Targeted Undersmoothing—Sensitivity Analysis for Sparse Estimators

Christian Hansen

The University of Chicago
Booth School of Business
5807 S. Woodlawn, Chicago, IL 60637
e-mail: chansen1@chicagobooth.edu

Damian Kozbur

University of Zürich
Department of Economics
Schönberggasse 1, 8001 Zürich
e-mail: damian.kozbur@econ.uzh.ch

Sanjog Misra

The University of Chicago
Booth School of Business
5807 S. Woodlawn, Chicago, IL 60637
e-mail: sanjog.misra@chicagobooth.edu

Abstract: This paper proposes a procedure for assessing sensitivity of inferential conclusions for functionals of sparse high-dimensional models following model selection. The proposed procedure is called targeted undersmoothing. Functionals considered include dense functionals that may depend on many or all elements of the high-dimensional parameter vector. The sensitivity analysis is based on systematic enlargements of an initially selected model. By varying the enlargements, one can conduct sensitivity analysis about the strength of empirical conclusions to model selection mistakes. We illustrate the procedure's performance through simulation experiments and two empirical examples.

JEL Codes: C12, C51, C55

Keywords and phrases: misspecification, model selection, sparsity, dense functionals, hypothesis testing.

* First version: August 2016. This version is of April 27, 2020. Sanjog Misra would like to thank the Kilts Center for Marketing at the University of Chicago Booth School of Business and the Neubauer Family Foundation for their financial support of this research. Christian Hansen would like to thank the National Science Foundation as well as The University of Chicago Booth School of Business for financial support of this research. Damian Kozbur would like to thank The University of Zürich for financial support of this research. In addition, the authors would like to thank Susan Athey, Alexandre Belloni, Victor Chernozhukov, Whitney Newey, as well as seminar participants at KU Leuven Post-Model Selection Colloquium (2016), The University of St. Gallen, Brigham Young University, Vanderbilt, University of Naples Federico II, Swiss Statistics Seminar (2017), and Asian Meetings of the Econometric Society (2017) for helpful comments.

1. Introduction

Rich data sets that contain information on a large number of variables for each observation are increasingly available to empirical researchers. Such data offer many opportunities for analyzing complex phenomena but pose practical and theoretical challenges. One key complication is that researchers may be uncertain about what specification to use when formulating a statistical model and that the complexity of choosing a specification increases rapidly in the number of available covariates.

There are a variety of model selection devices in the statistics and econometrics literature that can be used to aid empirical researchers in the choice of model specification; see, e.g. Hastie et al. (2009). A popular structure that underlies many statistical model selection procedures is sparsity. Sparsity is a general term for an assumption which states that the true model depends only on a small subset of the unknown parameters. A leading example is the sparse linear regression model

$$y_i = x_i'\beta_0 + \varepsilon_i, \quad s_0 = |support(\beta_0)| \ll n$$

where i indexes observations, n denotes sample size, y_i is an outcome, x_i are covariates, ε_i are conditional mean 0 idiosyncratic disturbance terms, and β_0 is an unknown parameter to be estimated with $support(\beta_0) = S_0$ and $s_0 = |S_0|$. Corresponding to the notion of sparsity, a sparse estimator returns a model in which only a small number of estimated parameters are nonzero. There are a variety of sensible sparse estimators in the literature, with a leading example being the Lasso estimator of Frank and Friedman (1993) and Tibshirani (1996).

The use of statistical model selection devices complicates drawing inferences for target parameters that depend on the selected model such as regression coefficients, individual specific treatment effects, elasticities, and other counterfactual objects. Heuristically, the difficulty arises due to the fact that model selectors will tend to

be imperfect in finite samples. Estimated parameters following model selection thus exhibit behavior akin to that of parameters estimated in misspecified models. See, for example, Leeb and Pötscher (2008) and Pötscher (2009) for formal development of the issues surrounding post-model-selection inference.

In this paper, we consider an intuitive and computationally simple way to assess the sensitivity of statistical conclusions to model selection mistakes in which relevant features are excluded from the selected model. We start with a scalar target parameter, ϑ_0, that may depend on the entirety of a high-dimensional parameter vector and additional inputs. For example, we may be interested in $\vartheta_0 = (x^*)'\beta_0$, the conditional mean for individuals with characteristics x^*, in the high dimensional linear model described above. Our proposal is to form sensitivity sets by starting with a typical confidence interval obtained from an initially selected model and then systematically enlarging the interval by perturbing the model to account for possible model selection mistakes. More formally, our proposed confidence set is constructed as the union of standard statistical confidence sets based on the convex hull of $\mathsf{CI}(\widehat{S}^{up}) \cup \mathsf{CI}(\widehat{S}^{low})$, where $\mathsf{CI}(S)$ denotes a confidence region for ϑ_0 based on a model S under the assumption that S is the correct model. $\widehat{S}^{up}$ and $\widehat{S}^{low}$ are in turn models selected from the data based on

1. An initially selected model $\widehat{S}^0$ chosen via a standard method targeting model fit to the data.
2. Two additionally selected models: an upper model, $\widehat{S}^{up} \supseteq \widehat{S}^0$, and a lower model, $\widehat{S}^{low} \supseteq \widehat{S}^0$, chosen to make the upper and lower bounds of the confidence set for the target parameter as large and small, respectively, as possible by adding a small number of variables to the model $\widehat{S}^0$.

Our focus is on situations with a large number of potential variables and where

the initial model selection is performed with a standard high dimensional estimator like the Lasso.[1] The subsequent model selection steps depend on the functional of interest and target the behavior of that functional accommodating model selection mistakes made in the first step. The subsequent steps are important since mistakes are inherent to all model selection procedures unless unrealistic conditions are imposed on the formal setting.[2]

We term the procedure outlined above 'targeted undersmoothing' (TU) and we term the resulting intervals 'targeted undersmoothing intervals' (TU intervals). This naming is due to a useful, though informal, heuristic analogy between high-dimensional estimation and nonparametric estimation. A key problem in nonparametric regression estimation is to choose a bandwidth (for kernel-based estimates) or a set of approximating functions (in series- or sieve-based methods). Sufficiently small bandwidths and more flexible sets of approximating functions each lead to undersmoothing in estimating the target function in the sense that bias may be taken to be small relative to sampling variation. Undersmoothing can thus be used to justify inference based on correctly-centered Gaussian approximations. For a review, see Li and Racine (2006). Choosing a bandwidth or set of approximating functions is not unlike choosing a penalty parameter in ℓ_1-penalized regression where smaller values of the penalty parameter result in more complex models.

Unfortunately, simply decreasing the penalty parameter in penalized estimation of a sparse high-dimensional model does not alleviate bias in the same way as decreasing a bandwidth in a traditional kernel problem due to the complexity of the model space in

[1]The proposed approach to sensitivity analysis could be used with any initial model, such as an intuitively selected baseline model, and in low-dimensional settings though we would recommend adopting the approach of Cattaneo et al. (2018) and Cattaneo et al. (2019) in low and moderate dimensional settings.

[2]In the linear model, such conditions include β-min conditions, which assert that nonzero unknown parameters must be bounded uniformly away from zero in absolute value, and conditions restricting the association between covariates.

high-dimensional problems. Heuristically, moderate strength signals whose exclusion leads to bias are hard to pick out from among the many irrelevant variables; and as the penalty parameter is lowered beyond theoretically justified levels, it is likely that the first variables to enter the model will be irrelevant signals that happen to be moderately correlated to the outcome in the sample at hand. Intuitively, the TU approach addresses this problem by undersmoothing in those directions that seem to be most likely to account for bias in the target parameter by directly focusing on the functional of interest rather than model fit.

