote estimators based on the use of generalized additive models (GAMs), probit regression (GLM), random forests (RF) or the Super Learner (SL), respectively. Blue and green symbols denote non-cross-fitted and cross-fitted estimators, respectively. This figure appears in color in the electronic version of this article.

We present results for AUC and for the accuracy-based importance of X_1 in Figures S1–S3. The results for both features and both importance measures are largely similar to those presented in Section 5.2 of the main manuscript. The need for cross-fitting is particularly striking in Figure S3, where we observed coverage near zero for intervals based on a non-cross-fitted random forest estimator of the oracle prediction functions. In Figure S4, we show the coverage of nominal 95% intervals based on the non-cross-fitted standard error estimator. Here, we observe reduced coverage in some cases compared to the results presented above. Taken together, these results highlight that when using simple estimators of the conditional mean functions (e.g., estimators based on correctly-specified parametric models), using cross-fitting appears to have minimal impact on the performance of the proposed inferential procedures and is therefore not needed. In contrast, when flexible nuisance estimators are used, it appears important to use cross-fitting when estimating VIM values and standard errors. The elimination of the constraint on nuisance estimator complexity (i.e., the Donsker class condition) achieved via cross-fitting does appear to translate into substantially improved practical performance when complex nuisance estimators are used.

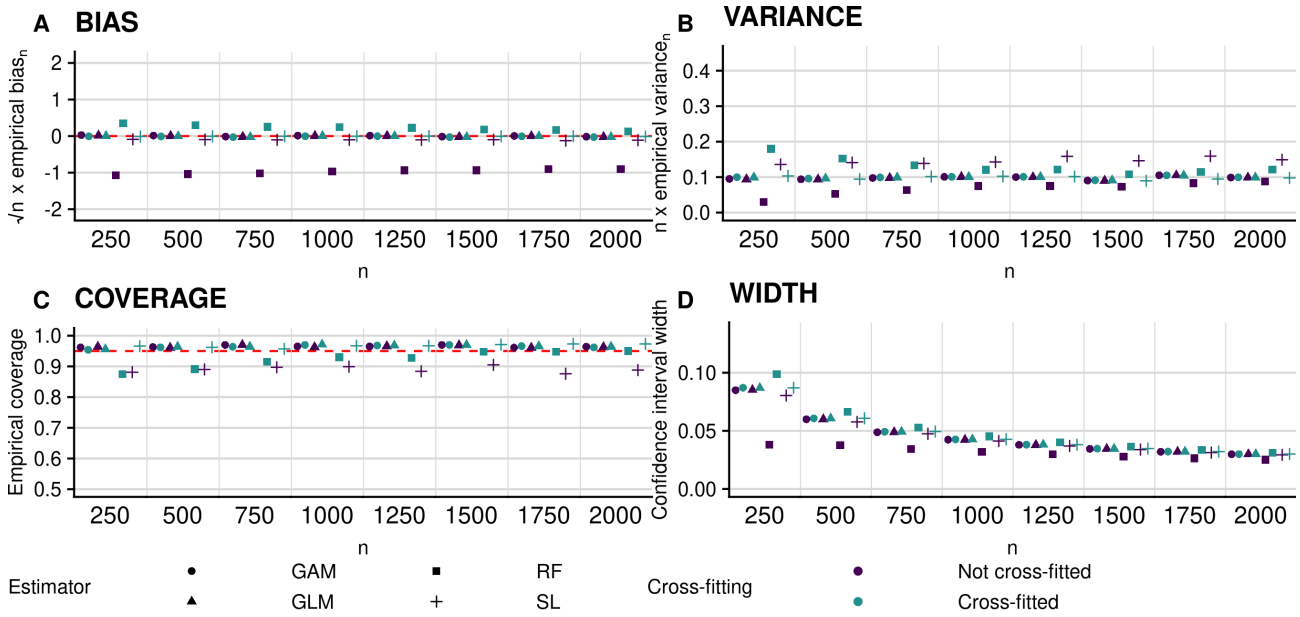

Figure S2: Performance of plug-in estimators for estimating (non-zero) importance of X_1 in terms of AUC under Scenario 1 (all features have non-zero importance). Clockwise from top left: empirical bias of the proposed plug-in estimator scaled by $n^{1/2}$; empirical variance scaled by n; empirical coverage of nominal 95% confidence intervals; and width of these intervals. Circles, triangles, squares, and plus symbols denote estimators based on the use of generalized additive models (GAMs), probit regression (GLM), random forests (RF) or the Super Learner (SL), respectively. Blue and green symbols denote non-cross-fitted and cross-fitted estimators, respectively. Coverage of intervals based on the non-cross-fitted RF-based estimator never exceeds 0.5 and is as low as zero in some cases. This figure appears in color in the electronic version of this article.

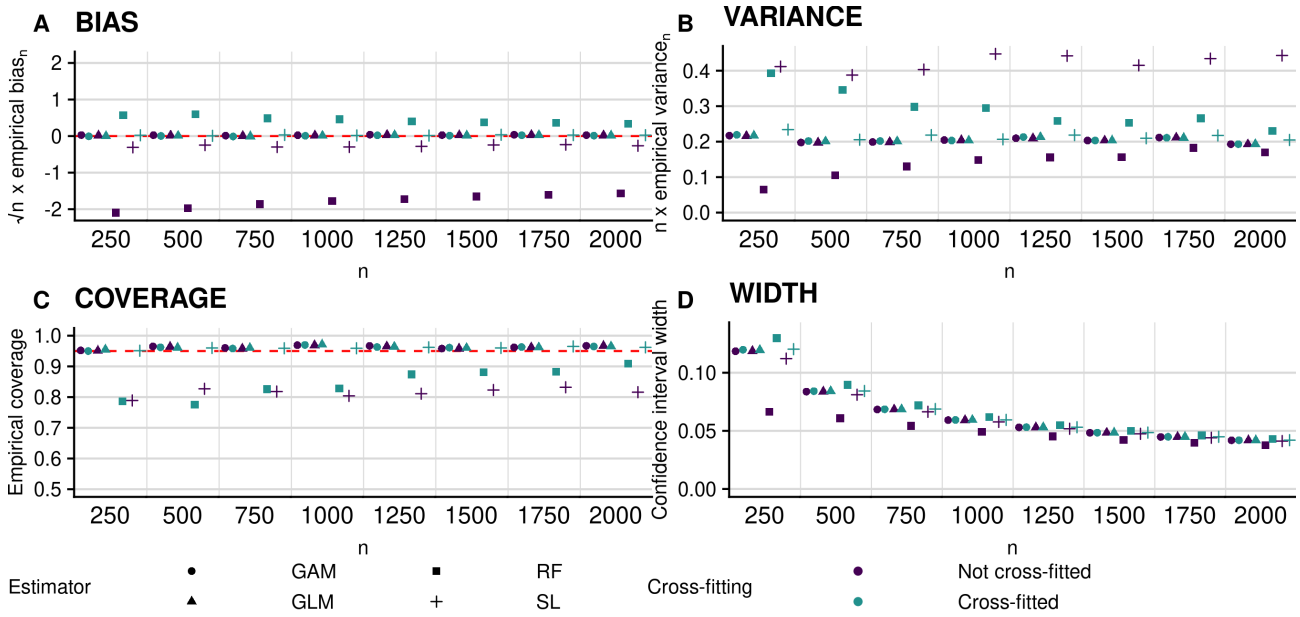

Figure S3: Performance of plug-in estimators for estimating (non-zero) importance of X_2 in terms of AUC under Scenario 1 (all features have non-zero importance). Clockwise from top left: empirical bias of the proposed plug-in estimator scaled by $n^{1/2}$; empirical variance scaled by n; empirical coverage of nominal 95% confidence intervals; and width of these intervals. Circles, triangles, squares, and plus symbols denote estimators based on the use of generalized additive models (GAMs), probit regression (GLM), random forests (RF) or the Super Learner (SL), respectively. Blue and green symbols denote non-cross-fitted and cross-fitted estimators, respectively. Coverage of intervals based on the non-cross-fitted RF-based estimator never exceeds 0.5 and is as low as zero in some cases. This figure appears in color in the electronic version of this article.

