SUPPLEMENTARY APPENDIX FOR "INFERENCE ON TREATMENT EFFECTS AFTER SELECTION AMONGST HIGH-DIMENSIONAL CONTROLS"

A. BELLONI, V. CHERNOZHUKOV, AND C. HANSEN

Abstract. In this supplementary appendix we provide additional results, omitted proofs and extensive simulations that complement the analysis of the main text (arXiv:1201.0224).

1. Intuition for the Importance of Double Selection

To build intuition, we discuss the case where there is only one control; that is, $p = 1$. This scenario provides the simplest possible setting where variable selection might be interesting. In this case, Lasso-type methods act like conservative t-tests which allows the properties of selection methods to be explained easily.

With $p = 1$, the model is

$$y_i = \alpha_0 d_i + \beta_g x_i + \zeta_i, \tag{1.1}$$

$$d_i = \beta_m x_i + v_i. \tag{1.2}$$

For simplicity, all errors and controls are taken as normal,

$$\begin{pmatrix} \zeta_i \\ v_i \end{pmatrix} \mid x_i \sim N\left(0, \begin{pmatrix} \sigma_\zeta^2 & 0 \\ 0 & \sigma_v^2 \end{pmatrix}\right), \quad x_i \sim N(0,1), \tag{1.3}$$

where the variance of x_i is normalized to be 1. The underlying probability space is equipped with probability measure P (referred to as "dgp" throughout the paper). Let $\mathbf{P}$ denote the collection of all dgps P where (1.1)-(1.3) hold with non-singular covariance matrices in (1.3). Suppose that we have an i.i.d. sample $(y_i, d_i, x_i)_{i=1}^n$ from the dgp $\mathrm{P}_n \in \mathbf{P}$. The subscript n signifies that the dgp and all true parameter values may change with n to better model finite-sample phenomena such coefficients being "close to zero". As in the rest of the paper, we keep the dependence of the true parameter values on n implicit. Under the stated assumption, x_i and d_i are jointly normal with variances $\sigma_x^2 = 1$ and $\sigma_d^2 = \beta_m^2 \sigma_x^2 + \sigma_v^2$ and correlation $\rho = \beta_m \sigma_x / \sigma_d$.

Date: First version: May 2010, This version of June 24, 2013.

The standard post-single-selection method for inference proceeds by applying model selection methods – ranging from standard t-tests to Lasso-type selectors – to the first equation only, followed by applying OLS to the selected model. In the model selection stage, standard selection methods would necessarily omit x_i wp $\to 1$ if

$$|\beta_g| \leqslant \frac{\ell_n}{\sqrt{n}} c_n, \ c_n := \frac{\sigma_\zeta}{\sigma_x \sqrt{1-\rho^2}}, \text{ for some } \ell_n \to \infty, \tag{1.4}$$

where ℓ_n is a slowly varying sequence depending only on $\mathbf{P}$. On the other hand, these methods would necessarily include x_i wp $\to 1$, if

$$|\beta_g| \geqslant \frac{\ell_n'}{\sqrt{n}} c_n, \text{ for some } \ell_n' > \ell_n, \tag{1.5}$$

where ℓ_n' is another slowly varying sequence in n depending only on $\mathbf{P}$. In most standard model selection devices with sensible tuning choices, we shall have $\ell_n = C\sqrt{\log n}$ and $\ell_n' = C'\sqrt{\log n}$ with constants C and C' depending only on $\mathbf{P}$. In the case of Lasso methods, we prove this in Section 5. This is also true in the case of the conservative t-test, which omits x_i if the t-statistic $|t| = |\widehat{\beta}_g|/\text{s.e.}(\widehat{\beta}_g) \geqslant \Phi^{-1}(1 - 1/(2n))$, where $\widehat{\beta}_g$ is the OLS estimator, and s.e.$(\widehat{\beta}_g)$ is the corresponding standard error. In this case, we have $\Phi^{-1}(1 - 1/(2n)) = \sqrt{2\log n}(1 + o(1))$ so the test will have power approaching 1 for alternatives of the form (1.5) with $\ell_n' = 2\sqrt{\log n}$ and power approaching 0 for alternatives of the form (1.5) with $\ell_n = \sqrt{\log n}$. [1]

A standard selection procedure would work with the first equation. Under "good" sequences of models P_n such that (1.5) holds, x_i is included wp $\to 1$, and the estimator becomes the standard OLS estimator with the standard large sample asymptotics under P_n

$$\sigma_n^{-1}\sqrt{n}(\widehat{\alpha} - \alpha_0) = \underbrace{\sigma_n^{-1}\mathbb{E}_n[v_i^2]^{-1}\sqrt{n}\mathbb{E}_n[v_i\zeta_i]}_{=:i} + o_P(1) \rightsquigarrow N(0,1),$$

where $\sigma_n^2 = \sigma_\zeta^2(\sigma_v^2)^{-1}$. On the other hand, when $\beta_g = 0$ or $\beta_g = o(\ell_n/\sqrt{n})$ and ρ is bounded away from 1, we have that

$$\sigma_n^{*-1}\sqrt{n}(\widehat{\alpha} - \alpha_0) = \underbrace{\sigma_n^{*-1}\mathbb{E}_n[d_i^2]^{-1}\sqrt{n}\mathbb{E}_n[d_i\zeta_i]}_{:=i^*} + o_P(1) \rightsquigarrow N(0,1),$$

where $\sigma_n^{*2} = \sigma_\zeta^2(\sigma_d^2)^{-1}$. The variance σ_n^{*2} is smaller than the variance σ_n^2 from estimation with x_i included if $\beta_m \neq 0$. The potential reduction in variance is often used as a "motivation" for the standard selection procedure. The estimator is super-efficient, achieving a variance smaller than the semi-parametric efficiency bound under homoscedasticity. That is, the estimator is "too good."

[1] This assumes that the canonical estimator of the standard error is used.

The "too good" behavior of the procedure that looks solely at the first equaiton has its price. There are plausible sequences of dgps P_n where $\beta_g = \frac{\ell_n'}{\sqrt{n}} c_n$, the coefficient on x_i is not zero but is close too zero,[2] in which the control x_i is dropped wp $\to 1$ and

$$|\sigma_n^{*-1} \sqrt{n}(\widehat{\alpha} - \alpha_0)| \rightsquigarrow \infty. \tag{1.6}$$

That is, the standard post-selection estimator is not asymptotically normal and even fails to be uniformly consistent at the rate of $\sqrt{n}$. This poor behavior occurs because the omitted variable bias created by dropping x_i may be large even when the magnitude of the regression coefficient, $|\beta_m|$, in the confounding equation (1.2) is small but is not exactly zero. To see this, note

$$\sigma_n^{*-1} \sqrt{n}(\widehat{\alpha} - \alpha_0) = \underbrace{\sigma_n^{*-1} \mathbb{E}_n[d_i^2]^{-1} \sqrt{n} \mathbb{E}_n[d_i \zeta_i]}_{=:i^*} + \underbrace{\sigma_n^{*-1} \mathbb{E}_n[d_i^2]^{-1} \sqrt{n} \mathbb{E}_n[d_i x_i] \beta_g}_{=:ii}.$$

The term i^* has standard behavior; namely $i^* \rightsquigarrow N(0,1)$. The term ii generates the *omitted variable bias*, and it may be arbitrarily large, since wp $\to 1$,

$$|ii| \geqslant \frac{1}{2} \frac{|\rho|}{\sqrt{1-\rho^2}} \ell_n \nearrow \infty,$$

if $\ell_n |\rho| \nearrow \infty$.[3] This yields the conclusion (1.6) by the triangle inequality.

In contrast to the standard approach, our post-double-selection method for inference proceeds by applying model selection methods, such as standard t-tests or Lasso-type selectors, to both equations and taking the selected controls as the union of controls selected from each equation. This selection is than followed by applying OLS to the selected controls. Thus, our approach drops x_i only if the omitted variable bias term ii is small. To see this, note that the double-selection-methods *include* x_i wp $\to 1$ if its coefficient in either (1.1) or (1.2) is not very small. Mathematically, x_i is included if

$$\text{either } |\beta_g| \geqslant \frac{\ell_n'}{\sqrt{n}} \left(\frac{\sigma_\zeta}{\sigma_x \sqrt{1-\rho^2}} \right) \text{ or } |\beta_m| \geqslant \frac{\ell_n'}{\sqrt{n}} \left(\frac{\sigma_v}{\sigma_x} \right) \tag{1.7}$$

where ℓ_n' is a slowly varying sequence in n. As already noted, $\ell_n \propto \ell_n' \propto \sqrt{\log n}$ would be standard for Lasso-type methods as well as for using simple t-tests to do model selection. Considering t-tests and using these rates, we would omit x_i if both $|t_g| = |\widehat{\beta}_g|/\text{s.e.}(\widehat{\beta}_g) \leqslant \Phi^{-1}(1 - 1/(2n))$ *and* $|t_m| = |\widehat{\beta}_g|/\text{s.e.}(\widehat{\beta}_g) \leqslant \Phi^{-1}(1 - 1/(2n))$ where $\widehat{\beta}_g$ and $\widehat{\beta}_m$ denote the OLS estimator from each equation and s.e. denotes the corresponding estimated standard errors.

[2]Such sequences are very relevant in that they are designed to generate approximations that better capture the fact that one cannot distinguish an estimated coefficient from 0 arbitrarily well in any given finite sample.

[3]Recall that $\rho = \beta_m \sigma_x / \sigma_d$, so $\ell_n |\rho| \nearrow \infty$ as long as $\ell_n |\beta_m| \nearrow \infty$ assuming that σ_x / σ_d is bounded away from 0 and ∞.

Note that the critical value used in the t-tests above is conservative in the sense that the false rejection probability is tending to zero because $\Phi^{-1}(1-1/(2n)) = \sqrt{2\log n}(1+o(1))$. We note that Lasso-type methods operate similarly.

Given the discussion in the preceding paragraph, it is immediate that the post-double selection estimator satisfies

$$\sigma_n^{-1}\sqrt{n}(\check{\alpha}-\alpha_0) = i + o_P(1) \rightsquigarrow N(0,1) \tag{1.8}$$

under any sequence of $\mathrm{P}_n \in \mathbf{P}$. We get this approximating distribution whether or not x_i is omitted. That this is the approximate distribution when x_i is included follows as in the single-selection case. To see that we get the same approximation when x_i is omitted, note that we drop x_i only if

$$\text{both } |\beta_g| \leqslant \frac{\ell_n}{\sqrt{n}} c_n \text{ and } |\beta_m'| \leqslant \frac{\ell_n}{\sqrt{n}}(\sigma_v/\sigma_x), \tag{1.9}$$

i.e. coefficients in front of x_i in both equations are small. In this case,

$$\sigma_n^{*-1}\sqrt{n}(\check{\alpha}-\alpha_0) = \underbrace{\sigma_n^{*-1}\mathbb{E}_n[d_i^2]^{-1}\sqrt{n}\mathbb{E}_n[d_i\zeta_i]}_{=i^*} + \underbrace{\sigma_n^{*-1}\mathbb{E}_n[d_i^2]^{-1}\sqrt{n}\mathbb{E}_n[d_i x_i]\beta_g}_{=ii}.$$

Once again, the term ii is due to *omitted variable bias*, and it obeys wp $\to 1$ under (1.9)

$$|ii| \leqslant 2\sigma_\zeta^{-1}\sigma_d\sigma_d^{-2}\sqrt{n}\sigma_x^2|\beta_m\beta_g| \leqslant 2\frac{\sigma_v/\sigma_d}{\sqrt{1-\rho^2}}\frac{\ell_n^2}{\sqrt{n}} = 2\frac{\ell_n^2}{\sqrt{n}} \to 0,$$

since $(\sigma_v/\sigma_d)^2 = 1-\rho^2$. Moreover, we can show $i^* - i = o_P(1)$ under such sequences, so the first order asymptotics of $\check{\alpha}$ is the same whether x_i is included or excluded.

To summarize, the post-single-selection estimator may not be root-n consistent in sensible models which translates into bad finite-sample properties. The potential poor finite-sample performance may be clearly seen in Monte-Carlo experiments. The estimator $\widehat{\alpha}$ is thus non-regular: its first-order asymptotic properties depend on the model sequence P_n in a strong way. In contrast, the post-double selection estimator $\check{\alpha}$ guards against omitted variables bias which reduces the dependence of the first-order behavior on P_n. This good behavior under sequences P_n translates into uniform with respect to $\mathrm{P} \in \mathbf{P}$ asymptotic normality.

We should note, of course, that the post-double-selection estimator is first-order equivalent to the long-regression in this model.[4] This equivalence disappears under approximating sequences with number of controls proportional to the sample size, $p \propto n$, or greater than the sample size, $p \gg n$. It is these scenarios that motivate the use of selection as a means of regularization. In these more complicated settings the intuition from this simple $p = 1$ example carries through, and the post-single selection method has a highly non-regular behavior while the post-double selection method continues to be regular.

It is also informative to consider semi-parametric efficiency in this simple example. The post-single-selection estimator is super-efficient when $\beta_m \neq 0$ and $\beta_g = 0$. The super-efficiency in this case is apparent upon noting that the estimator is root-n consistent and normal with asymptotic variance $\mathrm{E}[\zeta_i^2]\mathrm{E}[d_i^2]^{-1}$. This asymptotic variance is generally smaller than the semi-parametric efficiency bound $\mathrm{E}[\zeta_i^2]\mathrm{E}[v_i^2]^{-1}$. The price of this efficiency gain is the fact that the post-single-selection estimator breaks down when β_g may be small but non-zero. The corresponding confidence intervals therefore also break down. In contrast, the post-double-selection estimator remains well-behaved in any case, and confidence intervals based on the double-selection-estimator and are uniformly valid for this reason.

2. Extensions: Other Problems and Heterogeneous Treatment Effects

2.1. **Other Problems.** In order to discuss extensions in a very simple manner, we assume i.i.d sampling as well as assume away approximation errors, namely $g(z_i) = x_i'\beta_{g0}$ and $m(z_i) = x_i'\beta_{m0}$, where parameters β_{g0} and β_{m0} are high-dimensional and that $x_i = P(z_i)$ as before. In this paper we considered a moment condition:

$$\mathrm{E}[\varphi(y_i - d_i\alpha_0 - x_i'\beta)v_i] = 0, \tag{2.10}$$

where $\varphi(u) = u$ and v_i are measurable functions of z_i, and the target parameter is α_0. We selected the instrument v_i such that the equation is first-order insensitive to the parameter β at $\beta = \beta_{g0}$:

$$\frac{\partial}{\partial\beta}\mathrm{E}[\varphi(y_i - d_i\alpha_0 - x_i'\beta)v_i]\bigg|_{\beta=\beta_{g0}} = 0. \tag{2.11}$$

[4]This equivalence may be a reason double-selection was previously overlooked. There are higher-order differences between the long regression estimator and our estimator. In particular, our estimator has variance that is smaller than the variance of the long regression estimator by a factor that is proportional to $(1 - \ell_n/\sqrt{n})$ if $\beta_g = 0$; and when $\beta_g = \ell_n/\sqrt{n}$, there is a small reduction in variance that is traded against a small increase in bias.

Note that $\varphi(u) = u$ and $v_i = d_i - m(z_i)$ implement this condition. If (2.11) holds, the estimator of α_0 gets "immunized" against nonregular estimation of β_0, for example, via a post-selection procedure or other regularized estimators. Such immunization ideas are in fact behind the classical Frisch-Waugh-Robinson partialling out technique in the linear setting and the Neyman (1979)'s $C(\alpha)$ test in the nonlinear setting. Our contribution here is to recognize the importance of this immunization in the context of post-selection inference,[5] to develop a robust post-selection approach to inference on the target parameter, and characterize the uniformity regions of this procedure. Our approach uses modern selection methods to estimate g and the function m defining the intstrument v. In an ongoing work, we explore other regularization methods, such as the ridge method or combination of ridge method with Lasso methods, and characterize uniformity regions of the resulting procedures. Also, generalizations to nonlinear models, where φ is non-linear and can correspond to a likelihood score or quantile check function are given in Belloni, Chernozhukov, and Kato (2013) and Belloni, Chernozhukov, and Wei (2013); in these generalizations achieving (2.11) is also critical.

Within the context of this paper, a potentially important extension is to consider a general treatment effect model, where d_i is interacted with transformations of z_i. As long as the interest lies in a particular regression coefficient, the current framework covers this implicitly since x_i could contain interactions of d_i with transformations of controls z_i. In the case a fixed number of such regression coefficients is of interest, we can estimate each of the coefficients by re-labeling the corresponding regressor as d_i and other regressors as x_i and then applying our procedure the fixed number of times. Such component-wise procedure is valid as long as our regularity conditions hold for each of the resulting regression models in this manner.