Our paper complements many interesting papers that look at related problems. There is now a relatively large literature aimed at delivering uniformly valid inference for pre-specified target parameters where machine learning or model selection is used to estimate nuisance functions; see, for example, Chernozhukov et al. (2016) or Bickel et al. (1998) and references therein. Wager and Athey (2015) and Athey et al. (2016b) study asymptotically Gaussian inference for heterogeneous treatment effects using random forests in settings with low-dimensional controls. Athey and Imbens (2016) study estimation of heterogeneous treatment effects in conjunction with machine learning, relying on tree-based methods and a sample-splitting technique. Athey et al. (2016a) perform residual rebalancing to estimate average treatment effects with high dimensional control variables. Cai and Guo (2016) consider construction of confidence sets for dense functionals given by $a(\beta) = \|\beta\|_l$ for various $1 \leqslant l \leqslant \infty$. Both Zhu and Bradic (2016) and Zhu and Bradic (2017) construct hypothesis tests for objects similar to those considered in our paper via ℓ_1-projections of coefficient estimates to the set of coefficients consistent with the null. Zhu and Bradic (2016) only considers linear functionals while Zhu and Bradic (2017) considers general nonlinear functionals under strong sparsity conditions. Li and Müller (2020) consider inference about a single regression coefficient in a high-dimensional linear model subject to an

explicit constraint on the ℓ_2 norm of the coefficients on the remaining variables in the model. This approach provides an interesting complement to our approach as the explicit constraint on the ℓ_2 norm of coefficients is akin to a known level of sparsity but does not induce sparsity. Finally, Armstrong and Kolesar (2017) considers confidence intervals for functionals of a nonparametric regression function constructed given a researcher-specified upper bound on the true Lipschitz constants of the unknown nonparametric regression function in a low-dimensional setting. This approach is similar in spirit to our proposal where the known upper bound on the Lipschitz constant is analogous to a known upper bound on the level of sparsity s_0.

2. Targeted Undersmoothing for Sensitivity Analysis

This section describes the setting for TU and defines the TU algorithm. The discussion in this section is kept at a general level. Specific examples with sparse linear models and model selection via Lasso are presented and discussed in the sections that follow.

The setting is inference for an unknown parameter of interest $\vartheta_0 \in \mathbb{R}$, which is a real scalar defined by a functional

$$\vartheta_0 = a(\mathrm{P}_0).$$

We assume a is a known functional of a data generating process P_0. We assume that P_0 depends on a true unknown parameter β_0 with dimension $dim(\beta_0) = p$. We are primarily interested in high-dimensional applications and thus assume sparsity. Set

$$S_0 = support(\beta_0) \text{ and } s_0 = |S_0|$$

so that s_0 is the number of nonzero components of the vector β_0.[3]

[3]The setting and results in this paper and its supplement can be extended to the case that β_0 can be decomposed into a sparse component and a small component, so that $\beta_0 = \beta_0^{(1)} + \beta_0^{(2)}$, with $|support(\beta_0^{(1)})| \leqslant s_0$ and $\|\beta_0^{(2)}\|_2$ sufficiently small.

Consider a high-level setting where the researcher can construct the following objects.

1. A method for model selection, based on an appropriate measure of overall fit, giving an initial estimated support $\widehat{S}^0$ for S_0 with $\widehat{s} = |\widehat{S}^0|$; and, given $\widehat{S}^0$, an estimate $\widehat{\vartheta}$.
2. A method for estimating an asymptotically valid confidence region for ϑ_0 for known S_0. For each sufficiently sparse subset $K \subseteq \{1, ..., p\}$, $[\ell_K, u_K]$ is a feasible estimated confidence set for ϑ_0 with the additional property $\mathrm{P}([\ell_K, u_K] \ni \vartheta_0)$ is approximately $1 - \alpha$ for some predetermined $\alpha > 0$ whenever $K \supseteq S_0$.[4]
3. An upper bound $\bar{s}$ for assessing sensitivity to sparsity assumptions.

Algorithm 1 below defines TU. It takes an initially selected model $\widehat{S}^0$, and then searches over certain deviations that include $\widehat{S}^0$ and add no more than $\bar{s}$ extra variables. To choose how to add variables, we do not look at model fit but rather which deviation leads to the largest change in inferential statements about the parameter of interest. We do this separately for the upper and lower bound of the interval.

Algorithm 1. Targeted Undersmoothing: TU($\bar{s}$) Intervals.

Step 1. Select a model $\widehat{S}^0$ with $\widehat{s} = |\widehat{S}^0|$ by a fixed model selection procedure.

Step 2. Undersmoothing procedure.

Initialize: $\widehat{K}^{\text{low}}, \widehat{K}^{\text{up}} = \widehat{S}^0$

While $|\widehat{K}^{\text{low}}|, |\widehat{K}^{\text{up}}| \leqslant \bar{s} + \widehat{s}$

Set $\widehat{K}^{\text{low}} = \widehat{K}^{\text{low}} \cup \{\widehat{j}^{\text{low}}\}$ *with*

$\widehat{j}^{\text{low}} = \arg\min_{j \leqslant p} \ell_{\widehat{K}^{\text{low}} \cup \{j\}}$

Set $\widehat{K}^{\text{up}} = \widehat{K}^{\text{up}} \cup \{\widehat{j}^{\text{up}}\}$ *with*

$\widehat{j}^{\text{up}} = \arg\max_{j \leqslant p} u_{\widehat{K}^{\text{up}} \cup \{j\}}$

Set $\widehat{S}^{\text{low}} = \widehat{K}^{\text{low}}$ *and* $\widehat{S}^{\text{up}} = \widehat{K}^{\text{up}}$

Step 3. *Set* $[\ell, u] = [\ell_{\widehat{S}^{\text{low}}}, u_{\widehat{S}^{\text{lup}}}]$.

[4] Formally, $\sup_{K \supseteq S_0, |K| < r} |\mathrm{P}([\ell_K, u_K] \ni \vartheta_0) - (1 - \alpha)| < t$ for some tolerance t and range r.

An appealing feature of Algorithm 1 is that it produces a path for how upper and lower bounds change due to incrementally adding one variable at a time to a model as a byproduct. I.e. at each step from 1 to $\bar{s}$, $\min_{j\leqslant p} \ell_{\widehat{K}^{\text{low}}\cup\{j\}}$ and $\max_{j\leqslant p} u_{\widehat{K}^{\text{up}}\cup\{j\}}$ provide the smallest lower bound and largest upper bound on a confidence set for the parameter of interest that can be obtained by adding one variable to the model used in the previous step.

This path is analogous to the common empirical practice of presenting results for target parameters from a baseline model and then considering sensitivity of this target parameter by seeing how estimates change after successively adding regressors one at a time starting from the baseline model. The key difference is that relative to this common practice, variables in Algorithm 1 are added in a data-driven way where each additional variable is chosen adversarially to lead to the largest change in the confidence set relative to the previous model. As illustrated in the empirical examples, this path is easy to display and provides an assessment of the sensitivity of conclusions within a well-defined and easily understood class of perturbations from a baseline model.

By starting Algorithm 1 with a model selected through a high-quality model selection procedure, we guarantee that the baseline model does a good job fitting the data at hand.[5] Setting $\bar{s} = 0$ then corresponds to this procedure producing no model selection mistakes which can happen in scenarios where oracle model selection is possible; see, for example, Fan and Li (2001), Zou (2006), and Bunea et al. (2007) for sufficient conditions in the high-dimensional linear model setting. As one then considers increasing $\bar{s}$, one is considering scenarios where the initial selector is allowed to have made increasingly many selection mistakes.

[5]Note, TU starting from an intuitively selected baseline can be viewed as traditional sensitivity analysis without an initial model selection stage. A high-quality method for estimating $\widehat{S}^0$ can potentially lead to improved performance relative to an intuitive baseline that fails to capture important predictive features in the data.

Given the dependence of the proposed procedure on the choice of $\bar{s}$, we feel that the proposed approach will be most helpful when viewed through the lens of sensitivity analysis. Specifically, one may look at how confidence regions for objects of interest change as one varies $\bar{s}$ over sensible values, for example, $\bar{s} \in \{0, 1, ..., \bar{s}^*\}$. When feasible, one could also consider increasing $\bar{s}$ one unit at a time until confidence regions enlarge to the point that economic conclusions differ substantively from the initial model. By looking at several values for $\bar{s}$, one gains insight into how sensitive conclusions are to the number of model selection mistakes made by the initial selector. This approach is similar to applications of sensitivity analysis in treatment effects estimation where a variety of approaches to sensitivity analysis exist for gauging sensitivity of causal estimators to violations of underlying identifying assumptions; see, for example, Rosenbaum (2002) and Manski (2003) for textbook reviews of classic approaches and Small (2007), Conley et al. (2012), Rosenbaum (2015), Andrews et al. (2017), and Oster (2017) for some recent examples in the statistics and econometrics literature.