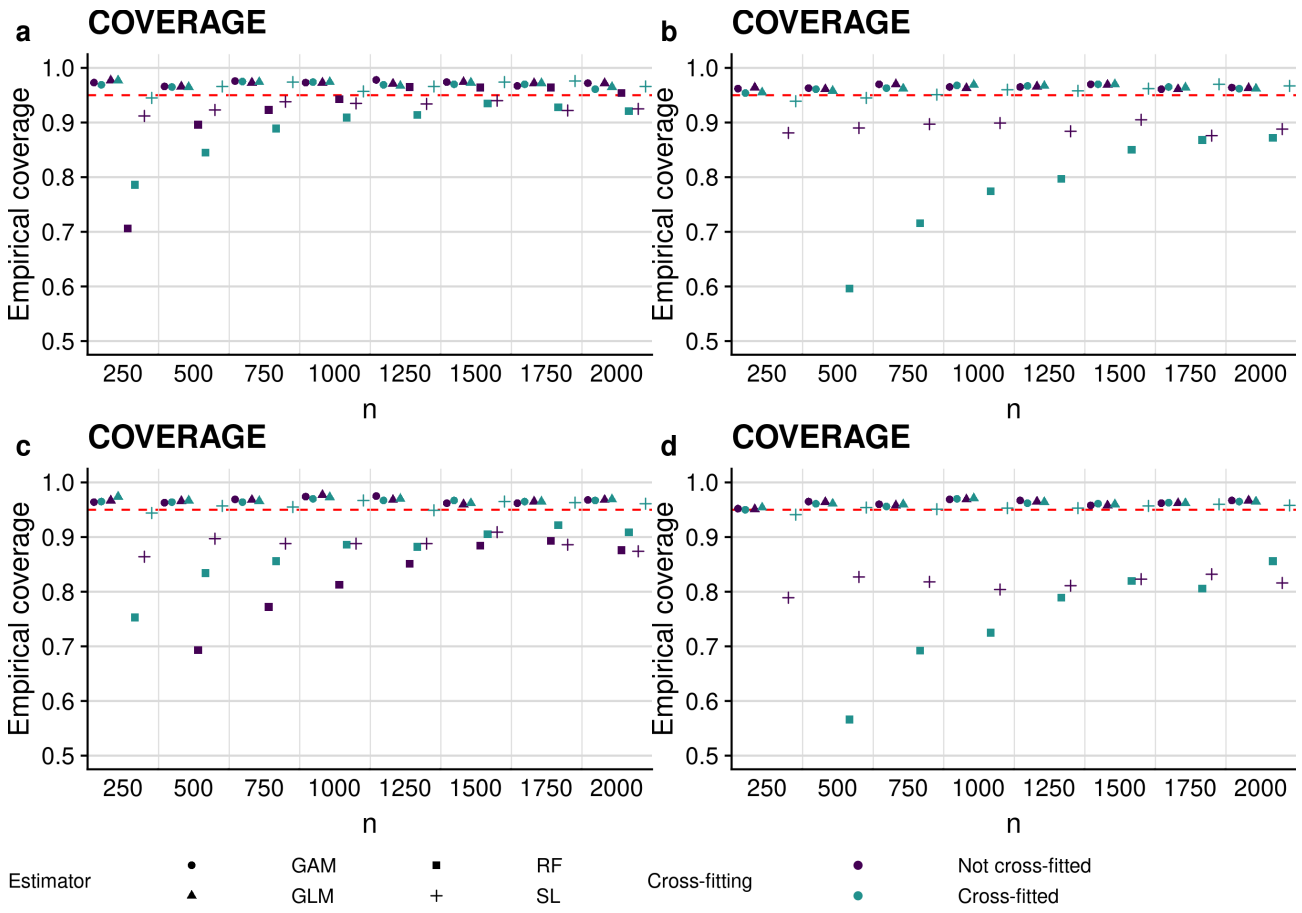

Figure S4: Empirical coverage of confidence intervals based on the non-cross-fitted standard error estimator under Scenario 1 (all features have non-zero importance). The rows correspond to the feature of interest, while the columns correspond to accuracy and AUC, respectively. Circles, triangles, squares, and plus symbols denote estimators based on the use of generalized additive models (GAMs), probit regression (GLM), random forests (RF) or the Super Learner (SL), respectively. Blue and green symbols denote non-cross-fitted and cross-fitted VIM estimators, respectively. Coverage of intervals based on the non-cross-fitted RF-based estimator (panels b and d) never exceeds 0.5 and is as low as zero in some cases. This figure appears in color in the electronic version of this article.

4.3 Properties of our proposal under the null hypothesis

In this section, we present additional results under Scenario 2. In this case, $p = 4$. We again generated 1000 random datasets of size $n \in \{100, 500, 1000, \ldots, 4000\}$, and considered the importance of both X_2 (a non-null feature) and X_3 (a null feature). Here, we highlight results for both features based on the AUC and for X_2 based on accuracy, and we provide the coverage of nominal 95% confidence intervals and proportion of tests rejected. We assess performance in the same way as in the main manuscript.

We present the results based on a cross-fitted standard error estimator in Figures S5–S7. In Figures S5 and S6, we observe high power across all sample sizes. We again observe residual bias for the non-cross-fitted VIM estimators based on flexible nuisance estimation (random forests and the Super Learner). In Figure S7, the cross-fitted VIM estimator based on random forests exhibits some residual bias but coverage and type I error are still near the nominal level. It is possible that this bias could be mitigated with cross-validation over a richer grid of tuning parameters. Similarly as in the main manuscript, since the bias for estimating the null feature appears to be small for the non-cross-fitted estimators, type I error is not inflated in these simulations. However, we expect in most cases that cross-fitting will yield a more adequate type I error control. Indeed, we see that this is the case by comparing the results for the cross-fitted estimator and cross-fitted versus non-cross-fitted standard error estimators (Figure S8). Here, we see a vastly inflated type I error for the cross-fitted random forests-based estimator, reflecting that in this case the non-cross-fitted standard error appears to be too small.

4.4 Using the bootstrap for interval estimation

In some cases, particularly those with limited sample sizes, it may be of interest to use a bootstrap scheme for interval estimation rather than a Wald construction using an influence function-based estimator of the asymptotic variance. Because estimation of f_0 and $f_{0,s}$ only contributes to the second-order behavior of the plug-in VIM estimator, a valid nonparametric bootstrap here would consist of bootstrapping the empirical distribution P_n but fixing the nuisance estimators f_n and $f_{n,s}$ across all bootstrap runs. Not having to re-fit estimators of the nuisance functions on each bootstrap sample makes this scheme particularly efficient to implement. Additionally, since we only use the bootstrap for interval estimation, we do not need to bootstrap the cross-fitting procedure. Our proposed bootstrap procedure in a case with no sample-splitting (i.e., under the alternative hypothesis) is as follows:

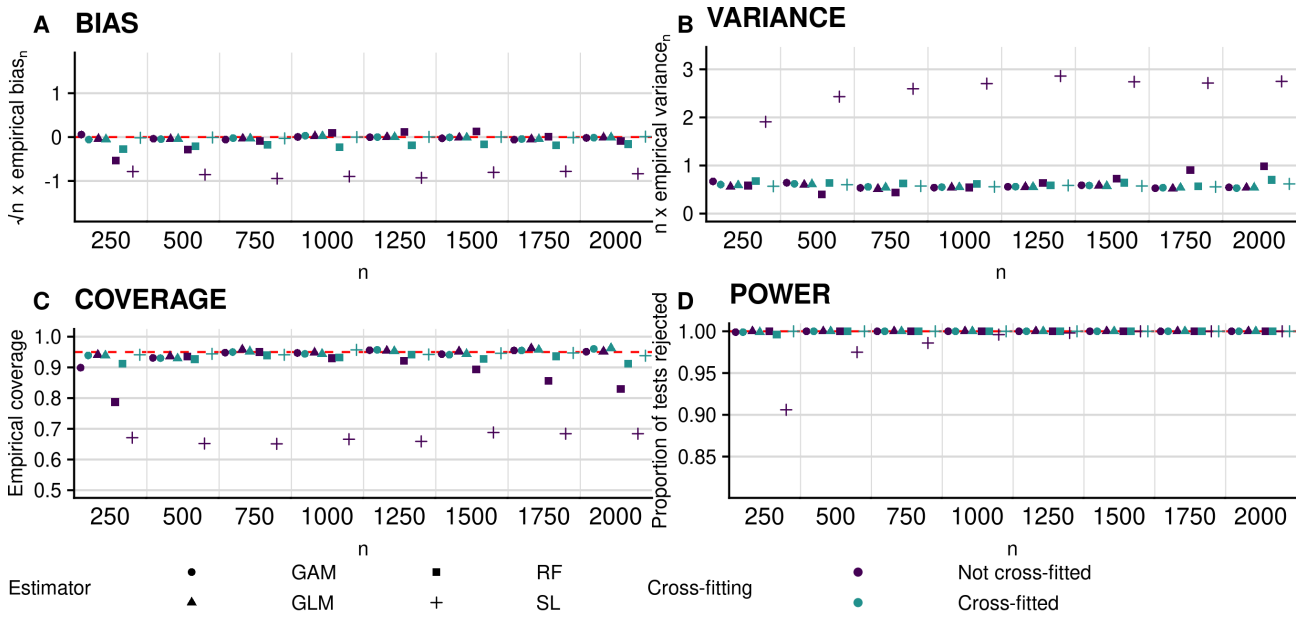

Figure S5: Performance of plug-in estimators for estimating (non-zero) importance of X_2 in terms of accuracy under Scenario 2. Clockwise from top left: empirical bias of the proposed plug-in estimator scaled by $n^{1/2}$; empirical variance scaled by n; empirical coverage of nominal 95% confidence intervals; and empirical power of the proposed hypothesis test. Circles, triangles, squares, and plus symbols denote estimators based on the use of generalized additive models (GAMs), probit regression (GLM), random forests (RF) or the Super Learner (SL), respectively. Blue and green symbols denote non-cross-fitted and cross-fitted estimators, respectively. This figure appears in color in the electronic version of this article.

