SUPPLEMENTARY APPENDIX FOR "PROGRAM EVALUATION WITH HIGH-DIMENSIONAL DATA"

A. BELLONI, V. CHERNOZHUKOV, I. FERNÁNDEZ-VAL, AND C. HANSEN

Abstract. This supplement contains details about implementing the estimation and inference procedure considered in the main text. We also provide a small set of simulation results to complement the theoretical development in the body of the paper.

1. Implementation Details

In this section, we provide details about how we implemented the methodology developed in the main body of the paper in the empirical example. We first discuss estimation of local average treatment effects (LATE) and then extend this discussion to estimation of local quantile treatment effects (LQTE). Estimation of local average treatment effects on the treated (LATE-T) and local quantile treatment effects on the treated (LQTE-T) proceed in a similar fashion and so are not discussed.

1.1. **Local Average Treatment Effects.** Recall that the LATE of treatment D on outcome Y is defined as

$$\Delta_{LATE} = \theta_Y(1) - \theta_Y(0) = \frac{\alpha_{1_1(D)Y}(1) - \alpha_{1_1(D)Y}(0)}{\alpha_{1_1(D)}(1) - \alpha_{1_1(D)}(0)} - \frac{\alpha_{1_0(D)Y}(1) - \alpha_{1_0(D)Y}(0)}{\alpha_{1_0(D)}(1) - \alpha_{1_0(D)}(0)}$$

for $\alpha_V(z)$ and $\theta_Y(d)$ defined in equations (3) and (5) in the text respectively. It then follows by plugging in the definition of $\alpha_V(z)$ that we can express the LATE as

$$\Delta_{LATE} = \frac{\alpha_Y(1) - \alpha_Y(0)}{\alpha_{1_1(D)}(1) - \alpha_{1_1(D)}(0)}.$$

To obtain an estimate of the LATE, we thus need estimates of $\alpha_Y(z)$ and $\alpha_{1_1(D)}(z)$. Using the low-bias moment function given in quation (42) of the text, estimates of these key quantities can be contructed from estimates of $\mathrm{E}[Y|Z=1,X]$, $\mathrm{E}[Y|Z=0,X]$, $\mathrm{E}[D|Z=1,X]$, $\mathrm{E}[D|Z=0,X]$, and $\mathrm{E}[Z|X]$ where Z is the binary instrument (401(k) eligibility); D is the binary treatment (401(k) participation); X is one of the pre-specified sets of variables corresponding to the Indicator, Indicator plus interactions, B-Spline, or B-Spline plus interactions specification with dimension p; and Y is either total net financial assets or total wealth. In our application, we have $\mathrm{E}[D|Z=0,X]=0$ since one cannot participate unless one is eligible. We estimate the remainder of the functions using post-LASSO to estimate $\mathrm{E}[Y|Z=1,X]$ and $\mathrm{E}[Y|Z=0,X]$ and post-ℓ_1-penalized logistic regression to estimate $\mathrm{E}[D|Z=1,X]$ and $\mathrm{E}[Z|X]$.

Date: January 31, 2014.

To estimate $\mathrm{E}[Y|Z=1,X]$, we postulate that $\mathrm{E}[Y|Z=1,X] \approx X'\beta_1$. Let $\mathcal{I}_1$ denote the indices of observations that have $z_i = 1$. To estimate the coefficients β_1, we apply the formulation of the post-LASSO estimator given in Belloni, Chen, Chernozhukov, and Hansen (2012) with outcomes $\{y_i\}_{i\in\mathcal{I}_1}$ and covariates $\{x_i\}_{i\in\mathcal{I}_1}$. We set $\lambda = 2.2\sqrt{n}\Phi^{-1}(1-(1/\log(n))/(2(2p)))$ where $\Phi(\cdot)$ is the standard normal distribution function. We calculate penalty loadings according to Algorithm A.1 of Belloni, Chen, Chernozhukov, and Hansen (2012) using post-LASSO coefficient estimates at each iteration and with a the maximum number of iterations set to 15.[1] Let $\widehat{\beta}_1$ denote the resulting post-LASSO estimates of the coefficients using λ given above and the final set of penalty loadings. We then estimate $\mathrm{E}[Y|Z=1,X=x_i]$ as $x_i'\widehat{\beta}_1$ for each $i = 1,...,n$. We follow the same procedure to obtain estimates of $\mathrm{E}[Y|Z=0,X=x_i]$ as $x_i'\widehat{\beta}_0$ for each $i = 1,...,n$ where $\widehat{\beta}_0$ are the post-LASSO estimates using only the observations with $z_i = 0$.