A related research direction being pursued is the study of estimation of average treatment effects when treatment effects are fully heterogeneous. When the treatment variable $d_i \in \{0, 1\}$ is binary (or discrete more generally) our approach is readily amenable to this problem. In this case the parameter of interest is the average treatment effect

$$\alpha_0 = \mathrm{E}[g(1, z_i) - g(0, z_i)],$$

where $g(d_i, z_i) = \mathrm{E}[y_i|d_i, z_i]$. We can write, assuming again no approximation errors for simplicity, $g(d_i, z_i) = d_i x_i'\beta_{g0,1} + (1 - d_i)x_i'\beta_{g0,0}$. Suppose that the propensity score $\mathrm{P}(d_i = 1 \mid z_i)$ is $m(z_i) = \Lambda(x_i'\beta_{m0})$, where $\Lambda(u)$ is a link such as logit or linear, then we can use the moment

[5] To the best of our knowledge, all prior theoretical and empirical work uses the standard post-selection approach based on the outcome equation alone, which is highly non-robust way of conducting inference, as shown in extensive monte-carlo, in Section 2.4, and in a sequence of fundamental critiques by Leeb and Potscher, see Leeb and Pötscher (2008) .

equations of Hahn (1998):

$$\mathrm{E}\left[\varphi(\alpha_0, y_i, d_i, g(0, z_i), g(1, z_i), m(z_i))\right] = 0, \tag{2.12}$$

where $\varphi(\alpha, y, d, g_0, g_1, m) = \alpha - \frac{d(y-g_1)}{m} + \frac{(1-d)(y-g_0)}{1-m} - (g_1 - g_0)$. It is straightforward to check that for each $j \in \{1, 2, 3\}$:

$$\frac{\partial}{\partial \beta_j} \mathrm{E}\left[\varphi_i(\alpha_0, y_i, d_i, x_i'\beta_1, x_i\beta_2, x_i'\beta_3)\right] = 0, \tag{2.13}$$

when $(\beta_1, \beta_2, \beta_3) = (\beta_{g0,0}, \beta_{g0,1}, \beta_{m0})$ (i.e., evaluated at the true values). Therefore, by using Hahn (1998)'s equation we obtain immunization against the crude estimation of either β_{g0} or β_{m0}, just like we do in the partially linear case. Hence we can use the selection approach to regularization and estimate the parameter of interest α_0. Note that in this case the resulting procedure is a double selection method, where terms explaining propensity score and the regression function are selected. Here too we can use the "union" approach in fitting each regression function involved, as it gives the best finite-sample performance in extensive computational experiments. Using the results of this paper, it is not difficult to show for the case that $\Lambda(u) = u$ that under the sparsity assumption imposed on both g and m, and additional assumptions needed to guarantee consistency of post-Lasso estimators $\widehat{g}$ and $\widehat{m}$ in the uniform norm, that the post-double-selection estimator $\check{\alpha}$ that solves $\mathbb{E}_n\left[\varphi(\check{\alpha}, y_i, d_i, \widehat{g}(0, z_i), \widehat{g}(1, z_i), \widehat{m}(z_i))\right] = 0$, has the following large sample behavior:

$$\sigma_n^{-1}\sqrt{n}(\check{\alpha} - \alpha_0) \rightsquigarrow N(0,1), \quad \sigma_n^2 = \mathrm{E}[\varphi^2(\alpha_0, y_i, d_i, g(0, z_i), g(1, z_i), m(z_i))], \tag{2.14}$$

where the latter is the semiparametric efficiency bound of Hahn (1998). Formal result is proven and stated below, and other formal results along these lines are given in an ongoing work that studies this as well as other types of effects.

2.2. **Theoretical Results on ATE with Heterogeneity.** Consider i.i.d. sample $(y_i, d_i, z_i)_{i=1}^n$ on the probability space $(\Omega, \mathfrak{F}, \mathrm{P})$, where we call P the data-generating process. Consider the case where treatment variable is binary $d_i \in \{0, 1\}$, and the outcome and propensity equations, as before,

$$y_i = g(d_i, z_i) + \zeta_i, \quad \mathrm{E}[\zeta_i \mid z_i, d_i] = 0, \tag{2.15}$$

$$d_i = m(z_i) + v_i, \quad \mathrm{E}[v_i \mid z_i] = 0. \tag{2.16}$$

The first target parameter is the average treatment effect: $\alpha_0 = \mathrm{E}[g(1, z_i) - g(0, z_i)]$, defined above, which is implicitly indexed by P, like other parameters. In this model d_i is *not* additively separable. The purpose of this section is to show that our analysis easily extends to this case, using our techniques.

The confounding factors z_i affect the policy variable via the propensity score $m(z_i)$ and the outcome variable via the function $g(d_i, z_i)$. Both of these functions are unknown and potentially complicated. As in the main text, we use linear combinations of control terms $x_i = P(z_i)$ to approximate $g(z_i)$ and $m(z_i)$, writing (2.15) and (2.16) as

$$y_i = \underbrace{\tilde{x}_i'\beta_{g0} + r_{gi}}_{g(d_i,z_i)} + \zeta_i, \tag{2.17}$$

$$d_i = \underbrace{\Lambda(x_i'\beta_{m0}) + r_{mi}}_{m(z_i)} + v_i, \tag{2.18}$$

where r_{gi} and r_{mi} are the approximation errors, and

$$\tilde{x}_i := (d_i x_i', (1-d_i)x_i')', \quad \beta_{g0} := (\beta_{g0,1}', \beta_{g0,0}'), \ \ x_i := P(z_i), \tag{2.19}$$

where $x_i'\beta_{g0,1}$, $x_i'\beta_{g0,0}$, and $x_i'\beta_{m0}$ are approximations to $g(1, z_i)$, $g(0, z_i)$, and $m(z_i)$, and $\Lambda(u) = u$ for the case of linear link and $\Lambda(u) = e^u/(1 + e^u)$ for the case of the logistic link. In order to allow for a flexible specification and incorporation of pertinent confounding factors, the vector of controls, $x_i = P(z_i)$, we can have a dimension $p = p_n$ which can be large relative to the sample size.

The efficient moment condition, derived by Hahn (1998), for parameter α_0 is as follows:

$$\mathrm{E}\left[\varphi(\alpha_0, y_i, d_i, g(0, z_i), g(1, z_i), m(z_i))\right] = 0, \tag{2.20}$$

where

$$\varphi(\alpha, y, d, g_0, g_1, m) = \alpha - \frac{d(y - g_1)}{m} + \frac{(1-d)(y - g_0)}{1 - m} - (g_1 - g_0).$$

The post-double-selection estimator $\check{\alpha}$ that solves

$$\mathbb{E}_n\left[\varphi(\check{\alpha}, y_i, d_i, \widehat{g}(0, z_i), \widehat{g}(1, z_i), \widehat{m}(z_i))\right] = 0, \tag{2.21}$$

where $\widehat{g}(d_i, z_i)$ and $\widehat{m}(z_i)$ are post-Lasso estimators of functions g and m based upon equations (2.17)-(2.18). In case of the logistic link Λ, Lasso for logistic regression is as defined in van de Geer (2008) and Bach (2010), and the associated post-Lasso estimators are as those defined in Belloni, Chernozhukov, and Wei (2013).

In what follows, we use $\|w_i\|_{\mathrm{P},q}$ to denote the $L^q(\mathrm{P})$ norm of a random variable w_i with law determined by P, and we $\|w_i\|_{\mathbb{P}_n,q}$ to denote the empirical $L^q(\mathbb{P}_n)$ norm of a random variable with law determined by the emprical measure $\mathbb{P}_n = n^{-1}\sum_{i=1}^n \delta_{w_i}$, i.e., $\|w_i\|_{\mathbb{P}_n,q} = (n^{-1}\sum_{i=1}^n \|w_i\|^q)^{1/q}$.

Consider fixed positive sequences $\delta_n \searrow 0$ and $\Delta_n \nearrow 0$ and constants $C > 0, c > 0, 1/2 > c' > 0$, which will not vary with P.

Condition HTE (P) . **Heterogeneous Treatment Effects**. *Consider i.i.d. sample $(y_i, d_i, z_i)_{i=1}^n$ on the probability space $(\Omega, \mathfrak{F}, \mathrm{P})$, where we shall call P the data-generating process, such that equations (2.17)-(2.18) holds, with $d_i \in \{0,1\}$. (i) Approximation errors satisfy $\|r_{gi}\|_{\mathrm{P},2} \leqslant \delta_n n^{-1/4}$, $\|r_{gi}\|_{\mathrm{P},\infty} \leqslant \delta_n$, and $\|r_{mi}\|_{\mathrm{P},2} \leqslant \delta_n n^{-1/4}$, $\|r_{mi}\|_{\mathrm{P},\infty} \leqslant \delta_n$. (ii) With P-probability no less than $1-\Delta_n$, estimation errors satisfy $\|\tilde{x}_i'(\widehat{\beta}_g - \beta_{g0})\|_{\mathbb{P}_n,2} \leqslant \delta_n n^{-1/4}$, $\|x_i'(\widehat{\beta}_m - \beta_{m0})\|_{\mathbb{P}_n,2} \leqslant \delta_n n^{-1/4}$, $K_n\|\widehat{\beta}_m - \beta_m\|_1 \leqslant \delta_n$, $K_n\|\widehat{\beta}_m - \beta_{m0}\|_1 \leqslant \delta_n$, estimators and approximations are sparse, namely $\|\widehat{\beta}_g\|_0 \leqslant Cs$, $\|\widehat{\beta}_m\|_0 \leqslant Cs$, and $\|\beta_{g0}\|_0 \leqslant Cs$, $\|\beta_{m0}\|_0 \leqslant Cs$ and the empirical and populations norms are equivalent on sparse subsets, namely $\sup_{\|\delta\|_0 \leqslant 2Cs} |\|\tilde{x}_i'\delta\|_{\mathbb{P}_n,2}/\|\tilde{x}_i'\delta\|_{\mathrm{P},2} - 1| \leqslant \delta_n$. (iii) The following boundedness conditions hold: $\|x_{ij}||_{\mathrm{P},\infty} \leqslant K_n$ for each j, $\|g\|_{\mathrm{P},\infty} \leqslant C$, $\|y_i\|_{\mathrm{P},\infty} \leqslant C$, $\mathrm{P}(c' \leqslant m(z_i) \leqslant 1-c') = 1$, and $\|\zeta_i^2\|_{\mathrm{P},2} \geqslant c$. (iv) The sparsity index obeys the following growth condition, $(s\log(p \vee n))^2/n \leqslant \delta_n$.*

These conditions are simple high-level conditions, which encode both the approximate sparsity of the models as well as impose some reasonable behavior on the post-selection estimators of m and g (or other sparse estimators). These conditions are implied by other more primitive conditions in the literature. Sufficient conditions for the equivalence between population and empirical sparse eigenvalues are given in Rudelson and Zhou (2011) and Rudelson and Vershynin (2008). The boundedness conditions are made to simplify arguments, and they could be dealt away with more complicated proofs, under more stringent side conditions.

Theorem 1 (**Uniform Post-Double Selection Inference on ATE**). *Consider the set $\mathbf{P}_n$ of data generating processes P such that equations (2.15)-(2.16) and Condition ATE (P) holds. (1) Then under any sequence $\mathrm{P} \in \mathbf{P}_n$,*

$$\sigma_n^{-1}\sqrt{n}(\check{\alpha} - \alpha_0) \rightsquigarrow N(0,1), \quad \sigma_n^2 = \mathrm{E}[\varphi^2(\alpha_0, y_i, d_i, g(0,z_i), g(1,z_i), m(z_i))], \tag{2.22}$$

(2) The result continues to hold with σ_n^2 replaced by $\widehat{\sigma}_n^2 := \mathbb{E}_n[\varphi^2(\alpha_0, y_i, d_i, \widehat{g}(0,z_i), \widehat{g}(1,z_i), \widehat{m}(z_i))]$. (3) Moreover, the confidence regions based upon post-double selection estimator $\check{\alpha}$ have the uniform asumptotic validity,

$$\lim_{n\to\infty} \sup_{\mathrm{P}\in\mathbf{P}_n} \left|\mathrm{P}\left(\alpha_0 \in [\check{\alpha} \pm \Phi^{-1}(1-\xi/2)\widehat{\sigma}_n/\sqrt{n}]\right) - (1-\xi)\right| = 0.$$

The next target parameter is the average treatment effect on the treated:

$$\gamma_0 = \mathrm{E}[g(1,z_i) - g(0,z_i)|d_i = 1].$$

The efficient moment condition, derived by Hahn (1998), for parameter γ_0 is as follows:

$$\mathrm{E}\left[\tilde{\varphi}(\gamma_0, y_i, d_i, g(0, z_i), g(1, z_i), m(z_i), \mu)\right] = 0, \tag{2.23}$$

where $\mu = \mathrm{E}[m(z_i)] = \mathrm{P}(d_i = 1)$, and

$$\tilde{\varphi}(\gamma, y, d, g_0, g_1, m, \mu) = \frac{d(y - g_1)}{\mu} - \frac{m(1-d)(y-g_0)}{(1-m)\mu} + \frac{d(g_1 - g_0)}{\mu} - \gamma\frac{d}{\mu}.$$

In this case the post-double-selection estimator $\check{\gamma}$ that solves

$$\mathbb{E}_n\left[\tilde{\varphi}(\check{\gamma}, y_i, d_i, \widehat{g}(0, z_i), \widehat{g}(1, z_i), \widehat{m}(z_i), \widehat{\mu})\right] = 0, \tag{2.24}$$

where $\widehat{g}(d_i, z_i)$ and $\widehat{m}(z_i)$ are post-Lasso estimators (or other sparse estimators obeying the regularity conditions posed in HTE) of functions g and m based upon equations (2.17)-(2.18), and $\widehat{\mu} = \mathbb{E}_n[d_i]$. As mentioned before, in case of the logistic link Λ, the post-Lasso estimators are as those defined in Belloni, Chernozhukov, and Wei (2013).

Theorem 2 (**Uniform Post-Double Selection Inference on ATT**). *Consider the set* $\mathbf{P}_n$ *of data generating processes* P *such that equations (2.15)-(2.16) and Condition HTE (P) holds. (1) Then under any sequence* $\mathrm{P} \in \mathbf{P}_n$,

$$\sigma_n^{-1}\sqrt{n}(\check{\gamma} - \gamma_0) \rightsquigarrow N(0,1), \quad \sigma_n^2 = \mathrm{E}[\tilde{\varphi}^2(\gamma_0, y_i, d_i, g(0, z_i), g(1, z_i), m(z_i), \mu)], \tag{2.25}$$

(2) The result continues to hold with σ_n^2 *replaced by* $\widehat{\sigma}_n^2 := \mathbb{E}_n[\tilde{\varphi}^2(\gamma_0, y_i, d_i, \widehat{g}(0, z_i), \widehat{g}(1, z_i), \widehat{m}(z_i), \mu)]$.
(3) Moreover, the confidence regions based upon post-double selection estimator $\check{\alpha}$ *have the uniform asumptotic validity,*

$$\lim_{n\to\infty} \sup_{\mathrm{P}\in\mathbf{P}_n} \left|\mathrm{P}\left(\gamma_0 \in [\check{\gamma} \pm \Phi^{-1}(1-\xi/2)\widehat{\sigma}_n/\sqrt{n}]\right) - (1-\xi)\right| = 0.$$

2.3. **Proof of Theorems 1 and 2.** The two results have identical structure and have nearly the same proof, and so we present the proof of the Proof of Theorems 1 only.

In the proof $a \lesssim b$ means that $a \leqslant Ab$, where the constant A depends on the constants in Condition HT only, but not on n once $n \geqslant n_0 = \min\{j : \delta_j \leqslant 1/2\}$, and not on $\mathrm{P} \in \mathbf{P}_n$. For the proof of claims (1) and (2) we consider a sequence P_n in $\mathbf{P}_n$, but for simplicity, we write P throughout the proof, omitting the index n. Since the argument is asymptotic, we can just assume that $n \geqslant n_0$ in what follows.

Step 1. In this step we establish claim (1).