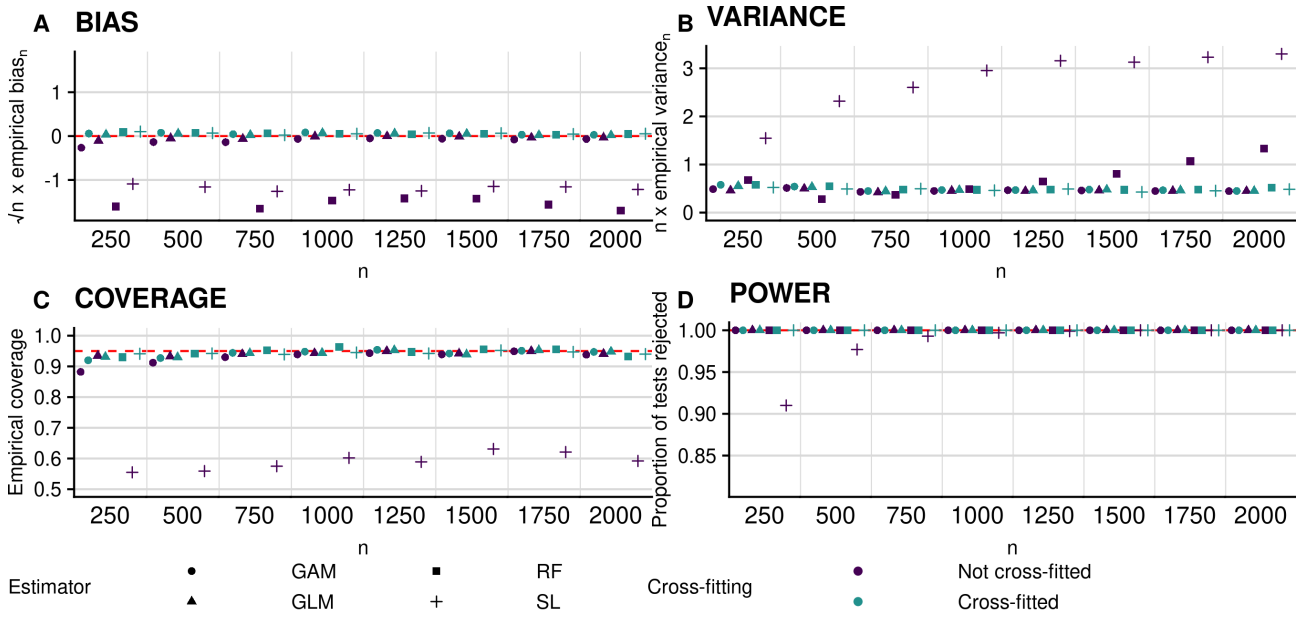

Figure S6: Performance of plug-in estimators for estimating (non-zero) importance of X_2 in terms of AUC under Scenario 2. Clockwise from top left: empirical bias of the proposed plug-in estimator scaled by $n^{1/2}$; empirical variance scaled by n; empirical coverage of nominal 95% confidence intervals; and empirical power of the proposed hypothesis test. Circles, triangles, squares, and plus symbols denote estimators based on the use of generalized additive models (GAMs), probit regression (GLM), random forests (RF) or the Super Learner (SL), respectively. Blue and green symbols denote non-cross-fitted and cross-fitted estimators, respectively. Coverage of intervals based on the non-cross-fitted RF-based estimator never exceeds 0.5 and is as low as zero in some cases. This figure appears in color in the electronic version of this article.

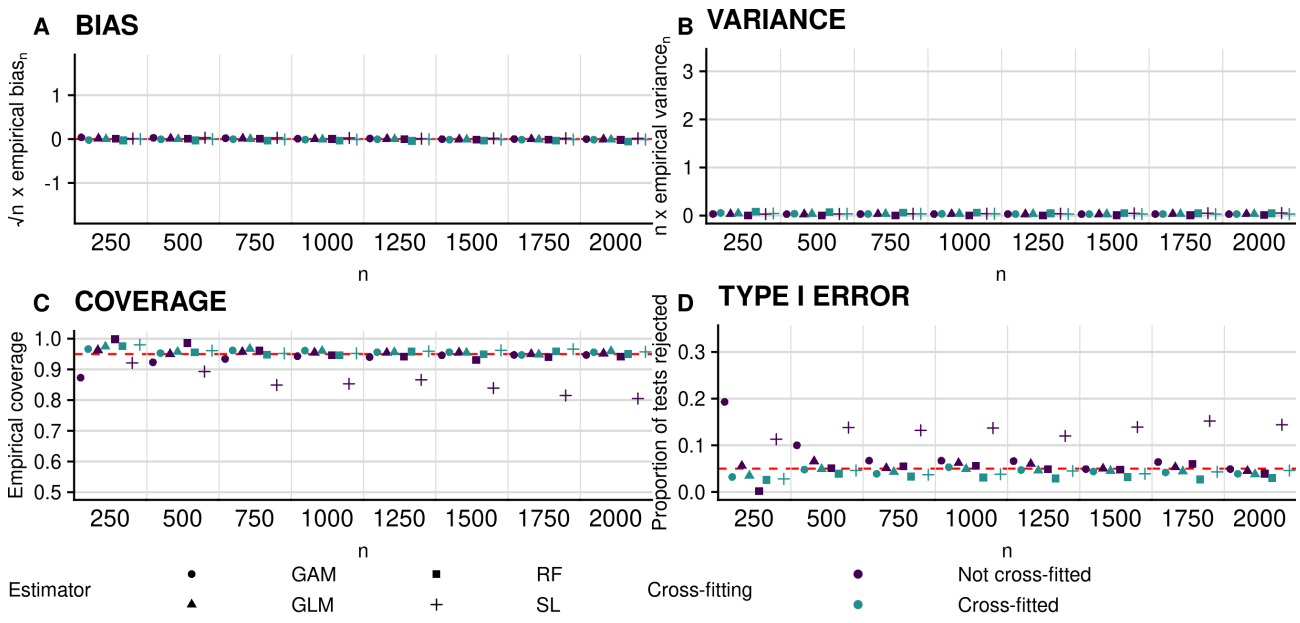

Figure S7: Performance of plug-in estimators for estimating (zero) importance of X_3 in terms of AUC under Scenario 2. Clockwise from top left: empirical bias of the proposed plug-in estimator scaled by $n^{1/2}$; empirical variance scaled by n; empirical coverage of nominal 95% confidence intervals; and empirical type I error of the proposed hypothesis test. Circles, triangles, squares, and plus symbols denote estimators based on the use of generalized additive models (GAMs), probit regression (GLM), random forests (RF) or the Super Learner (SL), respectively. Blue and green symbols denote non-cross-fitted and cross-fitted estimators, respectively. This figure appears in color in the electronic version of this article.

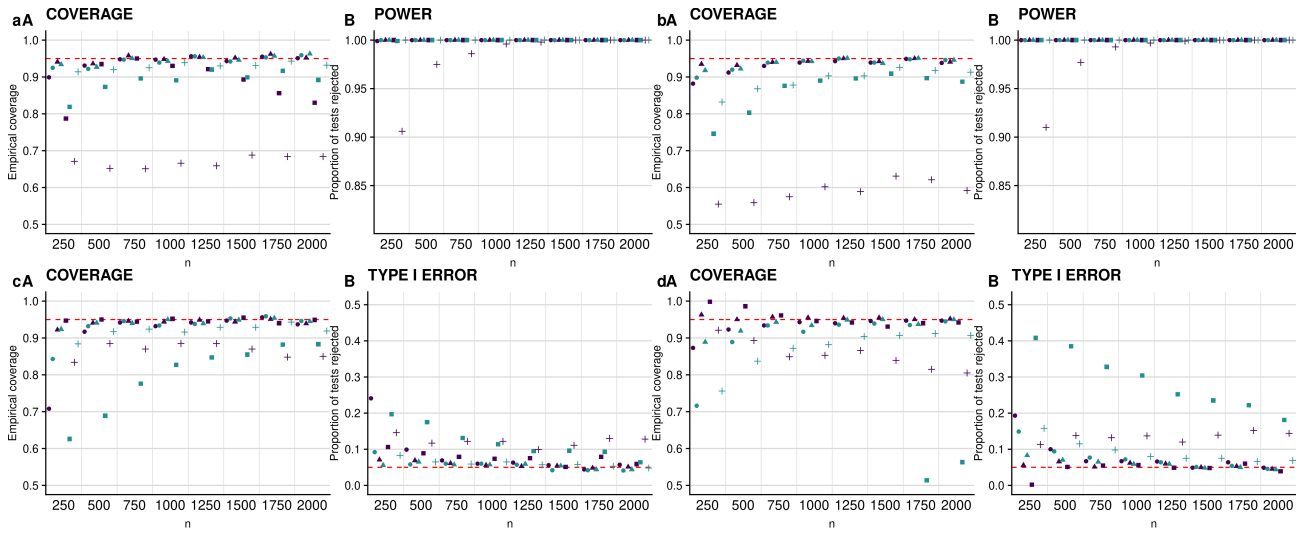

Figure S8: Empirical coverage of confidence intervals (A) and proportion of tests rejected (B) based on the non-cross-fitted standard error estimator under Scenario 2. The rows correspond to X_2 and X_3, respectively, while the columns correspond to accuracy (a,c) and AUC (b,d), respectively. Circles, triangles, squares, and plus symbols denote estimators based on the use of generalized additive models (GAMs), probit regression (GLM), random forests (RF) or the Super Learner (SL), respectively. Blue and green symbols denote non-cross-fitted and cross-fitted estimators, respectively. Coverage of intervals based on the non-cross-fitted RF-based estimator of importance of X_2 (panel bA) never exceeds 0.5 and is as low as zero in some cases. This figure appears in color in the electronic version of this article.

1. obtain estimator $\psi_{n,s}$ or $\psi^*_{n,s}$ of $\psi_{0,s}$;

2. obtain estimators f_n and $f_{n,s}$ of f_0 and $f_{0,s}$ based on the entire dataset;

3. create B bootstrap resamples of the original dataset;

4. For $b = 1, 2, \ldots, B$:

   (a) obtain $v_{n,b} := V(f_n, P_{n,b})$ and $v_{n,s,b} := V(f_{n,s}, P_{n,b})$ using the nuisance functions estimated on the entire dataset and the bootstrap empirical distribution $P_{n,b}$;

   (b) set $\psi_{n,s,b} := v_{n,b} - v_{n,s,b}$;

5. compute bootstrap variance estimator $\tau^2_{n,s,B} := \frac{1}{B}\sum_{b=1}^{B}\left(\psi_{n,s,b} - \frac{1}{B}\sum_{b=1}^{B}\psi_{n,s,b}\right)^2$ and resulting Wald-type confidence intervals (using $\psi^*_{n,s}$ or $\psi_{n,s}$), or form a percentile-based confidence interval with endpoints given by the 5th and 95th sample percentiles of $\{\psi_{n,s,1}, \psi_{n,s,2}, \ldots, \psi_{n,s,B}\}$.

