Manuscript submitted to *The Econometrics Journal*, pp. S1–S18.

Double Robustness for Complier Parameters and a Semiparametric Test for Complier Characteristics

Rahul Singh[†] and Liyang Sun[⋆]

[†] *MIT Economics, 50 Memorial Drive, Cambridge MA 02142, USA.*
[⋆] *CEMFI Economics, 5 Calle Casado del Alisal, Madrid 28014, Spain.*
E-mail: rahul.singh@mit.edu, lsun20@cemfi.es

Summary We propose a semiparametric test to evaluate (a) whether different instruments induce subpopulations of compliers with the same observable characteristics, on average; and (b) whether compliers have observable characteristics that are the same as the full population, treated subpopulation, or untreated subpopulation, on average. The test is a flexible robustness check for the external validity of instruments. To justify the test, we characterize the doubly robust moment for Abadie (2003)'s class of complier parameters, and we analyse a machine learning update to κ weighting that we call the automatic κ weight. We use the test to reinterpret the difference in local average treatment effect estimates that Angrist and Evans (1998a) obtain when using different instrumental variables.

Keywords: *Instrumental variable; kappa weight; machine learning; semiparametric efficiency*

S1. RATE CONDITIONS

In this section, we present assumptions to guarantee that the estimators $(\hat{\gamma}, \hat{\alpha})$ of the nonparametric functions (γ_0, α_0) satisfy the rate conditions in Assumption A.1. First, we place a weak assumption on the dictionary of basis functions b.

Assumption S1.1. (Bounded dictionary) *The dictionary is bounded. Formally, there exists some $C > 0$ such that $\max_j |b_j(Z, X)| \leq C$ almost surely.*

Next, we articulate assumptions required for convergence of $\hat{\alpha}$ under two regimes: the regime in which α_0 is dense and the regime in which α_0 is sparse.

Assumption S1.2. (Dense balancing weight) *The balancing weight α_0 is well approximated by the full dictionary b. Formally, assume there exist some $\rho_n \in \mathbb{R}^p$ and $C < \infty$ such that $|\rho_n|_1 \leq C$ and $\|\alpha_0 - b^\top \rho_n\|^2 = O\{(\log p/n)^{1/2}\}$.*

Assumption S1.2 is satisfied if, for example, α_0 is a linear combination of b.

Assumption S1.3. (Sparse balancing weight) *The balancing weight α_0 is well approximated by a sparse subset of the dictionary b. Formally, assume*

1 *There exist $C > 1$ and $\xi > 0$ such that for all $\bar{s} \leq C\,(\log p/n)^{-1/(1+2\xi)}$, there exists some $\bar{\rho} \in \mathbb{R}^p$ with $|\bar{\rho}|_1 \leq C$ and $\bar{s}$ nonzero elements such that $\|\alpha_0 - b^\top \bar{\rho}\|^2 \leq C\bar{s}^{-\xi}$.*
2 *$G = E\{b(Z, X)b(Z, X)^\top\}$ has largest eigenvalue uniformly bounded in n.*

3 Denote $\mathcal{J}_\rho = support(\rho)$. There exists $k > 3$ such that for $\rho = \rho_L, \bar{\rho}$

$$\text{RE}(k) = \inf_{\delta \in \Delta(\mathcal{J}_\rho)} \frac{\delta^\top G \delta}{\sum_{j \in \mathcal{J}_\rho} \delta_j^2} > 0, \quad \Delta(\mathcal{J}_\rho) = \left(\delta \in \mathbb{R}^p : \delta \neq 0, \sum_{j \in \mathcal{J}_\rho^c} |\delta_j| \leq k \sum_{j \in \mathcal{J}_\rho} |\delta_j| \right).$$

4 $\log p = O(\log n)$.

Assumption S1.3 is satisfied if, for example, α_0 is sparse or approximately sparse (Chernozhukov et al., 2022a). The uniform bound on the largest eigenvalue of G rules out the possibility that G is an equal correlation matrix. RE is the population version of the restricted eigenvalue condition (Bickel et al., 2009). It generalizes the familiar notion of no multicollinearity to the high dimensional setting. The final condition $\log p = O(\log n)$ rules out the possibility that $p = \exp(n)$; dimension cannot grow too much faster than sample size.

We adapt convergence guarantees from Chernozhukov et al. (2022a) for the balancing weight estimator $\hat{\alpha}$ in Algorithm A.2. We obtain a slow rate for dense α_0 and a fast rate for sparse α_0. In both cases, we require the data driven regularization parameter λ_n to approach 0 slightly slower than $(\log p/n)^{1/2}$.

ASSUMPTION S1.4. (REGULARIZATION) *$\lambda_n = a_n(\log p/n)^{1/2}$ for some $a_n \to \infty$.*

For example, one could set $a_n = \log\{\log(n)\}$ (Chatterjee and Jafarov, 2015). In Supplement S3, we provide and justify an iterative tuning procedure to determine data driven regularization parameter λ_n. The guarantees are as follows.

LEMMA S1.1. (DENSE BALANCING WEIGHT RATE) *Under Assumptions 2.1, S1.1, S1.2, and S1.4,*

$$\|\hat{\alpha} - \alpha_0\|^2 = O_p\left\{ a_n \left(\frac{\log p}{n} \right)^{1/2} \right\}, \quad |\hat{\rho}|_1 = O_p(1).$$

LEMMA S1.2. (SPARSE BALANCING WEIGHT RATE) *Under Assumptions 2.1, S1.1, S1.3, and S1.4,*

$$\|\hat{\alpha} - \alpha_0\|^2 = O_p\left\{ a_n^2 \left(\frac{\log p}{n} \right)^{2\xi/(1+2\xi)} \right\}, \quad |\hat{\rho}|_1 = O_p(1).$$

See Supplement S2 for the proofs. Whereas Lemma S1.1 does not require an explicit sparsity condition, Lemma S1.2 does. When $\xi > 1/2$, the rate in Lemma S1.2 is faster than the rate in Lemma S1.1 for a_n growing slowly enough. Interpreting the rate in Lemma S1.2, $n^{-2\xi/(1+2\xi)}$ is the well known rate of convergence if the identity of the nonzero components of $\bar{\rho}$ were known. The fact that their identity is unknown introduces a cost of $(\log p)^{2\xi/(1+2\xi)}$. The cost a_n^2 can be made arbitrarily small.

We place a rate assumption on the machine learning estimator $\hat{\gamma}$. It is a weak condition that allows $\hat{\gamma}$ to converge at a rate slower than $n^{-1/2}$. Importantly, it allows the analyst a broad variety of choices of machine learning estimators such as a neural network or lasso. Schmidt-Hieber (2020); Farrell et al. (2021) provide a rate for the former, while Lemmas S1.1 and S1.2 provide rates for the latter, using the functional $b \mapsto E\{b(Z,X)V^\top\}$ instead.

Assumption S1.5. (Regression rate) $\|\hat{\gamma}-\gamma_0\| = O_p(n^{-d_\gamma})$ *where*

1 *In the dense balancing weight regime,* $1/4 \leq d_\gamma \leq 1/2$*;*
2 *In the sparse balancing weight regime,* $1/2 - \xi/(1+2\xi) \leq d_\gamma \leq 1/2$.

These regime specific lower bounds on d_γ are sufficient conditions for the product rate condition.

Corollary S1.1. (Verifying rate condition) *Suppose the conditions of Lemma S1.1 or Lemma S1.2 hold as well as Assumption S1.5. Then the rate conditions of Assumption A.1 hold:* $|\hat{\alpha}|_\infty = O_p(1)$, $\|\hat{\alpha}-\alpha_0\| = o_p(1)$, $\|\hat{\gamma}-\gamma_0\| = o_p(1)$, *and* $\|\hat{\alpha}-\alpha_0\|\|\hat{\gamma}-\gamma_0\| = o_p(n^{-1/2})$.

The product rate condition in Corollary S1.1 formalizes the trade off in estimation error permitted in estimating (γ_0, α_0). In particular, faster convergence of $\hat{\alpha}$ permits slower convergence of $\hat{\gamma}$. Prior information about the balancing weight α_0 used to estimate $\hat{\alpha}$, encoded by sparsity or perhaps by additional moment restrictions, can be helpful in this way. We will appeal to this product condition while proving statistical guarantees for complier parameters.