(a) We begin with a preliminary observation. Define, for $t = (t_1, t_2, t_3)$,

$$\psi(y, d, t) = \frac{d(y - t_2)}{t_3} - \frac{(1-d)(y-t_1)}{1-t_3} + (t_2 - t_1).$$

The derivatives of this function with respect to t obey for all $k = (k_j)_{j=1}^3 \in \mathbb{N} : 0 \leqslant |k| \leqslant 3$,

$$|\partial_t^k \psi(y,d,t)| \leqslant L, \quad \forall (y,d,t) : |y| \leqslant C, |t_1| \leqslant C, |t_2| \leqslant C, c'/2 \leqslant |t_3| \leqslant 1 - c'/2, \tag{2.26}$$

where L depends only on c' and C, $|k| = \sum_{j=1}^3 k_j$, and

$$\partial_t^k := \partial_{t_1}^{k_1} \partial_{t_2}^{k_2} \partial_{t_3}^{k_3}.$$

(b). Let

$$\widehat{h}(z_i) := (\widehat{g}(0,z_i), \widehat{g}(1,z_i), \widehat{m}(z_i))', \quad h_0(z_i) := (g(0,z_i), g(1,z_i), m(z_i))',$$

$$f_{\widehat{h}}(y_i,d_i,z_i) := \psi(y_i,d_i,\widehat{h}(z_i)), \quad f_{h_0}(y_i,d_i,z_i) := \psi(y_i,d_i,h_0(z_i)).$$

We observe that with probability no less than $1-\Delta_n$,

$$\widehat{g}(0,\cdot) \in \mathcal{G}_0, \quad \widehat{g}(1,\cdot) \in \mathcal{G}_1 \text{ and } \widehat{m} \in \mathcal{M},$$

$$\mathcal{G}_d := \{ z \mapsto x'\beta : \|\beta\|_0 \leqslant sC, \|x_i'\beta - g(d,z_i)\|_{\mathrm{P},2} \lesssim \delta_n n^{-1/4}, \|x_i'\beta - g(d,z_i)\|_{\mathrm{P},\infty} \lesssim \delta_n \},$$

$$\mathcal{M} := \{ z \mapsto \Lambda(x'\beta) : \|\beta\|_0 \leqslant sC, \|\Lambda(x_i'\beta) - m(z_i)\|_{\mathrm{P},2} \lesssim \delta_n n^{-1/4}, \|\Lambda(x_i'\beta) - m(z_i)\|_{\mathrm{P},\infty} \lesssim \delta_n \}.$$

To see this note, that under assumption HT (P), under condition (i)-(ii), under the event occurring under condition (ii) of that assumption: for $n \geqslant n_0 = \min\{j : \delta_j \leqslant 1/2\}$:

$$\|\tilde{x}_i'\beta - g(d_i,z_i)\|_{\mathrm{P},2} \leqslant \|\tilde{x}_i'(\beta - \beta_{g0})\|_{\mathrm{P},2} + \|r_{gi}\|_{\mathrm{P},2} \leqslant 2\|\tilde{x}_i'(\beta - \beta_{g0})\|_{\mathbb{P}_n,2} + \|r_{gi}\|_{\mathrm{P},2} \leqslant 4\delta_n n^{-1/4},$$
$$\|\tilde{x}_i'\beta - g(d_i,z_i)\|_{\mathrm{P},\infty} \leqslant \|\tilde{x}_i'(\beta - \beta_{g0})\|_{\mathrm{P},\infty} + \|r_{gi}\|_{\mathrm{P},\infty} \leqslant K_n\|\beta - \beta_{0g}\|_1 + \delta_n \leqslant 2\delta_n,$$

for $\beta = \widehat{\beta}_g$, with evaluation after computing the norms, and noting that for any β

$$\|x_i'\beta - g(1,z_i)\|_{\mathrm{P},2} \vee \|x_i'\beta - g(0,z_i)\|_{\mathrm{P},2} \lesssim \|\tilde{x}_i'\beta - g(d_i,z_i)\|_{\mathrm{P},2}$$

under condition (iii). Furthermore, for $n \geqslant n_0 = \min\{j : \delta_j \leqslant 1/2\}$:

$$\begin{aligned}
\|\Lambda(x_i'\beta) - m(z_i)\|_{\mathrm{P},2} &\leqslant \|\Lambda(x_i'\beta) - \Lambda(x_i'\beta_{m0})\|_{\mathrm{P},2} + \|r_{mi}\|_{\mathrm{P},2} \\
&\lesssim \|\partial\Lambda\|_\infty \|\tilde{x}_i'(\beta - \beta_{m0})\|_{\mathrm{P},2} + \|r_{mi}\|_{\mathrm{P},2} \\
&\lesssim \|\partial\Lambda\|_\infty \|\tilde{x}_i'(\beta - \beta_{m0})\|_{\mathbb{P}_n,2} + \|r_{mi}\|_{\mathrm{P},2} \lesssim \delta_n n^{-1/4} \\
\|\Lambda(x_i'\beta) - m(z_i)\|_{\mathrm{P},\infty} &\leqslant \|\partial\Lambda\|_\infty \|\tilde{x}_i'(\beta - \beta_{g0})\|_{\mathrm{P},\infty} + \|r_{mi}\|_{\mathrm{P},\infty} \\
&\lesssim K_n\|\beta - \beta_{m0}\|_1 + \delta_n \leqslant 2\delta_n,
\end{aligned}$$

for $\beta = \widehat{\beta}_{m0}$, with evaluation after computing the norms.

Hence with probability at least $1-\Delta_n$,

$$\widehat{h} \in \mathcal{H}_n := \{ h = (\bar{g}(0,z), \bar{g}(1,z), \bar{m}(z)) \in \mathcal{G}_0 \times \mathcal{G}_1 \times \mathcal{M} \}.$$

(c) We have that

$$\alpha_0 = \mathrm{E}[f_{h_0}] \text{ and } \check{\alpha} = \mathbb{E}_n[f_{\widehat{h}}],$$

so that

$$\sqrt{n}(\check{\alpha} - \alpha_0) = \underbrace{\mathbb{G}_n[f_{h_0}]}_{i} + \underbrace{(\mathbb{G}_n[f_h] - \mathbb{G}_n[f_{h_0}])}_{ii} + \underbrace{\sqrt{n}(\mathrm{E}[f_h - f_{h_0}])}_{iii},$$

with h evaluated at $h = \widehat{h}$. By Liapunov central limit theorem,

$$\sigma_n^{-1} i \rightsquigarrow N(0,1).$$

(d) Note that for $\Delta_i = h(z_i) - h_0(z_i)$,

$$\begin{aligned} iii &= \sqrt{n} \sum_{|k|=1} \mathrm{E}[\partial_t^k \psi(y_i, d_i, h_0(z_i))\Delta_i^k] \\ &+ \sqrt{n} \sum_{|k|=2} 2^{-1}\mathrm{E}[\partial_t^k \psi(y_i, d_i, h_0(z_i))\Delta_i^k] \\ &+ \sqrt{n} \sum_{|k|=3} \int_0^1 6^{-1}\mathrm{E}[\partial_t^k \psi(y_i, d_i, h_0(z_i) + \lambda\Delta_i)\Delta_i^k] d\lambda, \\ &=: iii_a + iii_b + iii_c, \end{aligned}$$

(with h evaluated at $h = \widehat{h}$). By the law of iterated expectations and because

$$\mathrm{E}[\partial_t^k \psi(y_i, d_i, h_0(d_i, z_i))|d_i, z_i] = 0 \ \ \forall m \in \mathbb{N}^3 : |k| = 1,$$

we have that

$$iii_a = 0.$$

Moreover, uniformly for any $h \in \mathcal{H}_n$ we have that

$$|iii_b| \lesssim \sqrt{n}\|h - h_0\|_{\mathrm{P},2}^2 \lesssim \sqrt{n}(\delta_n n^{-1/4})^2 \leqslant \delta_n^2,$$

$$|iii_c| \lesssim \sqrt{n}\|h - h_0\|_{\mathrm{P},2}^2\|h - h_0\|_{\mathrm{P},\infty} \lesssim \sqrt{n}(\delta_n n^{-1/4})^2\delta_n \leqslant \delta_n^3.$$

Since $\widehat{h} \in \mathcal{H}_n$ with probability $1 - \Delta_n$, we have that once $n \geqslant n_0$,

$$\mathrm{P}(|iii| \lesssim \delta_n^2) \geqslant 1 - \Delta_n.$$

(e). Furthermore, we have that

$$|ii| \leqslant \sup_{h \in \mathcal{H}_n} |\mathbb{G}_n[f_h] - \mathbb{G}_n[f_{h_0}]|.$$

The class of functions $\mathcal{G}_d$ for $d \in \{0,1\}$ is a union of at most $\binom{p}{Cs}$ VC-subgraph classes of functions with VC indices bounded by $C's$. The class of functions $\mathcal{M}$ is a union of at most $\binom{p}{Cs}$ VC-subgraph classes of functions with VC indices bounded by $C's$ (monotone transformation Λ

preserve the VC-subgraph property). These classes are uniformly bounded and their entropies therefore satisfy

$$\log N(\varepsilon, \mathcal{M}, \|\cdot\|_{\mathbb{P}_n,2}) + \log N(\varepsilon, \mathcal{G}_0, \|\cdot\|_{\mathbb{P}_n,2}) + \log N(\varepsilon, \mathcal{G}_1, \|\cdot\|_{\mathbb{P}_n,2}) \lesssim s \log p + s \log(1/\varepsilon).$$

Finally, the class $\mathcal{F}_n = \{f_h - f_{h_0} : h \in \mathcal{H}_n\}$ is a Lipschitz transform of $\mathcal{H}_n$ with bounded Lipschitz coefficients and with a constant envelope. Therefore, we have that

$$\log N(\varepsilon, \mathcal{F}_n, \|\cdot\|_{\mathbb{P}_n,2}) \lesssim s \log p + s \log(1/\varepsilon).$$

We shall invoke the following lemma derived in Belloni and Chernozhukov (2011).

Lemma 1 (A Self-Normalized Maximal inequality)**.** *Let $\mathcal{F}$ be a measurable function class on a sample space. Let $F = \sup_{f\in\mathcal{F}} |f|$, and suppose that there exist some constants $\omega_n > 3$ and $\upsilon > 1$, such hat*

$$\log N(\epsilon \|F\|_{\mathbb{P}_n,2}, \mathcal{F}, \|\cdot\|_{\mathbb{P}_n,2}) \leqslant \upsilon m (\log(n \vee \omega_n) + \log(1/\epsilon)), \ \ 0 < \epsilon < 1.$$

Then for every $\delta \in (0, 1/6)$ we have

$$\sup_{f\in\mathcal{F}} |\mathbb{G}_n(f)| \leqslant C_\upsilon \sqrt{2/\delta} \sqrt{m \log(n \vee \omega_n)} (\sup_{f\in\mathcal{F}} \|f\|_{\mathrm{P},2} \vee \sup_{f\in\mathcal{F}} \|f\|_{\mathbb{P}_n,2}),$$

with probability at least $1-\delta$ for some constant that C_υ.

Then by Lemma 1 together and some simple calculations, we have that

$$\begin{aligned} |ii| &\leqslant \sup_{f\in\mathcal{F}_n} |\mathbb{G}_n(f)| = O_{\mathrm{P}}(1)\sqrt{s \log(p \vee n)} (\sup_{f\in\mathcal{F}_n} \|f\|_{\mathbb{P}_n,2} \vee \sup_{f\in\mathcal{F}_n} \|f\|_{\mathrm{P},2}) \\ &\leqslant O_{\mathrm{P}}(1)\sqrt{s \log(p \vee n)} (\sup_{h\in\mathcal{H}_n} \|h - h_0\|_{\mathbb{P}_n,2} \vee \sup_{h\in\mathcal{H}_n} \|h - h_0\|_{\mathrm{P},2}) = o_{\mathrm{P}}(1). \end{aligned}$$

The last conclusion follows because $\sup_{h\in\mathcal{H}_n} \|h - h_0\|_{\mathrm{P},2} \lesssim \delta_n n^{-1/4}$ by definition of the norm, and

$$\sup_{h\in\mathcal{H}_n} \|h - h_0\|_{\mathbb{P}_n,2} \qquad \leqslant O_{\mathrm{P}}(1) \cdot \left(\sup_{h\in\mathcal{H}_n} \|h - h_0\|_{\mathrm{P},2} + \|r_{gi}\|_{\mathrm{P},2} + \|r_{mi}\|_{\mathrm{P},2} \right),$$

where the last conclusion follows from the same argument as in step (b) but in a reverse order, switching from empirical norms to population norms, using equivalence of norms over sparse sets imposed in condition (ii) , and also using an application of Markov inequality to argue that $\|r_{gi}\|_{\mathbb{P}_n,2} + \|r_{mi}\|_{\mathbb{P}_n,2} = O_{\mathrm{P}}(1)(\|r_{gi}\|_{\mathrm{P},2} + \|r_{mi}\|_{\mathrm{P},2})$.

Step 2. Claim (2) follows from consistency: $\widehat{\sigma}_n/\sigma_n = 1 + o_{\mathrm{P}}(1)$, which follows from $\widehat{\sigma}_n$ being a Lipschitz transform of $\widehat{h}$ with respect to $\|\cdot\|_{\mathbb{P}_n,2}$, once $\widehat{h} \in \mathcal{H}_n$ and the consistency of $\widehat{h}$ for h under $\|\cdot\|_{\mathbb{P}_n,2}$.

Step 3. Claim (3) is immediate from claims (2) and (3) by the way of contradiction. $\square$

3. Deferred Proofs: Proof of Lemma 1

We establish the result for Lasso (the proof for other feasible Lasso estimators is similar).

By Lemma 7 in Belloni, Chen, Chernozhukov, and Hansen (2012), under our choice of penalty level and loadings, we have that the condition $\lambda/n \geqslant 2c\|\widehat{\Psi}^{-1}\mathbb{E}_n[\tilde{x}_i\epsilon_i]\|_\infty$ holds with probability $1-o(1)$. Thus, the conclusion of Lemma 11 of Belloni, Chen, Chernozhukov, and Hansen (2012) holds with probability $1-o(1)$, namely for $c_s = (\mathbb{E}_n[r_i^2])^{1/2}$

$$\widehat{s} \leqslant s + \left(\min_{m\in\mathcal{H}} \phi_{\max}(m)\right) \|\widehat{\Psi}^{-1}\|_\infty \left(\frac{2\bar{c}}{\kappa_{\bar{c}}} + \frac{4\bar{c}nc_s}{\lambda\sqrt{s}}\right)^2 \tag{3.27}$$

where $\bar{c} = (c+1)/(c-1)$,

$$\mathcal{H} = \left\{ m \in \mathbf{N} : m \geqslant 2s\phi_{\max}(m)\|\widehat{\Psi}^{-1}\|_\infty \left(\frac{2\bar{c}}{\kappa_{\bar{c}}} + \frac{4\bar{c}nc_s}{\lambda\sqrt{s}}\right)^2 \right\},$$

$$\kappa_{\bar{c}} \geqslant \max_{m\in\mathbf{N}} \frac{\sqrt{\phi_{\min}(m+s)}}{\|\widehat{\Psi}\|_\infty} \left(1 - \sqrt{\frac{\phi_{\max}(m+s)}{\phi_{\min}(m+s)}}\bar{c}\sqrt{s/m}\right).$$

By Condition SE, with probability $1-o(1)$ for n sufficiently large we have $\kappa_{\bar{c}} > \kappa'/2\|\widehat{\Psi}\|_\infty$ so that with the same probability

$$\frac{2\bar{c}}{\kappa_{\bar{c}}} \lesssim 1. \tag{3.28}$$

Moreover, by condition RF we have with probability $1-o(1)$ that

$$\max\{\|\widehat{\Psi}\|_\infty,\ \|\widehat{\Psi}^{-1}\|_\infty\} \lesssim 1. \tag{3.29}$$

Finally, since $\lambda \gtrsim \sqrt{n\log(p\vee n)}$ we have

$$\frac{4\bar{c}nc_s}{\lambda\sqrt{s}} \lesssim \frac{\sqrt{n}c_s}{\sqrt{s\log(p\vee n)}} \lesssim 1 \quad \text{with probability } 1-o(1) \tag{3.30}$$

since $c_s \lesssim_P \sqrt{s/n}$ by condition ASM and Chebyshev inequality.

Therefore, for some constant $\tilde{C}$, we have $\tilde{C}s \in \mathcal{H}$, so that $\min_{m\in\mathcal{H}} \phi_{\max}(m) \leqslant \kappa''$ for n sufficiently large with probability $1-o(1)$ by Condition SE. In turn combining this bound with (3.28), (3.29) and (3.30) into (3.27) we have that $\widehat{s} \lesssim s$ holds with probability $1-o(1)$ which is the first statement of (i).

To show the second statement in (i), note that

$$\min_{\beta\in\mathbb{R}^p:\ \beta_j=0\ \forall j\notin\widehat{T}} \sqrt{\mathbb{E}_n[f(\tilde{z}_i)-\tilde{x}_i'\beta]^2} \leqslant \sqrt{\mathbb{E}_n[f(\tilde{z}_i)-\tilde{x}_i'\widehat{\beta}]^2}$$

where $\widehat{\beta}$ is the Lasso estimator. Again by Lemma 7 in Belloni, Chen, Chernozhukov, and Hansen (2012) we have that the assumptions of Lemma 6 in Belloni, Chen, Chernozhukov, and Hansen (2012) hold with probability $1-o(1)$. Using Condition SE to bound $\kappa_{\bar{c}}$ from below and Condition RF to bound $\|\widehat{\Psi}\|_\infty$ from above with probability $1-o(1)$ as before, and $\lambda \lesssim \sigma\sqrt{n\log(p\vee n)}$, it follows from Lemma 6 in Belloni, Chen, Chernozhukov, and Hansen (2012) that with probability $1-o(1)$ that

$$\sqrt{\mathbb{E}_n[f(\tilde{z}_i)-\tilde{x}_i'\widehat{\beta}]^2} \lesssim \sigma\sqrt{\frac{s\log(p\vee n)}{n}}.$$

The results regarding Post-Lasso in (ii) follow similarly by invoking Lemma 8 in Belloni, Chen, Chernozhukov, and Hansen (2012).