Estimation of $\mathrm{E}[D|Z=1,X]$ and $\mathrm{E}[Z|X]$ proceed similarly replacing post-LASSO estimation with post-ℓ_1-penalized logistic regression. Specifically, we assume that $\mathrm{E}[D|Z=1,X] \approx \Lambda(X'\gamma_1)$ where $\Lambda(\cdot)$ is the logistic link function. We then obtain estimates of γ_1 by using the post-ℓ_1-penalized estimator defined in equations (29) and (30) in the text based on the logistic link function and with outcomes $\{d_i\}_{i\in\mathcal{I}_1}$ and covariates $\{x_i\}_{i\in\mathcal{I}_1}$ for $\mathcal{I}_1$ defined as above. We set $\lambda = 2.2\sqrt{n}\Phi^{-1}(1-(1/\log(n))/(2(2p)))$ where $\Phi(\cdot)$ is the standard normal distribution function. We calculate penalty loadings using Algorithm 1 from Section 6.1 of the main text with a maximum of 15 iterations.[2] Let $\widehat{\gamma}_1$ denote the resulting post-ℓ_1-penalized estimates of the coefficients using λ given above and the final set of penalty loadings. We estimate $\mathrm{E}[D|Z=1,X=x_i]$ as $\Lambda(x_i'\widehat{\gamma}_1)$ for each $i = 1,...,n$. We follow this procedure to obtain estimates of $\mathrm{E}[Z|X]$ as $\Lambda(x_i'\widehat{\tau})$ for each $i = 1,...,n$ where $\widehat{\tau}$ are the post-ℓ_1-penalized coefficient estimates obtained with $\{z_i\}_{i=1}^n$ as the outcome and $\{x_i\}_{i=1}^n$ as covariates using $\lambda = 2.2\sqrt{n}\Phi^{-1}(1-(1/\log(n))/(2p))$.

Using these baseline quantities, we obtain estimates

$$\begin{aligned}
\widehat{\alpha}_Y(1) &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{z_i(y_i - x_i'\widehat{\beta}_1)}{\Lambda(x_i'\widehat{\tau})} + x_i'\widehat{\beta}_1\right) = \frac{1}{n}\sum_{i=1}^{n}\psi_{1,i} \\
\widehat{\alpha}_Y(0) &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{(1-z_i)(y_i - x_i'\widehat{\beta}_0)}{1-\Lambda(x_i'\widehat{\tau})} + x_i'\widehat{\beta}_0\right) = \frac{1}{n}\sum_{i=1}^{n}\psi_{0,i} \\
\widehat{\alpha}_{1_1(D)}(1) &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{z_i(d_i - \Lambda(x_i'\widehat{\gamma}_1))}{\Lambda(x_i'\widehat{\tau})} + \Lambda(x_i'\widehat{\gamma}_1)\right) = \frac{1}{n}\sum_{i=1}^{n}\upsilon_{1,i} \\
\widehat{\alpha}_{1_1(D)}(0) &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{(1-z_i)d_i}{1-\Lambda(x_i'\widehat{\tau})}\right) = \frac{1}{n}\sum_{i=1}^{n}\upsilon_{0,i} = 0.
\end{aligned}$$

[1] We stop iterating before reaching the maximum number of iterations if the ℓ_2-norm of the difference in penalty loadings calculated across consecutive iterations is less than 10^{-6}.

[2] We stop iterating before reaching the maximum number of iterations if the ℓ_2-norm of the difference in penalty loadings calculated across consecutive iterations is less than 10^{-6}.