S2. PROOF OF CONSISTENCY AND ASYMPTOTIC NORMALITY FOR AUTO-κ

S2.1. Lemmas from previous work

In this section, we prove consistency and asymptotic normality. For simplicity, we focus on the affine complier parameters of Definition A.1. Corollary A.1 shows that this class that includes several popular complier parameters, including the leading case of average complier characteristics. The inference arguments can be generalized to the entire class in Definition 2.1, including moments that are nonlinear in θ, by introducing heavier notation and additional sample splitting for the nonlinear cases; see Chernozhukov et al. (2022) for details.

We present the results in two subsections. In this subsection, we quote lemmas from previous work. In the next subsection, we present original arguments to prove consistency and asymptotic normality for our instrumental variable setting.

Consider the notation

$$\begin{aligned}
\psi(w,\gamma,\alpha,\theta) &= m(w,\gamma,\theta) + \phi(w,\gamma,\alpha,\theta);\\
m(w,\gamma,\theta) &= A(\theta)\tilde{m}(w,\gamma);\\
\tilde{m}(w,\gamma) &= \gamma(1,x) - \gamma(0,x);\\
\phi(w,\gamma,\alpha,\theta) &= \alpha(z,x)A(\theta)\{v-\gamma(z,x)\}.
\end{aligned}$$

Definition S2.1. *Define the following matrix* $G \in \mathbb{R}^{p\times p}$ *and the vector* $M \in \mathbb{R}^p$*:*

$$\begin{aligned}
G &= E\{b(Z,X)b(Z,X)^\top\},\\
M &= E\{m(W,b,\theta_0)\}.
\end{aligned}$$

Proposition S2.1. (Lemma A10 of Chernozhukov et al. (2022a)) *Under Assumption S1.1, we have* $|\hat{G}-G|_\infty = O_p\{(\log p/n)^{1/2}\}$.

Proposition S2.2. (Lemma 8 of Chernozhukov et al. (2022a)) *Under Assumptions 2.1 and S1.1, we have* $|\hat{M} - M|_\infty = O_p\{(\log p/n)^{1/2}\}$.

Proof. (Proof of Lemma S1.1) Applying Proposition S2.1 and Proposition S2.2, the proof follows Chernozhukov et al. (2022a, Theorem 2).

Proof. (Proof of Lemma S1.2) Applying Proposition S2.1 and Proposition S2.2, the proof follows Chernozhukov et al. (2022a, Theorem 1). The argument that $|\hat{\rho}|_1 = O_p(1)$ is analogous to Chernozhukov et al. (2022a, Lemma A9).

Lemma S2.1. (Proof of Corollary 7 of Chernozhukov et al. (2022a)) *Under Assumptions 2.1 and A.1, the following results hold.*

1 $E\{\tilde{m}(W,\gamma_0)^2\} < \infty$,
2 $E[\{\tilde{m}(W,\gamma) - \tilde{m}(W,\gamma_0)\}^2] \le C\|\gamma - \gamma_0\|^2$,
3 $\max_j |\tilde{m}(W,b_j) - \tilde{m}(W,0)| \le C$.

Lemma S2.2. (Theorem 2.1 Newey and McFadden (1994)) *Consider* $\hat{\theta}$ *defined as* $\text{argmin}_{\theta\in\Theta} \hat{Q}(\theta)$, *where* $\hat{Q} : \Theta \to \mathbb{R}$ *estimates* $Q_0 : \Theta \to \mathbb{R}$. *If*

1 Θ *is compact,*
2 Q_0 *is continuous in* θ *over* Θ,
3 Q_0 *is uniquely maximized at* θ_0,
4 $\sup_{\theta\in\Theta} |\hat{Q}(\theta) - Q_0(\theta)| = o_p(1)$,

then $\hat{\theta} = \theta_0 + o_p(1)$.

S2.2. Consistency and asymptotic normality

Proposition S2.3. *Suppose the conditions of Theorem A.1 hold. Then for each fold* I_ℓ *the following holds:*

1 $E[\{m(W,\hat{\gamma}_{-\ell},\theta_0) - m(W,\gamma_0,\theta_0)\}^2 \mid I_{-\ell}] = o_p(1)$,
2 $E[\{\phi(W,\hat{\gamma}_{-\ell},\alpha_0,\theta_0) - \phi(W,\gamma_0,\alpha_0,\theta_0)\}^2 \mid I_{-\ell}] = o_p(1)$,
3 $E[\{\phi(W,\gamma_0,\hat{\alpha}_{-\ell},\theta_0) - \phi(W,\gamma_0,\alpha_0,\theta_0)\}^2 \mid I_{-\ell}] = o_p(1)$.

The notation $E(\cdot \mid I_{-\ell})$ *means conditional on* $W_{-\ell} = (W_i)_{i\notin I_\ell}$, *i.e. observations not in fold* I_ℓ.

Proof. First observe that

$$\phi(W,\hat{\gamma}_{-\ell},\alpha_0,\theta_0) - \phi(W,\gamma_0,\alpha_0,\theta_0) = \alpha_0(z,x)A(\theta_0)\{\gamma_0(z,x) - \hat{\gamma}_{-\ell}(z,x)\},$$
$$\phi(W,\gamma_0,\hat{\alpha}_{-\ell},\theta_0) - \phi(W,\gamma_0,\alpha_0,\theta_0) = \{\hat{\alpha}_{-\ell}(z,x) - \alpha_0(z,x)\}A(\theta_0)\{v - \gamma_0(z,x)\}.$$

To lighten the proof, we slightly abuse notation as follows:

$$\|\gamma_0 - \hat{\gamma}_{-\ell}\|^2 = E[\{\gamma_0(Z,X) - \hat{\gamma}_{-\ell}(Z,X)\}^2 \mid I_\ell];$$
$$\|\alpha_0 - \hat{\alpha}_{-\ell}\|^2 = E[\{\alpha(Z,X) - \hat{\alpha}_{-\ell}(Z,X)\}^2 \mid I_\ell].$$

1 By Lemma S2.1, the convergence holds due to $\|\gamma_0 - \hat{\gamma}_{-\ell}\| = o_p(1)$.

2 By Assumption S1.5 and Assumption A.1, we have

$$\|\alpha_0 A(\theta_0)(\gamma_0 - \hat{\gamma}_{-\ell})\| \leq CA(\theta_0)\|\gamma_0 - \hat{\gamma}_{-\ell}\| = o_p(1).$$

3 By Lemma S1.1 or Lemma S1.2, Assumption A.1, and law of iterated expectations with respect to $I_{-\ell}$, we have

$$\|(\hat{\alpha}_{-\ell} - \alpha_0)A(\theta_0)\{v - \gamma_0(z,x)\}\| \leq \|\hat{\alpha}_{-\ell} - \alpha_0\|A(\theta_0)C\vec{1} = o_p(1)$$

where $\vec{1}$ is the vector of ones.

Proposition S2.4. *Suppose the conditions of Theorem A.1 hold. Then*

$$n^{-1/2}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\{\phi(W_i,\hat{\gamma}_{-\ell},\hat{\alpha}_{-\ell},\theta_0) - \phi(W_i,\hat{\gamma}_{-\ell},\alpha_0,\theta_0)$$
$$- \phi(W_i,\gamma_0,\hat{\alpha}_{-\ell},\theta_0) + \phi(W_i,\gamma_0,\alpha_0,\theta_0)\} = o_p(1).$$

Proof. To begin, write

$$\begin{aligned}&\phi(w,\hat{\gamma}_{-\ell},\hat{\alpha}_{-\ell},\theta_0) - \phi(w,\hat{\gamma}_{-\ell},\alpha_0,\theta_0) - \phi(w,\gamma_0,\hat{\alpha}_{-\ell},\theta_0) + \phi(w,\gamma_0,\alpha_0,\theta_0)\\ &= -\{\hat{\alpha}_{-\ell}(z,x) - \alpha_0(z,x)\}A(\theta_0)\{\hat{\gamma}_{-\ell}(z,x) - \gamma_0(z,x)\}.\end{aligned}$$