4. Split-Sample Estimation and Inference

In this section we discuss a variant of the double selection estimator based on sample splitting. The motivation for the split-sample estimator is that its use allows us to relax the requirement $s^2\log^2(p\vee n)=o(n)$ that is assumed in the full-sample counterpart to the milder condition

$$s\log(p\vee n)=o(n).$$

To define the estimator, divide the sample randomly into (approximately) equal parts a and b with sizes $n_a=\lceil n/2\rceil$ and $n_b=n-n_a$. We use superscripts a and b for variables in the first and second subsample respectively. We let the index $k=a,b$ refer to one of the subsamples and let $k^c=\{a,b\}\setminus\{k\}$ refer to the other.

For each subsample $k=a,b$, the model $\widehat{I}^k$ is selected based on the subsample k independently from the subsample k^c. In what follows the model $\widehat{I}^k$ is used to fit the subsample k^c. A constructive way to obtain $\widehat{I}^a$ and $\widehat{I}^b$ is to apply the double selection method for each subsample to select the sets of controls $\widehat{I}^a := \widehat{I}^a_1\cup\widehat{I}^a_2\cup\widehat{I}^a_3$ and $\widehat{I}^b := \widehat{I}^b_2\cup\widehat{I}^b_2\cup\widehat{I}^b_3$.

Then we form estimates in the two subsamples

$$(\check{\alpha}^a,\check{\beta}^a) = \underset{\alpha\in\mathbb{R},\beta\in\mathbb{R}^p}{\operatorname{argmin}}\{\mathbb{E}_{n_a}[(y_i-d_i\alpha-x_i'\beta)^2]\ :\ \beta_j=0,\forall j\notin\widehat{I}^b\}, \text{and}$$

$$(\check{\alpha}^b,\check{\beta}^b) = \underset{\alpha\in\mathbb{R},\beta\in\mathbb{R}^p}{\operatorname{argmin}}\{\mathbb{E}_{n_b}[(y_i-d_i\alpha-x_i'\beta)^2]\ :\ \beta_j=0,\forall j\notin\widehat{I}^a\}.$$

For an index i in the subsample k, we define the residuals

$$\widehat{\zeta}^o_i := [y_i - d_i\check{\alpha}_k - x_i'\check{\beta}_k]\{n_k/(n_k - \widehat{s}^{k^c} - 1)\}^{1/2} \tag{4.31}$$

$$\widehat{v}_i := d_i - x_i'\widehat{\beta}_k \text{ and} \tag{4.32}$$

$$\widehat{\zeta}_i := \widehat{\zeta}^o_i \; 1\{|\widehat{\zeta}^o_i| \vee |\widehat{v}_i| \leqslant Cn^{1/2}/[(\widehat{s}^{k^c} \vee n^{1/2})\log n]^{1/2}\} \tag{4.33}$$

where $\widehat{\beta}_k \in \arg\min_\beta\{\mathbb{E}_{n_k}[(d_i - x_i'\beta)^2] : \beta_j = 0, \forall j \notin \widehat{I}^{k^c}\}$ and $\widehat{s}^{k^c} = |\widehat{I}^{k^c}|$.

Finally, we combine the estimates into the split-sample estimator based on $\widehat{I}^a$ and $\widehat{I}^b$ is defined as

$$\check{\alpha}_{ab} = \{(n_a/n)\Upsilon^a + (n_b/n)\Upsilon^b\}^{-1}\{(n_a/n)\Upsilon^a\check{\alpha}_a + (n_b/n)\Upsilon^b\check{\alpha}_b\}, \tag{4.34}$$

where $\Upsilon^k = D^{k\prime}\mathcal{M}_{\widehat{I}^{k^c}}D^k/n_k$.

We state below sufficient conditions for the analysis of the split-sample method.

Condition ASTESS (P). *(i) $\{(y_i, d_i, z_i), i = 1, \ldots, n\}$ are i.n.i.d. vectors on $(\Omega, \mathcal{F}, \mathrm{P})$ that obey the model (2.2)-(2.3), and the vector $x_i = P(z_i)$ is a dictionary of transformations of z_i, which may depend on n but not on* P. *(ii) The true parameter value α_0, which may depend on* P, *is bounded, $\|\alpha_0\| \leqslant C$. (iii) Functions m and g admit an approximately sparse form. Namely there exists $s \geqslant 1$ and β_{m0} and β_{g0}, which depend on n and* P, *such that*

$$m(z_i) = x_i'\beta_{m0} + r_{mi}, \quad \|\beta_{m0}\|_0 \leqslant s, \quad \{\bar{\mathrm{E}}[r_{mi}^2]\}^{1/2} \leqslant C\sqrt{s/n}, \tag{4.35}$$

$$g(z_i) = x_i'\beta_{g0} + r_{gi}, \quad \|\beta_{g0}\|_0 \leqslant s, \quad \{\bar{\mathrm{E}}[r_{gi}^2]\}^{1/2} \leqslant C\sqrt{s/n}. \tag{4.36}$$

(iv) The sparsity index obeys $s\log(p \vee n)/n \leqslant C\delta_n$. (v) For each subsample $k = a, b$, the model $\widehat{I}^{k^c}$ satisfies condition HLMS. (vi) We have $\bar{\mathrm{E}}[|v_i^q| + |\zeta_i^q|] \leqslant C$ for some $q > 4$ and $n^{2/q}s\log(p \vee n)/n \leqslant C\delta_n$.

The Conditions ASTESS(i)-(iii) agree with the corresponding conditions in ASTE. The remaining conditions ASTESS(iv)-(v) are implied by Condition ASTE. We note that Condition ASTESS(vi) is needed only for obtaining consistent estimates of the asymptotic variance. Such conditions are mild since they do not require uniform estimation of the functions g and m.

The next result establishes that the split-sample estimator $\check{\alpha}_{ab}$ has similar large sample properties to the full-sample double-selection estimator under weaker growth condition.

Theorem 3 (Inference on Treatment Effects, Split Sample)**.** *Let $\{\mathrm{P}_n\}$ be a sequence of data-generating processes. Assume conditions ASTESS*(P)*(i-v), SM*(P)*, and SE*(P) *hold for $\mathbb{P}_n = \mathrm{P}_n$ for each n and each subsample. The split sample estimator $\check{\alpha}_{ab}$ based on $\widehat{I}^a$ and $\widehat{I}^b$ obeys,*

$$([\bar{\mathrm{E}}v_i^2]^{-1}\bar{\mathrm{E}}[v_i^2\zeta_i^2][\bar{\mathrm{E}}v_i^2]^{-1})^{-1/2}\sqrt{n}(\check{\alpha}_{ab} - \alpha_0) \rightsquigarrow N(0,1).$$

Moreover, if Condition ASTESS(P*)(vi) also holds, the result continues to apply if* $\bar{\mathrm{E}}[v_i^2]$ *and* $\bar{\mathrm{E}}[v_i^2\zeta_i^2]$ *are replaced by* $\mathbb{E}_n[\widehat{v}_i^2]$ *and* $\mathbb{E}_n[\widehat{v}_i^2\widehat{\zeta}_i^2]$ *for* $\widehat{\zeta}_i$ *and* $\widehat{v}_i$ *defined in (4.33) and (4.32).*

Proof. We use the same notation as in the proof of Theorem 1 with the addition of sub/superscripts indicating the appropriate subsample $k = a, b$, where $k^c = \{a, b\} \setminus \{k\}$.

Step 0.(Combining) In this step we combine both subsample estimators. Letting $\Upsilon^k = D^{k\prime}\mathcal{M}_{\widehat{I}^{k^c}} D^k/n_k$, for $k = a, b$, so that we have

$$\begin{aligned}
\sqrt{n}(\check{\alpha}_{ab} - \alpha_0) &= ((n_a/n)\Upsilon^a + (n_b/n)\Upsilon^b)^{-1} \times \\
&\times \ ((n_a/n)\Upsilon^a\sqrt{n}(\check{\alpha}_a - \alpha_0) + (n_b/n)\Upsilon^b\sqrt{n}(\check{\alpha}_b - \alpha_0)) \\
&= (V'V/n + o_P(1))^{-1} \times \\
&\times \ ((n_a/n)\Upsilon^a\sqrt{n}(\check{\alpha}_a - \alpha_0) + (n_b/n)\Upsilon^b\sqrt{n}(\check{\alpha}_b - \alpha_0)) + o_P(1) \\
&= \{V'V/n\}^{-1} \times \{(1/\sqrt{2}) \times \mathbb{G}_{na}[v_i\zeta_i] + (1/\sqrt{2})\mathbb{G}_{nb}[v_i\zeta_i]\} + o_P(1) \\
&= \{V'V/n\}^{-1} \times \mathbb{G}_n[v_i\zeta_i] + o_P(1)
\end{aligned}$$

where we are also using the fact that

$$\mathbb{E}_{n_k}[\widehat{v}_i^2] - \mathbb{E}_{n_k}[v_i^2] = o_P(1), \quad k = a, b$$

which follows similarly to the proofs given in Step 5.

For $\sigma_n^2 := [\bar{\mathrm{E}}v_i^2]^{-1}\bar{\mathrm{E}}[v_i^2\zeta_i^2][\bar{\mathrm{E}}v_i^2]^{-1}$, define

$$Z_n = \sigma_n^{-1}\sqrt{n}(\check{\alpha}_{ab} - \alpha_0) = \mathbb{G}_n[z_{i,n}] + o_{\mathrm{P}}(1),$$

where $z_{i,n} = \sigma_n^{-1}v_i\zeta_i/\sqrt{n}$ are i.n.i.d. with mean zero. We have that for some small enough $\delta > 0$

$$\bar{\mathrm{E}}|z_{i,n}|^{2+\delta} \lesssim \bar{\mathrm{E}}\left[|v_i|^{2+\delta}|\zeta_i|^{2+\delta}\right] \lesssim \sqrt{\bar{\mathrm{E}}|v_i|^{4+2\delta}}\sqrt{\bar{\mathrm{E}}|\zeta_i|^{4+2\delta}} \lesssim 1,$$

by Condition SM(ii).

This condition verifies the Lyapunov condition and thus implies that $Z_n \to_d N(0, 1)$.

Step 1.(Main) For the subsample $k = a, b$ write $\check{\alpha}_k = \left[D^{k\prime}\mathcal{M}_{\widehat{I}^{k^c}} D^k/n_k\right]^{-1}[D^{k\prime}\mathcal{M}_{\widehat{I}^{k^c}} Y^k/n_k]$ so that

$$\sqrt{n_k}(\check{\alpha}_k - \alpha_0) = \left[D^{k\prime}\mathcal{M}_{\widehat{I}^{k^c}} D^k/n_k\right]^{-1}[D^{k\prime}\mathcal{M}_{\widehat{I}^{k^c}}(g^k + \zeta^k)/\sqrt{n_k}] =: ii_k^{-1} \cdot i_k.$$

By Steps 2 and 3, $ii_k = V^{k\prime}V^k/n_k + o_P(1)$ and $i_k = V^{k\prime}\zeta^k/\sqrt{n_k} + o_P(1)$. Next note that $V^{k\prime}V^k/n_k = \mathrm{E}[V^{k\prime}V^k/n_k] + o_P(1)$ by Chebyshev, and we have that $\bar{\mathrm{E}}_k[v_i^2\zeta_i^2]$ and $\mathrm{E}[V^{k\prime}V^k/n_k]$ are bounded from above and away from zero by assumption.

Step 2. (Behavior of i_k.) Decompose

$$i_k = V^{k\prime}\zeta^k/\sqrt{n_k} + \underbrace{m^{k\prime}\mathcal{M}_{\widehat{I}^{k^c}}g^k/\sqrt{n_k}}_{=:i_{k,a}} + \underbrace{m^{k\prime}\mathcal{M}_{\widehat{I}^{k^c}}\zeta^k/\sqrt{n_k}}_{=:i_{k,b}} + \underbrace{V^{k\prime}\mathcal{M}_{\widehat{I}^{k^c}}g^k/\sqrt{n_k}}_{=:i_{k,c}} - \underbrace{V^{k\prime}\mathcal{P}_{\widehat{I}^{k^c}}\zeta^k/\sqrt{n_k}}_{=:i_{k,d}}.$$

First, note that by Condition ASTESS we have

$$|i_{k,a}| = |m^{k\prime}\mathcal{M}_{\widehat{I}^{k^c}}g^k/\sqrt{n_k}| \leqslant \|\mathcal{M}_{\widehat{I}^{k^c}}m^k\| \; \|\mathcal{M}_{\widehat{I}^{k^c}}g^k\|/\sqrt{n_k} = o_P(1).$$

Second, by the split sample construction, we have that $\widehat{I}^{k^c}$ is independent from ζ^k, and by assumption of the model m^k is also independent of ζ^k. Thus by Chebyshev inequality

$$|i_{k,b}| \lesssim_P \|\mathcal{M}_{\widehat{I}^{k^c}}m^k/\sqrt{n_k}\| = o_P(1),$$

where the last relation follows by ASTESS.

Third, using similar independence arguments, by Chebyshev and Condition ASTESS, conclude

$$|i_{k,c}| \lesssim_P \|\mathcal{M}_{\widehat{I}^{k^c}}g^k/\sqrt{n_k}\| = o_P(1).$$

Fourth, using that $\widehat{s}^{k^c} \lesssim_P s$ by ASTESS so that $\phi^{-1}_{\min}(\widehat{s}^{k^c}) \lesssim_P 1$ by condition SE, we have that

$$|i_{k,d}| \leqslant |\tilde{\beta}_{V^k}(\widehat{I}^{k^c})' X^{k\prime}\zeta^k/\sqrt{n_k}| \lesssim_P \sqrt{s/n} = o_P(1),$$

by Chebyshev since $\|X^k\tilde{\beta}_{V^k}(\widehat{I}^{k^c})/\sqrt{n_k}\| \lesssim_P \sqrt{s/n_k}$ because of the independence of the two subsamples k and k^c.

Step 3.(Behavior of ii_k.) Since $ii_k = (m^k + V^k)'\mathcal{M}_{\widehat{I}^{k^c}}(m^k + V^k)/n_k$, decompose

$$ii_k = V^{k\prime}V^k/n_k + \underbrace{m^{k\prime}\mathcal{M}_{\widehat{I}^{k^c}}m^k/n_k}_{=:ii_{k,a}} + \underbrace{2m^{k\prime}\mathcal{M}_{\widehat{I}^{k^c}}V^k/n_k}_{=:ii_{k,b}} - \underbrace{V^{k\prime}\mathcal{P}_{\widehat{I}^{k^c}}V^k/n_k}_{=:ii_{k,c}}.$$

Then $|ii_{k,a}| = o_P(1)$ by Condition ASTESS, $|ii_{k,b}| = o_P(1)$ by reasoning similar to deriving the bound for $|i_{k,b}|$, and $|ii_{k,c}| = o_P(1)$ by reasoning similar to deriving the bound for $|i_{k,d}|$.

Step 4.(Auxiliary Bounds.) Note that

$$\begin{aligned}\|g^k - X^k\check{\beta}_k\| &= \|g^k - \mathcal{P}_{\widehat{I}^{k^c}}(Y^k - D^k\check{\alpha}_k)\| \\ &\leqslant \|\mathcal{M}_{\widehat{I}^{k^c}}g^k\| + |\check{\alpha}_k - \alpha_0|\|\mathcal{P}_{\widehat{I}^{k^c}}D^k\| + \|\mathcal{P}_{\widehat{I}^{k^c}}\zeta^k\|.\end{aligned}$$

By condition ASTESS $\|\mathcal{M}_{\widehat{I}^{k^c}}g^k\| = o_P(n^{1/4})$ and by condition SM(ii) we have $\|\mathcal{P}_{\widehat{I}^{k^c}}D^k/\sqrt{n_k}\| \leqslant \|D^k/\sqrt{n_k}\| \lesssim_P 1$, and by Step 1 we have $|\check\alpha_k - \alpha_0| \lesssim_P n^{-1/2}$. Moreover,

$$\begin{aligned}\|\mathcal{P}_{\widehat{I}^{k^c}}\zeta^k\| &= \|X^k[\widehat{I}^{k^c}](X^k[\widehat{I}^{k^c}]'X^k[\widehat{I}^{k^c}])^{-1}X^k[\widehat{I}^{k^c}]'\zeta^k\|\\ &\leqslant [\sqrt{\phi_{\max,k}(\widehat{s}^{k^c})}/\phi_{\min,k}(\widehat{s}^{k^c})]\|X^k[\widehat{I}^{k^c}]'\zeta^k/\sqrt{n_k}\|.\end{aligned}$$

We have $\sqrt{\phi_{\max,k}(\widehat{s}^{k^c})}/\phi_{\min,k}(\widehat{s}^{k^c}) \lesssim_P 1$ by condition SE, and $\|X^k[I^{k^c}]'\zeta^k/\sqrt{n_k}\| \lesssim_P \sqrt{\widehat{s}^{k^c}}$ by condition SM(ii), the independence between the selected components $\widehat{I}^{k^c}$ and ζ^k since they are based on different subsamples, and applying Chebyshev inequality.

Finally, collecting terms we have

$$\|g^k - X^k\check\beta_k\|/\sqrt{n^k} \lesssim_P o(n^{-1/4}) + \sqrt{\widehat{s}^{k^c}/n_k}$$

Similarly, we have $\|m^k - X^k\widehat\beta_k\|/\sqrt{n^k} \lesssim_P o(n^{-1/4}) + \sqrt{\widehat{s}^{k^c}/n_k}$.