We then plug these estimates in to obtain

$$\widehat{\Delta}_{LATE} = \frac{\widehat{\alpha}_Y(1) - \widehat{\alpha}_Y(0)}{\widehat{\alpha}_{1_1(D)}(1) - \widehat{\alpha}_{1_1(D)}(0)}.$$

In the paper, we report both analytic and bootstrap standard error estimates for the LATE. The analytic standard errors are calculated as

$$\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{\psi_{1,i} - \psi_{0,i}}{\widehat{\alpha}_{1_1(D)}(1) - \widehat{\alpha}_{1_1(D)}(0)} - \widehat{\Delta}_{LATE}\right)^2 /n}.$$

We use wild bootstrap weights for obtaining the multiplier bootstrap estimates of the standard errors with 500 bootstrap replications. Specifically, for each $b = 1, ..., 500$, we calculate a bootstrap estimate of the LATE as

$$\widehat{\Delta}^b_{LATE} = \frac{\frac{1}{n}\sum_{i=1}^{n}(\psi_{1,i} - \psi_{0,i})\xi^b_i}{\frac{1}{n}\sum_{i=1}^{n}(\upsilon_{1,i} - \upsilon_{0,i})\xi^b_i}$$

where $\xi^b_i = 1 + r^b_{1,i}/\sqrt{2} + ((r^b_{2,i})^2 - 1)/2$ is the bootstrap draw for multiplier weight for observation i in bootstrap repetition b where $r^b_{1,i}$ and $r^b_{2,i}$ are random numbers generated as iid draws from two independent standard normal random variables. The boostrap standard error estimate is then the boostrap sample standard deviation of the $\widehat{\Delta}^b_{LATE}$: $\sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\widehat{\Delta}^b_{LATE} - \frac{1}{B}\sum_{b=1}^{B}\widehat{\Delta}^b_{LATE}\right)^2}$.

1.2. **Local Quantile Treatment Effects.** Calculation and inference for LQTE is more cumbersome than for the LATE and closely follows the procedure outlined in the paper using Strategy 1. We begin by choosing the set over which we would like to look at the LQTE. In our example, we chose to look at quantiles in the interval $[0.1, 0.9]$.

To calculate the LQTE, we first calculate the local average structural function for outcomes $Y_u = 1(Y \leq u)$ for a set of u and then invert to obtain estimates of the LQTE. In our example, we chose to look at $u \in [q_Y(.05), q_Y(.95)]$ where $q_Y(.05)$ and $q_Y(.95)$ are respectively the sample 5^{th} and 95^{th} percentiles of the outcome of interest Y. Since looking at the continuum of values in this interval is infeasible in practice, we discretize the interval and look at $Y_u = 1(Y \leq u)$ for $u \in \{q_Y(.05), q_Y(.06), q_Y(.07), ..., q_Y(.93), q_Y(.94), q_Y(.95)\}$. I.e. we set u equal to each percentile of Y between the 5^{th} and 95^{th} percentiles for a total of 91 different values of u to be considered. For each value of u, we need an estimate of the local average structural function defined in (5) in the text for $d \in \{0, 1\}$:

$$\theta_{1(Y\leq u)}(d) = \frac{\alpha_{1_d(D)1(Y\leq u)}(1) - \alpha_{1_d(D)1(Y\leq u)}(0)}{\alpha_{1_d(D)}(1) - \alpha_{1_d(D)}(0)}.$$