While we focused the discussion on sensitivity intervals, TU can be used to target hypothesis testing as well. Suppose the null hypothesis of interest is $H_0 : \vartheta_0 = \bar{\vartheta}$ for a prespecified $\bar{\vartheta}$ and that, given a model $S \subseteq \{1, ..., p\}$, that $\widehat{W}_S$ is an observable test statistic with associated p-value $\widehat{p}_S$. Then TU can be used by choosing $\widehat{K} \supseteq \widehat{S}^0$ and taking the set $|\widehat{K}| \leqslant \bar{s} + \widehat{s}$ which makes the test most conservative (equivalently the set of variables that maximizes $\widehat{p}_{\widehat{S}}$).

In an online supplement, we provide a set of high-level conditions under which a variant of Algorithm 1 that replaces forward stepwise selection in Step *2* with full best subsets provides uniformly valid inference for the target parameter in the scenario where one has a known upper bound on s_0. In this respect, one may use the idea underlying TU intervals to obtain uniformly valid confidence intervals in cases

where the researcher believes the required conditions, including a known upper bound on s_0, are satisfied.[6]

3. Application I: Heterogeneous Treatment Effects from JPTA

The impact of job training programs on the earnings of trainees, especially those with low income, is of interest to both policy makers and academic economists. Evaluating heterogeneous causal effect of training programs on earnings is difficult due to the fact that individual characteristics vary across the sample; it is unlikely that many individuals share exactly the same values of observed covariates. The problem is made worse the higher the dimension of the collected covariates.

We consider data from a randomized training experiment conducted under the Job Training Partnership Act (JTPA). In the experiment, people were randomly assigned the offer of JTPA training services. Given the random assignment of the offer of treatment, we focus on estimating the average treatment effect of the offer of treatment, or the intention to treat effect, conditional on individual characteristics.

To capture the effects of training on earnings, we estimate a model of the form

$$y_i = x_i'\xi_0 + (d_i \cdot x_i)'\gamma_0 + \varepsilon_i$$

where d_i indicates whether training was offered, the outcomes y_i are earnings, x_i is a vector of covariates which includes a constant, ε_i is an unobservable, and $(\xi_0, \gamma_0) \equiv \beta_0$ are parameters. In this example, we limit the analysis to the sample of adult males. Earnings are measured as total earnings over the 30 month period following the assignment into the treatment or control group, and average earnings in the sample are \$19,147. Observed control variables are dummies for black and Hispanic persons, a dummy indicating high-school graduates and GED holders, five age-group dummies, a

[6]The online supplement also describes asymptotic properties of functionals $\widehat{\vartheta}$ estimated by plugging in $\widehat{\beta}$, an estimated high-dimensional parameter. This may be helpful as TU intervals are built around $\widehat{\vartheta}$.

marital status dummy, a dummy indicating whether the applicant worked 12 or more weeks in the 12 months prior to the assignment, a dummy signifying that earnings data are from a second follow-up survey, and dummies for the recommended service strategy. See Abadie et al. (2002) for detailed information regarding data collection procedures, sample selection criteria, and institutional details of the JTPA along with additional facts and discussion about the JTPA training experiment. The dataset has 5102 observations.

In this example, we are interested in estimating TU intervals for a fixed, individual-specific treatment effect. We form estimates by first calculating a Lasso-based model selection and Post-Lasso estimator of the coefficients

$$\widehat{S}^0,\ (\widehat{\xi}_{PL}, \widehat{\gamma}_{PL})$$

using the procedure detailed in Appendix A. Then, given an individual with covariates x^*, we calculate a corresponding functional of interest, the individual-specific intent to treat effect,

$$\vartheta_0 = \vartheta_0(x^*) = x^{*\prime}\widehat{\gamma}_{PL}.$$

There are many ways to construct regressors from the set of dummy variables available. As we are working in the sparse linear model framework, it is important that we believe that sparsity in the set of regressors is a reasonable approximation to the underlying process. With this consideration in mind, we construct the set of regressors over which model selection will occur by combining two different basis expansions of the raw control variables. Using different bases provides additional robustness by increasing the plausibility of a sparse representation within the union of the set of considered regressors.

To obtain the first representation for x_i, we consider all products of the discrete variables available. That is, we adopt the common convention of including the dummy

variables themselves, all first order interactions between the main dummy variables, all second order interactions, and all further higher order interactions. Excluding empty and small cells, the dimension of this representation of the covariate space is 313.[7]

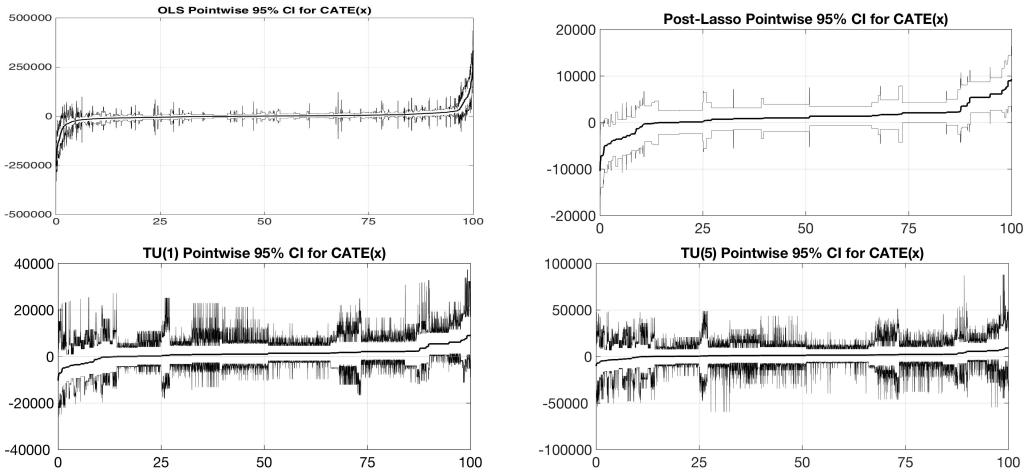
For the second representation, we consider the Hadamard-Walsh basis defined as follows. Let $v_{i1}, ..., v_{ik}$ denote the original set of indicator variables. Let each subset $A \subseteq \{1, ..., k\}$ index a transformation of $(v_{i1}, ..., v_{ik})$ given by $\psi_A(v_{i1}, ..., v_{ik}) = (-1)^{|A \cap \{j: v_{ij}=1\}|}$. The terms $\psi_A(v_{i1}, ..., v_{ik})$ then define the regressors used in the second representation of the raw input variables. The Hadamard-Walsh expansion is both a set-Fourier basis as well as a generalization of the logical function xor- "exclusive or." In particular, if the cardinality of A is 2 so that $A = \{k_1, k_2\}$ then $\psi_{\{k_1,k_2\}}(v_{i1}, ..., v_{ik}) = \mathsf{xor}(v_{ik_1}, v_{ik_2})$.[8]

After appending the two representations, the result is that $\dim(x_i) = 2927$, including the constant term. Interacting x_i with the d_i, the total dimensionality of the model parameters is 5854, which exceeds $n = 5102$. The assumption of sparsity over β_0, which implies sparsity over both ξ_0 and γ_0, is thus important in this example. No researcher would include the 2927 main effects, and very few would believe that including only the baseline 313 main effects would be sensible in this setting. Rather, we are relying on a model selector to pick out the main effects that are important for predicting the outcome in the control state and thus likely useful for reducing variance and for understanding baseline outcomes and to pick out important variables that help us understand treatment effect heterogeneity. As such selectors will likely be imperfect, we then wish to gauge whether conclusions about objects of interest are

[7] Specifically, we start by eliminating all variables with $\leqslant 5$ nonzero entries in *either* the control or treated subsample. After these deletions, we then remove any variables if the corresponding diagonal R term in QR decomposition of the design matrix was $< 10^{-6}$ over *either* the control or treated subsample.

[8] We choose to only include ψ_A terms as potential covariates for $1 < |A| < 6$. Note that for $|A| = 1$, the resulting transformations are perfectly correlated to the original indicator variables and are excluded.

FIG 1. *JTPA CATE Estimates: Hadamard-Walsh Specification*



Note: These figures report estimates of the treatment effect for each individual in the JTPA sample along with pointwise 95% confidence intervals where the set of controls is constructed by taking all possible interactions of the baseline dummy variables and augmenting with the Hadamard-Walsh basis as described in the main text. Individuals are sorted on the horizontal axis according to the percentile rank of their estimated effect $\widehat{\vartheta}_i = x_i'\widehat{\gamma}_{PL}$. Estimates based on OLS and Post-Lasso are reported in the top panel. The bottom panel presents results based on TU with $\bar{s} = 1$ ("TU(1)") and with $\bar{s} = 5$ ("TU(5)") respectively. It is important to note that vertical axis is different in each figure.

sensitive to the possibility that a small number of variables - either main effects or interactions - have been excluded from the model.