Step 5.(Variance Estimation.) Since $\widehat{s}^k \lesssim_P s = o(n)$, $(n_k - \widehat{s}^k - 1)/n_k = o_P(1)$, so we can use n as the denominator. Recall the definitions $\widehat\zeta_i^o = y_i - d_i\check\alpha_k - x_i'\check\beta_k$, $\widehat v_i = d_i - x_i'\widehat\beta_k$ and $\widehat\zeta_i = \widehat\zeta_i^o 1\{|\widehat\zeta_i^o| \vee |\widehat v_i| \leqslant H_k\}$ if i belongs to subsample k where $H_k = C\sqrt{n/[(\widehat{s}^{k^c} \vee n^{1/2})\log n]}$. For notational convenience let $A_i = \{|\widehat\zeta_i^o| \vee |\widehat v_i| \leqslant H_k\}$. Since $q > 4$, $\widehat{s}^{k^c} \lesssim_P s$, and $n^{2/q}s\log(n \vee p) = o(n)$, we have $n^{1/q} = o_P(H_k)$. Hence consider

$$\begin{aligned}\mathbb{E}_n[\widehat v_i^2] &= (n_a/n)D^{a\prime}\mathcal{M}_{\widehat{I}^b}D^a/n_a + (n_b/n)D^{b\prime}\mathcal{M}_{\widehat{I}^a}D^b/n_b\\ &= (n_a/n)ii_a + (n_b/n)ii_b = V'V/n + o_P(1) = \bar{\mathrm{E}}[v_i^2] + o_P(1)\end{aligned}$$

by Step 3 and $\bar{\mathrm{E}}[|v_i|^q] \lesssim 1$ for some $q > 4$ by condition SM(ii).

By Condition ASTESS(vi), for each subsample $k = a, b$, we have $\mathbb{E}_{n_k}[v_i^2\zeta_i^2] - \bar{\mathrm{E}}_k[v_i^2\zeta_i^2] \to_P 0$ by Vonbahr-Esseen's inequality in von Bahr and Esseen (1965) since $\bar{\mathrm{E}}_k[|v_i\zeta_i|^{2+\delta}] \leqslant (\bar{\mathrm{E}}_k[|v_i|^{4+2\delta}]\bar{\mathrm{E}}_k[|\,\zeta_i|^{4+2\delta}])^{1/2}$ is uniformly bounded for $4 + 2\delta \leqslant q$. Thus it suffices to show that $\mathbb{E}_{n_k}[\widehat v_i^2\widehat\zeta_i^2] - \mathbb{E}_{n_k}[v_i^2\zeta_i^2] \to_P 0$. By the triangular inequality

$$\begin{aligned}|\mathbb{E}_{n_k}[\widehat v_i^2\widehat\zeta_i^2 - v_i^2\zeta_i^2]| &\leqslant |\mathbb{E}_{n_k}[(\widehat v_i^2\widehat\zeta_i^2 - v_i^2\zeta_i^2)1\{A_i\}]| + |\mathbb{E}_{n_k}[(\widehat v_i^2\widehat\zeta_i^2 - v_i^2\zeta_i^2)1\{A_i^c\}]|\\ &\leqslant |\mathbb{E}_{n_k}[(\widehat v_i^2 - v_i^2)\zeta_i^2 1\{A_i\}]| + |\mathbb{E}_{n_k}[v_i^2(\widehat\zeta_i^2 - \zeta_i^2)1\{A_i\}]| +\\ &\quad + |\mathbb{E}_{n_k}[(\widehat v_i^2 - v_i^2)(\widehat\zeta_i^2 - \zeta_i^2)1\{A_i\}]| + o_P(1)\end{aligned}$$

since $|\mathbb{E}_{n_k}[(\widehat v_i^2\widehat\zeta_i^2 - v_i^2\zeta_i^2)1\{A_i^c\}]| = o_P(1)$ by Step 6. Then,

$$\begin{aligned}|\mathbb{E}_{n_k}[v_i^2(\widehat\zeta_i^2 - \zeta_i^2)1\{A_i\}]| &\leqslant 2\underbrace{\mathbb{E}_{n_k}[\{d_i(\alpha_0 - \check\alpha_k)\}^2 v_i^2]}_{=:iii_1} + 2\underbrace{\mathbb{E}_{n_k}[\{x_i'\check\beta_k - g_i\}^2 v_i^2]}_{=:iii_2}\\ &\quad + 2\underbrace{\max_{i\leqslant n}|v_i|\{\mathbb{E}_{n_k}[\zeta_i^2 v_i^2]\}^{1/2}\{\mathbb{E}_{n_k}[d_i^2(\alpha_0 - \check\alpha_k)^2]\}^{1/2}}_{=:iii_3}\\ &\quad + 2\underbrace{\max_{i\leqslant n}|v_i|\{\mathbb{E}_{n_k}[\zeta_i^2 v_i^2]\}^{1/2}\{\mathbb{E}_{n_k}[(g_i - x_i'\check\beta_k)^2]\}^{1/2}}_{=:iii_4}\end{aligned}$$

As a consequence of Condition SM(ii) we have $\mathrm{E}[\max_{i\leqslant n} d_i^2] \lesssim n^{2/q}$, $\mathrm{E}[\max_{i\leqslant n} \zeta_i^2] \lesssim n^{2/q}$, $\mathrm{E}[\max_{i\leqslant n} v_i^2] \lesssim n^{2/q}$, thus by Markov inequality we have $\max_{i\leqslant n} |d_i| + |\zeta_i| + |v_i| \lesssim_P n^{1/q}$.

We have the following relations:

$$
\begin{aligned}
iii_1 &\leqslant |\alpha_0 - \check{\alpha}_k|^2 \mathbb{E}_{n_k}[d_i^2] \max_{i\leqslant n} v_i^2 \lesssim_P n^{-1} n^{2/q} = o_P(1),\\
iii_2 &\leqslant \max_{i\leqslant n} v_i^2 \mathbb{E}_{n_k}[\{x_i'\check{\beta}_k - g_i\}^2] \lesssim_P n^{2/q}\{o(n^{-1/4}) + \sqrt{\widehat{s}^{k^c}/n}\}^2 = o_P(1),\\
iii_3 &\lesssim_P n^{1/q}\sqrt{1/n} = o_P(1),\\
iii_4 &\lesssim_P n^{1/q}\{o(n^{-1/4}) + \sqrt{\widehat{s}^{k^c}/n}\} = o_P(1),
\end{aligned}
$$

since $\mathbb{E}_{n_k}[\zeta_i^2 v_i^2] \lesssim_P 1$, $\mathbb{E}_{n_k}[\{x_i'\check{\beta}_k - g_i\}^2] \lesssim_P \{o(n^{-1/4}) + \sqrt{\widehat{s}^{k^c}/n}\}^2$ by Step 4, $\widehat{s}^{k^c} \lesssim_P s$, and $|\check{\alpha}_k - \alpha_0|^2 \lesssim_P 1/n$ by Step 1.

Similarly, $\mathbb{E}_{n_k}[(\widehat{v}_i^2 - v_i^2)\zeta_i^2] = o_P(1)$.

Finally, since $\max_{i\leqslant n} \|1\{A_i\}(\widehat{v}_i, \widehat{\zeta}_i, \zeta_i, v_i)'\|_\infty^2 \lesssim_P (H_k^2 \vee n^{2/q}) \lesssim_P H_k^2$, we have

$$
\begin{aligned}
|\mathbb{E}_{n_k}[(\widehat{v}_i^2 - v_i^2)(\widehat{\zeta}_i^2 - \zeta_i^2)1\{A_i\}]| &\leqslant \{\mathbb{E}_{n_k}[(\widehat{v}_i^2 - v_i^2)^2 1\{A_i\}]\mathbb{E}_{n_k}[(\widehat{\zeta}_i^2 - \zeta_i^2)^2 1\{A_i\}]\}^{1/2}\\
&\leqslant \{\mathbb{E}_{n_k}[2(\widehat{v}_i^2 + v_i^2)(\widehat{v}_i - v_i)^2 1\{A_i\}]\mathbb{E}_{n_k}[2(\widehat{\zeta}_i^2 + \zeta_i^2)(\widehat{\zeta}_i - \zeta_i)^2 1\{A_i\}]\}^{1/2}\\
&\lesssim_P (H_k^2 \vee n^{2/q})\{\mathbb{E}_{n_k}[(\widehat{v}_i - v_i)^2]\mathbb{E}_{n_k}[(\widehat{\zeta}_i - \zeta_i)^2]\}^{1/2}\\
&\lesssim_P H_k^2\{o(n^{-1/4}) + \sqrt{\widehat{s}^{k^c}/n}\}^2\\
&\lesssim \tfrac{n}{(\widehat{s}^{k^c} \vee n^{1/2})\log n}\{o(n^{-1/2}) + \widehat{s}^{k^c}/n\} = o(1).
\end{aligned}
$$

Step 6.(Controlling large terms) By definition of the event A_i we have

$$
\begin{aligned}
H_k^2 \mathbb{E}_{n_k}[1\{A_i^c\}] &\leqslant \mathbb{E}_{n_k}[\widehat{\zeta}_i^{o2} 1\{A_i^c\}]\\
&\leqslant 4\mathbb{E}_{n_k}[\zeta_i^2 1\{A_i^c\}] + 4\mathbb{E}_{n_k}[d_i^2(\check{\alpha}_k - \alpha_0)^2 1\{A_i^c\}] + 4\mathbb{E}_{n_k}[\{x_i'\check{\beta}_k - g_i\}^2 1\{A_i^c\}]\\
&\lesssim_P n^{2/q}\mathbb{E}_{n_k}[1\{A_i^c\}] + n^{2/q-1}\mathbb{E}_{n_k}[1\{A_i^c\}] + \mathbb{E}_{n_k}[\{x_i'\check{\beta}_k - g_i\}^2].
\end{aligned}
$$

Since $n^{1/q} = o_P(H_k)$, and $\mathbb{E}_{n_k}[\{x_i'\check{\beta}_k - g_i\}^2] \lesssim_P o(n^{-1/2}) + \widehat{s}^{k^c}/n$, we have

$$
\mathbb{E}_{n_k}[1\{A_i^c\}] \lesssim_P \{o(n^{-1/2}) + \widehat{s}^{k^c}/n\}/H_k^2.
$$

Therefore,

$$
\mathbb{E}_{n_k}[\tilde{\zeta}_i^2 \tilde{v}_i^2 1\{A_i^c\}] \lesssim_P n^{4/q}\mathbb{E}_{n_k}[1\{A_i^c\}] \lesssim_P n^{4/q}\{o(n^{-1/2}) + \widehat{s}^{k^c}/n\}/H_k^2.
$$

Finally note that

$$
\frac{n^{4/q}\{o(n^{-1/2}) + \widehat{s}^{k^c}/n\}}{H_k^2} \lesssim \frac{n^{2/q}}{n^{1/2}}\frac{n^{2/q}(\widehat{s}^{k^c} \vee n^{1/2})\log n}{n} + \frac{n^{2/q}\widehat{s}^{k^c}\log n}{n}\frac{n^{2/q}(\widehat{s}^{k^c} \vee n^{1/2})}{n} = o_P(1)
$$

since $q > 4$, $\widehat{s}^{k^c} \lesssim_P s$, and $n^{2/q} s \log(n \vee p) = o(n)$ by ASTESS. Also, by construction, we have $\mathbb{E}_{n_k}[\widehat{\zeta}_i^2 \widehat{v}_i^2 1\{A_i^c\}] = 0$.

□

5. Additional Simulation Results

In this section, we present additional simulation results. All of the simulation results are based on the structural model

$$y_i = d_i'\alpha_0 + x_i'(c_y\beta_0) + \sigma_y(d_i, x_i)\zeta_i, \quad \zeta_i \sim N(0,1) \tag{5.37}$$

where $p = \dim(x_i) = 200$, the covariates $x \sim N(0, \Sigma)$ with $\Sigma_{kj} = (0.5)^{|j-k|}$, $\alpha_0 = .5$, and the sample size n is set to 100. In each design, we generate

$$d_i^* = x_i'(c_d\beta_1) + \sigma_d(x_i)v_i, \quad v_i \sim N(0,1) \tag{5.38}$$

with $\mathrm{E}[\zeta_i v_i] = 0$. Inference results for all designs are based on conventional t-tests with standard errors calculated using the heteroscedasticity consistent jackknife variance estimator discussed in MacKinnon and White (1985). We set λ according to the algorithm outlined in Appendix A with $1 - \gamma = .95$. We draw new x's, ζ's and v's at every replication and draw new β_0's and β_1's at every replication in the random coefficient designs.

In the first thirteen designs, $\beta_1 = \beta_0$. We set the constants c_y and c_d to generate desired population values for the reduced form R^2's, i.e. the R^2's for equations (5.37) and (5.38). Let R_y^2 be the desired R^2 for the regression of y on x and R_d^2 be the desired R^2 from the regression of d on x. For each equation, we choose c_y and c_d to generate $R^2 = 0, .2, .4, .6,$ and $.8$. In the heteroscedastic and binary designs discussed below, we choose c_y and c_d based on R^2 as if (5.37) held with $d_i = d_i^*$ and v_i and ζ_i were homoscedastic with variance equal to the average variance and label the results by R^2 as in the other cases. In the homoscedastic cases, we set $\sigma_y = \sigma_d = 1$; and in the heteroscedastic cases, the average of $\sigma_d(x_i)$ and the average of $\sigma_y(d_i, x_i)$ are both one. We set

$$c_d = \sqrt{\frac{R_d^2}{(1 - R_d^2)\beta_0'\Sigma\beta_0}}$$

$$c_y = \frac{-(1 - R_y^2)\alpha_0 c_d \beta_0'\Sigma\beta_0 + \sqrt{(1 - R_y^2)R_y^2\beta_0'\Sigma\beta_0(\alpha_0^2 + 1)}}{(1 - R_y^2)\beta_0'\Sigma\beta_0}$$

- Design 1. $d_i = d_i^*$, $\beta_0 = (1, 1/2, 1/3, 1/4, 1/5, 0, 0, 0, 0, 0, 1, 1/2, 1/3, 1/4, 1/5, 0, ..., 0)'$, $\sigma_y = \sigma_d = 1$.
- Design 2. $d_i = d_i^*$, $\beta_0 = (1, 1/4, 1/9, 1/16, 1/25, 0, 0, 0, 0, 0, 1, 1/4, 1/9, 1/16, 1/25, 0, ..., 0)'$, $\sigma_y = \sigma_d = 1$.
- Design 22. $d_i = d_i^*$, $\beta_{0,j} = (1/j)^2$, $\sigma_y = \sigma_d = 1$.

- Design 3. $d_i = d_i^*$, $\beta_0 = (1, 1/2, 1/3, 1/4, 1/5, 0, 0, 0, 0, 0, 1, 1/2, 1/3, 1/4, 1/5, 0, ..., 0)'$, $\sigma_d = \sqrt{\frac{(1+x_i'\beta_0)^2}{\frac{1}{n}\sum_{i=1}^n (1+x_i'\beta_0)^2}}$, $\sigma_y = \sqrt{\frac{(1+\alpha_0 d_i + x_i'\beta_0)^2}{\frac{1}{n}\sum_{i=1}^n (1+\alpha_0 d_i + x_i'\beta_0)^2}}$.
- Design 4. $d_i = d_i^*$, $\beta_0 = (1, 1/4, 1/9, 1/16, 1/25, 0, 0, 0, 0, 0, 1, 1/4, 1/9, 1/16, 1/25, 0, ..., 0)'$, $\sigma_d = \sqrt{\frac{(1+x_i'\beta_0)^2}{\frac{1}{n}\sum_{i=1}^n (1+x_i'\beta_0)^2}}$, $\sigma_y = \sqrt{\frac{(1+\alpha_0 d_i + x_i'\beta_0)^2}{\frac{1}{n}\sum_{i=1}^n (1+\alpha_0 d_i + x_i'\beta_0)^2}}$.
- Design 44. $d_i = d_i^*$, $\beta_{0,j} = (1/j)^2$, $\sigma_d = \sqrt{\frac{(1+x_i'\beta_0)^2}{\frac{1}{n}\sum_{i=1}^n (1+x_i'\beta_0)^2}}$, $\sigma_y = \sqrt{\frac{(1+\alpha_0 d_i + x_i'\beta_0)^2}{\frac{1}{n}\sum_{i=1}^n (1+\alpha_0 d_i + x_i'\beta_0)^2}}$.
- Design 5. $d_i = \mathbf{1}\{d_i^* > 0\}$, $\beta_0 = (1, 1/2, 1/3, 1/4, 1/5, 0, 0, 0, 0, 0, 1, 1/2, 1/3, 1/4, 1/5, 0, ..., 0)'$, $\sigma_y = \sigma_d = 1$.
- Design 6. $d_i = d_i^*$, $\beta_{0,j} \sim N(0, 1)$, $\sigma_y = \sigma_d = 1$.
- Design 7. $d_i = d_i^*$, $\widetilde{\beta}_0 = (1, 1/2, 1/3, 1/4, 1/5, 0, 0, 0, 0, 0, 1, 1/2, 1/3, 1/4, 1/5, 0, ..., 0)'$, $\beta_{0,j} \sim N(0, \widetilde{\beta}_{0,j}^2)$, $\sigma_y = \sigma_d = 1$.
- Design 72. $d_i = d_i^*$, $\widetilde{\beta}_0 = (1, 1/4, 1/9, 1/16, 1/25, 0, 0, 0, 0, 0, 1, 1/4, 1/9, 1/16, 1/25, 0, ..., 0)'$, $\beta_{0,j} \sim N(0, \widetilde{\beta}_{0,j}^2)$, $\sigma_y = \sigma_d = 1$.
- Design 722. $d_i = d_i^*$, $\widetilde{\beta}_{0,j} = (1/j)^2$, $\beta_{0,j} \sim N(0, \widetilde{\beta}_{0,j}^2)$, $\sigma_y = \sigma_d = 1$.
- Design 8. $d_i = d_i^*$, $\widetilde{\beta}_{0,j} = u_j z_{1,j} + (1 - u_j) z_{2,j}$, $u_j \sim$ Bernoulli(.05), $z_{1,j} \sim N(0, 25)$, $z_{2,j} \sim N(0, .0025)$, $\sigma_y = \sigma_d = 1$
- Design 1001. $d_i = d_i^*$, $\beta_{0,j} = \mathbf{1}\{j \in \{2, 4, 6, ..., 38, 40\}\}$, $\sigma_y = \sigma_d = 1$.