As with the LATE, we need estimates of $\mathrm{E}[D|Z = 1, X]$ and $\mathrm{E}[Z|X]$. We estimate these quantities as we did for the LATE but change the value of the penalty parameter used to reflect the fact that we are now interested in a large set, in theory a continuum, of model selection problems. Specifically, we assume that $\mathrm{E}[D|Z = 1, X] \approx \Lambda(X'\gamma_1)$ where $\Lambda(\cdot)$ is the logistic link function. We then obtain estimates of γ_1 by using the post-ℓ_1-penalized estimator defined in equations (29) and (30) in the text based on the logistic link function and with outcomes $\{d_i\}_{i\in\mathcal{I}_1}$

and covariates $\{x_i\}_{i\in\mathcal{I}_1}$ for $\mathcal{I}_1$ defined as above. We set $\lambda = 2.2\sqrt{n}\Phi^{-1}(1-(1/\log(n))/(2n(2p)))$ where $\Phi(\cdot)$ is the standard normal distribution function. We calculate penalty loadings using Algorithm 1 from Section 6.1 of the main text with a maximum of 15 iterations.[3] Let $\widehat{\gamma}_1$ denote the resulting post-ℓ_1-penalized estimates of the coefficients using λ given above and the final set of penalty loadings. We estimate $\mathrm{E}[D|Z=1,X=x_i]$ as $\Lambda(x_i'\widehat{\gamma}_1)$ for each $i=1,...,n$. We follow this procedure to obtain estimates of $\mathrm{E}[Z|X]$ as $\Lambda(x_i'\widehat{\tau})$ for each $i=1,...,n$ where $\widehat{\tau}$ are the post-ℓ_1-penalized coefficient estimates obtained with $\{z_i\}_{i=1}^n$ as the outcome and $\{x_i\}_{i=1}^n$ as covariates and $\lambda = 2.2\sqrt{n}\Phi^{-1}(1-(1/\log(n))/(2np))$. We also still have $\mathrm{E}[D|Z=0,X]=0$ in our application since one cannot participate in a 401(k) unless one is eligible. We then plug-in these estimates to obtain

$$\begin{aligned}
\widehat{\alpha}_{1_1(D)}(1) &= \frac{1}{n}\sum_{i=1}^n\left(\frac{z_i(d_i-\Lambda(x_i'\widehat{\gamma}_1))}{\Lambda(x_i'\widehat{\tau})}+\Lambda(x_i'\widehat{\gamma}_1)\right) = \frac{1}{n}\sum_{i=1}^n \upsilon_{1,1,i}\\
\widehat{\alpha}_{1_1(D)}(0) &= \frac{1}{n}\sum_{i=1}^n\left(\frac{(1-z_i)d_i}{1-\Lambda(x_i'\widehat{\tau})}\right) = \frac{1}{n}\sum_{i=1}^n \upsilon_{1,0,i} = 0\\
\widehat{\alpha}_{1_0(D)}(1) &= 1-\widehat{\alpha}_{1_1(D)}(1)\\
\widehat{\alpha}_{1_0(D)}(0) &= 1-\widehat{\alpha}_{1_1(D)}(0).
\end{aligned}$$

We also need to obtain estimates of $\alpha_{1_d(D)1(Y\le u)}(z)$ for each value of u and for $(z,d)\in\{(0,0),(0,1),(1,0),(1,1)\}$. These estimates will depend on the propensity score, $\mathrm{E}[Z|X]$, estimated above and quantities of the form $\mathrm{E}[1(D=d)1(Y\le u)|Z=z,X]$. We again approximate this function with $\mathrm{E}[1(D=d)1(Y\le u)|Z=z,X]\approx\Lambda(X'\beta_{u,d,z})$ and estimate the coefficients $\beta_{u,d,z}$ for each combination of d and z and each u using the post-ℓ_1-penalized estimator defined in equations (29) and (30) in the text based on the logistic link function. We set $\lambda = 2.2\sqrt{n}\Phi^{-1}(1-(1/\log(n))/(2n(2p)))$ where $\Phi(\cdot)$ is the standard normal distribution function. We calculate penalty loadings using Algorithm 1 from Section 6.1 of the main text with a maximum of 15 iterations.[4] We follow this procedure for each u with $\{1(y_i\le u)1(d_i=1)\}_{i\in\mathcal{I}_1}$ as the outcome and covariates $\{x_i\}_{i\in\mathcal{I}_1}$, with $\{1(y_i\le u)1(d_i=0)\}_{i\in\mathcal{I}_1}$ as the outcome and covariates $\{x_i\}_{i\in\mathcal{I}_1}$, and with $\{1(y_i\le u)1(d_i=0)\}_{i\in\mathcal{I}_0}$ as the outcome and covariates $\{x_i\}_{i\in\mathcal{I}_0}$ for $\mathcal{I}_1$ and $\mathcal{I}_0$ defined as above to obtain point estimates $\widehat{\beta}_{u,1,1}$, $\widehat{\beta}_{u,0,1}$, and $\widehat{\beta}_{u,0,0}$ respectively. We then estimate $\mathrm{E}[1(D=1)1(Y\le u)|Z=1,X]$ as $\Lambda(x_i'\widehat{\beta}_{u,1,1})$ for each $i=1,...,n$ and obtain estimates of $\mathrm{E}[1(D=0)1(Y\le u)|Z=1,X]$, and $\mathrm{E}[1(D=0)1(Y\le u)|Z=0,X]$ analogously. As before, we have $\mathrm{E}[1(D=1)1(Y\le u)|Z=0,X]=0$ since one cannot participate unless one is eligible.