Evaluating $\vartheta_0 = \vartheta_0(x^*)$ requires the specification of a fixed covariate x^*. As noted above, $\vartheta_0(x^*)$ is intent to treat for an individual with given observable characteristics x^*. For simplicity and illustration, we set $x^* = x_i$ for each individual i in the dataset. The values x_i are treated as fixed (nonrandom). We thus calculate and perform TU for n functionals in the set $\{\vartheta_0(x^*) = \vartheta_0(x_i)\}_{i=1}^n$ treating the functional for each x_i as the object of interest in turn.

Figure 1 presents pointwise intervals for the individual specific effects $\vartheta_{0,i}$ for all individuals $i = 1, ..., n$ under four different inferential methods using the variables defined above.[9] Therefore, in each panel, an inferential procedure is carried out a total of n times for n different functionals indexed by i. The first panel presents OLS-based

[9]We provide results using only the first set of variables formed from the usual construction of interactions in a supplementary appendix.

confidence intervals, using only the standard multiplicative interaction expansion. The second panel presents oracle-style confidence intervals, which ignore first stage model selection. Interestingly, the initial model selection procedure picks terms from both the interaction expansion and the Hadamard-Walsh expansion. The third panel presents TU(1) estimates using $\bar{s} = 1$, and the fourth panel presents TU(5) estimates using $\bar{s} = 5$. The TU intervals are calculated according Algorithm 1. In each panel, the horizontal axis indexes individuals, sorted according to the percentile rank of their estimated effect $\widehat{\vartheta}_i = x_i'\widehat{\gamma}_{PL}$. The vertical axis represents estimates and intervals for $\widehat{\vartheta}_i$.

Using $\bar{s} = 1$ we see that the interval lengths increase relative to the oracle-style intervals. There still remains a set of individuals for whom the corresponding TU interval excludes zero. With $\bar{s} = 5$, for all individuals, the corresponding intervals contain zero. Though not pictured in Figure 1, we note that all intervals for individual-specific treatment effects include 0 as soon as $\bar{s} = 2$. Note that certain OLS-based intervals have very wide range relative to all TU(5) intervals; some individuals have upper OLS-based interval bounds exceeding \$250,000 and lower OLS-based interval bounds exceeding \$-250,000.

We report results for testing the null hypothesis of no treatment heterogeneity, $H_0 : \gamma_0 = 0$, in Table 1. The procedure is implemented using the standard Wald test. The null hypothesis is rejected for $\bar{s} \leqslant 1$ at the 5% level but not to rejected for $\bar{s} \geqslant 2$.

Taken together, the results in this section suggest there is mild evidence for treatment effect heterogeneity in this example. We would reject the hypothesis of no heterogeneity and also obtain some evidence for individual specific treatment effects that differ from zero when using oracle model selection results. However, we cannot rule out the possibility of no treatment effect heterogeneity after allowing for a modest number of model selection mistakes. Thus, to draw strong conclusions about treat-

TABLE 1
Testing the Null Hypothesis of No Treatment Effect Heterogeneity: Hadamard-Walsh Specification

Estimator	W-statistic	df	p-value
PL	20.6884	9	0.0141
TU(1)	19.4059	10	0.0354
TU(2)	18.1018	10	0.0533
TU(3)	17.5105	10	0.0638
TU(4)	16.8746	10	0.0772
TU(5)	16.3060	10	0.0912
TU(6)	15.7466	10	0.1071
TU(7)	15.2801	10	0.1222
TU(8)	14.8188	10	0.1388
TU(9)	14.3024	10	0.1596
TU(10)	13.9031	10	0.1775

Note: This table presents results for testing the null hypothesis of no treatment effect heterogeneity when the set of controls is constructed by taking all possible interactions of the baseline dummy variables and augmenting with the Hadamard-Walsh basis as described in the main text. We report the value of the Wald statistic ("W-statistic"), degrees of freedom ("df"), and associated p-value ("p-value"). Results for testing this hypothesis based on OLS and Post-Lasso estimates are provided in the first two rows of the table. Rows labeled "TU(j)" correspond to TU with $\bar{s} = j$.

ment effect heterogeneity, one must believe that the initial model selection procedure is very close to perfect in this example.

4. Application II : Heterogeneous Treatment Effects in Direct Mail

The targeting of individuals with appropriate interventions that induce preferred outcomes is a relevant problem in various application areas including business, political science and economics. In the field of marketing, such targeting has been the key instrument of retailers that use direct mail as the focal intervention to inform and persuade their customers to purchase from their catalogs. These catalogs are often relatively expensive to produce and firms spend significant amounts in this endeavor.

Our data for this example comes from a large multi-product retailer that sells directly to consumers online but also via mail, phone and retail channels. The firm's budget for direct-mailed catalogs is over $120M and net sales per year are in excess of $1.5B. The firm routinely runs experiments to evaluate the effectiveness of its catalog mailing strategy. Typically, these experiments have two conditions (mail, no-mail) that are randomized across customers. Our data focuses on one such experiment that involved over 290,000 customers. The data also include a list of 486 descriptors of

the individual customers. These descriptors include demographic characteristics (age, income, gender, state), details of past promotional activity they may have received as well as their past consumption behavior data including purchases, the timing of such purchases, the number of orders in the past year, and the extent of their expenditures with the firm. The design matrix in our analysis contains 2139 columns once categorical variables are expanded.

In our analysis, we estimate the following simple specification of a model with heterogeneous treatment effects:

$$y_i = f_0(d_i, x_i, \varepsilon_i) = x_i'\xi_0 + (d_i \cdot x_i)'\gamma_0 + \varepsilon_i.$$

In the above, d_i is an indicator that a consumer has been randomly assigned to receive a direct mail marketing instrument (a catalog), and the x_i are customer characteristics. y_i are dollar expenditures by the customer over a 3-month horizon following the mailing of the marketing instrument. We assume that $(x_i, \varepsilon_i)_{i=1}^n$ are n i.i.d. draws, having the same distribution as the generic pair of random variables (x, ε).

We assume that the firm is interested in evaluating a marketing strategy formed from targeting individuals based on their individual-specific treatment effects versus one of two simple baseline strategies - either mailing to no one or mailing to everyone. A mailing strategy $\tilde{d} = \tilde{d}(x)$ assigns customers with characteristics x to either receive the mailing or not. We adopt TU to provide a simple mechanism for the firm to statistically evaluate the difference between two competing mailing strategies on the basis of average expected profits. The average expected profit from implementing $\tilde{d}$ is

$$\mathrm{E}[\pi(\tilde{d})] = \mathrm{E}\left[\nu f_0\left(\tilde{d}(x), x, \varepsilon\right) - \tilde{d}(x)c\right].$$

In the above quantity, the firm has a known margin $(0 < \nu < 1)$ that applies to sales generated by its customers. For simplicity, we assume that the cost to the firm

of targeting each consumer, c, is constant and known *ex ante*.[10] Within the model, the only remaining source of uncertainty are the unanticipated demand shocks ε, which are only observed via outcomes and are assumed to be conditionally mean zero.

We examine two extremal mailing strategies where either no customers receive a catalog ('no-mailings') by setting $\tilde{d}(x) = 0$ uniformly or a 'blanket-mailing' strategy wherein all customers receive a catalog (i.e. $\tilde{d}(x) = 1$ for all x). The no-mailings strategy expected profits are

$$\mathrm{E}[\pi^0] = \mathrm{E}[\nu f_0(0, x, \varepsilon)] = \nu \mathrm{E}[x'\xi_0].$$

Similarly, the expected profit for the blanket mailing strategy can be written as

$$\mathrm{E}[\pi^1] = \mathrm{E}[\nu f_0(1, x, \varepsilon) - c] = \nu \mathrm{E}[x'(\xi_0 + \gamma_0)] - c.$$

A sophisticated firm might be interested in optimizing the mailing strategy based on expected consumer response.[11] One simple, sensible mailing strategy would be to mail to a consumer with characteristics x whenever the expected increment in profits for that customer exceeds costs. The rule can be described by

$$d^*(x) = \mathbf{1}\{\nu(x'\xi_0 + x'\gamma_0) - \nu(x'\xi_0) > c\} = \mathbf{1}\{\nu(x'\gamma_0) > c\}.$$

Using this strategy, we then have expected per consumer profit of

$$\begin{aligned}\mathrm{E}[\pi^*] &= \mathrm{E}[\nu f_0(d^*(x), x, \varepsilon) - c d^*(x)] \\ &= \nu \mathrm{E}[x\xi_0] + \nu \mathrm{E}[(d^*(x) \cdot x)'\gamma_0] - c\Pr(d^*(x) = 1).\end{aligned}$$

[10] A more general approach would be to write costs as functions of x. Implementing this approach would require specific data about individual mailing costs which we currently do not have. We could also assume that costs are drawn from some known distribution where the exact realization is unknown by the firm until after the mailings have been sent out and calculate expected profits integrating over this cost distribution.