We consider again Scenario 1, where $p = 2$. For each scenario presented here, we generated 1000 random datasets of size $n \in \{100, 500, 1000, \ldots, 4000\}$, and considered the importance of both X_1

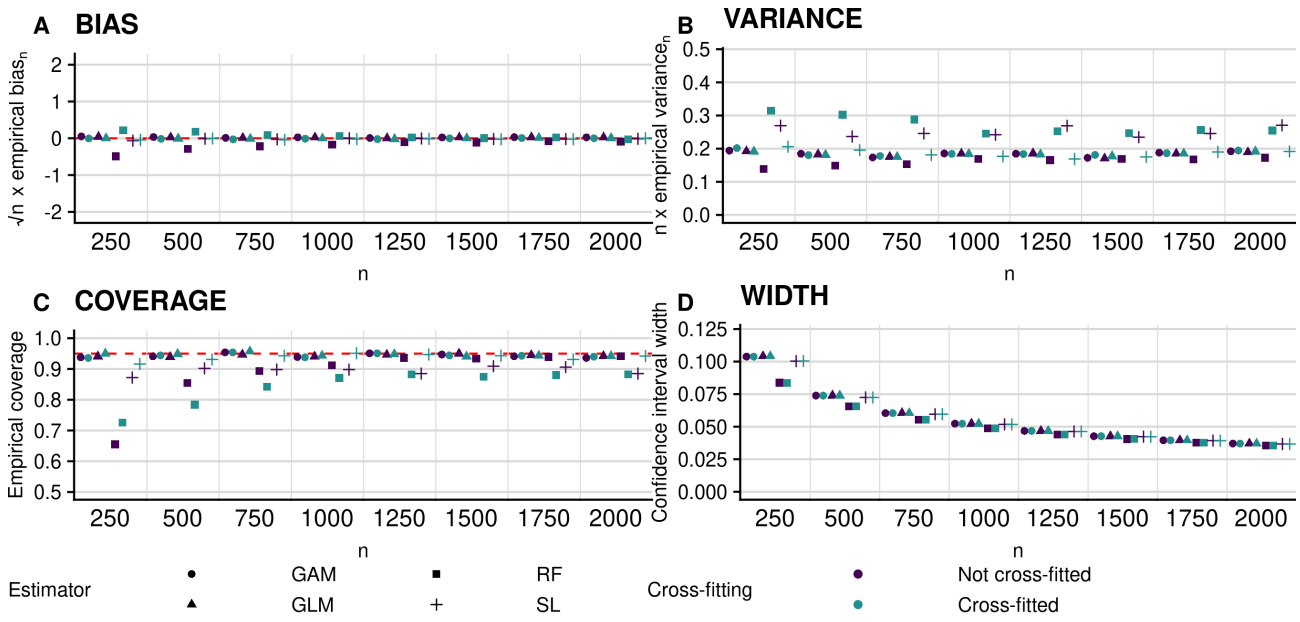

Figure S9: Performance of plug-in estimators for estimating (non-zero) importance of X_1 in terms of accuracy under Scenario 1, using the bootstrap for interval estimation. Clockwise from top left: empirical bias of the proposed plug-in estimator scaled by $n^{1/2}$; empirical variance scaled by n; empirical coverage of nominal 95% confidence intervals; and average width of these intervals. Circles, triangles, squares and plus symbols denote estimators based on the use of generalized additive models (GAMs), probit regression (GLM), random forests (RF), and the Super Learner (SL), respectively. Blue and green symbols denote non-cross-fitted and cross-fitted estimators, respectively. This figure appears in color in the electronic version of this article.

and X_2. We assess performance in the same way as in the main manuscript, though we use the bootstrap-based intervals in place of those based on the influence function. We present the results of this experiment in Figures S9–S12. The results for bias and variance are unchanged from the previous experiments. Encouragingly, both coverage and width for the bootstrap-based intervals are similar to the coverage and width of the IF-based intervals, though in the smaller sample size settings the bootstrap-based intervals are slightly narrower than the IF-based intervals.

4.5 Higher dimensions and correlated features

We now consider two scenarios under increasing dimension, both with and without correlated features. Here, $p \in \{50, 100, 200\}$ and Σ is either a $p \times p$ identity matrix (Scenario 3) or a $p \times p$ diagonal matrix with 1 on the diagonal and all off-diagonal elements equal to zero except $\Sigma_{1,3} = \Sigma_{3,1} = 0.7$ and $\Sigma_{2,4} = \Sigma_{4,2} = 0.2$ (Scenario 4). Thus, in Scenario 4, X_3 and X_4 are not directly important for predicting the outcome, but might be found to be important in isolation due to their correlation with the important features X_1 and X_2. In these experiments, we considered $n \in \{500, 3000\}$ for each

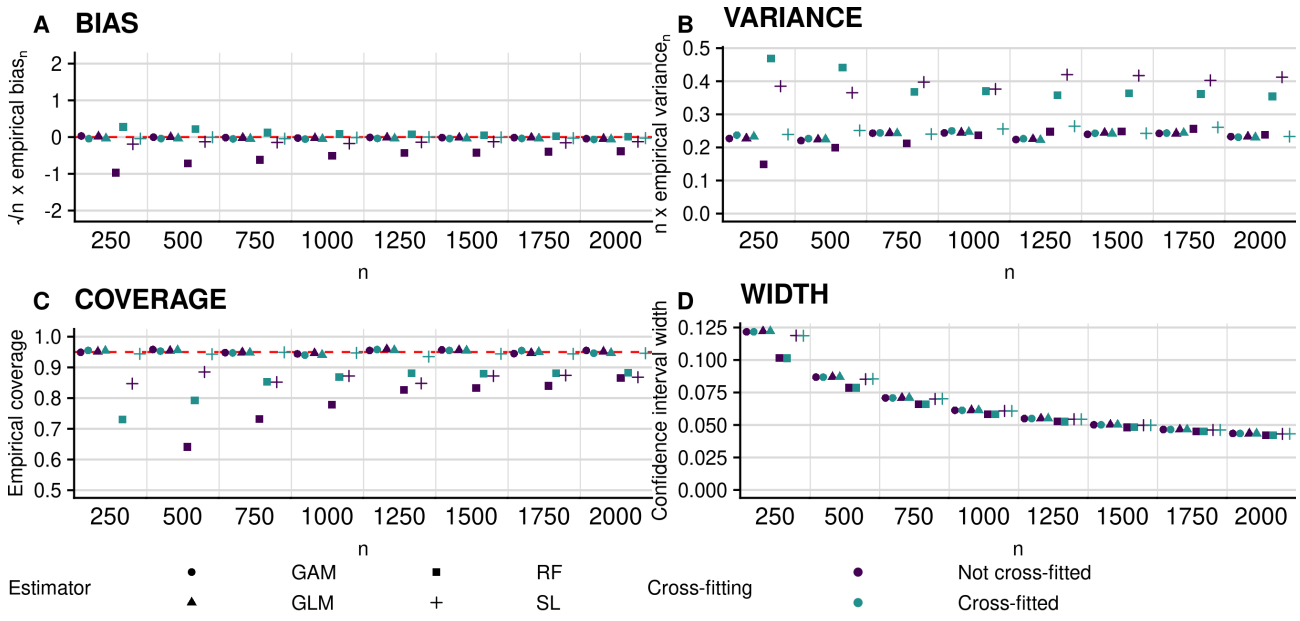

Figure S10: Performance of plug-in estimators for estimating (non-zero) importance of X_2 in terms of accuracy under Scenario 1, using the bootstrap for interval estimation. Clockwise from top left: empirical bias of the proposed plug-in estimator scaled by $n^{1/2}$; empirical variance scaled by n; empirical coverage of nominal 95% confidence intervals; and average width of these intervals. Circles, triangles, squares and plus symbols denote estimators based on the use of generalized additive models (GAMs), probit regression (GLM), random forests (RF), and the Super Learner (SL), respectively. Blue and green symbols denote non-cross-fitted and cross-fitted estimators, respectively. This figure appears in color in the electronic version of this article.