Because convergence in first mean implies convergence in probability, it suffices to analyse

$$\begin{aligned}&E\left[\left|n^{-1/2}\sum_{\ell=1}^{L}\sum_{i\in I_\ell} -\{\hat{\alpha}_{-\ell}(Z_i,X_i) - \alpha_0(Z_i,X_i)\}A(\theta_0)\{\hat{\gamma}_{-\ell}(Z_i,X_i) - \gamma_0(Z_i,X_i)\}\right|\right]\\
&\leq \sum_{\ell=1}^{L} E\left[n^{1/2}\frac{1}{n}\sum_{i\in I_\ell}|-\{\hat{\alpha}_{-\ell}(Z_i,X_i) - \alpha_0(Z_i,X_i)\}A(\theta_0)\{\hat{\gamma}_{-\ell}(Z_i,X_i) - \gamma_0(Z_i,X_i)\}|\right]\\
&= \sum_{\ell=1}^{L} E\left(E\left[n^{1/2}\frac{1}{n}\sum_{i\in I_\ell}|\{\hat{\alpha}_{-\ell}(Z_i,X_i) - \alpha_0(Z_i,X_i)\}A(\theta_0)\{\hat{\gamma}_{-\ell}(Z_i,X_i) - \gamma_0(Z_i,X_i)\}| \mid I_{-\ell}\right]\right)\\
&= \sum_{\ell=1}^{L} E\left(E\left[\left|n^{1/2}\frac{n_\ell}{n}\{\hat{\alpha}_{-\ell}(Z_i,X_i) - \alpha_0(Z_i,X_i)\}A(\theta_0)\{\hat{\gamma}_{-\ell}(Z_i,X_i) - \gamma_0(Z_i,X_i)\}\right| \mid I_{-\ell}\right]\right).\end{aligned}$$

Applying Hölder's inequality elementwise and Corollary S1.1, we have convergence for each summand as follows:

$$\begin{aligned}&E\left[|n^{1/2}\frac{n_\ell}{n}\{\hat{\alpha}_{-\ell}(Z_i,X_i) - \alpha_0(Z_i,X_i)\}A(\theta_0)\{\hat{\gamma}_{-\ell}(Z_i,X_i) - \gamma_0(Z_i,X_i)\}| \mid I_{-\ell}\right]\\
&\leq E\left[|n^{1/2}\{\hat{\alpha}_{-\ell}(Z_i,X_i) - \alpha_0(Z_i,X_i)\}A(\theta_0)\{\hat{\gamma}_{-\ell}(Z_i,X_i) - \gamma_0(Z_i,X_i)\}| \mid I_{-\ell}\right]\\
&\leq n^{1/2}\|\hat{\alpha}_{-\ell} - \alpha_0\|A(\theta_0)\|\hat{\gamma}_{-\ell} - \gamma_0\|\\
&= o_p(1).\end{aligned}$$

In the penultimate step, we slightly abuse notation, using

$$\|\gamma_0 - \hat{\gamma}_{-\ell}\|^2 = E[\{\gamma_0(Z,X) - \hat{\gamma}_{-\ell}(Z,X)\}^2 \mid I_\ell];$$
$$\|\alpha_0 - \hat{\alpha}_{-\ell}\|^2 = E[\{\alpha(Z,X) - \hat{\alpha}_{-\ell}(Z,X)\}^2 \mid I_\ell].$$

PROPOSITION S2.5. *Under Assumption 2.1, for each fold I_ℓ, the following holds:*

1 $n^{1/2}E\{\psi(W,\hat{\gamma}_{-\ell},\alpha_0,\theta_0)\}=o_p(1)$;
2 $n^{1/2}E\{\phi(W,\gamma_0,\hat{\alpha}_{-\ell},\theta_0)\}=o_p(1)$.

PROOF. To begin, write

$$E\{\psi(W,\hat{\gamma}_{-\ell},\alpha_0,\theta_0)\}=E[A(\theta_0)\{\hat{\gamma}_{-\ell}(1,X)-\hat{\gamma}_{-\ell}(0,X)\}+\alpha_0(Z,X)A(\theta_0)\{V-\hat{\gamma}_{-\ell}(Z,X)\}];$$
$$E\{\phi(W,\gamma_0,\hat{\alpha}_{-\ell},\theta_0)\}=E[\hat{\alpha}_{-\ell}(Z,X)A(\theta_0)\{V-\gamma_0(Z,X)\}].$$

1 By Proposition 3.2, $E\left\{\psi(W,\hat{\gamma}_{-\ell},\alpha_0,\theta_0)\mid I_{-\ell}\right\}=0$. Applying the law of iterated expectations, we have $E\{\psi(W,\hat{\gamma}_{-\ell},\alpha_0,\theta_0)\}=0$.
2 By law of iterated expectations, $E\left\{\phi(W,\gamma_0,\hat{\alpha}_{-\ell},\theta_0)\mid I_{-\ell}\right\}=0$. Applying the law of iterated expectations, we have $E\{\psi(W,\hat{\gamma}_{-\ell},\alpha_0,\theta_0)\}=0$.

PROPOSITION S2.6. *Suppose the conditions of Theorem A.1 hold. Then*

1 *The Jacobian J exists.*
2 *There exists a neighborhood $\mathcal{N}$ of θ_0 with respect to $|\cdot|_2$ such that*
  (a) $\|\hat{\gamma}_{-\ell}-\gamma_0\|=o_p(1)$*;*
  (b) $\|\hat{\alpha}_{-\ell}-\alpha_0\|=o_p(1)$*;*
  (c) *For $\|\gamma-\gamma_0\|$ and $\|\alpha-\alpha_0\|$ small enough, $\psi(W_i,\gamma,\alpha,\theta)$ is differentiable in θ with probability approaching one;*
  (d) *There exists $\zeta>0$ and $d(W)$ such that $E\{d(W)\}<\infty$ and for $\|\gamma-\gamma_0\|$ small enough,*
$$\left|\frac{\partial\psi(w,\gamma,\alpha,\theta)}{\partial\theta}-\frac{\partial\psi(w,\gamma,\alpha,\theta_0)}{\partial\theta}\right|_2\le d(w)|\theta-\theta_0|_2^{\zeta}.$$
3 *For any fold I_ℓ and any components (j,k) ,*
$$E\left\{\left|\frac{\partial\psi_j(W,\hat{\gamma}_{-\ell},\hat{\alpha}_{-\ell},\theta_0)}{\partial\theta_k}-\frac{\partial\psi_j(W,\gamma_0,\alpha_0,\theta_0)}{\partial\theta_k}\right|\right\}=o_p(1).$$

PROOF. To begin, write
$$\frac{\partial\psi(w,\gamma,\alpha,\theta)}{\partial\theta}=\frac{\partial A(\theta)}{\partial\theta}\{\gamma(1,x)-\gamma(0,x)\}+\alpha(z,x)\frac{\partial A(\theta)}{\partial\theta}\{v-\gamma(z,x)\}$$
where $\partial A(\theta)/\partial\theta$ is a tensor consisting of 1s and 0s.

To lighten the proof, we slightly abuse notation as follows:
$$\|\gamma_0-\hat{\gamma}_{-\ell}\|^2=E[\{\gamma_0(Z,X)-\hat{\gamma}_{-\ell}(Z,X)\}^2\mid I_\ell];$$
$$\|\alpha_0-\hat{\alpha}_{-\ell}\|^2=E[\{\alpha(Z,X)-\hat{\alpha}_{-\ell}(Z,X)\}^2\mid I_\ell].$$

1 It suffices to show the second moment of the derivative is finite. By triangle inequality and Assumption A.1 we have
$$\left\|\frac{\partial A(\theta_0)}{\partial\theta}\{\gamma_0(1,x)-\gamma_0(0,x)\}+\alpha_0(z,x)\frac{\partial A(\theta)}{\partial\theta}\{v-\gamma_0(z,x)\}\right\|$$
$$\le\frac{\partial A(\theta_0)}{\partial\theta}\left\{\|\gamma_0(1,x)-\gamma_0(0,x)\|+CC'\right\}.$$

To bound the right hand side, by Lemma S2.1 we have

$$\|\gamma_0(1,x)-\gamma_0(0,x)\| \le \|\gamma_0(1,x)\| + \|\gamma_0(0,x)\| \le C\|\gamma_0\| < \infty.$$

2 (a) The convergence holds due to Assumption S1.5.
(b) The convergence holds due to Lemma S1.1 or Lemma S1.2.
(c) Differentiability holds since $\partial\psi(w,\gamma,\alpha,\theta)/\partial\theta$ does not depend on θ.
(d) The left hand side is exactly $\vec{0}$ since $\partial\psi(w,\gamma,\alpha,\theta)/\partial\theta$ does not depend on θ.