[11] See Athey and Wager (2017) and van der Laan and Luedtke (2016) for approaches to estimating and performing inference for optimal treatment strategies.

Next, compare the targeted strategy to the 'blanket' or 'no-mailing' strategies. The difference in profit between the targeted and no-mailing strategies is

$$\begin{aligned}\mathrm{E}[\Delta\pi^{*0}] &= \mathrm{E}\left[\pi^{*}\right] - \mathrm{E}\left[\pi^{0}\right] \\ &= \nu \mathrm{E}[(d^{*}(x)\cdot x)'\gamma_0] - c\Pr(d^{*}(x)=1).\end{aligned}$$

Similary, the difference between the targeted and blanket strategies would be

$$\begin{aligned}\mathrm{E}[\Delta\pi^{*1}] &= \mathrm{E}\left[\pi^{*}\right] - \mathrm{E}\left[\pi^{1}\right] \\ &= \nu \mathrm{E}[(d^{*}(x)-1)\cdot x'\gamma_0] - c(\Pr(d^{*}(x)=1)-1).\end{aligned}$$

Natural estimators exist for $\mathrm{E}[\Delta\pi^{*0}]$ and $\mathrm{E}[\Delta\pi^{*1}]$. An estimator for $\mathrm{E}[\Delta\pi^{*0}]$ is

$$\widehat{\Delta\pi^{*0}} = \frac{\nu}{n}\sum_{i=1}^{n}\left[\mathbf{1}\{\nu\left(x_i'\widehat{\gamma}_0\right) > c\}\left(x_i'\widehat{\gamma}_0 - c/\nu\right)\right]$$

for some estimator $\widehat{\gamma}_0$. Similarly, a natural estimator of $\mathrm{E}[\Delta\pi^{*1}]$ is

$$\widehat{\Delta\pi^{*1}} = \frac{\nu}{n}\sum_{i=1}^{n}\left[\left(\mathbf{1}\{\nu\left(x_i'\widehat{\gamma}_0\right) > c\} - 1\right)\left(x_i'\widehat{\gamma}_0 - c/\nu\right)\right]$$

for an estimator $\widehat{\gamma}_0$. Under the sparsity assumptions on the true model maintained in this paper and conventional regularity conditions, $\widehat{\Delta\pi^{*0}}$ and $\widehat{\Delta\pi^{*1}}$ will by asymptotically normal with standard error that can be readily estimated when γ_0 is estimated from the true model. Based on this observation, we can apply the TU approach to conduct inference on potential profit improvements from targeting based on the rule $d^{*}(x)$ relative to the two simple baseline strategies.

In this example, we form estimates by first calculating a Lasso-based model selection and Post-Lasso estimator of the coefficients $\widehat{S}^0$, $(\widehat{\xi}_{PL}, \widehat{\gamma}_{PL})$, using the procedure described in detail in Appendix A. TU intervals for $\mathrm{E}[\Delta\pi^{*0}]$ and $\mathrm{E}[\Delta\pi^{*1}]$ are printed in Table 2 for $\bar{s} \leqslant 10$.[12] The margin parameter is set to $\nu = 0.30$ and the cost pa-

[12] Before estimation, variables with a small number of nonzero observations are excluded. In the first pass, variables with $\leqslant 100$ nonzero entries in the *entire* sample were eliminated. In the second pass, variables with corresponding diagonal R term in the design matrix QR decomposition $< 10^{-6}$ in *either* control or treated subsample were eliminated.

rameter is set at $c = 0.70$ based on input from the firm. Also reported are OLS-based estimates. Intervals are based on heteroskedasticity-consistent standard errors.

TABLE 2
Estimates for Average Profit Differential

	Relative to no Mailing: $E[\Delta\pi^{*0}]$				Relative to Blanket Mailing: $E[\Delta\pi^{*1}]$			
Estimator	Estimate	S.E.	Lower	Upper	Estimate	S.E.	Lower	Upper
OLS	1.1514	0.0655	1.0229	1.2798	0.6332	0.0789	0.4785	0.7879
PL	0.6984	0.0441	0.6119	0.7849	0.1811	0.0497	0.0837	0.2784
TU(1)			0.6099	0.7960			0.0821	0.2905
TU(2)			0.6083	0.8063			0.0807	0.3001
TU(3)			0.6070	0.8131			0.0798	0.3076
TU(4)			0.6062	0.8188			0.0788	0.3132
TU(5)			0.6054	0.8269			0.0779	0.3205
TU(6)			0.6045	0.8323			0.0773	0.3261
TU(7)			0.6036	0.8375			0.0767	0.3309
TU(8)			0.6029	0.8430			0.0762	0.3361
TU(9)			0.6023	0.8476			0.0758	0.3401
TU(10)			0.6018	0.8514			0.0754	0.3437

Note: This table presents estimates of the average profit differential between the targeted mailing strategy and the strategy that mails to no one, $E[\Delta\pi^{*0}]$, and estimates of the average profit differential between the targeted mailing strategy and the strategy that mails to everyone, $E[\Delta\pi^{*1}]$. OLS and Post-Lasso estimates of the average profit differential and associated standard errors are provided in the "Estimate" and "S.E." columns in the first two rows. The "Lower" and "Upper" columns respectively report the lower and upper bounds of 95% confidence intervals. Rows labeled "TU(j)" correspond to TU with $\bar{s} = j$.

The TU intervals for $E[\Delta\pi^{*0}]$ and $E[\Delta\pi^{*1}]$ are robust to different assumptions about the true underlying sparsity level $\bar{s}$. Interestingly, the OLS-based intervals are completely different from the TU intervals for every value of $\bar{s}$. This difference is likely due to a failure of OLS in this example. In the setting of the simulation study below, we find that OLS intervals achieve poor coverage probabilities with coverages as low as 0.00% in some settings. The poor performance of OLS in the simulation study is due to biases arising from taking a nonlinear transformation of the estimated coefficient vector and a failure of the standard delta method with a large number of covariates.[13] The OLS estimates seem to overstate both $E[\Delta\pi^{*0}]$ and $E[\Delta\pi^{*1}]$.

[13] Bias corrections for the delta method in settings with many covariates are described in Cattaneo et al. (2019). For simplicity, we report the estimates and intervals which correspond to common practice.

5. Simulation Study

In this section, we present a simulation study to demonstrate properties of TU in finite samples. The simulation is designed based on the example in Section 4. We generate data for each simulation replication as iid draws for $i = 1, ..., n$ from

$$y_i = \xi_0^0 + x_i'\xi_0 + d_i\gamma_0^0 + d_i \cdot x_i'\gamma_0 + \varepsilon_i,$$

$$p = 2 + 2\dim(x_i) = 2(1+k), \quad w_{ij} \sim N(0,1) \text{ with } \text{corr}(w_{ij_1}, w_{ij_2}) = 0.8^{|j_1 - j_2|},$$

$$x_{ij} = (w_{ij} - \tau_j)\mathbf{1}\{w_{ij} \geqslant \tau_j\},\ \tau_j \sim \text{unif}(0, 1.28)\ iid,\ d_i \sim \text{Bernoulli}(0.5),\ \varepsilon_i \sim N(0,1),$$

$$(\xi_0^0, \xi_0') = c_{.25}(1/\sqrt{s_0}, (2/\sqrt{s_0})\iota_{s_0/4}', (2/\sqrt{ns_0})\iota_{s_0/4}', 0_{k-s_0/2}') \odot (1, \upsilon'),$$

$$(\gamma_0^0, \gamma_0') = c_{.25}(1/(2\sqrt{s_0}), (4/\sqrt{ns_0})\iota_{s_0/4}', (4/\sqrt{s_0})\iota_{s/4}', 0_{k-s_0/2}') \odot (1, \upsilon'),$$

where $c_{.25}$ is a constant that is chosen so that the population R^2 of the regression of y_i onto $(1, x_i', d_i, d_i x_i')$ is 0.25.[14]

Two simulation designs are based on varying $p \in \{202, 602\}$ and setting $s_0 = 8$.[15] In all simulations, $n = 400$. Estimates and intervals are constructed for three functionals: (1) the value of a single coefficient (specifically $[\gamma_0]_1$, the first component of γ_0), (2) an individual treatment effect for a fixed hypothetical subject (with $x^* = .5\iota_{\dim(x_i)}$), and (3) the average per-person profit differential from a targeting rule based on estimated individual specific treatment effects and a rule which treats no one ($\text{E}[\Delta\pi^{*0}]$ defined in Section 4).