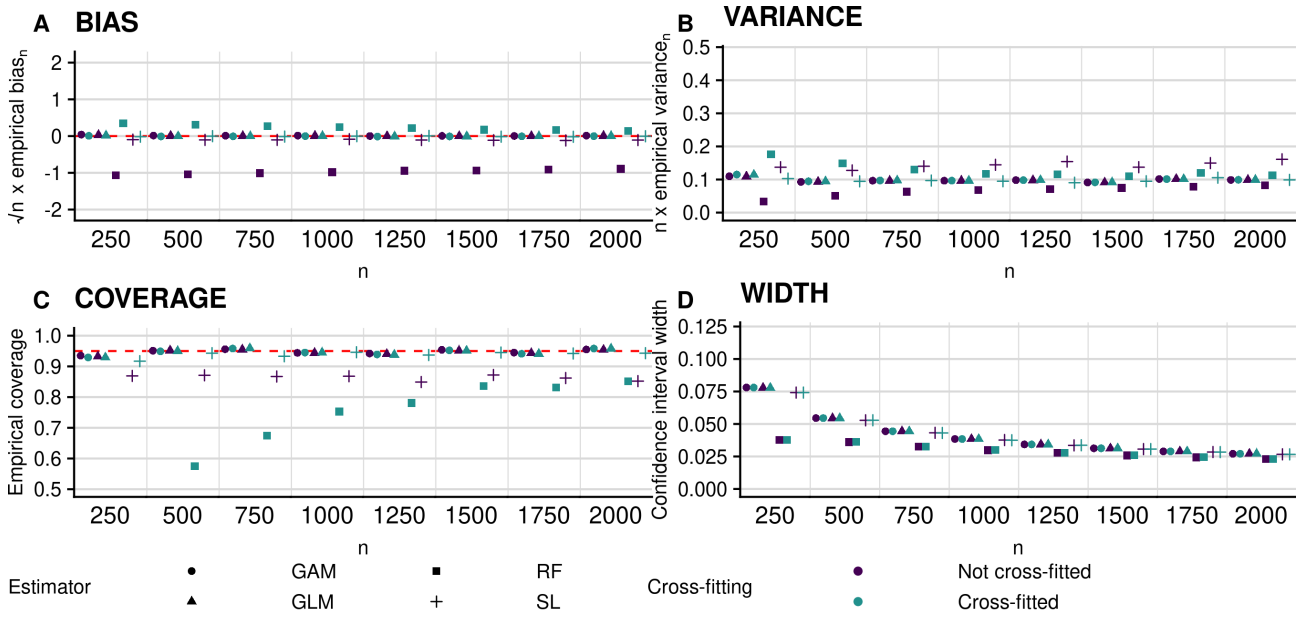

Figure S11: Performance of plug-in estimators for estimating (non-zero) importance of X_1 in terms of AUC under Scenario 1, using the bootstrap for interval estimation. Clockwise from top left: empirical bias of the proposed plug-in estimator scaled by $n^{1/2}$; empirical variance scaled by n; empirical coverage of nominal 95% confidence intervals; and average width of these intervals. Circles, triangles, squares and plus symbols denote estimators based on the use of generalized additive models (GAMs), probit regression (GLM), random forests (RF), and the Super Learner (SL), respectively. Blue and green symbols denote non-cross-fitted and cross-fitted estimators, respectively. Coverage of intervals based on the non-cross-fitted RF-based estimator never exceeds 0.5 and is as low as zero in some cases. This figure appears in color in the electronic version of this article.

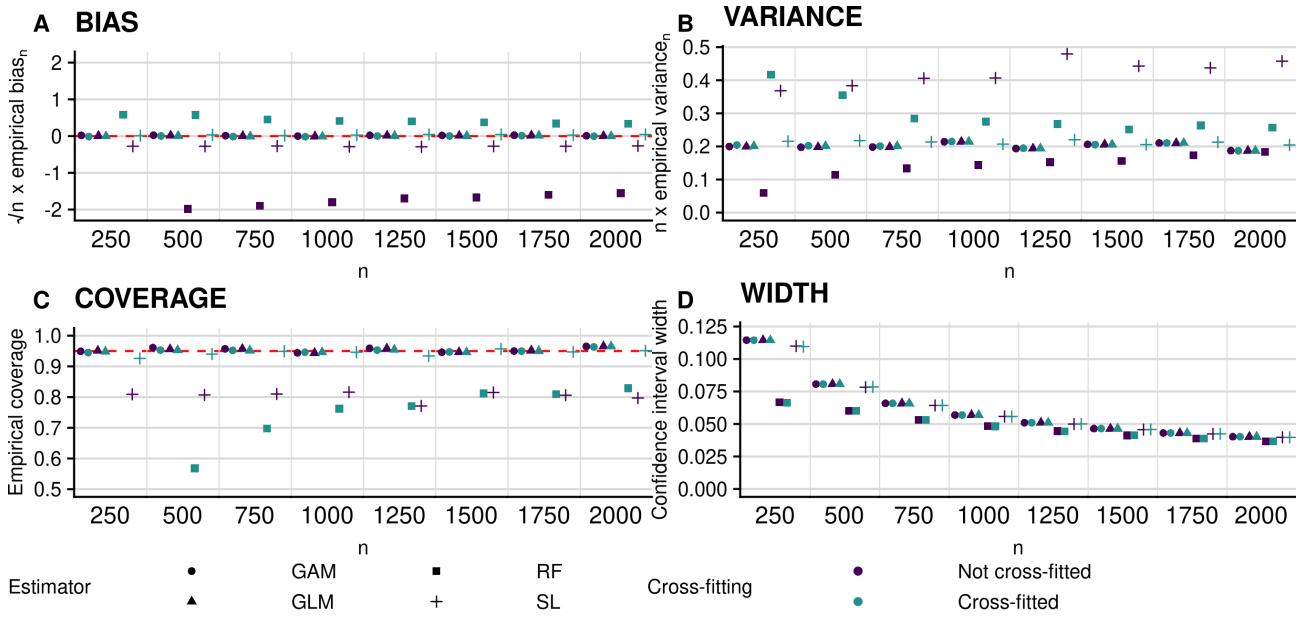

Figure S12: Performance of plug-in estimators for estimating (non-zero) importance of X_2 in terms of AUC under Scenario 1, using the bootstrap for interval estimation. Clockwise from top left: empirical bias of the proposed plug-in estimator scaled by $n^{1/2}$; empirical variance scaled by n; empirical coverage of nominal 95% confidence intervals; and average width of these intervals. Circles, triangles, squares and plus symbols denote estimators based on the use of generalized additive models (GAMs), probit regression (GLM), random forests (RF), and the Super Learner (SL), respectively. Blue and green symbols denote non-cross-fitted and cross-fitted estimators, respectively. In this experiment, the coverage of non-cross-fitted RF was never above 0.5, and was as low as zero. This figure appears in color in the electronic version of this article.

p, and assessed the importance of each individual feature as well as the feature groups (X_1, X_3) and (X_2, X_4), again using both accuracy and AUC. We use cross-fitting to estimate the VIM value in all cases, and we use the Super Learner with candidate library consisting of boosted trees, random forests, and the lasso to estimate f_0 and $f_{0,s}$. We then compute the empirical bias scaled by $n^{1/2}$, the empirical variance scaled by n, the empirical coverage of nominal 95% confidence intervals, and the proportion of tests rejected.

We display the results under Scenario 3 in Figures S13 and S14. Here, we find that at the smaller sample size ($n = 500$), there is some excess bias for the features with non-null importance, and that this bias increases with increasing p; this is accompanied by a decrease in coverage. However, with a larger sample size ($n = 3000$), we recover similar performance to that observed in Section **??** of the main manuscript and the preceeding sections of this supplement. Type I error is controlled at the nominal level in all cases.

We display the results under Scenario 4 in Figures S15 and S16. We find similar results overall to those from Scenario 3. In smaller samples, it appears to be advantageous to consider groups of correlated features rather than the features alone; this is particularly striking in Figure S16. As the sample size grows, the difference in performance diminishes.

Overall, the statistical performance of our procedure appear to be impacted more strongly by noise covariates in small samples than in large samples, regardless of the level of correlation among covariates. It is possible that this performance could be improved in small samples by including more aggressive sparsity-inducing algorithms in our ensemble. Indeed, the performance of our estimator of each VIM value depends on the rate at which the nuisance functions can be estimated, and this rate certainly slows down as the number of covariates grows, unless we can leverage stronger structure. We note that, while perhaps minimally impacting the statistical performance of our procedure, correlated features nevertheless render the interpretation of individual-variable importance more challenging: the population-level importance value itself changes in the presence of correlation. This difficulty can be partially mitigated by assessing group variable importance instead; however, this requires groups to either be known a priori (as in Section 6 of the main manuscript) or estimated, and in this latter case, further work must be done to ensure that the desired inferential properties (e.g., correct coverage) are preserved.