3 It suffices to analyse the difference

$$\begin{aligned}\xi &= \hat{\gamma}_{-\ell}(1,x)-\hat{\gamma}_{-\ell}(0,x)+\hat{\alpha}_{-\ell}(z,x)\{v-\hat{\gamma}_{-\ell}(z,x)\}\\ &\quad -[\gamma_0(1,x)-\gamma_0(0,x)+\alpha_0(z,x)\{v-\gamma_0(z,x)\}]\\ &= \hat{\gamma}_{-\ell}(1,x)-\gamma_0(1,x)\\ &\quad -\hat{\gamma}_{-\ell}(0,x)+\gamma_0(0,x)\\ &\quad +\hat{\alpha}_{-\ell}(z,x)\{v-\hat{\gamma}_{-\ell}(z,x)\}-\alpha_0(z,x)\{v-\hat{\gamma}_{-\ell}(z,x)\}\\ &\quad +\alpha_0(z,x)\{v-\hat{\gamma}_{-\ell}(z,x)\}-\alpha_0(z,x)\{v-\gamma_0(z,x)\}\\ &= \hat{\gamma}_{-\ell}(1,x)-\gamma_0(1,x)\\ &\quad -\hat{\gamma}_{-\ell}(0,x)+\gamma_0(0,x)\\ &\quad +\{\hat{\alpha}_{-\ell}(z,x)-\alpha_0(z,x)\}\{v-\gamma_0(z,x)\}\\ &\quad +\{\hat{\alpha}_{-\ell}(z,x)-\alpha_0(z,x)\}\{\gamma_0(z,x)-\hat{\gamma}_{-\ell}(z,x)\}\\ &\quad +\alpha_0(z,x)\{\gamma_0(z,x)-\hat{\gamma}_{-\ell}(z,x)\}\end{aligned}$$

where we use the decomposition

$$\begin{aligned}&\hat{\alpha}_{-\ell}(z,x)\{v-\hat{\gamma}_{-\ell}(z,x)\}-\alpha_0(z,x)\{v-\hat{\gamma}_{-\ell}(z,x)\}\\ &= \{\hat{\alpha}_{-\ell}(z,x)-\alpha_0(z,x)\}\{v-\gamma_0(z,x)+\gamma_0(z,x)-\hat{\gamma}_{-\ell}(z,x)\}.\end{aligned}$$

Hence

$$\begin{aligned}E\left(|\xi|\right) &\le E\left\{|\hat{\gamma}_{-\ell}(1,X)-\gamma_0(1,X)|\right\}\\ &\quad + E\left\{|\hat{\gamma}_{-\ell}(0,X)-\gamma_0(0,X)|\right\}\\ &\quad + E\left[|\{\hat{\alpha}_{-\ell}(Z,X)-\alpha_0(Z,X)\}\{V-\gamma_0(Z,X)\}|\right]\\ &\quad + E\left[|\{\hat{\alpha}_{-\ell}(Z,X)-\alpha_0(Z,X)\}\{\gamma_0(Z,X)-\hat{\gamma}_{-\ell}(Z,X)\}|\right]\\ &\quad + E\left[|\alpha_0(Z,X)\{\gamma_0(Z,X)-\hat{\gamma}_{-\ell}(Z,X)\}|\right].\end{aligned}$$

Consider the first term. Under Assumption S1.5, applying law of iterated expectation, Jensen's inequality, and Lemma S2.1, we have

$$\begin{aligned}E\left\{|\hat{\gamma}_{-\ell}(1,X)-\gamma_0(1,X)|\right\} &= E\left[E\left\{|\hat{\gamma}_{-\ell}(1,X)-\gamma_0(1,X)| \mid I_{-\ell}\right\}\right]\\ &\le E\{\|\hat{\gamma}_{-\ell}(1,x)-\gamma_0(1,x)\|\}\\ &\le CE(\|\hat{\gamma}_{-\ell}-\gamma_0\|)\\ &= o_p(1).\end{aligned}$$

Likewise for the second term. Consider the third term. Under Assumption A.1, applying law of iterated expectation, Lemma S1.1 or Lemma S1.2, and Hölder's inequality we have

$$\begin{aligned}
&E[|\{\hat{\alpha}_{-\ell}(Z,X)-\alpha_0(Z,X)\}\{V-\gamma_0(Z,X)\}|] \\
&= E\left(E[|\{\hat{\alpha}_{-\ell}(Z,X)-\alpha_0(Z,X)\}\{V-\gamma_0(Z,X)\}| \mid I_{-\ell}]\right) \\
&\leq E\left\{\|\hat{\alpha}_{-\ell}-\alpha_0\|\|v-\gamma_0(z,x)\|\right\} \\
&\leq CE(\|\hat{\alpha}_{-\ell}-\alpha_0\|) \\
&= o_p(1).
\end{aligned}$$

Consider the fourth term. By law of iterated expectations, Hölder's inequality, and Corollary S1.1 we have

$$\begin{aligned}
&E\left[|\{\hat{\alpha}_{-\ell}(Z,X)-\alpha_0(Z,X)\}\{\gamma_0(Z,X)-\hat{\gamma}_{-\ell}(Z,X)\}|\right] \\
&= E\left(E\left[|\{\hat{\alpha}_{-\ell}(Z,X)-\alpha_0(Z,X)\}\{\gamma_0(Z,X)-\hat{\gamma}_{-\ell}(Z,X)\}| \mid I_{-\ell}\right]\right) \\
&\leq E\left(\|\hat{\alpha}_{-\ell}-\alpha_0\|\|\gamma_0-\hat{\gamma}_{-\ell}\|\right) \\
&= o_p(1).
\end{aligned}$$

Consider the fifth term. By law of iterated expectations, Assumptions S1.5 and A.1, and Jensen's inequality, we have

$$\begin{aligned}
E\left[|\alpha_0(Z,X)\{\gamma_0(Z,X)-\hat{\gamma}_{-\ell}(Z,X)\}|\right] &= E\left(E\left[|\alpha_0(Z,X)\{\gamma_0(Z,X)-\hat{\gamma}_{-\ell}(Z,X)\}| \mid I_{-\ell}\right]\right) \\
&\leq CE\left[E\left\{|\gamma_0(Z,X)-\hat{\gamma}_{-\ell}(Z,X)| \mid I_{-\ell}\right\}\right] \\
&\leq CE(\|\gamma_0-\hat{\gamma}_{-\ell}\|) \\
&= o_p(1).
\end{aligned}$$

Proposition S2.7. *Suppose the conditions of Theorem A.1 hold. Then $\hat{\theta}=\theta_0+o_p(1)$.*

Proof. We verify the four conditions of Lemma S2.2 with

$$\begin{aligned}
Q_0(\theta) &= E\{\psi_0(\theta)\}^\top E\{\psi_0(\theta)\}, \\
\hat{Q}(\theta) &= \left\{\frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\hat{\psi}_i(\theta)\right\}^\top \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\hat{\psi}_i(\theta), \\
\psi_0(\theta) &= \psi(W,\gamma_0,\alpha_0,\theta), \\
\hat{\psi}_i(\theta) &= \psi(W_i,\hat{\gamma}_{-\ell},\hat{\alpha}_{-\ell},\theta).
\end{aligned}$$

1 The first condition follows from Assumption A.1,
2 The second condition follows from Corollary A.1.
3 The third condition follows from Corollary A.1.
4 Define

$$\begin{aligned}
\eta_0(w) &= \gamma_0(1,x)-\gamma_0(0,x)+\alpha_0(z,x)\{v-\gamma_0(z,x)\} \\
\hat{\eta}_{-\ell}(w) &= \hat{\gamma}_{-\ell}(1,x)-\hat{\gamma}_{-\ell}(0,x)+\hat{\alpha}_{-\ell}(z,x)\{v-\hat{\gamma}_{-\ell}(z,x)\}.
\end{aligned}$$