In the last thirteen designs, we set the constants c_y and c_d according to

$$c_d = \sqrt{\frac{R_d^2}{(1 - R_d^2)\beta_1'\Sigma\beta_1}}$$

$$c_y = \sqrt{\frac{R_d^2}{(1 - R_d^2)\beta_0'\Sigma\beta_0}}$$

for $R_d^2 = 0, .2, .4, .6$, and .8 and $R_y^2 = 0, .2, .4, .6$, and .8.

- Design 1a. $d_i = d_i^*$, $\beta_0 = (1, 1/2, 1/3, 1/4, 1/5, 0, 0, 0, 0, 0, 1, 1/2, 1/3, 1/4, 1/5, 0, ..., 0)'$, $\beta_1 = (1, 1/2, 1/3, 1/4, 1/5, 1/6, 1/7, 1/8, 1/9, 1/10, 0, ..., 0)'$, $\sigma_y = \sigma_d = 1$.
- Design 2a. $d_i = d_i^*$, $\beta_0 = (1, 1/4, 1/9, 1/16, 1/25, 0, 0, 0, 0, 0, 1, 1/4, 1/9, 1/16, 1/25, 0, ..., 0)'$, $\beta_1 = (1, 1/4, 1/9, 1/16, 1/25, 1/36, 1/49, 1/64, 1/81, 1/100, 0, ..., 0)'$, $\sigma_y = \sigma_d = 1$.
- Design 22a. $d_i = d_i^*$, $\beta_{0,j} = (1/j)^2$, $\beta_{1,j} = (1/j)^2$, $\sigma_y = \sigma_d = 1$.
- Design 3a. $d_i = d_i^*$, $\beta_0 = (1, 1/2, 1/3, 1/4, 1/5, 0, 0, 0, 0, 0, 1, 1/2, 1/3, 1/4, 1/5, 0, ..., 0)'$, $\beta_1 = (1, 1/2, 1/3, 1/4, 1/5, 1/6, 1/7, 1/8, 1/9, 1/10, 0, ..., 0)'$, $\sigma_d = \sqrt{\frac{(1+x_i'\beta_1)^2}{\frac{1}{n}\sum_{i=1}^n (1+x_i'\beta_1)^2}}$, $\sigma_y = \sqrt{\frac{(1+\alpha_0 d_i + x_i'\beta_0)^2}{\frac{1}{n}\sum_{i=1}^n (1+\alpha_0 d_i + x_i'\beta_0)^2}}$.

- Design 4a. $d_i = d_i^*$, $\beta_0 = (1, 1/4, 1/9, 1/16, 1/25, 0, 0, 0, 0, 0, 1, 1/4, 1/9, 1/16, 1/25, 0, ..., 0)'$, $\beta_1 = (1, 1/4, 1/9, 1/16, 1/25, 1/36, 1/49, 1/64, 1/81, 1/100, 0, ..., 0)'$, $\sigma_d = \sqrt{\frac{(1+x_i'\beta_1)^2}{\frac{1}{n}\sum_{i=1}^n (1+x_i'\beta_1)^2}}$, $\sigma_y = \sqrt{\frac{(1+\alpha_0 d_i + x_i'\beta_0)^2}{\frac{1}{n}\sum_{i=1}^n (1+\alpha_0 d_i + x_i'\beta_0)^2}}$.
- Design 44a. $d_i = d_i^*$, $\beta_{0,j} = (1/j)^2$, $\beta_{1,j} = (1/j)^2$, $\sigma_d = \sqrt{\frac{(1+x_i'\beta_1)^2}{\frac{1}{n}\sum_{i=1}^n (1+x_i'\beta_1)^2}}$, $\sigma_y = \sqrt{\frac{(1+\alpha_0 d_i + x_i'\beta_0)^2}{\frac{1}{n}\sum_{i=1}^n (1+\alpha_0 d_i + x_i'\beta_0)^2}}$.
- Design 5a. $d_i = \mathbf{1}\{d_i^* > 0\}$, $\beta_0 = (1, 1/2, 1/3, 1/4, 1/5, 0, 0, 0, 0, 0, 1, 1/2, 1/3, 1/4, 1/5, 0, ..., 0)'$, $\beta_1 = (1, 1/2, 1/3, 1/4, 1/5, 1/6, 1/7, 1/8, 1/9, 1/10, 0, ..., 0)'$, $\sigma_y = \sigma_d = 1$.
- Design 6a. $d_i = d_i^*$, $\beta_{0,j} \sim N(0,1)$, $\beta_{1,j} \sim N(0,1)$, $\mathrm{E}[\beta_{0,j}\beta_{1,j}] = .8$, $\sigma_y = \sigma_d = 1$.
- Design 7a. $d_i = d_i^*$, $\widetilde{\beta}_0 = (1, 1/2, 1/3, 1/4, 1/5, 0, 0, 0, 0, 0, 1, 1/2, 1/3, 1/4, 1/5, 0, ..., 0)'$, $\widetilde{\beta}_1 = (1, 1/2, 1/3, 1/4, 1/5, 1/6, 1/7, 1/8, 1/9, 1/10, 0, ..., 0)'$, $\beta_{0,j} = \widetilde{\beta}_{0,j} z_{0,j}$, $\beta_{1,j} = \widetilde{\beta}_{1,j} z_{1,j}$, $z_{0,j} \sim N(0,1)$, $z_{1,j} \sim N(0,1)$, $\mathrm{E}[z_{0,j} z_{1,j}] = .8$, $\sigma_y = \sigma_d = 1$.
- Design 72a. $d_i = d_i^*$, $\widetilde{\beta}_0 = (1, 1/4, 1/9, 1/16, 1/25, 0, 0, 0, 0, 0, 1, 1/4, 1/9, 1/16, 1/25, 0, ..., 0)'$, $\widetilde{\beta}_1 = (1, 1/4, 1/9, 1/16, 1/25, 1/36, 1/49, 1/64, 1/81, 1/100, 0, ..., 0)'$, $\beta_{0,j} = \widetilde{\beta}_{0,j} z_{0,j}$, $\beta_{1,j} = \widetilde{\beta}_{1,j} z_{1,j}$, $z_{0,j} \sim N(0,1)$, $z_{1,j} \sim N(0,1)$, $\mathrm{E}[z_{0,j} z_{1,j}] = .8$, $\sigma_y = \sigma_d = 1$.
- Design 722a. $d_i = d_i^*$, $\widetilde{\beta}_{0,j} = (1/j)^2$, $\widetilde{\beta}_{1,j} = (1/j)^2$, $\beta_{0,j} = \widetilde{\beta}_{0,j} z_{0,j}$, $\beta_{1,j} = \widetilde{\beta}_{1,j} z_{1,j}$, $z_{0,j} \sim N(0,1)$, $z_{1,j} \sim N(0,1)$, $\mathrm{E}[z_{0,j} z_{1,j}] = .8$, $\sigma_y = \sigma_d = 1$.
- Design 8a. $d_i = d_i^*$, $\widetilde{\beta}_{0,j} = 5u_j z_{11,j} + .05(1-u_j) z_{12,j}$, $\widetilde{\beta}_{1,j} = 5u_j z_{21,j} + .05(1-u_j) z_{22,j}$, $u_j \sim$ Bernoulli(.05), $z_{11,j} \sim N(0,1)$, $z_{12,j} \sim N(0,1)$, $z_{21,j} \sim N(0,1)$, $z_{22,j} \sim N(0,1)$, $\sigma_y = \sigma_d = 1$
- Design 1001a. $d_i = d_i^*$, $\beta_{0,j} = \mathbf{1}\{j \in \{2, 4, 6, ..., 38, 40\}\}$, $\beta_{1,j} = \mathbf{1}\{j \in \{1, 3, 5, ..., 37, 39\}\}$, $\sigma_y = \sigma_d = 1$.

Results are summarized in figures and tables below. In the tables, we report results for the four estimators considered in the main text (Oracle, Double-Selection Oracle, Post-Lasso, and Double-Selection). We also report results for regular Lasso (Lasso), the union of the Double-Selection interval with the Post-Lasso interval (Double-Selection Union ADS), using the union of the set of variables selected by Double-Selection and the set of variables selected by running Lasso of y on d and x without penalizing d (Double-Selection + I3), and the split-sample procedure discussed in the text (Split-Sample). For Double-Selection Union ADS, the point estimate is taken as the midpoint of the union of the intervals.

Appendix Table 1. Simulation Results for Selected R^2 Values

	First Stage R^2 = .2		First Stage R^2 = .2		First Stage R^2 = .8		First Stage R^2 = .8	
	Structure R^2 = 0		Structure R^2 = .8		Structure R^2 = 0		Structure R^2 = .4	
	RMSE	Coverage	RMSE	Coverage	RMSE	Coverage	RMSE	Coverage
	Design 1 - Linear Decay with Cut-Off							
Oracle	0.104	0.068	0.106	0.049	0.106	0.052	0.106	0.049
Double-Selection Oracle	0.103	0.051	0.106	0.049	0.106	0.049	0.106	0.049
Lasso	0.137	0.198	0.430	0.886	0.405	1.000	0.496	1.000
Post-Lasso	0.135	0.191	0.164	0.166	0.403	1.000	0.493	1.000
Double-Selection	0.125	0.122	0.114	0.061	0.127	0.118	0.114	0.078
Double-Selection Union ADS	0.123	0.119	0.117	0.057	0.126	0.115	0.117	0.077
Double-Selection + I3	0.123	0.121	0.117	0.069	0.126	0.116	0.117	0.087
Split-Sample	0.137	0.194	0.309	0.386	0.229	0.632	0.235	0.554
	Design 2 - Quadratic Decay with Cut-Off							
Oracle	0.101	0.059	0.105	0.045	0.105	0.064	0.107	0.053
Double-Selection Oracle	0.101	0.059	0.105	0.045	0.104	0.051	0.104	0.042
Lasso	0.137	0.193	0.348	0.780	0.405	1.000	0.496	1.000
Post-Lasso	0.136	0.197	0.121	0.096	0.404	1.000	0.493	1.000
Double-Selection	0.120	0.112	0.108	0.051	0.113	0.078	0.107	0.062
Double-Selection Union ADS	0.119	0.106	0.110	0.045	0.113	0.078	0.110	0.062
Double-Selection + I3	0.119	0.109	0.110	0.058	0.113	0.080	0.110	0.069
Split-Sample	0.135	0.191	0.206	0.195	0.154	0.270	0.153	0.230
	Design 22 - Quadratic Decay							
Oracle	0.100	0.051	0.103	0.051	0.106	0.073	0.103	0.050
Double-Selection Oracle	0.101	0.051	0.103	0.051	0.102	0.050	0.102	0.052
Lasso	0.138	0.211	0.263	0.564	0.405	1.000	0.496	1.000
Post-Lasso	0.137	0.205	0.110	0.064	0.402	0.987	0.489	0.974
Double-Selection	0.107	0.063	0.107	0.058	0.109	0.074	0.104	0.062
Double-Selection Union ADS	0.108	0.063	0.108	0.055	0.108	0.072	0.106	0.061
Double-Selection + I3	0.108	0.068	0.107	0.060	0.109	0.074	0.106	0.068
Split-Sample	0.121	0.138	0.124	0.087	0.119	0.116	0.123	0.118
	Design 3 - Linear Decay with Cut-Off and Heteroscedasticity							
Oracle	0.143	0.070	0.150	0.075	0.145	0.068	0.150	0.075
Double-Selection Oracle	0.144	0.074	0.150	0.075	0.150	0.075	0.150	0.075
Lasso	0.168	0.142	0.536	0.746	0.411	0.990	0.500	0.999
Post-Lasso	0.167	0.140	0.257	0.236	0.410	0.990	0.500	0.999
Double-Selection	0.156	0.108	0.164	0.089	0.159	0.129	0.158	0.102
Double-Selection Union ADS	0.156	0.107	0.170	0.080	0.158	0.125	0.158	0.101
Double-Selection + I3	0.156	0.108	0.166	0.107	0.158	0.125	0.158	0.109
Split-Sample	0.175	0.198	0.398	0.411	0.260	0.609	0.274	0.567
	Design 4 - Quadratic Decay with Cut-Off and Heteroscedasticity							
Oracle	0.142	0.062	0.147	0.070	0.141	0.066	0.146	0.080
Double-Selection Oracle	0.142	0.062	0.147	0.070	0.145	0.067	0.146	0.070
Lasso	0.165	0.139	0.447	0.642	0.410	0.995	0.501	1.000
Post-Lasso	0.166	0.138	0.173	0.113	0.410	0.994	0.500	1.000
Double-Selection	0.152	0.092	0.150	0.075	0.147	0.086	0.147	0.082
Double-Selection Union ADS	0.152	0.090	0.155	0.070	0.147	0.085	0.149	0.080
Double-Selection + I3	0.152	0.093	0.154	0.078	0.147	0.085	0.149	0.088
Split-Sample	0.172	0.185	0.287	0.260	0.195	0.349	0.197	0.295

Appendix Table 1. Simulation Results for Selected R^2 Values

	First Stage R^2 = .2		First Stage R^2 = .2		First Stage R^2 = .8		First Stage R^2 = .8	
	Structure R^2 = 0		Structure R^2 = .8		Structure R^2 = 0		Structure R^2 = .4	
	RMSE	Coverage	RMSE	Coverage	RMSE	Coverage	RMSE	Coverage
	Design 44 - Quadratic Decay and Heteroscedasticity							
Oracle	0.163	0.080	0.166	0.084	0.158	0.080	0.164	0.088
Double-Selection Oracle	0.164	0.078	0.166	0.084	0.162	0.082	0.164	0.091
Lasso	0.173	0.131	0.382	0.460	0.410	0.996	0.503	0.999
Post-Lasso	0.175	0.139	0.178	0.097	0.409	0.994	0.501	0.993
Double-Selection	0.165	0.098	0.167	0.081	0.162	0.082	0.165	0.083
Double-Selection Union ADS	0.167	0.098	0.170	0.078	0.162	0.082	0.166	0.083
Double-Selection + I3	0.166	0.103	0.169	0.086	0.162	0.083	0.166	0.087
Split-Sample	0.177	0.170	0.205	0.160	0.177	0.168	0.183	0.172
	Design 5 - Binary Treatment							
Oracle	0.228	0.063	0.230	0.053	0.309	0.062	0.306	0.054
Double-Selection Oracle	0.225	0.055	0.230	0.053	0.306	0.054	0.306	0.054
Lasso	0.278	0.144	0.717	0.724	1.435	0.999	1.712	1.000
Post-Lasso	0.278	0.142	0.296	0.117	1.300	0.966	1.447	0.924
Double-Selection	0.259	0.101	0.239	0.055	0.364	0.109	0.339	0.080
Double-Selection Union ADS	0.260	0.100	0.246	0.053	0.364	0.109	0.349	0.079
Double-Selection + I3	0.260	0.105	0.246	0.061	0.374	0.123	0.349	0.094
Split-Sample	0.271	0.124	0.562	0.247	0.795	0.672	0.823	0.596
	Design 6 - Gaussian Random Coefficients							
Oracle	0.134	0.180	0.452	0.091	0.306	0.916	0.313	0.723
Double-Selection Oracle	0.134	0.180	0.452	0.091	0.481	0.125	0.476	0.136
Lasso	0.133	0.185	0.805	0.987	0.399	1.000	0.497	1.000
Post-Lasso	0.134	0.182	0.772	0.980	0.398	1.000	0.496	1.000
Double-Selection	0.139	0.191	0.646	0.899	0.389	1.000	0.464	1.000
Double-Selection Union ADS	0.137	0.188	0.659	0.899	0.387	1.000	0.465	1.000
Double-Selection + I3	0.137	0.189	0.653	0.913	0.387	1.000	0.464	1.000
Split-Sample	0.137	0.206	0.795	0.983	0.397	1.000	0.496	1.000
	Design 7 - Gaussian Random Coefficients, Linear Decay in Std. Dev. with Cut-Off							
Oracle	0.101	0.056	0.104	0.047	0.105	0.069	0.105	0.055
Double-Selection Oracle	0.101	0.056	0.104	0.047	0.103	0.050	0.103	0.050
Lasso	0.134	0.192	0.337	0.749	0.403	1.000	0.500	1.000
Post-Lasso	0.135	0.188	0.119	0.078	0.401	1.000	0.496	1.000
Double-Selection	0.119	0.106	0.109	0.053	0.112	0.083	0.108	0.063
Double-Selection Union ADS	0.117	0.102	0.112	0.050	0.112	0.082	0.112	0.062
Double-Selection + I3	0.118	0.103	0.112	0.061	0.112	0.082	0.111	0.074
Split-Sample	0.131	0.170	0.197	0.184	0.149	0.262	0.152	0.235
	Design 72 - Gaussian Random Coefficients, Quadratic Decay in Std. Dev. with Cut-Off							
Oracle	0.101	0.054	0.102	0.053	0.102	0.051	0.102	0.056
Double-Selection Oracle	0.101	0.054	0.102	0.053	0.102	0.051	0.102	0.056
Lasso	0.134	0.189	0.305	0.686	0.403	1.000	0.499	1.000
Post-Lasso	0.134	0.183	0.108	0.065	0.402	1.000	0.496	1.000
Double-Selection	0.117	0.095	0.105	0.060	0.105	0.066	0.103	0.057
Double-Selection Union ADS	0.116	0.092	0.107	0.057	0.105	0.065	0.106	0.056
Double-Selection + I3	0.116	0.094	0.107	0.065	0.104	0.066	0.106	0.062
Split-Sample	0.129	0.168	0.186	0.171	0.140	0.201	0.133	0.152