[3] We stop iterating before reaching the maximum number of iterations if the ℓ_2-norm of the difference in penalty loadings calculated across consecutive iterations is less than 10^{-6}.

[4] We stop iterating before reaching the maximum number of iterations if the ℓ_2-norm of the difference in penalty loadings calculated across consecutive iterations is less than 10^{-6}.

We then plug-in these estimates to obtain

$$\widehat{\alpha}_{1_1(D)1(Y\leq u)}(1) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{z_i(d_i 1(y_i \leq u) - \Lambda(x_i'\widehat{\beta}_{u,1,1}))}{\Lambda(x_i'\widehat{\tau})} + \Lambda(x_i'\widehat{\beta}_{u,1,1})\right) = \frac{1}{n}\sum_{i=1}^{n}\kappa_{u,1,1,i}$$

$$\widehat{\alpha}_{1_1(D)1(Y\leq u)}(0) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{(1-z_i)(d_i 1(y_i \leq u))}{1-\Lambda(x_i'\widehat{\tau})}\right) = \frac{1}{n}\sum_{i=1}^{n}\kappa_{u,1,0,i} = 0$$

$$\widehat{\alpha}_{1_0(D)1(Y\leq u)}(1) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{z_i((1-d_i)1(y_i \leq u) - \Lambda(x_i'\widehat{\beta}_{u,0,1}))}{\Lambda(x_i'\widehat{\tau})} + \Lambda(x_i'\widehat{\beta}_{u,0,1})\right) = \frac{1}{n}\sum_{i=1}^{n}\kappa_{u,0,1,i}$$

$$\widehat{\alpha}_{1_0(D)1(Y\leq u)}(0) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{(1-z_i)((1-d_i)1(y_i \leq u) - \Lambda(x_i'\widehat{\beta}_{u,0,0}))}{1-\Lambda(x_i'\widehat{\tau})} + \Lambda(x_i'\widehat{\beta}_{u,0,0})\right) = \frac{1}{n}\sum_{i=1}^{n}\kappa_{u,0,0,i}.$$

Estimates of the local average structural (distribution) functions are formed using the estimators defined in the previous two paragraphs as

$$\widehat{\theta}_{1(Y\leq u)}(d) = \frac{\widehat{\alpha}_{1_d(D)1(Y\leq u)}(1) - \widehat{\alpha}_{1_d(D)1(Y\leq u)}(0)}{\widehat{\alpha}_{1_d(D)}(1) - \widehat{\alpha}_{1_d(D)}(0)}.$$

To obtain LQTE estimates, we then need to invert these local average structural functions. Since we only have the estimated distribution for each d evaluated on the finite grid of points $u \in \{q_Y(.05), q_Y(.06), q_Y(.07), ..., q_Y(.93), q_Y(.94), q_Y(.95)\}$, we do this inversion by linearly interpolating the value of the distribution function between these points to find the value of the outcome associated with each quantile in the set $q \in [0.1, 0.11, .0, 12, ..., 0.89, .0.9]$ which we denote as $\widehat{\theta}_Y^{\leftarrow}(q,d)$. The LQTE at point q is then estimated as $\widehat{\Delta}(q) = \widehat{\theta}_Y^{\leftarrow}(q,1) - \widehat{\theta}_Y^{\leftarrow}(q,0)$.