It follows that for $i\in I_\ell$,

$$\begin{aligned}
&\psi_0(\theta)=A(\theta)\eta_0(W), \quad E\{\psi_0(\theta)\}=A(\theta)E\{\eta_0(W)\}; \\
&\hat{\psi}_i(\theta)=A(\theta)\hat{\eta}_{-\ell}(W_i), \quad \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\hat{\psi}_i(\theta)=A(\theta)\frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\hat{\eta}_{-\ell}(W_i).
\end{aligned}$$

It suffices to show $n^{-1}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\hat{\eta}_{-\ell}(W_i) = E\{\eta_0(W)\} + o_p(1)$ since by continuous mapping theorem this implies that for all θ in Θ, $n^{-1}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\hat{\psi}_i(\theta) = E\{\psi_0(\theta)\} + o_p(1)$ and hence $\hat{Q}(\theta) = Q_0(\theta) + o_p(1)$ uniformly.
We therefore turn to proving the sufficient condition. Write

$$\begin{aligned} &\frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\hat{\eta}_{-\ell}(W_i) - E\{\eta_0(W)\} \\ =\ & \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\{\hat{\eta}_{-\ell}(W_i) - \eta_0(W_i)\} + \frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\eta_0(W_i) - E\{\eta_0(W)\}. \end{aligned}$$

Consider the initial terms. Denote $\xi_i = \hat{\eta}_{-\ell}(W_i) - \eta_0(W_i)$ as in Proposition S2.6 item 3. We prove convergence in mean by

$$\begin{aligned} E\left(\left|\frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\xi_i\right|\right) &\le \sum_{\ell=1}^{L} E\left(\frac{1}{n}\sum_{i\in I_\ell}|\xi_i|\right) \\ &= \sum_{\ell=1}^{L} E\left\{E\left(\frac{1}{n}\sum_{i\in I_\ell}|\xi_i| \mid I_{-\ell}\right)\right\} \\ &= \sum_{\ell=1}^{L} E\left\{\frac{n_\ell}{n}E(|\xi_i| \mid I_{-\ell})\right\} \\ &\le \sum_{\ell=1}^{L} E\left\{E(|\xi_i| \mid I_{-\ell})\right\} \\ &= o_p(1) \end{aligned}$$

where the first inequality is due to triangle inequality, the second equality is due to the law of iterated expectations, and the rest echoes the proof of Proposition S2.6 item 3.
Consider the latter terms. By the weak law of large numbers, if $E\{\eta_0(W)^2\} < \infty$ then

$$\frac{1}{n}\sum_{\ell=1}^{L}\sum_{i\in I_\ell}\eta_0(W_i) - E\{\eta_0(W)\} = \frac{1}{n}\sum_{i=1}^{n}\eta_0(W_i) - E\{\eta_0(W)\} = o_p(1).$$

To finish the argument, we verify $E\{\eta_0(W)^2\} = \|\eta_0\|^2 < \infty$. By triangle inequality, Assumption A.1, and Lemma S2.1,

$$\|\eta_0\| = \|\gamma_0(1,x) - \gamma_0(0,x) + \alpha_0(z,x)\{v - \gamma_0(z,x)\}\| \le \|\gamma_0(1,x) - \gamma_0(0,x)\| + CC'.$$

To bound the right hand side, appeal to Lemma S2.1:

$$\|\gamma_0(1,x) - \gamma_0(0,x)\| \le \|\gamma_0(1,x)\| + \|\gamma_0(0,x)\| \le C\|\gamma_0\| < \infty.$$

Proposition S2.8. *Suppose the conditions of Theorem A.1 hold. Then the following holds.*

1 $\hat{\theta} = \theta_0 + o_p(1)$,
2 $J^\top J$ *is nonsingular,*

3 $E\{\psi_0(W)^2\} < \infty$,
4 $E[\{\phi(W,\hat{\gamma}_{-\ell},\hat{\alpha}_{-\ell},\theta_0) - \phi(W,\hat{\gamma}_{-\ell},\alpha_0,\theta_0) - \phi(W,\gamma_0,\hat{\alpha}_{-\ell},\theta_0) + \phi(W,\gamma_0,\alpha_0,\theta_0)\}^2] = o_p(1)$.

Proof. As in the proof of Proposition S2.4, we can write

$$\begin{aligned}&\phi(w,\hat{\gamma}_{-\ell},\hat{\alpha}_{-\ell},\theta_0) - \phi(w,\hat{\gamma}_{-\ell},\alpha_0,\theta_0) - \phi(w,\gamma_0,\hat{\alpha}_{-\ell},\theta_0) + \phi(w,\gamma_0,\alpha_0,\theta_0)\\ &= -\{\hat{\alpha}_{-\ell}(z,x) - \alpha_0(z,x)\}A(\theta_0)\{\hat{\gamma}_{-\ell}(z,x) - \gamma_0(z,x)\}.\end{aligned}$$

To lighten the proof, we slightly abuse notation as follows:

$$\begin{aligned}\|\gamma_0 - \hat{\gamma}_{-\ell}\|^2 &= E[\{\gamma_0(Z,X) - \hat{\gamma}_{-\ell}(Z,X)\}^2 \mid I_\ell];\\ \|\alpha_0 - \hat{\alpha}_{-\ell}\|^2 &= E[\{\alpha(Z,X) - \hat{\alpha}_{-\ell}(Z,X)\}^2 \mid I_\ell].\end{aligned}$$

1 Convergence holds due to Proposition S2.7.
2 Nonsingularity holds due to Assumption A.1.
3 $E\{\psi_0(W)^2\} < \infty$ is immediate from $E\{\eta_0(W)^2\}$, which is proved in Proposition S2.7 item 4.
4 It suffices to analyse

$$\begin{aligned}&E\left([\{\hat{\alpha}_{-\ell}(z,x) - \alpha_0(z,x)\}A(\theta_0)\{\hat{\gamma}_{-\ell}(z,x) - \gamma_0(z,x)\}]^2\right)\\ &= E\left\{E\left([\{\hat{\alpha}_{-\ell}(z,x) - \alpha_0(z,x)\}A(\theta_0)\{\hat{\gamma}_{-\ell}(z,x) - \gamma_0(z,x)\}]^2 \mid I_{-\ell}\right)\right\}\\ &= E\left\{\|(\hat{\alpha}_{-\ell} - \alpha_0)A(\theta_0)(\hat{\gamma}_{-\ell} - \gamma_0)\|^2\right\}\\ &\le 2E\left\{\|\hat{\alpha}_{-\ell}A(\theta_0)(\hat{\gamma}_{-\ell} - \gamma_0)\|^2 + \|\alpha_0 A(\theta_0)(\hat{\gamma}_{-\ell} - \gamma_0)\|^2\right\}.\end{aligned}$$

Consider the first term. By Hölder's inequality, Assumption S1.1, and either Lemma S1.1 or Lemma S1.2, we have

$$|\hat{\alpha}_{-\ell}(z,x)| = |\hat{\rho}_{-\ell}^\top b(z,x)| \le |\hat{\rho}_{-\ell}|_1 |b(z,x)|_\infty = O_p(1).$$

It follows by Assumption S1.5 that

$$\|\hat{\alpha}_{-\ell}A(\theta_0)(\hat{\gamma}_{-\ell} - \gamma_0)\| = O_p(1)\|\hat{\gamma}_{-\ell} - \gamma_0\| = O_p(n^{-d_\gamma}) = o_p(1).$$

Consider the second term. By Assumption S1.5 and Assumption A.1, we have

$$\|\alpha_0 A(\theta_0)(\hat{\gamma}_{-\ell} - \gamma_0)\| \le CA(\theta_0)\|\hat{\gamma}_{-\ell} - \gamma_0\| = o_p(1).$$

Proof. (Proof of Theorem A.1) The proof now follows from Chernozhukov et al. (2022, Theorem 9). In particular, Proposition S2.3 verifies Chernozhukov et al. (2022, Assumption 1), Proposition S2.4 verifies Chernozhukov et al. (2022, Assumption 2), Proposition S2.5 verifies Chernozhukov et al. (2022, Assumption 3), Proposition S2.6 verifies Chernozhukov et al. (2022, Assumption 5), and Proposition S2.8 verifies Chernozhukov et al. (2022, Assumption 4). Finally, the parameter θ_0 is exactly identified; J is a square matrix, the GMM weighting can be taken as the identity matrix, so the formula for the asymptotic covariance matrix simplifies.