Appendix Table 1. Simulation Results for Selected R^2 Values

	First Stage R^2 = .2		First Stage R^2 = .2		First Stage R^2 = .8		First Stage R^2 = .8	
	Structure R^2 = 0		Structure R^2 = .8		Structure R^2 = 0		Structure R^2 = .4	
	RMSE	Coverage	RMSE	Coverage	RMSE	Coverage	RMSE	Coverage
	Design 722 - Gaussian Random Coefficients, Quadratic Decay in Std. Dev.							
Oracle	0.101	0.051	0.101	0.050	0.101	0.053	0.101	0.050
Double-Selection Oracle	0.101	0.051	0.101	0.050	0.101	0.052	0.101	0.050
Lasso	0.134	0.176	0.224	0.446	0.404	1.000	0.498	1.000
Post-Lasso	0.135	0.182	0.102	0.055	0.403	0.998	0.492	0.976
Double-Selection	0.106	0.070	0.103	0.053	0.103	0.062	0.102	0.057
Double-Selection Union ADS	0.106	0.070	0.105	0.053	0.103	0.062	0.104	0.053
Double-Selection + I3	0.106	0.073	0.104	0.059	0.103	0.063	0.104	0.060
Split-Sample	0.120	0.135	0.111	0.071	0.106	0.072	0.106	0.067
	Design 8 - Mixture of Normals							
Oracle	0.106	0.074	0.105	0.056	0.109	0.061	0.105	0.055
Double-Selection Oracle	0.103	0.053	0.105	0.056	0.106	0.055	0.105	0.051
Lasso	0.134	0.186	0.585	0.945	0.402	1.000	0.495	1.000
Post-Lasso	0.134	0.185	0.355	0.510	0.400	1.000	0.493	1.000
Double-Selection	0.129	0.138	0.216	0.214	0.231	0.576	0.228	0.500
Double-Selection Union ADS	0.128	0.135	0.226	0.213	0.230	0.574	0.230	0.499
Double-Selection + I3	0.128	0.139	0.223	0.252	0.230	0.577	0.230	0.507
Split-Sample	0.136	0.185	0.498	0.703	0.329	0.933	0.377	0.913
	Design 1001 - 20 Non-Overlapping Constant Coefficients							
Oracle	0.135	0.193	0.115	0.056	0.115	0.056	0.115	0.056
Double-Selection Oracle	0.135	0.193	0.115	0.056	0.115	0.056	0.115	0.056
Lasso	0.135	0.196	0.755	0.992	0.403	1.000	0.496	1.000
Post-Lasso	0.134	0.194	0.608	0.905	0.401	1.000	0.494	1.000
Double-Selection	0.139	0.185	0.329	0.399	0.327	0.910	0.330	0.861
Double-Selection Union ADS	0.136	0.177	0.354	0.399	0.325	0.910	0.333	0.861
Double-Selection + I3	0.137	0.179	0.349	0.492	0.325	0.911	0.332	0.864
Split-Sample	0.137	0.200	0.731	0.952	0.392	1.000	0.475	1.000

Appendix Table 2. Simulation Results for Selected R^2 Values

	First Stage R^2 = .4		First Stage R^2 = .4		First Stage R^2 = .8		First Stage R^2 = .8	
	Structure R^2 = .4		Structure R^2 = .8		Structure R^2 = .4		Structure R^2 = .4	
	RMSE	Coverage	RMSE	Coverage	RMSE	Coverage	RMSE	Coverage
	Design 1a - Linear Decay with Cut-Off							
Oracle	0.105	0.047	0.105	0.049	0.097	0.079	0.099	0.054
Double-Selection Oracle	0.104	0.047	0.105	0.049	0.110	0.051	0.110	0.049
Lasso	0.294	0.852	0.493	0.943	0.239	0.992	0.564	1.000
Post-Lasso	0.261	0.684	0.280	0.377	0.231	0.989	0.549	0.999
Double-Selection	0.110	0.058	0.112	0.060	0.112	0.061	0.111	0.062
Double-Selection Union ADS	0.112	0.049	0.115	0.052	0.112	0.050	0.111	0.055
Double-Selection + I3	0.112	0.065	0.114	0.069	0.109	0.066	0.111	0.069
Split-Sample	0.168	0.221	0.228	0.241	0.131	0.141	0.194	0.206
	Design 2a - Quadratic Decay with Cut-Off							
Oracle	0.103	0.045	0.105	0.045	0.101	0.053	0.105	0.043
Double-Selection Oracle	0.103	0.045	0.105	0.045	0.103	0.050	0.105	0.050
Lasso	0.289	0.812	0.408	0.878	0.238	0.997	0.570	1.000
Post-Lasso	0.258	0.648	0.190	0.145	0.235	0.999	0.567	0.999
Double-Selection	0.107	0.050	0.108	0.054	0.108	0.054	0.106	0.056
Double-Selection Union ADS	0.108	0.047	0.110	0.049	0.108	0.048	0.107	0.053
Double-Selection + I3	0.107	0.060	0.110	0.056	0.105	0.062	0.107	0.059
Split-Sample	0.130	0.102	0.160	0.112	0.125	0.076	0.155	0.122
	Design 22a - Quadratic Decay							
Oracle	0.103	0.053	0.103	0.051	0.104	0.063	0.104	0.047
Double-Selection Oracle	0.103	0.053	0.103	0.051	0.102	0.049	0.104	0.047
Lasso	0.344	0.845	0.404	0.901	0.330	1.000	0.770	1.000
Post-Lasso	0.201	0.255	0.124	0.092	0.330	0.999	0.668	0.835
Double-Selection	0.104	0.055	0.107	0.057	0.103	0.056	0.107	0.059
Double-Selection Union ADS	0.106	0.053	0.109	0.055	0.104	0.054	0.112	0.056
Double-Selection + I3	0.106	0.060	0.109	0.060	0.104	0.061	0.112	0.070
Split-Sample	0.117	0.096	0.132	0.103	0.112	0.089	0.137	0.144
	Design 3a - Linear Decay with Cut-Off and Heteroscedasticity							
Oracle	0.144	0.081	0.146	0.073	0.120	0.103	0.130	0.074
Double-Selection Oracle	0.144	0.074	0.146	0.073	0.153	0.080	0.153	0.079
Lasso	0.314	0.676	0.580	0.828	0.245	0.944	0.574	0.999
Post-Lasso	0.297	0.637	0.419	0.462	0.240	0.942	0.558	0.999
Double-Selection	0.153	0.081	0.157	0.084	0.149	0.085	0.150	0.087
Double-Selection Union ADS	0.157	0.076	0.162	0.071	0.151	0.073	0.153	0.073
Double-Selection + I3	0.155	0.089	0.159	0.093	0.146	0.086	0.150	0.096
Split-Sample	0.226	0.313	0.291	0.301	0.168	0.210	0.233	0.277
	Design 4a - Quadratic Decay with Cut-Off and Heteroscedasticity							
Oracle	0.144	0.065	0.147	0.063	0.138	0.067	0.145	0.062
Double-Selection Oracle	0.144	0.065	0.147	0.063	0.143	0.066	0.146	0.066
Lasso	0.309	0.671	0.527	0.736	0.242	0.946	0.576	1.000
Post-Lasso	0.298	0.634	0.380	0.295	0.241	0.945	0.572	1.000
Double-Selection	0.148	0.071	0.149	0.069	0.148	0.067	0.148	0.074
Double-Selection Union ADS	0.150	0.062	0.151	0.066	0.149	0.054	0.150	0.070
Double-Selection + I3	0.148	0.075	0.151	0.082	0.145	0.070	0.149	0.077
Split-Sample	0.193	0.216	0.215	0.186	0.167	0.144	0.201	0.163

Appendix Table 2. Simulation Results for Selected R^2 Values

	First Stage R^2 = .4		First Stage R^2 = .4		First Stage R^2 = .8		First Stage R^2 = .8	
	Structure R^2 = .4		Structure R^2 = .8		Structure R^2 = .4		Structure R^2 = .4	
	RMSE	Coverage	RMSE	Coverage	RMSE	Coverage	RMSE	Coverage
	Design 44a - Quadratic Decay and Heteroscedasticity							
Oracle	0.166	0.097	0.166	0.091	0.161	0.099	0.166	0.093
Double-Selection Oracle	0.166	0.097	0.166	0.091	0.163	0.088	0.166	0.093
Lasso	0.408	0.707	0.563	0.691	0.336	0.982	0.803	0.997
Post-Lasso	0.358	0.541	0.220	0.140	0.336	0.982	0.785	0.955
Double-Selection	0.166	0.087	0.167	0.084	0.164	0.084	0.167	0.077
Double-Selection Union ADS	0.167	0.086	0.172	0.080	0.164	0.084	0.169	0.077
Double-Selection + I3	0.166	0.089	0.172	0.091	0.164	0.085	0.168	0.081
Split-Sample	0.215	0.234	0.206	0.163	0.176	0.157	0.193	0.185
	Design 5a - Binary Treatment							
Oracle	0.243	0.058	0.242	0.047	0.288	0.055	0.288	0.045
Double-Selection Oracle	0.241	0.058	0.242	0.047	0.306	0.050	0.305	0.043
Lasso	0.597	0.637	0.848	0.789	0.857	0.933	1.709	0.995
Post-Lasso	0.478	0.415	0.390	0.170	0.797	0.840	1.308	0.695
Double-Selection	0.244	0.047	0.251	0.049	0.297	0.056	0.305	0.054
Double-Selection Union ADS	0.251	0.047	0.259	0.046	0.304	0.053	0.308	0.045
Double-Selection + I3	0.251	0.062	0.259	0.059	0.303	0.057	0.308	0.053
Split-Sample	0.385	0.187	0.490	0.185	0.407	0.162	0.539	0.184
	Design 6a - Gaussian Random Coefficients							
Oracle	0.231	0.487	0.445	0.090	0.215	0.878	0.370	0.110
Double-Selection Oracle	0.205	0.286	0.539	0.107	0.534	0.138	0.518	0.000
Lasso	0.347	0.928	0.825	0.999	0.279	0.999	0.672	1.000
Post-Lasso	0.343	0.920	0.806	0.999	0.278	0.999	0.667	1.000
Double-Selection	0.301	0.761	0.672	0.953	0.254	0.968	0.603	1.000
Double-Selection Union ADS	0.304	0.761	0.687	0.952	0.256	0.968	0.611	1.000
Double-Selection + I3	0.302	0.775	0.682	0.958	0.255	0.969	0.607	1.000
Split-Sample	0.346	0.909	0.814	0.995	0.278	0.997	0.669	1.000
	Design 7a - Gaussian Random Coefficients, Linear Decay in Std. Dev. with Cut-Off							
Oracle	0.102	0.054	0.104	0.053	0.099	0.056	0.103	0.052
Double-Selection Oracle	0.102	0.054	0.104	0.053	0.103	0.059	0.105	0.049
Lasso	0.318	0.768	0.356	0.790	0.303	0.999	0.708	1.000
Post-Lasso	0.214	0.282	0.113	0.065	0.303	0.998	0.672	0.830
Double-Selection	0.107	0.065	0.110	0.054	0.105	0.061	0.108	0.060
Double-Selection Union ADS	0.109	0.058	0.113	0.043	0.106	0.053	0.111	0.042
Double-Selection + I3	0.109	0.069	0.112	0.063	0.106	0.069	0.110	0.074
Split-Sample	0.117	0.076	0.136	0.079	0.110	0.073	0.132	0.085
	Design 72a - Gaussian Random Coefficients, Quadratic Decay in Std. Dev. with Cut-Off							
Oracle	0.102	0.053	0.102	0.053	0.101	0.052	0.102	0.055
Double-Selection Oracle	0.102	0.053	0.102	0.053	0.102	0.052	0.102	0.053
Lasso	0.325	0.775	0.349	0.781	0.314	0.999	0.732	1.000
Post-Lasso	0.205	0.255	0.108	0.061	0.313	0.999	0.692	0.833
Double-Selection	0.106	0.064	0.109	0.069	0.106	0.065	0.109	0.064
Double-Selection Union ADS	0.108	0.059	0.110	0.059	0.107	0.060	0.109	0.046
Double-Selection + I3	0.108	0.071	0.109	0.071	0.107	0.068	0.108	0.063
Split-Sample	0.114	0.082	0.129	0.084	0.112	0.076	0.127	0.081

Appendix Table 2. Simulation Results for Selected R^2 Values

	First Stage R^2 = .4		First Stage R^2 = .4		First Stage R^2 = .8		First Stage R^2 = .8	
	Structure R^2 = .4		Structure R^2 = .8		Structure R^2 = .4		Structure R^2 = .4	
	RMSE	Coverage	RMSE	Coverage	RMSE	Coverage	RMSE	Coverage
	Design 722a - Gaussian Random Coefficients, Quadratic Decay in Std. Dev.							
Oracle	0.101	0.051	0.101	0.050	0.100	0.051	0.101	0.049
Double-Selection Oracle	0.101	0.051	0.101	0.050	0.101	0.050	0.101	0.050
Lasso	0.323	0.781	0.335	0.787	0.332	1.000	0.735	1.000
Post-Lasso	0.176	0.167	0.102	0.053	0.332	1.000	0.605	0.567
Double-Selection	0.103	0.054	0.103	0.053	0.102	0.055	0.102	0.056
Double-Selection Union ADS	0.105	0.052	0.104	0.052	0.103	0.055	0.105	0.045
Double-Selection + I3	0.105	0.062	0.104	0.056	0.103	0.060	0.105	0.059
Split-Sample	0.110	0.081	0.109	0.064	0.104	0.062	0.105	0.059
	Design 8a - Mixture of Normals							
Oracle	0.102	0.053	0.108	0.062	0.093	0.051	0.106	0.065
Double-Selection Oracle	0.105	0.049	0.108	0.059	0.104	0.056	0.109	0.058
Lasso	0.321	0.836	0.630	0.945	0.267	0.966	0.617	0.991
Post-Lasso	0.283	0.692	0.446	0.710	0.263	0.944	0.570	0.974
Double-Selection	0.150	0.165	0.207	0.217	0.144	0.244	0.231	0.337
Double-Selection Union ADS	0.155	0.158	0.222	0.210	0.147	0.232	0.248	0.308
Double-Selection + I3	0.153	0.181	0.218	0.272	0.145	0.255	0.241	0.387
Split-Sample	0.255	0.597	0.489	0.700	0.215	0.694	0.449	0.806
	Design 1001a - 20 Non-Overlapping Constant Coefficients							
Oracle	0.100	0.053	0.100	0.053	0.072	0.050	0.072	0.050
Double-Selection Oracle	0.131	0.050	0.131	0.050	0.131	0.050	0.131	0.050
Lasso	0.324	0.891	0.741	0.997	0.261	0.997	0.626	1.000
Post-Lasso	0.305	0.837	0.617	0.962	0.257	0.994	0.592	1.000
Double-Selection	0.188	0.319	0.281	0.328	0.150	0.343	0.262	0.466
Double-Selection Union ADS	0.191	0.319	0.298	0.324	0.152	0.341	0.269	0.463
Double-Selection + I3	0.190	0.337	0.294	0.379	0.151	0.348	0.265	0.485
Split-Sample	0.306	0.822	0.683	0.968	0.243	0.971	0.567	0.997