For the LQTE, we only report inference based on the multiplier bootstrap using 500 bootstrap replications. For each $b = 1, ..., 500$, we generate bootstrap weights as $\xi_i^b = 1 + r_{1,i}^b/\sqrt{2} + ((r_{2,i}^b)^2 - 1)/2$ for observation i in bootstrap repetition b where $r_{1,i}^b$ and $r_{2,i}^b$ are random numbers generated as iid draws from two independent standard normal random variables. We then use these weights to form bootstrap estimates of the local average structural functions

$$\widehat{\theta}^b_{1(Y\leq u)}(d) = \frac{\widehat{\alpha}^b_{1_d(D)1(Y\leq u)}(1) - \widehat{\alpha}^b_{1_d(D)1(Y\leq u)}(0)}{\widehat{\alpha}^b_{1_d(D)}(1) - \widehat{\alpha}^b_{1_d(D)}(0)}$$

where

$$\widehat{\alpha}^b_{1_1(D)}(1) = \frac{1}{n}\sum_{i=1}^{n} \xi^b_i \upsilon_{1,1,i},$$
$$\widehat{\alpha}^b_{1_1(D)}(0) = \frac{1}{n}\sum_{i=1}^{n} \xi^b_i \upsilon_{1,0,i},$$
$$\widehat{\alpha}^b_{1_0(D)}(1) = 1 - \widehat{\alpha}^b_{1_1(D)}(1),$$
$$\widehat{\alpha}^b_{1_0(D)}(0) = 1 - \widehat{\alpha}^b_{1_1(D)}(0),$$
$$\widehat{\alpha}^b_{1_1(D)1(Y\leq u)}(1) = \frac{1}{n}\sum_{i=1}^{n} \xi^b_i \kappa_{u,1,1,i},$$
$$\widehat{\alpha}^b_{1_1(D)1(Y\leq u)}(0) = \frac{1}{n}\sum_{i=1}^{n} \xi^b_i \kappa_{u,1,0,i} = 0,$$
$$\widehat{\alpha}^b_{1_0(D)1(Y\leq u)}(1) = \frac{1}{n}\sum_{i=1}^{n} \xi^b_i \kappa_{u,0,1,i},$$
$$\widehat{\alpha}^b_{1_0(D)1(Y\leq u)}(0) = \frac{1}{n}\sum_{i=1}^{n} \xi^b_i \kappa_{u,0,0,i}.$$

From these bootstrap estimates of the average structural distribution functions, we obtain bootstrap LQTE estimates as above through inversion by linearly interpolating the value of the distribution function between the finite set of points at which we have estimated values to find the value of the outcome associated with each quantile in the set $q \in [0.1, 0.11, .0, 12, ..., 0.89, .0.9]$, denoted $(\widehat{\theta^{\leftarrow}_Y}(q,d))^b$. The bootstrap estimate of the LQTE for bootstrap replication b at point q is then $\widehat{\Delta}^b(q) = (\widehat{\theta^{\leftarrow}_Y}(q,1))^b - (\widehat{\theta^{\leftarrow}_Y}(q,0))^b$. We form bootstrap standard error estimates for the LQTE at each quantile q as $s(q) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\widehat{\Delta}^b(q) - \frac{1}{B}\sum_{b=1}^{B}\widehat{\Delta}^b(q)\right)^2}$

We also use the bootstrap LQTE estimates to obtain the critical values we use when plotting the uniform confidence bands in our example. We form bootstrap t-statistics for each quantile q as $t^b(q) = (\widehat{\Delta}^b(q) - \widehat{\Delta}(q))/s(q)$. We then take $t^b_{\max} = \max_q\{|t^b(q)|\}$ and use the 95^{th} percentile of the bootstrap distribution of $t^b_{\max}$ as the critical value in constructing the confidence intervals for our figures.