S3. TUNING

Algorithm A.2 takes as given the value of regularization parameter λ_n. For practical use, we provide an iterative tuning procedure to empirically determine λ_n. This is precisely the tuning procedure of Chernozhukov et al. (2022a), adapted from Chernozhukov et al. (2022b). Due to its iterative nature, the tuning procedure is most clearly stated as a replacement for Algorithm A.2.

Recall that the inputs to Algorithm A.2 are observations in $I_{-\ell}$, i.e. excluding fold ℓ. The analyst must also specify the p dimensional dictionary b. For notational convenience, we assume b includes the intercept in its first component: $b_1(z,x) = 1$. In this tuning procedure, the analyst must further specify a low dimensional subdictionary b^{low} of b. As in Algorithm A.2, the output of the tuning procedure is $\hat{\alpha}_{-\ell}$, an estimator of the balancing weight trained only on observations in $I_{-\ell}$.

The tuning procedure is as follows.

Algorithm S3.1. (Regularized balancing weight with tuning) *For observations in $I_{-\ell}$,*

Step *1. Initialize $\hat{\rho}_{-\ell}$ using b^{low}:*

$$\hat{G}^{low}_{-\ell} = \frac{1}{n-n_\ell}\sum_{i\in I_{-\ell}} b^{low}(Z_i,X_i)b^{low}(Z_i,X_i)^\top;$$

$$\hat{M}^{low}_{-\ell} = \frac{1}{n-n_\ell}\sum_{i\in I_{-\ell}} b^{low}(1,X_i) - b^{low}(0,X_i);$$

$$\hat{\rho}_{-\ell} = \left\{\begin{matrix}\left(\hat{G}^{low}_{-\ell}\right)^{-1}\hat{M}^{low}_{-\ell}\\ 0\end{matrix}\right\}.$$

Step *2. Calculate moments*

$$\hat{G}_{-\ell} = \frac{1}{n-n_\ell}\sum_{i\in I_{-\ell}} b(Z_i,X_i)b(Z_i,X_i)^\top;$$

$$\hat{M}_{-\ell} = \frac{1}{n-n_\ell}\sum_{i\in I_{-\ell}} b(1,X_i) - b(0,X_i).$$

Step *3. While $\hat{\rho}_{-\ell}$ has not converged,*

1 Update normalization

$$\hat{D}_{-\ell} = \left(\frac{1}{n-n_\ell}\sum_{i\in I_{-\ell}}[b(Z_i,X_i)b(Z_i,X_i)^\top\hat{\rho}_{-\ell} - \{b(1,X_i)-b(0,X_i)\}]^2\right)^{1/2}.$$

2 Update $(\lambda_n, \hat{\rho}_{-\ell})$

$$\lambda_n = \frac{c_1}{(n-n_\ell)^{1/2}}\Phi^{-1}\left(1-\frac{c_2}{2p}\right);$$

$$\hat{\rho}_{-\ell} = \operatorname*{argmin}_{\rho}\, \rho^\top\hat{G}_{-\ell}\rho - 2\rho^\top\hat{M}_{-\ell} + 2\lambda_n c_3|\hat{D}_{-\ell,11}\rho_1| + 2\lambda_n\sum_{j=2}^{p}|\hat{D}_{-\ell,jj}\rho_j|;$$

where ρ_j is the jth coordinate of ρ and $\hat{D}_{-\ell,jj}$ is the jth diagonal entry of $\hat{D}_{-\ell}$.

Step *4. Set $\hat{\alpha}_{-\ell}(z,x) = b(z,x)^\top\hat{\rho}_{-\ell}$.*

In step 1, b^{low} is sufficiently low dimensional that $\hat{G}^{\mathrm{low}}_{-\ell}$ is invertible. In practice, we take $dim(b^{low}) = dim(b)/40$.

In step 3, (c_1, c_2, c_3) are hyperparameters taken as $(1/2, 0.1, 0.1)$ in practice. We implement the optimization via generalized coordinate descent with soft thresholding. See Chernozhukov et al. (2022a) for a detailed derivation of this soft thresholding routine. In the optimization, we initialize at the previous value of $\hat{\rho}_{-\ell}$. For numerical stability, we use $\hat{D}_{-\ell} + 0.2I$ instead of $\hat{D}_{-\ell}$, and we cap the maximum number of iterations at 10.

We justify Algorithm S3.1 in the same manner as Chernozhukov et al. (2022b, Section 5.1). Specifically, we appeal to Belloni and Chernozhukov (2013, Theorem 8) for the homoscedastic case and Belloni et al. (2012, Theorem 1) for the heteroscedastic case.

S4. SIMULATIONS

S4.1. Simultaneous confidence band

Suppose we wish to form a simultaneous confidence band for the components of $\hat{\theta}$, which may be the complier counterfactual outcome distribution based on a finite grid $\mathcal{U}$, which is a subset of $\mathcal{Y}$. The following procedure allows us to do so from some estimator $\hat{C}$ for the asymptotic variance C of $\hat{\theta}$. Let $\hat{S} = diag(\hat{C})$ and $S = diag(C)$ collect the diagonal elements of these matrices.

ALGORITHM S4.1. (SIMULTANEOUS CONFIDENCE BAND) *Given $\hat{C}$,*

STEP *1. Calculate $\hat{\Sigma} = \hat{S}^{-1/2}\hat{C}\hat{S}^{-1/2}$.*

STEP *2. Sample Q from $\mathcal{N}(0, \hat{\Sigma})$ and compute the value $\hat{c}_a$ as the $(1-a)$ quantile of sampled $|Q|_\infty$.*

STEP *3. Form the confidence band*

$$(l_j, u_j) = \left\{\hat{\theta}_j - \hat{c}_a(\hat{C}_{jj}/n)^{1/2},\ \hat{\theta}_j + \hat{c}_a(\hat{C}_{jj}/n)^{1/2}\right\}$$

where $\hat{C}_{jj}$ is the diagonal entry of $\hat{C}$ corresponding to jth element $\hat{\theta}_j$ of θ.

COROLLARY S4.1. (SIMULTANEOUS CONFIDENCE BAND) *Suppose the conditions of Theorem A.1 hold. Then for a fixed and finite grid $\mathcal{U}$, the confidence band in Algorithm S4.1 jointly covers the true counterfactual distributions θ_0 at all grid points y in $\mathcal{U}$ with probability approaching the nominal level, i.e.* $\mathrm{pr}\{(\theta_0)_j \in (l_j, u_j)$ *for all* $j\} \to 1 - a$.

PROOF. Let c_a be the $(1-a)$ quantile of $|\mathcal{N}(0, \Sigma)|_\infty$ where $\Sigma = S^{-1/2}CS^{-1/2}$ and $S = diag(C)$. We first show that this critical value ensures correct asymptotic simultaneous coverage of confidence bands in the form of the rectangle

$$\{(l_0)_j, (u_0)_j\} = \left\{\hat{\theta}_j - c_a\left(\frac{C_{jj}}{n}\right)^{1/2}, \hat{\theta}_j + c_a\left(\frac{C_{jj}}{n}\right)^{1/2}\right\}$$

where C_{jj} is the diagonal entry of C corresponding to jth element $\hat{\theta}_j$ of θ.

The argument is as follows. Denote $(l_0, u_0) = \times_{j=1}^{2d}\{(l_0)_j, (u_0)_j\}$ where $d = dim(\mathcal{U})$.

Then the simultaneous coverage probability is

$$\begin{aligned}\mathrm{pr}\{\theta_0 \text{ is in } (l_0, u_0)\} &= \mathrm{pr}\{n^{1/2}(\hat{\theta} - \theta_0) \text{ is in } S^{1/2}(-c_a, c_a)^{2d}\} \\ &\to \mathrm{pr}\{\mathcal{N}(0, C) \text{ is in } S^{1/2}(-c_a, c_a)^{2d}\} \\ &= \mathrm{pr}\{S^{-1/2}\mathcal{N}(0, C) \text{ is in } (-c_a, c_a)^{2d}\} \\ &= \mathrm{pr}\{|\mathcal{N}(0, \Sigma)|_\infty \le c_a\} \\ &= 1 - a.\end{aligned}$$

Gaussian multiplier bootstrap is operationally equivalent to approximating c_a with $\hat{c}_a$, calculated in Algorithm S4.1, which is based on the consistent estimator $\hat{C}$.