We simulate 500 replications and present properties of several estimators:

1. **True.** Infeasible estimator based on OLS on the correct support.
2. **All.** Estimator based on OLS using all covariates.
3. **Double.** The post-double selection estimator as described in Belloni et al. (2014)

[14] ι_m is an $m \times 1$ vector of ones, 0_m is an $m \times 1$ vector of zeros, υ is a $k \times 1$ vector with j^{th} element given by $\upsilon_j = (-1)^{j-1}$, and $\odot$ denotes the Hadamard product.

[15] A supplementary appendix provides additional results with $s_0 \in \{4, 16\}$.

4. **Lasso.** Estimator based on Lasso.
   – Standard errors computed using Lasso residuals.
5. **PL.** Estimator based on the Post-Lasso of Belloni and Chernozhukov (2013).
   – Standard errors computed using Post-Lasso residuals.
6. **LCV.** Estimator based on Lasso with penalty chosen by 10-fold cross validation.
   – Standard errors computed using Lasso residuals.
7. **ZB.** Confidence intervals based on inverting the hypothesis test proposed in Zhu and Bradic (2016).
8. **TU(1).** Targeted undersmoothing with $\overline{s} = 1$ using Algorithm 1.
   – Initial model $\widehat{S}^0$ selected using method described in Appendix A.
9. **TU(10).** Targeted undersmoothing with $\overline{s} = 10$ using Algorithm 1.
   – Initial model $\widehat{S}^0$ selected using method described in Appendix A.

Further details about the estimators considered are provided in Appendix A.

Performance measures of the nine procedures, including estimates of bias, standard deviation, root mean-square error (RMSE), coverage probability for a 95% interval, and corresponding interval length are printed in Tables 3-4 and Figures 1-2. The figures provide average interval lengths and coverage probabilities along the 10-steps of the TU forward selection path produced in the simulation.

When $p < n$, a simple feasible option is to estimate the full-model without any model selection ('All'). In our simulation which is based on a linear model with conditional mean zero errors, this approach clearly results in small bias for the individual regression parameter and for the individual-specific treatment effect. The cost of estimating the full model is decreased estimation precision as evidenced by relatively large standard deviation, RMSE, and interval lengths relative to the other point estimators. Importantly, for the estimation of profit differential functional, the estimator is dominated by bias due to this functional's nonlinear dependence on the model parameters. This bias then results in extremely poor coverage properties for the true profit differential. This behavior can be viewed as a failure of the delta-method in

TABLE 3. Simulation Results: $n = 400$, $p = 202$, $s_0 = 8$

	True	All	Double	Lasso	PL	LCV	ZB	TU(1)	TU(10)
					A. RegCoef				
Bias	0.04	0.01	0.84	-0.09	-0.08	-0.14			
Std. Dev.	0.63	0.74	0.67	0.02	0.18	0.55			
RMSE	0.63	0.74	1.07	0.09	0.19	0.57			
Coverage	0.94	0.92	0.67	0.02	0.01	0.64	0.79	0.99	1.00
Int. Length	2.25	2.61	2.33	0.04	0.03	1.51	1.22	2.12	4.25
					B. TE				
Bias	0.02	0.01		0.12	0.13	0.13			
Std. Dev.	0.21	1.57		0.12	0.27	0.45			
RMSE	0.21	1.57		0.17	0.30	0.47			
Coverage	0.94	0.92		0.87	0.76	0.97	0.91	0.99	1.00
Int. Length	0.78	5.79		0.56	0.58	1.88	19.57	2.16	6.72
					C. PI				
Bias	0.02	0.31		-0.09	-0.07	-0.02			
Std. Dev.	0.10	0.10		0.11	0.11	0.11			
RMSE	0.10	0.33		0.14	0.13	0.11			
Coverage	0.95	0.06		0.86	0.87	0.90		0.95	1.00
Int. Length	0.40	0.36		0.44	0.43	0.39		0.50	0.74

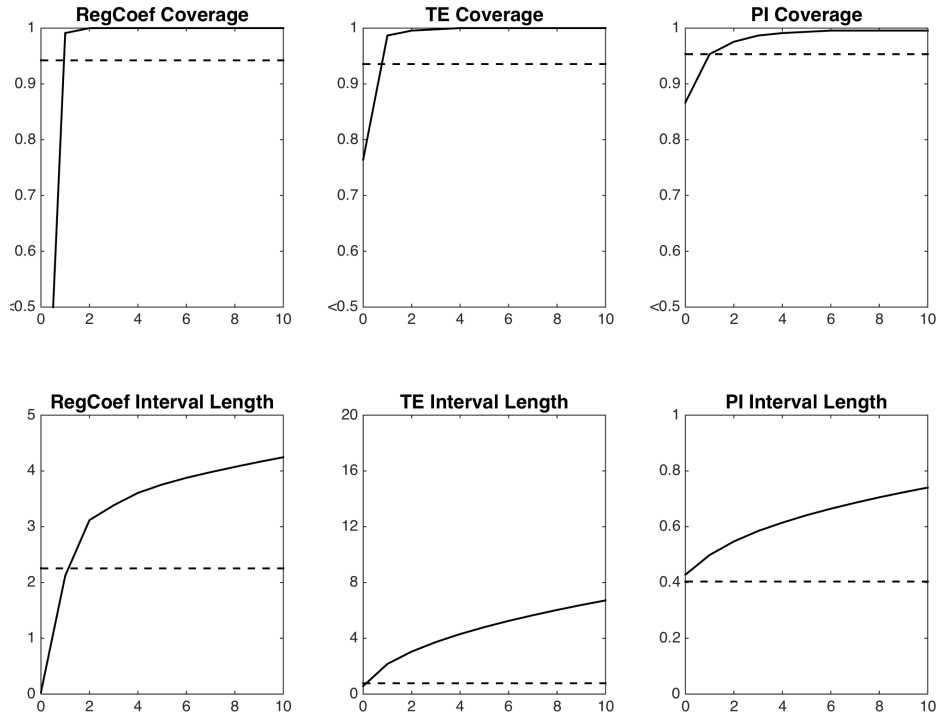
Note: This table presents point estimation and inferential results from a simulation designed to mimic the empirical example in Section 4 in a setting with $p < n$. Results are provided for estimation and inference on a pre-specified coefficient of interest ('RegCoef'), an individual specific treatment effect ('TE'), and the difference in expected profits between a targeted and non-targeted treatment strategy as defined in Section 4 ('PI'). Coverage and interval length report coverage and average length of 95% intervals.

moderate or high-dimensional models; see Cattaneo et al. (2019).

Next examine the performance of 'Lasso' and 'PL'. The Lasso penalty parameter in this case is set in a manner that theoretically provides Lasso with an optimal rate of convergence and guarantees that the $\widehat{s} = O(1)s_0$. We conduct inference in these cases by relying on oracle-type results that ignore the first step model selection. In general, the resulting estimators are competitive in terms of RMSE for all objects considered across all different designs. However, their bias also tends to be comparable to their standard deviation due to regularization and model selection mistakes. Oracle-style approximations do not account explicitly for this bias and as a result do not achieve correct coverage rates. Coverage for these procedures is generally far from the nominal 95% and is, in some cases, 0%.