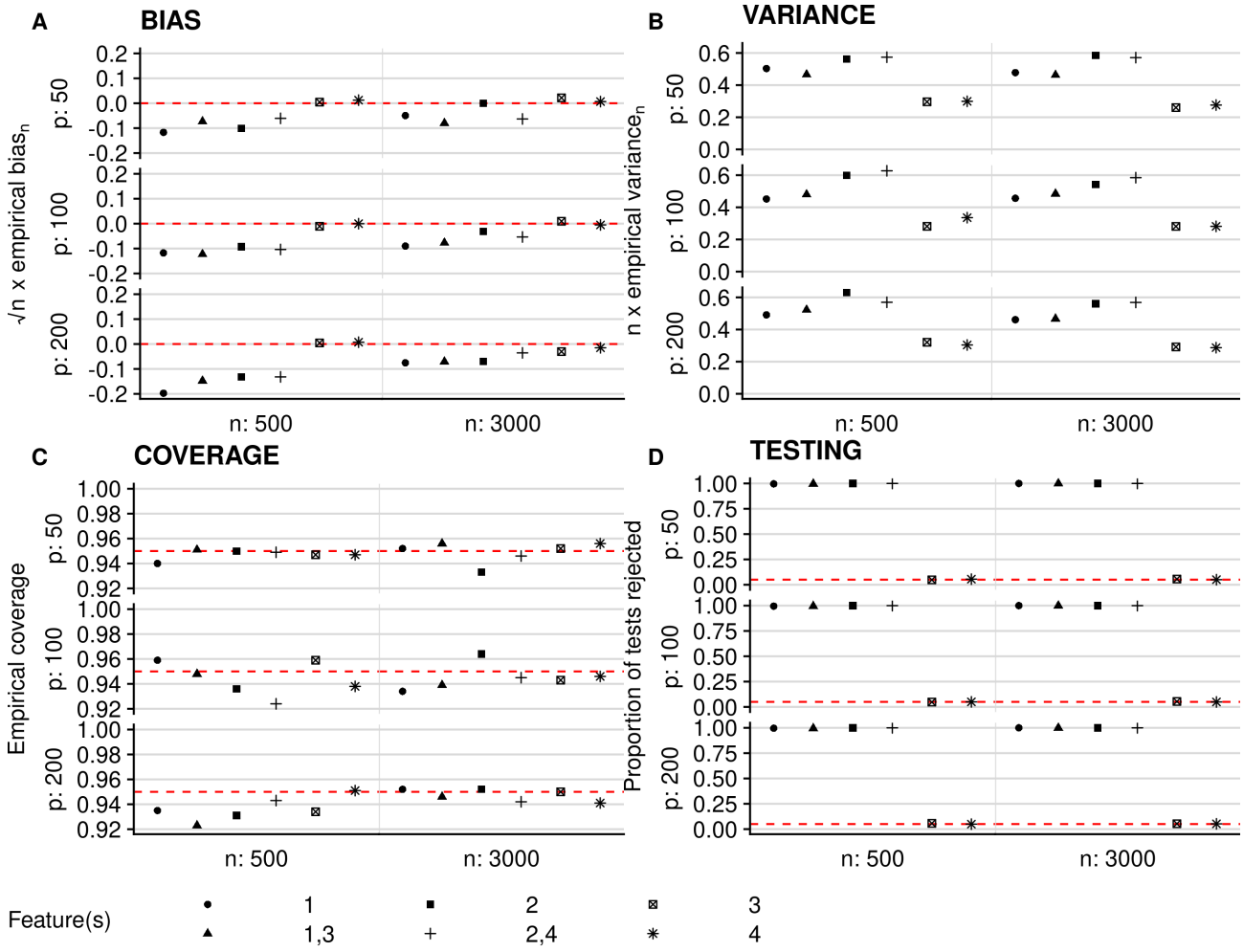

Figure S13: Performance of plug-in estimators for estimating importance in terms of accuracy under Scenario 3 (all features are independent). Clockwise from top left: empirical bias for the proposed plug-in estimator scaled by $n^{1/2}$; empirical variance scaled by n; empirical coverage of nominal 95% confidence intervals for the true importance; and empirical type I error of the proposed hypothesis test. The different symbols denote the feature(s) of interest.

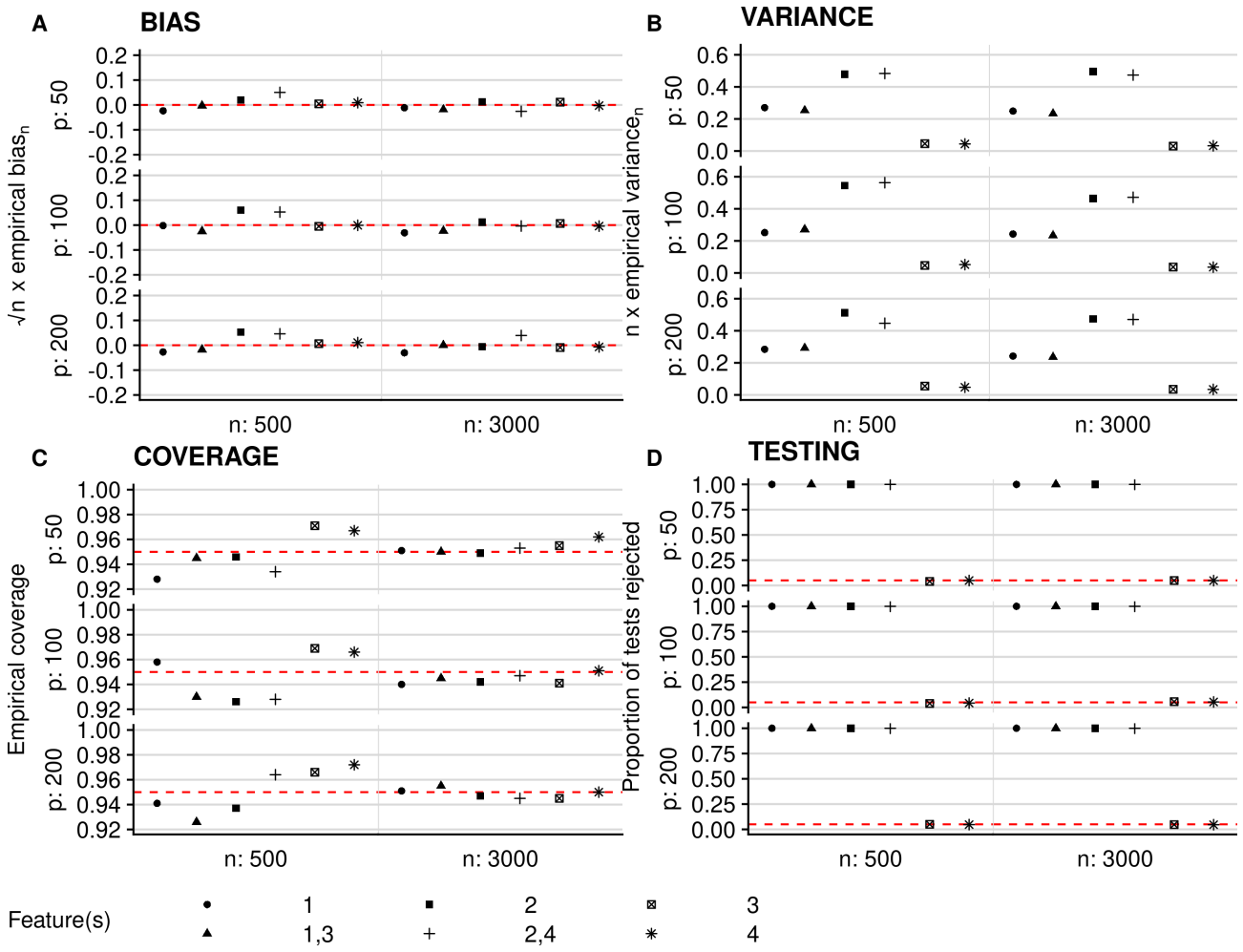

Figure S14: Performance of plug-in estimators for estimating importance in terms of AUC under Scenario 3 (all features are independent). Clockwise from top left: empirical bias for the proposed plug-in estimator scaled by $n^{1/2}$; empirical variance scaled by n; empirical coverage of nominal 95% confidence intervals for the true importance; and empirical type I error of the proposed hypothesis test. The different symbols denote the feature(s) of interest.

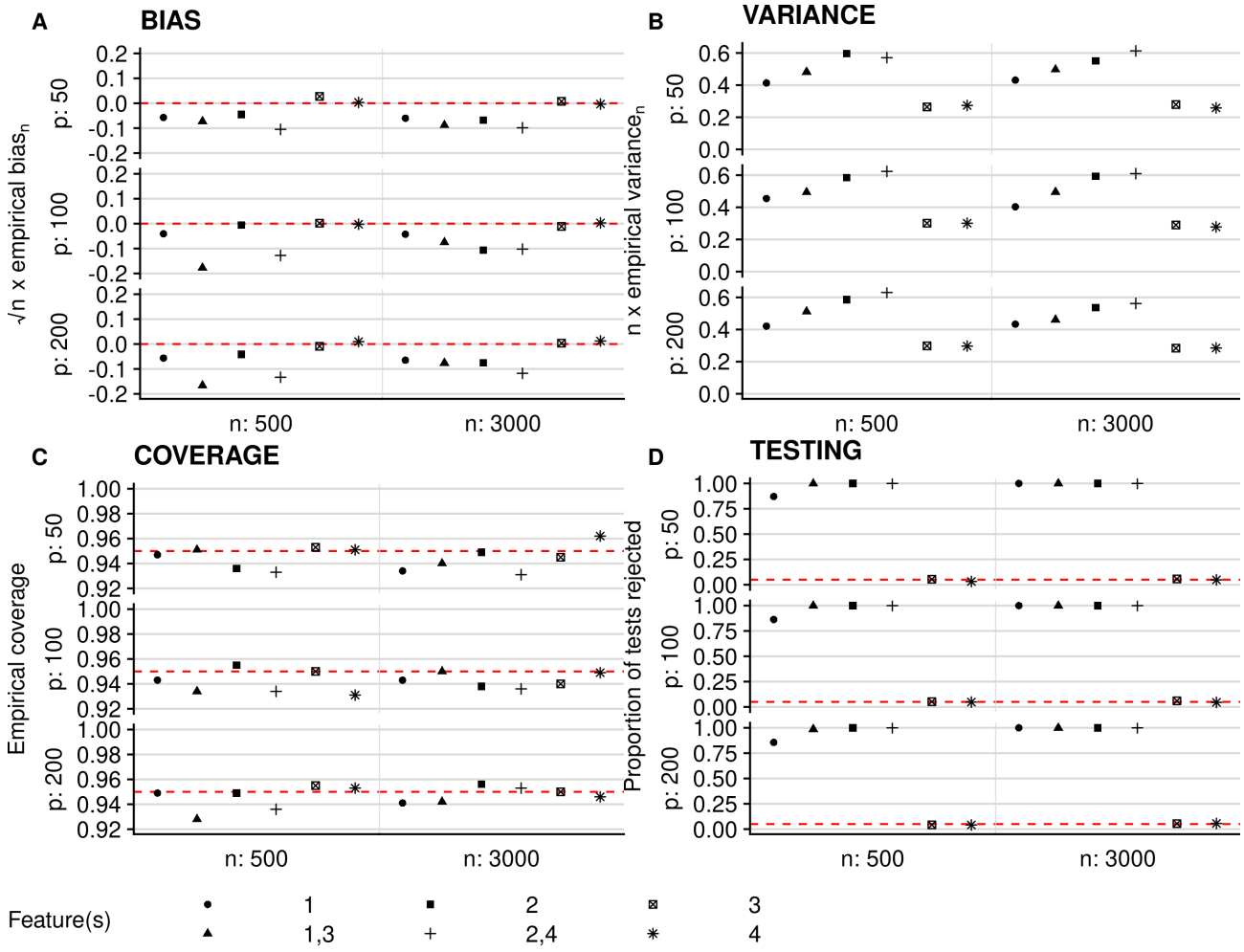

Figure S15: Performance of plug-in estimators for estimating importance in terms of accuracy under Scenario 4 (some features are correlated). Clockwise from top left: empirical bias for the proposed plug-in estimator scaled by $n^{1/2}$; empirical variance scaled by n; empirical coverage of nominal 95% confidence intervals for the true importance; and empirical type I error of the proposed hypothesis test. The different symbols denote the feature(s) of interest.