S4.2. Results

We compare the performance of our proposed Auto-κ estimator with κ weighting (Abadie, 2003) and the original debiased machine learning with explicit propensity scores (Chernozhukov et al., 2018) in simulations. We focus on counterfactual distributions as our choice of complier parameter θ_0 over the grid $\mathcal{U}$ specified on the horizontal axis of Figure 1.

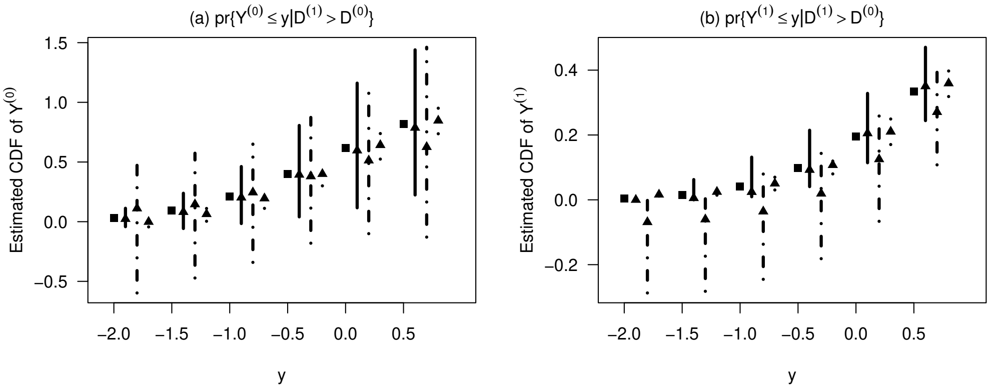

Figure 1. Numerical stability simulation. Simulation performance of κ weight (line), debiased machine learning (dot dash), and Auto-κ (dots) estimators for the counterfactual distribution, where the grid point is specified on the horizontal axis. The true values are solid squares. The vertical lines mark the 10% and 90% quantiles of the estimates across simulation draws and the solid triangles mark the median.

We consider a simple simulation design where Y is a continuous outcome, D is a binary treatment, Z is a binary instrumental variable, and X is a continuous covariate. We provide more details on the simulation design below. Each simulation consists of $n = 1000$ observations. We conduct 1000 such simulations and implement each estimator as follows.

For the κ weight, we estimate the propensity score $\hat{\pi}$ by logistic regression, which we then use in the weights $\hat{\kappa}^{(0)}(W), \hat{\kappa}^{(1)}(W)$ and subsequently the estimator $\hat{\theta}$. For debiased machine learning, we use five folds. We estimate the propensity score $\hat{\pi}$ by ℓ_1 regularized logistic regression, using a dictionary of basis functions $b(X)$ consisting of fourth order polynomials of X. We estimate $\hat{\gamma}$ by lasso, using a dictionary of basis functions $b(Z, X)$ consisting of fourth order polynomials of X and interactions between Z and the polynomials.

For Auto-κ, the key difference is that instead of estimating the propensity score, we directly estimate the balancing weight $\hat{\alpha}$ as described in Appendix A, using a dictionary of basis functions $b(Z, X)$ consisting of fourth order polynomials of X and interactions between Z and the polynomials. Subsequently, we estimate $\hat{\theta}$ and construct simultaneous confidence bands by steps outlined above. Since the true propensity scores $\pi_0(X)$ are highly nonlinear, we expect κ weighting and debiased machine learning to encounter issues of numerical instability. Furthermore, κ weighting might not be as efficient as the debiased machine learning and Auto-κ estimators, which have the semiparametrically efficient asymptotic variance.

Table 1. Bias and RMSE simulation for $\mathrm{pr}\{Y^{(0)} \leq y \mid D^{(1)} > D^{(0)}\}$

		Bias			RMSE	
y	κ weight	DML	Auto-κ	κ weight	DML	Auto-κ
-2.0	-3	-138	-37	99	3070	75
-1.5	-1	-119	-32	172	2576	76
-1.0	3	-45	-20	250	2040	79
-0.5	2	-35	2	384	1953	80
0.0	-17	18	21	556	1738	92
0.5	-12	3	34	638	3072	98
overall	-5	-53	-5	350	2391	83

Note: RMSE, root mean square error; DML, debiased machine learning; Auto-κ, automatic κ weighting. All entries have been multiplied by 10^3.

Table 2. Bias and RMSE simulation for $\mathrm{pr}\{Y^{(1)} \leq y \mid D^{(1)} > D^{(0)}\}$

		Bias			RMSE	
y	κ weight	DML	Auto-κ	κ weight	DML	Auto-κ
-2.0	2	-115	13	28	444	15
-1.5	4	-114	12	39	441	16
-1.0	8	-110	11	57	432	20
-0.5	16	-103	11	78	410	26
0.0	21	-93	16	90	379	35
0.5	21	-79	27	92	315	44
overall	12	-102	15	64	403	26

Note: RMSE, root mean square error; DML, debiased machine learning; Auto-κ, automatic κ weighting. All entries have been multiplied by 10^3.

For each value in the grid $\mathcal{U}$, Tables 1 and 2 present the bias and the root mean square error (RMSE) of each estimator across simulation draws. The last row averages the performance across grid points. Figure 1 visualizes the median as well as the 10% and 90% quantiles across simulation draws. Auto-κ outperforms debiased machine learning by a large margin due to numerical stability. Even though Auto-κ uses regularized machine learning to estimate $(\hat{\gamma}, \hat{\alpha})$, regularization bias does not translate into bias for estimating the counterfactual distribution due to the doubly robust moment function. In terms of efficiency, Auto-κ substantially outperforms κ weighting. Lastly, the simultaneous confidence bands based on the Auto-κ estimator have coverage probability of 98.4% for the counterfactual distribution of $Y^{(0)}$ and 93.6% for the counterfactual distribution of $Y^{(1)}$, which are quite close to the nominal level of 95%.

Numerical instability from inverting $\hat{\pi}$ is a known issue. In practice, researchers may try trimming and censoring. Trimming means excluding observations for which $\hat{\pi}$ is extreme. We trim according to Belloni et al. (2017), dropping observations with $\hat{\pi}$ not in $(10^{-12}, 1-10^{-12})$. Censoring means imposing bounds on $\hat{\pi}$ for such observations. We censor by setting $\hat{\pi} < 10^{-12}$ to be 10^{-12} and $\hat{\pi} > 1 - 10^{-12}$ to be $1 - 10^{-12}$. Auto-κ without trimming or censoring outperforms κ weighting and debiased machine learning even with trimming and censoring in this simulation design. Compare Figure 1, which has no preprocessing, with Figure 2, which has trimming, and Figure 3, which has censoring, to see this phenomenon. This property is convenient, since ad hoc trimming and censoring have limited theoretical justification (Crump et al., 2009).

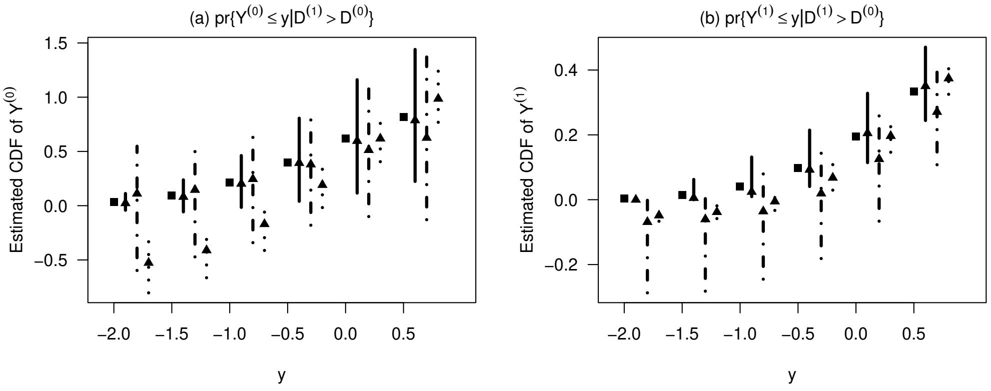

Figure 2. Numerical stability simulation: Trimming. Simulation performance of κ weight (line), debiased machine learning (dot dash), and Auto-κ (dots) estimators for the counterfactual distribution, where the grid point is specified on the horizontal axis. The true values are solid squares. The vertical lines mark the 10% and 90% quantiles of the estimates across simulation draws and the solid triangles mark the median. Observations with extreme propensity scores $\hat{\pi}$ not in $(10^{-12}, 1 - 10^{-12})$ are dropped.