2. Simulation Experiment

In this section, we present results from a brief simulation experiment. The results illustrate the performance of our proposed treatment effect estimator that makes use of estimating equations satisfying the key orthogonality condition given in equation (2) in the main text and variable selection relative to an estimator that uses variable selection but is based on a "naive" estimating equation that does not satisfy the orthogonality condition. We find that inference based on the naive estimator can suffer from substantial size distortions and that the performance of this estimator is strongly dependent on features of the data generating process (DGP). We also find

that tests based on the estimator constructed using our procedure have size close to the nominal level uniformly across all DGPs we consider consistent with the theory developed in the paper.

For simplicity, we consider the case where the treatment, d_i, is exogenous conditional on control variables x_i. In this case, we can apply the results of the paper substituting d_i for z_i in each instance where instruments z_i are used since d_i is conditionally exogenous and thus a valid instrument for itself. All of the simulation results are based on data generated as

$$d_i = \mathbf{1}\left\{\frac{\exp\{x_i'(c_d\theta_0)\}}{1+\exp\{x_i'(c_d\theta_0)\}} > v_i\right\}$$
$$y_i = d_i[x_i'(c_y\theta_0)] + \zeta_i$$

where $v_i \sim U(0,1)$, $\zeta_i \sim N(0,1)$, v_i and ζ_i are independent, $p = \dim(x_i) = 250$, the covariates $x_i \sim N(0,\Sigma)$ with $\Sigma_{kj} = (0.5)^{|j-k|}$, and the sample size $n = 200$. θ_0 is a $p\times 1$ vector with elements set as $\theta_{0,j} = (1/j)^2$ for $j = 1, ..., p$. c_d and c_y are scalars that control the strength of the relationship between the controls, the outcome, and the treatment variable. We use several different combinations of c_d and c_y, setting $c_d = \sqrt{\frac{(\pi^2/3)R_d^2}{(1-R_d^2)\theta_0'\Sigma\theta_0}}$ and $c_y = \sqrt{\frac{R_d^2}{(1-R_d^2)\theta_0'\Sigma\theta_0}}$ for all combinations of $R_d^2 \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$ and $R_y^2 \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$.

We report results for two different inference procedures in Figure 2. The right panel of the figure shows size of 5% level t-tests for the average treatment effect where the point estimate is formed using our proposed estimator based on model selection and orthogonal estimating equations and the standard error is estimated using a plug-in estimator of the asymptotic variance. The left panel shows size of 5% level t-tests for the average treatment effect estimated as

$$\widehat{\theta}_{naive} = \frac{1}{n}\sum_{i=1}^{n}(\widehat{g}_y(1,x_i) - \widehat{g}_y(0,x_i))$$

where $\widehat{g}_y(d,x_i)$ is a post-model-selection estimator of $\mathrm{E}[Y|D=d, X=x_i]$ and the standard error is estimated using a plug-in estimator of the asymptotic variance of $\widehat{\theta}_{naive}$.

Both procedures rely on post-model-selection estimates of the conditional expectations $\mathrm{E}[Y|D = d, X = x_i]$, and we use exactly the same estimator of this quantity in both cases. Specifically, we apply the Square-Root LASSO of Belloni, Chernozhukov, and Wang (2011) with outcome Y and covariates $(D, D*X_1, ..., D*X_p, (1-D), (1-D)*X_1, ..., (1-D)*X_p)$ to select variables. We set the penalty level in the Square-Root LASSO using the "exact" option of Belloni, Chernozhukov, and Wang (2011) under the assumption of homoscedastic, Gaussian errors ζ_i with the tuning confidence level required in Belloni, Chernozhukov, and Wang (2011) set equal to 95%. After running the square-root-LASSO, we then estimate regression coefficients by regressing Y onto only those variables that were estimated to have non-zero coefficients by the square-root LASSO. We then form estimates of $\mathrm{E}[Y|D=1, X=x_i]$ by plugging in $(1, x_i')'$ into the estimated model for $i = 1, ..., n$ and form estimates of $\mathrm{E}[Y|D=0, X=x_i]$ by plugging in $(0, x_i')'$ into the estimated model for $i = 1, ..., n$.