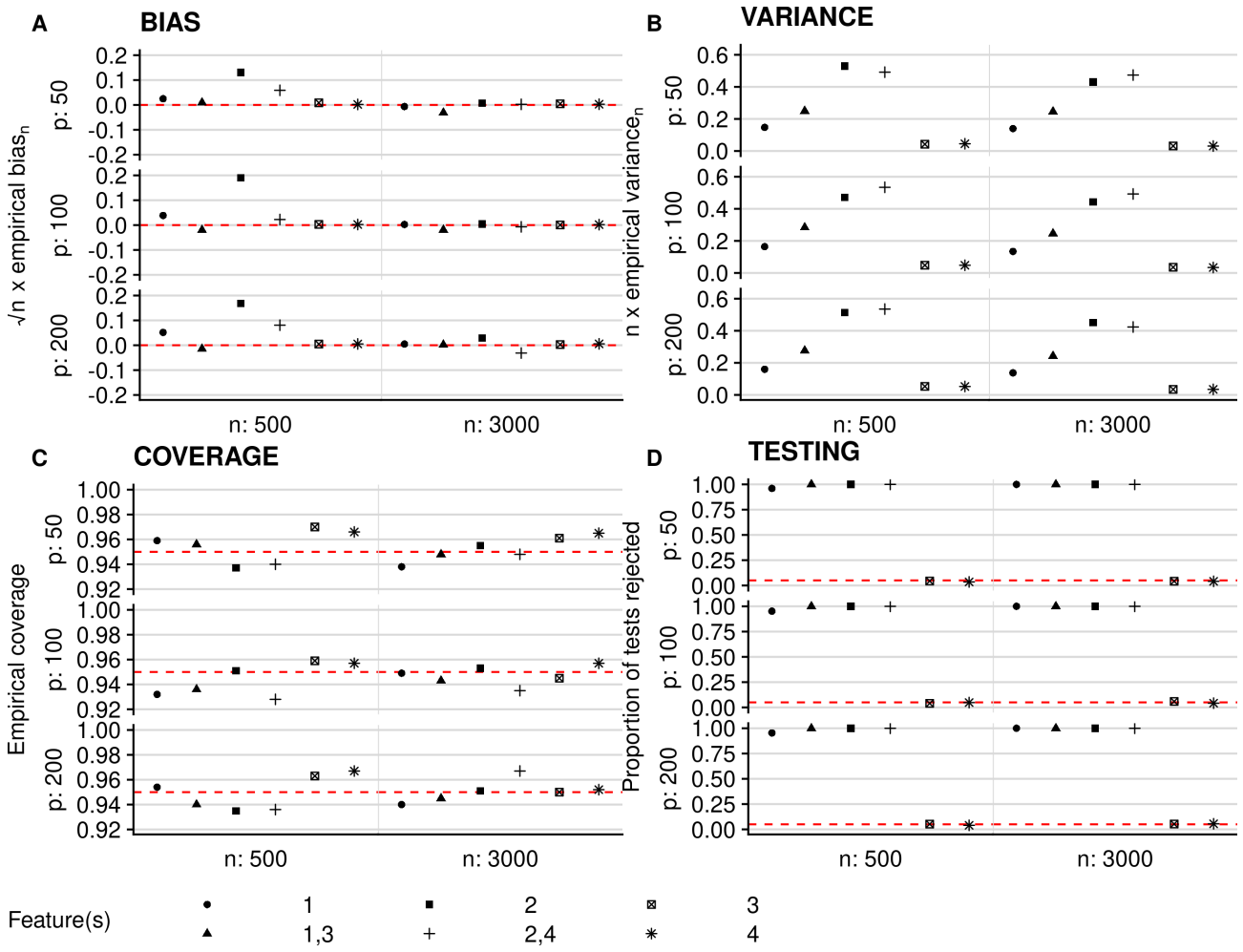

Figure S16: Performance of plug-in estimators for estimating importance in terms of AUC under Scenario 4 (some features are correlated). Clockwise from top left: empirical bias for the proposed plug-in estimator scaled by $n^{1/2}$; empirical variance scaled by n; empirical coverage of nominal 95% confidence intervals for the true importance; and empirical type I error of the proposed hypothesis test. The different symbols denote the feature(s) of interest.

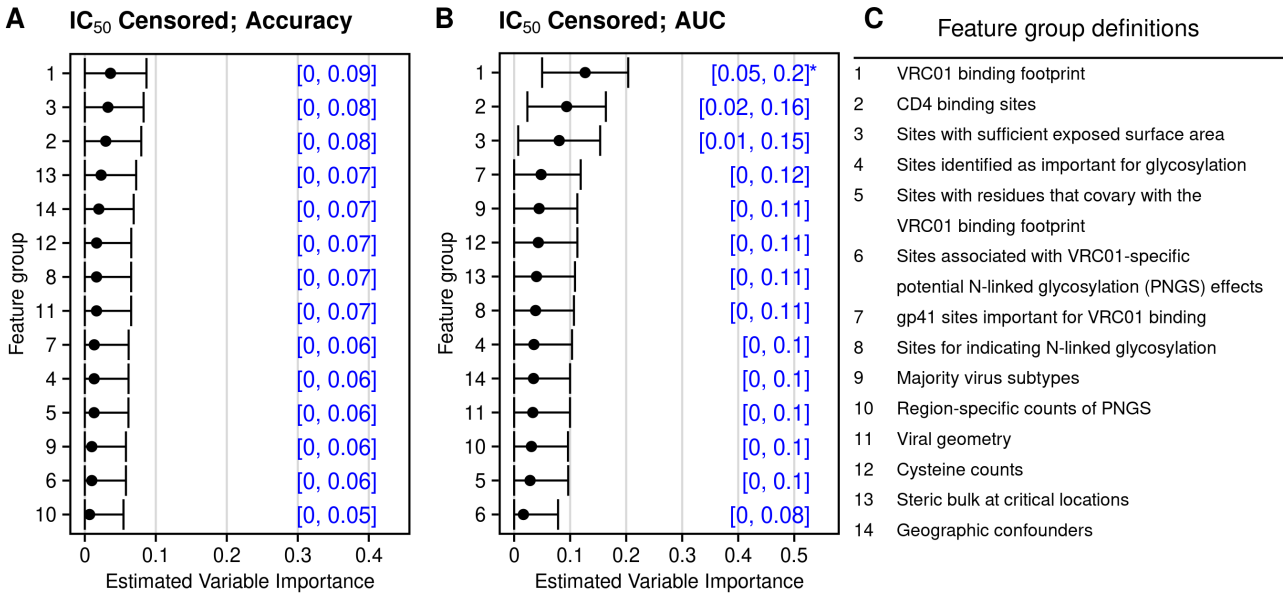

Figure S17: Variable importance measured by accuracy (panel A) and AUC (panel B) for the groups defined in panel C. Stars denote importance deemed statistically significantly different from zero at the 0.0038 (0.05 / 13) level.

5 Additional details for the study of an antibody against HIV-1

5.1 Harmonized analysis with Magaret et al. (2019)

In Figure S17, we display the results of an analysis harmonized to use the same outcome as in Magaret et al. (2019). This sensitivity outcome is the indicator of whether or not the IC_{50} value was right-censored. Viruses with right-censored IC_{50} values are thought to be resistant to VRC01, while viruses with non-censored IC_{50} values may instead be more sensitive to VRC01. In this case, we consider the *conditional* importance of each group of features relative to the remaining features. Overall, these results are largely in line with both Magaret et al. (2019) and with the results presented in the main manuscript. However, we see here that only the VRC01 binding footprint has p-value less than 0.0038 (denoted by stars in Figure S17; this value results from a Bonferroni correction from testing 13 groups and an initial level of 0.05), and only for the AUC measure. The exact p-value is given by 6.1×10^{-4}.

5.2 Library of candidate learning algorithms

In this section, we describe the library of candidate learning algorithms used in our analysis replicating the results of Magaret et al. (2019). We used a wide array of flexible machine learning-based algorithms in the hope that this large library would yield a cross-validated algorithm with good predictive perfor-

Table S4: Library of canididate learners for the Super Learner with descriptions.

Function name	Description
SL.mean	intercept only regression
SL.xgboost1	boosted regression trees with maximum depth of 1
SL.xgboost2	boosted regression trees with maximum depth of 2
SL.xgboost4	boosted regression trees with maximum depth of 4
SL.xgboost6	boosted regression trees with maximum depth of 6
SL.xgboost8	boosted regression trees with maximum depth of 8
SL.ranger.small	random forest with mtry equal to one-half times square root of number of predictors
SL.ranger.reg	random forest with mtry equal to square root of number of predictors
SL.ranger.large	random forest with mtry equal to two times square root of number of predictors
SL.glmnet.0	GLMNET with lambda selected by 5-fold CV and alpha equal to 0
SL.glmnet.25	GLMNET with lambda selected by 5-fold CV and alpha equal to 0.25
SL.glmnet.50	GLMNET with lambda selected by 5-fold CV and alpha equal to 0.5
SL.glmnet.75	GLMNET with lambda selected by 5-fold CV and alpha equal to 0.75
SL.glmnet.1	GLMNET with lambda selected by CV and alpha equal to 1

mance. The particular machine learning techniques included were: the lasso with logit link function (implemented in the `glmnet` R package), random forests (implemented in the `ranger` R package), and gradient boosted decision trees (implemented in the `xgboost` R package), each with a variety of choices for the tuning parameters. In Table S4, we provide a description of each candidate learning algorithm in our library. Our final estimator is the convex combination of these algorithms chosen to minimize the ten-fold cross-validated negative log likelihood. In all cases, we adjusted for geographic region as a potential confounding variable.