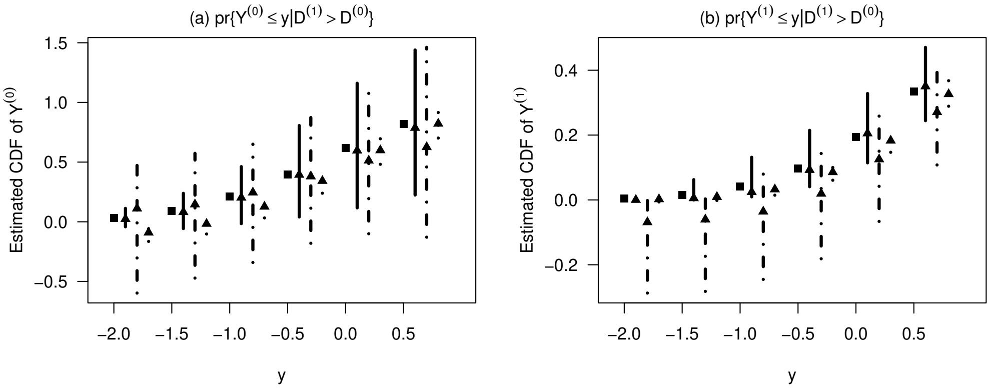

Figure 3. Numerical stability simulation: Censoring. Simulation performance of κ weight (line), debiased machine learning (dot dash), and Auto-κ (dots) estimators for the counterfactual distribution, where the grid point is specified on the horizontal axis. The true values are solid squares. The vertical lines mark the 10% and 90% quantiles of the estimates across simulation draws and the solid triangles mark the median. Observations with extreme propensity scores are censored by setting $\hat{\pi} < 10^{-12}$ to be 10^{-12} and $\hat{\pi} > 1 - 10^{-12}$ to be $1 - 10^{-12}$.

S4.3. Design

Each simulation consists of a sample of $n = 1000$ observations. A given observation is generated from the following model.

1 Draw X from $\mathcal{U}[0, 1]$.
2 Draw $Z \mid X = x$ from Bernoulli$\{\pi_0(x)\}$, where $\pi_0(x) = (0.05)1_{x \leq 0.5} + (0.95)1_{x > 0.5}$.
3 Draw $D \mid Z = z, X = x$ from Bernoulli(zx).
4 Draw $Y \mid Z = z, X = x$ from $\mathcal{N}(2zx^2, 1)$.

From observations of $W = (Y, D, Z, X^\top)^\top$, we estimate complier counterfactual outcome distributions $\hat{\theta} = (\hat{\beta}^\top, \hat{\delta}^\top)^\top$ at a few grid points y in $(-2.0, -1.5, -1.0, -0.5, 0, 0.5)$. The true parameter values are

$$\beta_0^y = \frac{\int_0^1 \{\Phi(y - 2x^2)(x - 1) + \Phi(y)\}\mathrm{d}x}{\int_0^1 x\mathrm{d}x}, \quad \delta_0^y = \frac{\int_0^1 \{\Phi(y - 2x^2)x\}\mathrm{d}x}{\int_0^1 x\mathrm{d}x}.$$

S5. APPLICATION DETAILS

Angrist and Evans (1998a) estimate the impact of childbearing D on female labour supply Y in a sample of 394,840 mothers, aged 21–35 with at least two children, from the 1980 Census (Angrist and Evans, 1998b; Angrist and Fernández-Val, 2013b). The first instrument Z_1 is twin births: Z_1 indicates whether the mother's second and third children were twins. The second instrument Z_2 is same-sex siblings: Z_2 indicates whether the mother's initial two children were siblings with the same sex. The authors reason that both (Z_1, Z_2) are quasi random events that induce having a third child such that the independence assumption holds unconditionally. However, the instruments are not independent of X, and therefore $\pi_0(X)$ still depends on X and may be estimated.

Angrist and Fernández-Val (2013a) use parametric κ weights to estimate two complier characteristics: (a) the average age of the mother's second child; and (b) the years of schooling of the mother. For a given characteristic $f(X) = X$, the authors specify the instrument propensity score model as

$$\pi_0(X) = [1 + \exp\{-(\beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4)\}]^{-1}.$$

As discussed in Section 3, such an approach is only valid when the parametric assumption on $\pi_0(X)$ is correct.

The semiparametric Auto-κ approach we propose combines the doubly robust moment function from Theorems 4.1 and 4.2 with the meta procedure in Algorithm A.1 and the regularized balancing weights in Algorithm A.2. We expand the dictionary of basis functions to include sixth order polynomials of X, and interactions between Z and polynomials of X. We directly estimate and regularize both the regression $\hat{\gamma}$ and the balancing weights $\hat{\alpha}$, tuning the regularization according to Algorithm S3.1. We set the hyperparameters (c_1, c_2, c_3) as $(0.5, 0.1, 0.1)$. In sample splitting, we partition the sample into five folds.

Finally, as a robustness check, we verify that κ weighting and Auto-κ yield similar estimated shares of compliers, i.e. similar estimates of $\text{pr}\{D^{(1)} > D^{(0)}\}$. These shares are typically reported in empirical research to interpret the strength and relevance of an instrumental variable. In the language of two stage least squares, these estimates correspond to the first stage. Table 3 reports the complier share estimates underlying the results of Table 1. Auto-κ produces similar complier share estimates to the κ weight approach of Angrist and Fernández-Val (2013a) while allowing for more flexible models and regularization.

Table 3. Comparison of complier shares

	Average age of second child		Average schooling of mother	
	Twins	Same-sex	Twins	Same-sex
κ weight	0.60	0.06	0.60	0.06
Auto-κ	0.73	0.06	0.77	0.06

REFERENCES

Abadie, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics 113*(2), 231–263.

Angrist, J. D. and W. N. Evans (1998a). Children and their parents' labor supply: Evidence from exogenous variation in family size. *American Economic Review 88*(3), 450–477.

Angrist, J. D. and W. N. Evans (1998b). Children and their parents' labor supply: Evidence from exogenous variation in family size: Replication package. *American Economic Review 88*(3), Data deposited at `https://economics.mit.edu/people/faculty/josh--angrist/angrist--data--archive`.

Angrist, J. D. and I. Fernández-Val (2013a). ExtrapoLATE-ing: External validity and overidentification in the LATE framework. In *Advances in Economics and Econometrics*, pp. 401–434.

Angrist, J. D. and I. Fernández-Val (2013b). ExtrapoLATE-ing: External validity and overidentification in the LATE framework: Replication package. In *Advances in Economics and Econometrics*, pp. Data deposited at `https://sites.bu.edu/ivanf/research/`.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica 80*(6), 2369–2429.

Belloni, A. and V. Chernozhukov (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli 19*(2), 521–547.

Belloni, A., V. Chernozhukov, I. Fernández-Val, and C. Hansen (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica 85*(1), 233–298.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics 37*(4), 1705–1732.

Chatterjee, S. and J. Jafarov (2015). Prediction error of cross-validated lasso. *arXiv:1502.06291*.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. K. Newey, and J. M. Robins (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal 21*(1), C1–C68.

Chernozhukov, V., J. C. Escanciano, H. Ichimura, W. K. Newey, and J. M. Robins (2022). Locally robust semiparametric estimation. *Econometrica 90*(4), 1501–1535.

Chernozhukov, V., W. K. Newey, and R. Singh (2022a). Automatic debiased machine learning of causal and structural effects. *Econometrica 90*(3), 967–1027.

Chernozhukov, V., W. K. Newey, and R. Singh (2022b). De-biased machine learning of global and local parameters using regularized Riesz representers. *The Econometrics Journal 25*(3), 576–601.

Crump, R. K., V. J. Hotz, G. W. Imbens, and O. A. Mitnik (2009, March). Dealing with limited overlap in estimation of average treatment effects. *Biometrika 96*(1), 187–199.

Farrell, M. H., T. Liang, and S. Misra (2021). Deep neural networks for estimation and inference. *Econometrica 89*(1), 181–213.

Newey, W. K. and D. McFadden (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics 4*, 2111–2245.

Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics 48*(4), 1875–1897.