The 'LCV' estimator is similar to 'Lasso' and 'PL' in that it applies oracle-style inference after selecting a model from the data. The difference is that cross-validation

FIGURE 1. Simulation Results: $n = 400$, $p = 202$, $s_0 = 8$



Note: This figure provides coverage and average length of 95% intervals for 10 steps in the TU path starting from the baseline Post-Lasso model in the $p < n$ simulation. Results are provided for estimation and inference on a pre-specified coefficient of interest ('RegCoef'), an individual specific treatment effect ('TE'), and the difference in expected profits between a targeted and non-targeted treatment strategy as defined in Section 4 ('PI').

tends to produce penalty parameters that are much smaller than the theoretically motivated values used in 'Lasso' and 'PL'. This reduction in the penalty parameter allows extra variables to enter the model relative to the case where the larger penalty parameters are used. In this sense, such a procedure can also be thought of as an undersmoothing procedure, though the "undersmoothing" is targeted toward model fit. In these simulations, 'LCV' tends to produce estimates of the regression coefficient and individual-specific treatment effect with bias similar to that obtained with 'Lasso' and 'PL', though 'LCV' also tends to have a larger standard deviation than these estimators as well. The similar bias and larger standard deviation results in 'LCV' tending to be outperformed in terms of RMSE for these objects but also results

TABLE 4. Simulation Results: $n = 400$, $p = 602$, $s_0 = 8$

	True	All	Double	Lasso	PL	LCV	ZB	TU(1)	TU(10)
					A. RegCoef				
Bias	-0.03		0.78	-0.09	-0.08	-0.15			
Std. Dev.	0.65		0.68	0.01	0.12	0.43			
RMSE	0.65		1.03	0.09	0.15	0.45			
Coverage	0.93		0.72	0.02	0.01	0.55	0.77	1.00	1.00
Int. Length	2.25		2.34	0.02	0.03	1.13	1.26	2.04	4.68
					B. TE				
Bias	-0.01			0.11	0.13	0.15			
Std. Dev.	0.22			0.12	0.25	0.44			
RMSE	0.22			0.17	0.28	0.46			
Coverage	0.93			0.87	0.74	0.99	0.98	1.00	1.00
Int. Length	0.77			0.56	0.58	2.27	24.82	2.66	9.98
					C. PI				
Bias	0.01			-0.09	-0.08	-0.04			
Std. Dev.	0.10			0.11	0.11	0.11			
RMSE	0.10			0.15	0.13	0.12			
Coverage	0.95			0.85	0.88	0.89		0.95	1.00
Int. Length	0.40			0.44	0.43	0.40		0.51	0.85

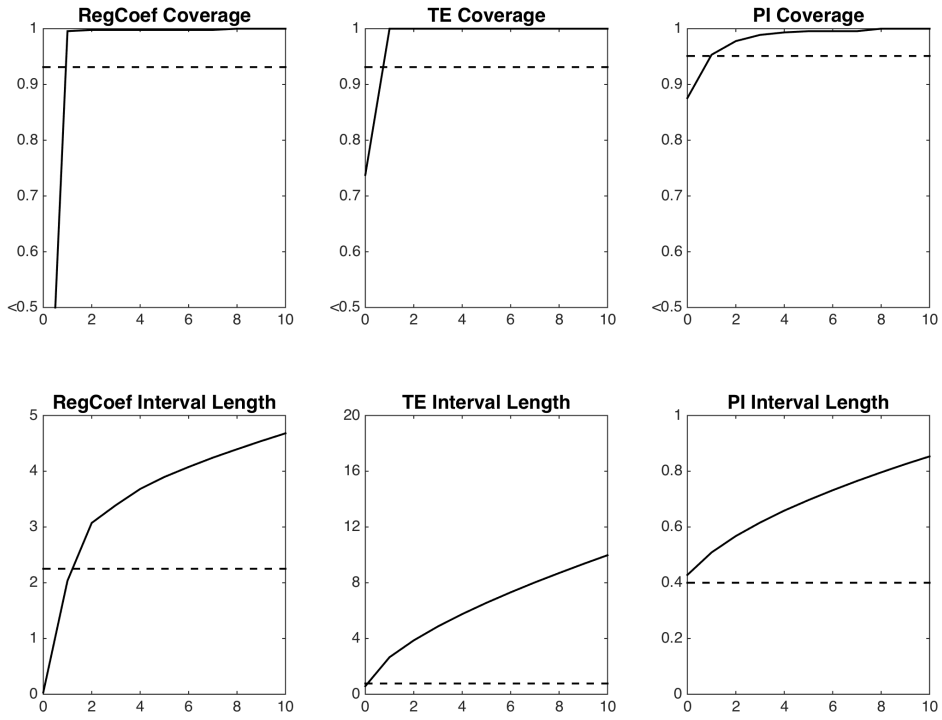
Note: This table presents point estimation and inferential results from a simulation designed to mimic the empirical example in Section 4 in a setting with $p > n$. Results are provided for estimation and inference on a pre-specified coefficient of interest ('RegCoef'), an individual specific treatment effect ('TE'), and the difference in expected profits between a targeted and non-targeted treatment strategy as defined in Section 4 ('PI'). Coverage and interval length report coverage and average length of 95% confidence intervals. The column 'All' is included and blank to emphasize that using all regressors is infeasible in this scenario.

in better coverage of the 'LCV' intervals relative to the 'Lasso' or 'PL' intervals. For the profit differential, 'LCV' is generally less-biased than 'Lasso' and 'PL' while having a similar standard deviation. Thus, 'LCV' is competitive in terms of RMSE for this object. However, sufficient bias remains for coverage of intervals to remain substantively distorted.

Next, we turn to 'Double' which, in terms of our objects of interest, is only readily available for inference about a single regression coefficient. We see that the 'Double' point estimator has a large bias which translates into relatively poor coverage properties in our simulation.[16] We conjecture that this behavior may be improved by considering double machine learning as in Chernozhukov et al. (2016). We note that TU offers an approach to gauging the sensitivity of conclusions to model selection mistakes and could be applied directly to semiparametric targets using orthogonal

[16] In the appendix, we consider the $s_0 = 4$ case in which 'Double' delivers performance comparable to the infeasible oracle.

FIGURE 2. Simulation Results: $n = 400$, $p = 602$, $s_0 = 8$



Note: This figure provides coverage and average length of 95% intervals for 10 steps in the TU path starting from the baseline Post-Lasso model in the $p > n$ simulation. Results are provided for estimation and inference on a pre-specified coefficient of interest ('RegCoef'), an individual specific treatment effect ('TE'), and the difference in expected profits between a targeted and non-targeted treatment strategy as defined in Section 4 ('PI').

estimating equations as in Belloni et al. (2014) or Chernozhukov et al. (2016). We do not pursue this for brevity.

The 'ZB' method does not achieve 95% coverage for the regression coefficient $\zeta_{0,1}$ in any of our simulations. The 'ZB' method gives better coverage probabilities for the individual treatment effect with near or above 95% coverage in all simulation designs. However, the 'ZB' intervals for the individual specific effects are very long.[17]

Finally, we examine the TU approach. For TU, we set the initial model and point estimates to be those underlying the 'PL' results. An interesting feature of the presented simulations is that 'TU(1)' achieves nearly correct coverage uniformly across

[17]Looking at the expanded set of simulations in the Supplementary Appendix, we also see that the lengths of the 'ZB' confidence intervals grow considerably with the underlying value of s_0.

the simulation designs - achieving higher than 90% coverage in every design. While not reported in the table, 'TU(2)' achieves higher than 95% coverages in all cases. We see the inherent conservativeness in sensitivity analysis considering a large class of models in that 'TU(10)' uniformly has nearly 100% coverage in all cases. Importantly, the good coverage properties are uniform across all designs and all parameters considered. As must be the case, the intervals produced by the TU approach are relatively wide and become wider as one allows for more selection mistakes. The losses relative to the infeasible optimum are modest for small $\bar{s}$, and the intervals are still potentially informative even in the most extreme case we consider.

Overall, we believe these results are favorable to the TU approach. Of the considered feasible alternatives, it is the only procedure that produces uniformly good coverage properties. The cost is increased imprecision about what conclusions can be drawn from the data. This increase in imprecision seems honest as it reflects the potential for substantive biases resulting from model selection mistakes. The procedure is also anchored on initial point estimates that have relatively good properties for estimating the parameters of interest.