5.3 Super Learner performance

We now describe the empirical performance of the Super Learner in this application for both the outcome considered in the main manuscript ($IC_{50} < 1$) and the IC_{50} censored outcome described above. In Table S5, we show the coefficients of each candidate learner in the final Super Learner ensemble for each outcome. The rows of this table are each of the ten cross-validation folds broken down by outcome, while the columns are the individual learners. Here, we see that for the IC_{50} censored outcome, the most commonly chosen algorithms in the final ensemble were boosted trees with maximum depth of 2 or 4, random forests with a large number of features chosen at each split, and the elastic net with various values of α. For the $IC_{50} < 1$ outcome, the most commonly chosen algorithms were again boosted trees with maximum depth of 2, 4, or 6, random forests with a medium and large number of features chosen at each split; the elastic net was often not chosen by the Super Learner.

Table S5: Table of Super Learner weights for each outcome, candidate learner and cross-validation fold. We have removed 'SL.' from the name of each learner.

mean	xgboost1	xgboost2	xgboost4	xgboost6	xgboost8	ranger.small	ranger.reg	ranger.large	glmnet.0	glmnet.25	glmnet.50	glmnet.75	glmnet.1	fold
IC_{50} censored														
0	0	0.05	0.00	0.00	0.00	0	0.00	0.74	0	0.00	0.21	0.00	0.00	1
0	0	0.08	0.00	0.00	0.00	0	0.00	0.62	0	0.00	0.30	0.00	0.00	2
0	0	0.01	0.00	0.00	0.00	0	0.00	0.50	0	0.48	0.00	0.00	0.00	3
0	0	0.00	0.08	0.00	0.00	0	0.00	0.51	0	0.40	0.00	0.00	0.00	4
0	0	0.00	0.04	0.00	0.00	0	0.00	0.58	0	0.00	0.38	0.00	0.00	5
0	0	0.11	0.00	0.00	0.00	0	0.00	0.62	0	0.27	0.00	0.00	0.00	6
0	0	0.11	0.00	0.00	0.00	0	0.00	0.51	0	0.07	0.00	0.31	0.00	7
0	0	0.05	0.00	0.00	0.00	0	0.00	0.74	0	0.00	0.14	0.00	0.08	8
0	0	0.01	0.01	0.00	0.00	0	0.00	0.62	0	0.23	0.00	0.12	0.00	9
0	0	0.07	0.00	0.00	0.00	0	0.00	0.36	0	0.27	0.00	0.00	0.31	10
$IC_{50} < 1$														
0	0	0.00	0.13	0.00	0.00	0	0.00	0.87	0	0.00	0.00	0.00	0.00	1
0	0	0.00	0.18	0.00	0.00	0	0.00	0.82	0	0.00	0.00	0.00	0.00	2
0	0	0.00	0.00	0.16	0.00	0	0.00	0.84	0	0.00	0.00	0.00	0.00	3
0	0	0.00	0.06	0.06	0.00	0	0.00	0.89	0	0.00	0.00	0.00	0.00	4
0	0	0.02	0.12	0.05	0.00	0	0.00	0.82	0	0.00	0.00	0.00	0.00	5
0	0	0.11	0.00	0.00	0.03	0	0.00	0.86	0	0.00	0.00	0.00	0.00	6
0	0	0.05	0.00	0.00	0.00	0	0.11	0.84	0	0.00	0.00	0.00	0.00	7
0	0	0.00	0.00	0.07	0.03	0	0.00	0.90	0	0.00	0.00	0.00	0.00	8
0	0	0.12	0.00	0.00	0.00	0	0.39	0.41	0	0.00	0.00	0.07	0.00	9
0	0	0.00	0.00	0.07	0.00	0	0.00	0.93	0	0.00	0.00	0.00	0.00	10

In Figure S18, we display the cross-validated AUC and 95% confidence intervals (obtained on the logit scale and then inverted; thus, the intervals may not be symmetric about the point estimate of AUC) for both outcomes and each of the candidate learning algorithms in the Super Learner, along with the Super Learner ensemble algorithm and the classical cross-validated selector (the "discrete Super Learner"). We used the R package `cvAUC` to compute these point and interval estimates. Similarly to Magaret et al. (2019), we see that, of all the individual algorithms, random forests have the best performance in this application for both outcomes, followed by the lasso and boosted trees (for the IC_{50} censored outcome) and the reverse for the $IC_{50} < 1$ outcome. Additionally, we estimate the cross-validated AUC of the overall Super Learner to be 0.90 for the IC_{50} censored outcome, with a 95% confidence interval of (0.87, 0.94). For the $IC_{50} < 1$ outcome, we estimate the cross-validated AUC of the overall Super Learner to be 0.83 (0.80, 0.86). Magaret et al. (2019) performed an analysis for IC_{50} censored separately on two independent splits of these data, and obtained cross-validated AUCs of 0.86 (0.81, 0.92) and 0.87 (0.81, 0.93) on these two subsets.

In Figure S19, we display cross-validated ROC curves for the Super Learner, discrete Super Learner, and the top-performing individual algorithm. These ROC curves are similar to those presented in Magaret et al. (2019) — in both analyses, we see a large cross-validated true positive rate for each chosen cross-validated false positive rate. These results suggest that for both outcomes, our predictor is well-calibrated for discriminating between the outcome classes.

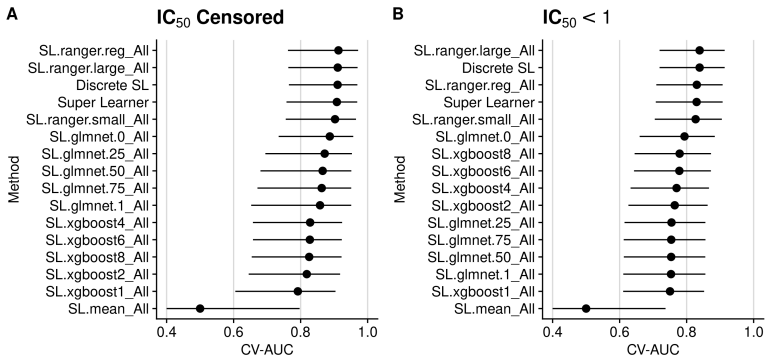

Figure S18: Point estimates of cross-validated AUC with 95% confidence intervals for each candidate learning algorithm in the Super Learner for each outcome.

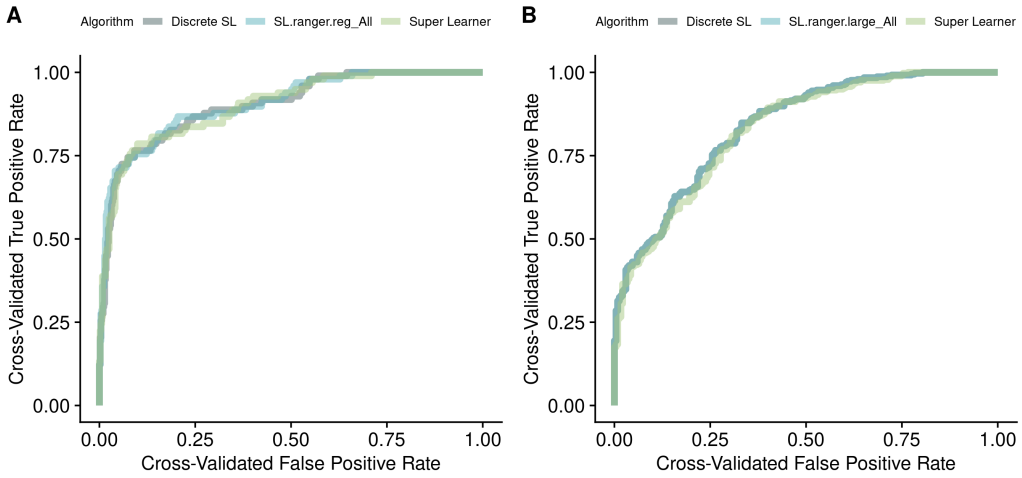

Figure S19: Cross-validated ROC curves for each outcome for the Super Learner (light green), discrete Super Learner (gray), and top-performing individual algorithm (random forests). IC_{50} censored is displayed in panel A, while $IC_{50} < 1$ is displayed in panel B.

References

Chen, T., T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, T. Zhou, M. Li, J. Xie, M. Lin, Y. Geng, and Y. Li (2019). *xgboost: Extreme Gradient Boosting*. R package version 0.82.1.

Frangakis, C. E., T. Qian, Z. Wu, and I. Diaz (2015). Deductive derivation and turing-computerization of semiparametric efficient estimation. *Biometrics 71*(4), 867–874.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*(1), 1–22.

Hastie, T. (2019). *gam: Generalized Additive Models*. R package version 1.16.1.

Luedtke, A. and M. van der Laan (2016). Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *The Annals of Statistics 44*(2), 713–742.

Luedtke, A. R., M. Carone, and M. J. van der Laan (2015). Discussion of "deductive derivation and turing-computerization of semiparametric efficient estimation" by frangakis et al. *Biometrics 71*(4), 875.

Magaret, C., D. Benkeser, B. Williamson, B. Borate, L. Carpp, et al. (2019). Prediction of VRC01 neutralization sensitivity by HIV-1 gp160 sequence features. *PLoS Computational Biology 15*(4), e1006952.

van der Vaart, A. (2000). *Asymptotic Statistics*, Volume 3. Cambridge University Press.

Wood, S. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society Series B (Statistical Methodology) 73*(1), 3–36.

Wright, M. N. and A. Ziegler (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software 77*(1), 1–17.