For our proposed method, we also need an estimate of the propensity score. We obtain our estimates of the propensity score by using ℓ_1-penalized logistic regression with D as the

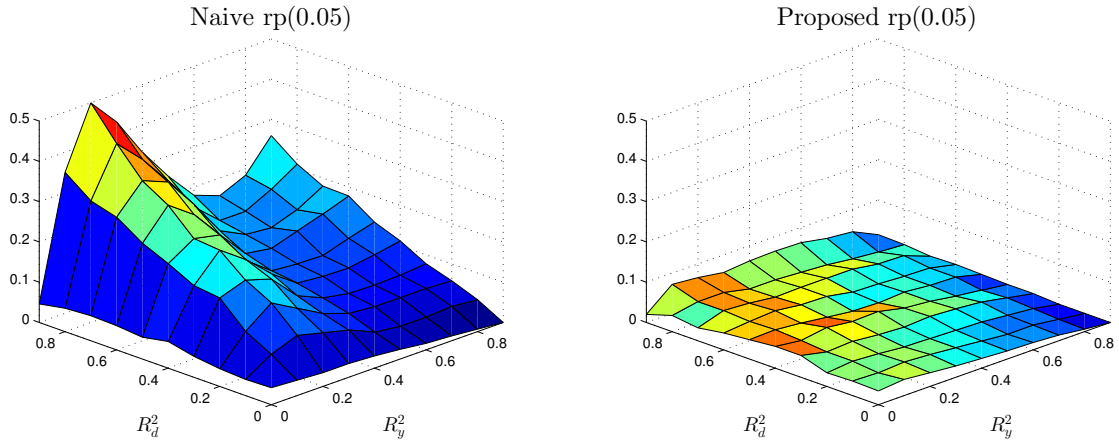

FIGURE 1. Rejection frequencies of 5% level tests for average treatment effect estimators following model selection. The left panel shows size of a test based on a "naive" estimator (Naive rp(0.05)), and the right panel shows size of a test based on our proposed procedure (Proposed rp(0.05)).

outcome and X as the covariates with penalty level set equal to $.5\sqrt{n}\Phi^{-1}(1-1/2p)/n$ where $\Phi(\cdot)$ is the standard normal distribution function using the MATLAB function "glmlasso".[5] We standardize the variables in X and set penalty loadings equal to 1. After running the ℓ_1-penalized logistic regression, we estimate the propensity score by taking fitted values from the conventional logistic regression of D onto only those variables that had non-zero estimated coefficients in the ℓ_1-penalized logistic regression.

Looking at the results, we see the behavior of the naive testing procedure depends heavily on the underlying coefficient sequence used to generate the data. There are substantial size distortions for many of the coefficient designs considered with good performance, size close to the nominal level, only occuring in a handful of cases. It is worth noting that in practice one does not know the underlying DGP and even estimation of the quantities necessary to know where one is in the figure may be infeasible even in this simple scenario. Our proposed procedure does a much better job at delivering accurate inference, producing tests with size close to the nominal level across all designs considered. That is, the simulation illustrates the uniformity derived in the theoretical development of our estimator illustrating that its performance is relatively good uniformly across a variety of coefficient sequences. While simply illustrative, these simulation results reinforce the theoretical development of the main paper which prove that our proposed estimation and inference procedures have good properties uniformly across a variety of DGPs where approximate sparsity holds.

[5]This penalty level is equivalent to that discussed in the main paper since "glmlasso" scales the problem in a slightly different way.

References

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, 2369–2429, Arxiv, 2010.

Belloni, A., V. Chernozhukov, and L. Wang (2011): "Square-Root-LASSO: Pivotal Recovery of Sparse Signals via Conic Programming," *Biometrika*, 98(4), 791–806, Arxiv, 2010.