Appendix A: Implementation Details

This appendix additional implementation details for computations performed in the main text.[18]

A.1. Model Selection Implementation

In this paper, the procedure for selecting $\widehat{S}^0$ for models defined by

$$y_i = x_i'\xi_0 + (d_i \cdot x_i)'\gamma_0 + \varepsilon_i$$

[18] There are many choices about how to implement the different procedures, e.g. whether to split into treatment and control observations and which penalty parameters to use. The choices below were based on initial simulations where they seemed to produce the most favorable performance for the non-TU approaches.

as in Sections 3–5 is given in Algorithm A1 below.

Algorithm A1. Initial model selection in heterogeneous effects linear model.

Step 1. Divide the sample into two sets: $A_0 = \{i : d_i = 0\}$ and $A_1 = \{i : d_i = 1\}$.

Step 2. Within each sample, demean the observations.

Step 3. Using the demeaned observations, run the modified heteroskedastic Lasso regression (described in Algorithm A2 in Appendix A.3) of y_i on x_i over subset A_0 and let $\widehat{S}^{0,0}$ be the set of covariates selected. Again using the demeaned observations, run the modified heteroskedastic Lasso regression of y_i on x_i over subset A_1 and let $\widehat{S}^{0,1}$ be the set of covariates selected.

Step 4. The final model $\widehat{S}^0$ consists of the constant term, the main effect of d_i, the ξ_0 components corresponding to covariate indexes in $\widehat{S}^{0,0} \cup \widehat{S}^{0,1}$, and the interaction terms (γ_0 terms) corresponding to covariate indexes in $\widehat{S}^{0,0} \cup \widehat{S}^{0,1}$.

A.2. Additional Simulation Details

In the simulation study, all standard errors are computed using conventional heteroskedasticity consistent standard errors (e.g. White (1980)) using the estimated residuals indicated above. We give details on implementation specifics in the following paragraph.

For 'True,' 'All,' and 'Double,' we directly estimate the linear model defined in Section 5 using only the variables with non-zero coefficients ('True'), all the variables ('All'), or variables selected by Lasso ('Double'). For 'Double,' we apply Belloni et al. (2014) with a minor modification to select variables. Specifically, we implement the relevant Lasso regressions from Belloni et al. (2014) using the modified heteroskedastic Lasso outlined in Appendix A.3. To implement 'Lasso' and 'PL,' we use Algorithm A1 to select a model. 'PL' then obtains final estimates by re-estimating coefficients by OLS with only the variables selected by Lasso. For 'LCV,' we use a modification of the

procedure in Appendix A.1, where 10-fold cross-validation within each subset is used to choose the tuning parameter to use in that subset. We then apply the conventional Lasso within each subset based on these estimated tuning parameters. For these methods, we then can obtain estimates and standard errors for the functionals of interest in the obvious manner. We implement 'ZB' using the method of inference for dense linear functionals of a parameter vector from Zhu and Bradic (2016). Finally, the PL model serves as our initial model when applying TU. We apply TU for $\bar{s} = 1, ..., 10$.

A.3. Lasso Implementation

The implementation of Lasso in this paper is performed according to Algorithm A2.

Algorithm A2. Modified Heteroskedastic Lasso: Marginal Correlation-Based Initial Penalty Loadings. The modified heteroskedastic Lasso is identical to Belloni et al. (2012) with a small modification. Belloni et al. (2012) relies on 'initial penalty loadings,' which require preliminary estimates of individual specific residuals. To obtain these initial estimates of residuals, $e_i^{initial}$, we regress y_i on the 5 covariates with the highest marginal correlation with y_i and use the resulting residuals. This approach can be shown to be formally valid when the number of covariates with high marginal correlations to y_i used is bounded by a constant which does not depend on n. In contrast, note that Belloni et al. (2012) suggest $e_i^{initial} = y_i - \bar{y}$. Finally, the penalty loadings are updated with one iteration as described in Belloni et al. (2012).

References

A. Abadie, J. Angrist, and G. Imbens. Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica*, 70(1):91–117, 2002. ISSN 1468-0262. . URL http://dx.doi.org/10.1111/1468-0262.00270.

I. Andrews, M. Gentzkow, and J. M. Shapiro. Measuring the sensitivity of parameter estimates to estimation moments. *Quarterly Journal of Economics*, 132(4):1553–1592, 2017.

T. Armstrong and M. Kolesar. Finite-sample optimal estimation and inference on averate treatment effects under unconfoundedness. *Working Paper*, 2017.

S. Athey and G. Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016. . URL http://www.pnas.org/content/113/27/7353.abstract.

S. Athey and S. Wager. Efficient policy learning. *ArXiv e-prints*, Feb. 2017.

S. Athey, G. W. Imbens, and S. Wager. Approximate Residual Balancing: De-Biased Inference of Average Treatment Effects in High Dimensions. *ArXiv e-prints*, Apr. 2016a.

S. Athey, J. Tibshirani, and S. Wager. Generalized Random Forests. *ArXiv e-prints*, Oct. 2016b.

A. Belloni and V. Chernozhukov. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547, 2013. ArXiv, 2009.

A. Belloni, D. Chen, V. Chernozhukov, and C. Hansen. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80:2369–2429, 2012. Arxiv, 2010.

A. Belloni, V. Chernozhukov, and C. Hansen. Inference on treatment effects after selection amongst high-dimensional controls with an application to abortion on crime. *Review of Economic Studies*, 81(2):608–650, 2014.

P. J. Bickel, C. A. Klaassen, Y. Ritov, J. A. Wellner, et al. Efficient and adaptive estimation for semiparametric models. 1998.

F. Bunea, A. Tsybakov, and M. H. Wegkamp. Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics*, 1:169–194, 2007.

T. T. Cai and Z. Guo. Accuracy Assessment for High-dimensional Linear Regression. *ArXiv e-prints*, Mar. 2016.

M. Cattaneo, M. Jansson, and W. Newey. Inference in linear regression models with many covariates and heteroscedasticity. *Journal of the American Statistical Association*, 113:1350–1361, 2018.

M. D. Cattaneo, M. Jansson, and X. Ma. Two-step estimation and inference with possibly many included covariates. *Review of Economic Studies*, 86:1095–1122, 2019.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. K. Newey, and J. Robins. Double Machine Learning for Treatment and Causal Parameters. *ArXiv e-prints*, July 2016.

T. G. Conley, C. B. Hansen, and P. E. Rossi. Plausibly exogenous. *Review of Economics and Statistics*, 94(1):260–272, 2012.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of American Statistical Association*, 96(456):1348–1360, 2001.

I. E. Frank and J. H. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993.

T. Hastie, R. Tibshirani, and J. Friedman. *Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, NY, 2009.

H. Leeb and B. M. Pötscher. Can one estimate the unconditional distribution of post-model-selection

estimators? *Econometric Theory*, 24(2):338–376, 2008. ISSN 0266-4666. . URL http://dx.doi.org/10.1017/S0266466608080158.

C. Li and U. K. Müller. Linear regression with many controls of limited explanatory power. Working Paper, 2020.

Q. Li and J. S. Racine. *Nonparametric Econometrics: Theory and Practice*. Princeton University Press: Princeton, NJ, 2006.

C. F. Manski. *Partial Identification of Probability Distributions*. Springer-Verlag, 2003.

E. Oster. Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business and Economic Statistics*, 0(0):1–18, 2017. . URL https://doi.org/10.1080/07350015.2016.1227711.

B. M. Pötscher. Confidence sets based on sparse estimators are necessarily large. *Sankhyā*, 71(1, Ser. A):1–18, 2009. ISSN 0972-7671.

P. R. Rosenbaum. *Observational Studies*. Springer-Verlag, 2002.

P. R. Rosenbaum. Bahadur efficiency of sensitivity analyses in observational studies. *Journal of the American Statistical Association*, 110(509):205–217, 2015. . URL https://doi.org/10.1080/01621459.2014.960968.

D. S. Small. Sensitivity analysis for instrumental variables regression with overidentifying restrictions. *Journal of the American Statistical Association*, 102(479):1049–1058, 2007.

R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58: 267–288, 1996.

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

S. Wager and S. Athey. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *ArXiv e-prints*, Oct. 2015.

H. White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838, 1980.

Y. Zhu and J. Bradic. Linear hypothesis testing in dense high-dimensional linear models. *ArXiv e-prints*, Oct. 2016.

Y. Zhu and J. Bradic. A projection pursuit framework for testing general high-dimensional hypothesis. *ArXiv e-prints*, May 2017.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, 2006.