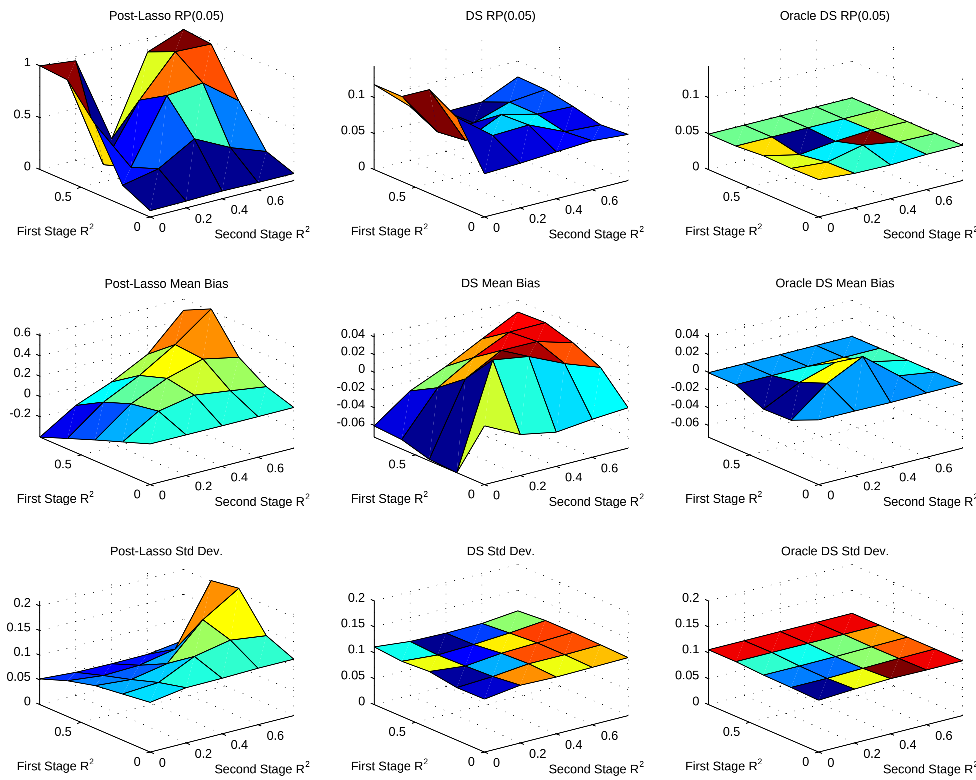

FIGURE 1. Design 1

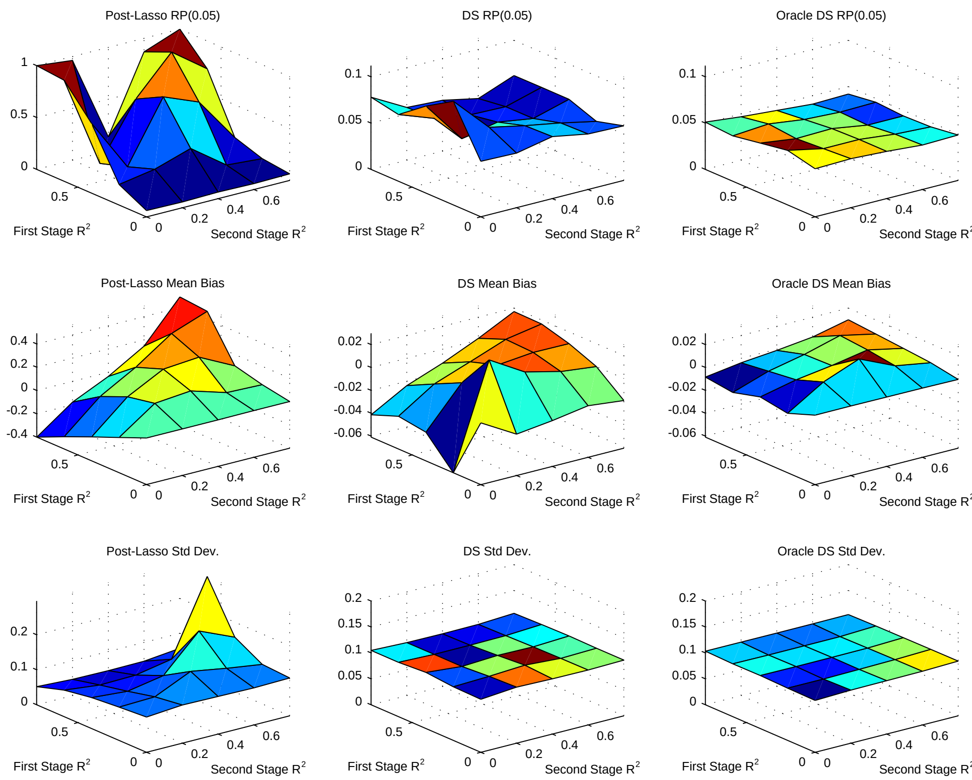

FIGURE 2. Design 2

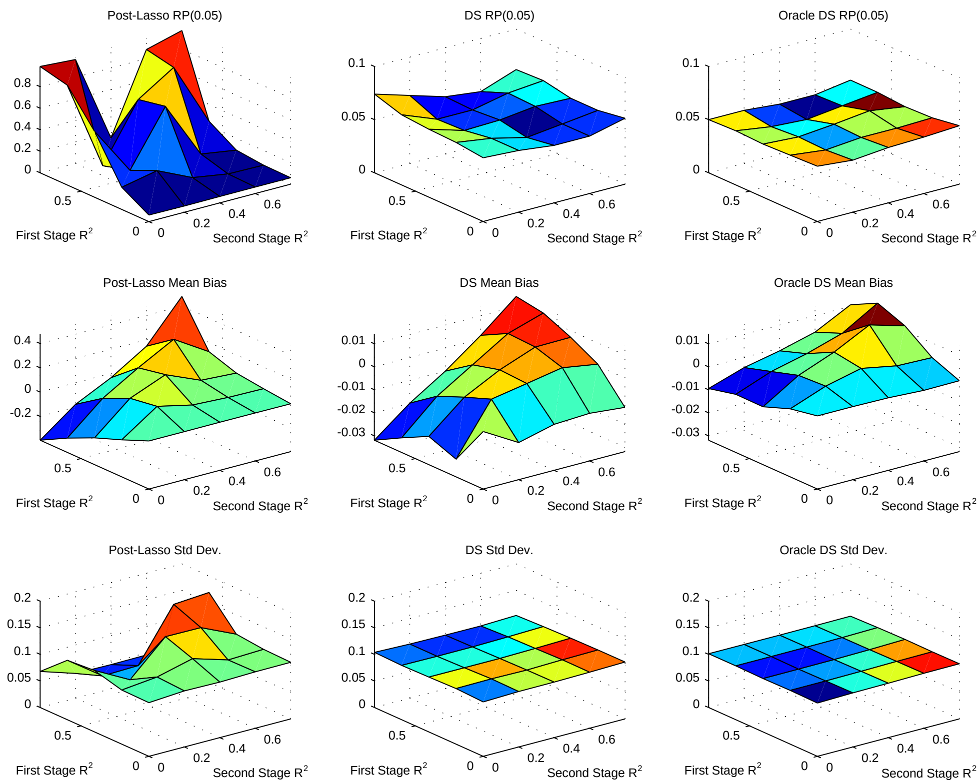

FIGURE 3. Design 22

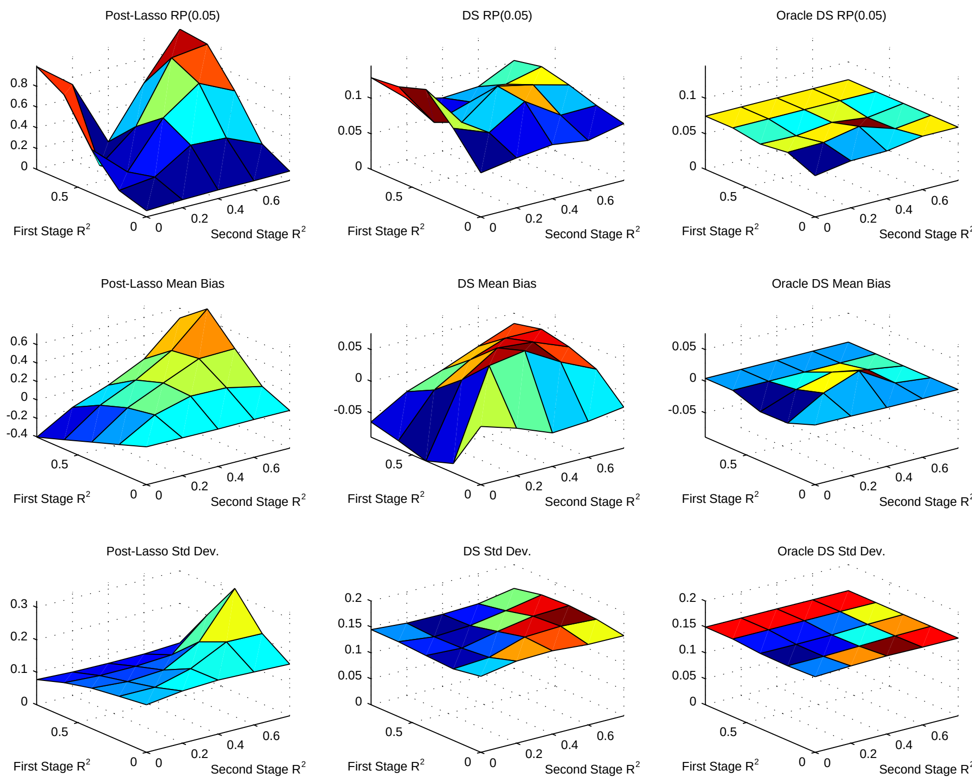

Figure 4. Design 3

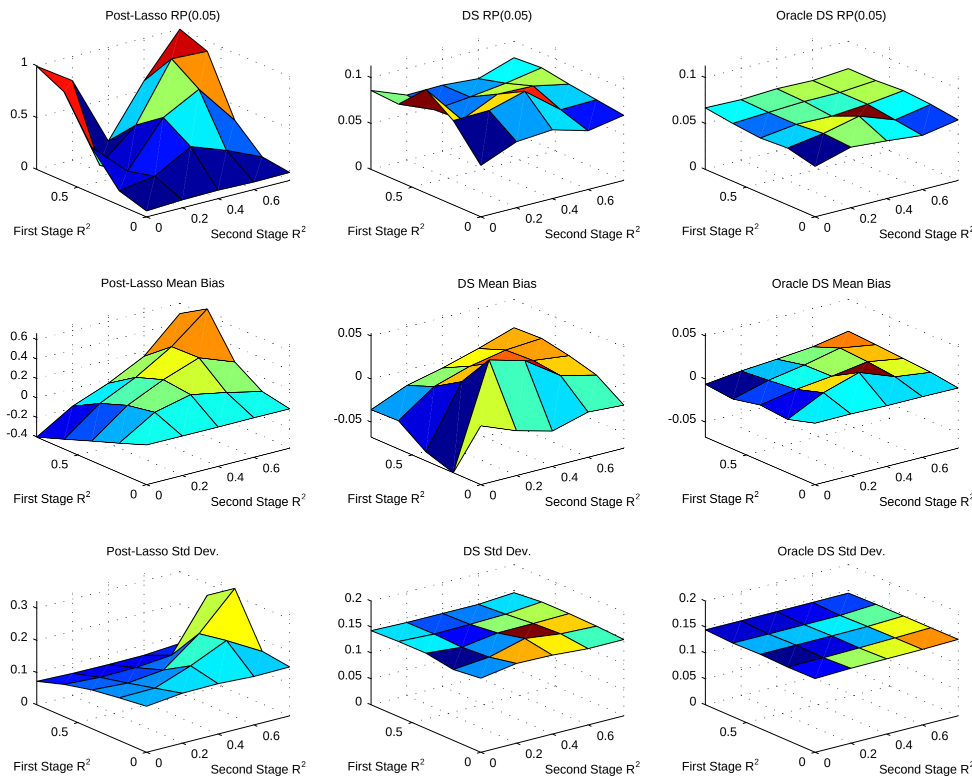

Figure 5. Design 4

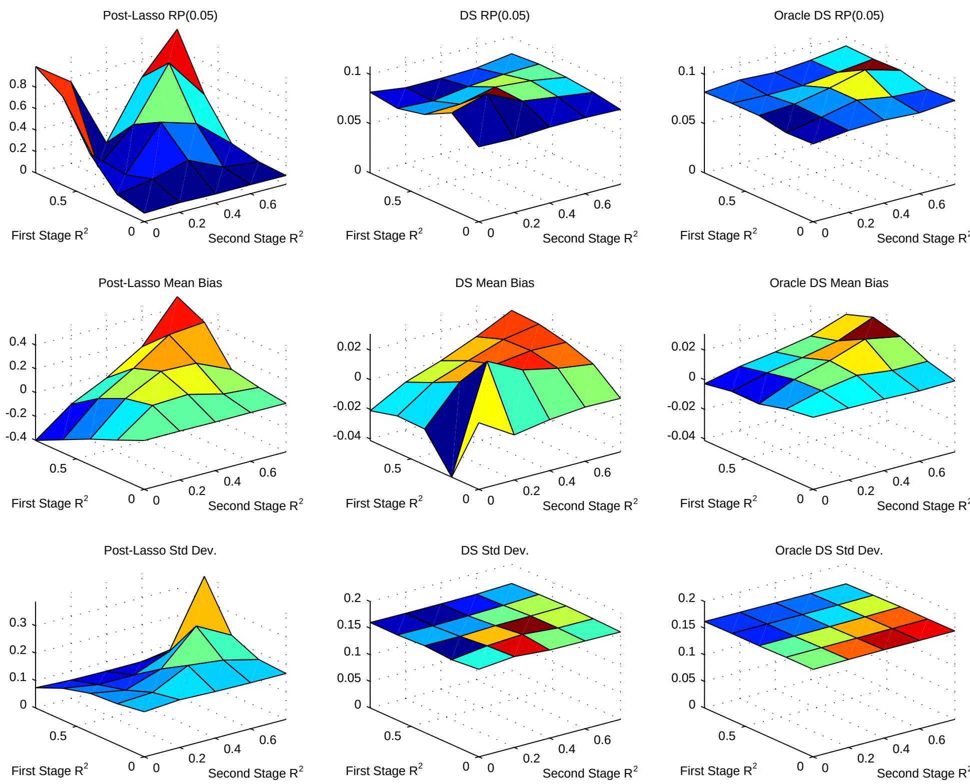

FIGURE 6. Design 44

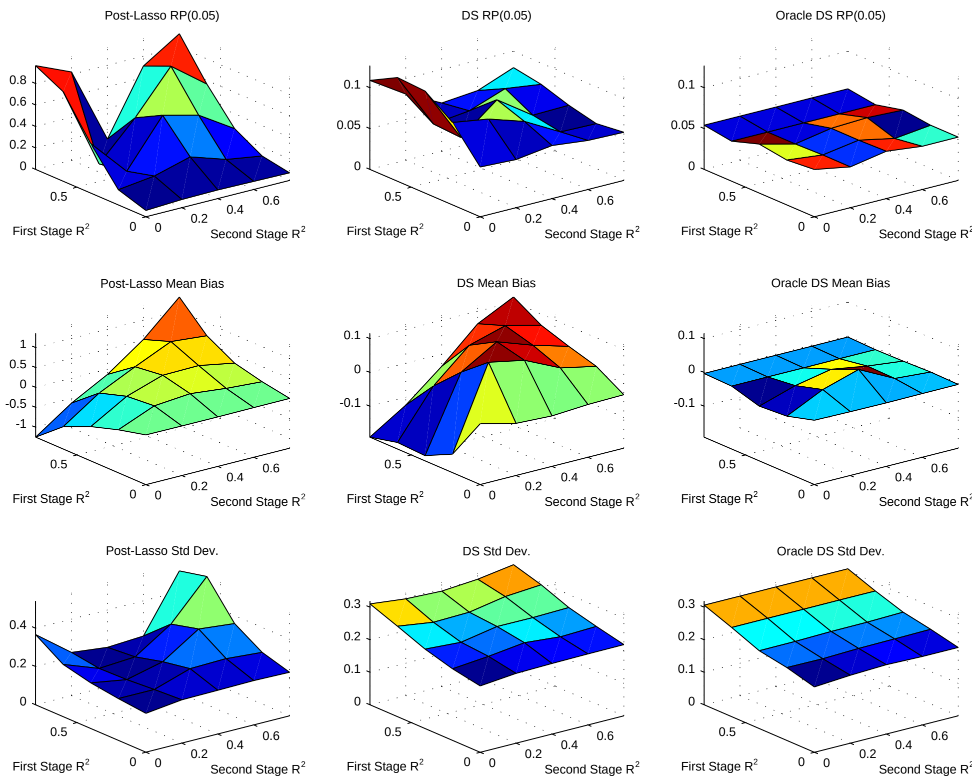

FIGURE 7. Design 5

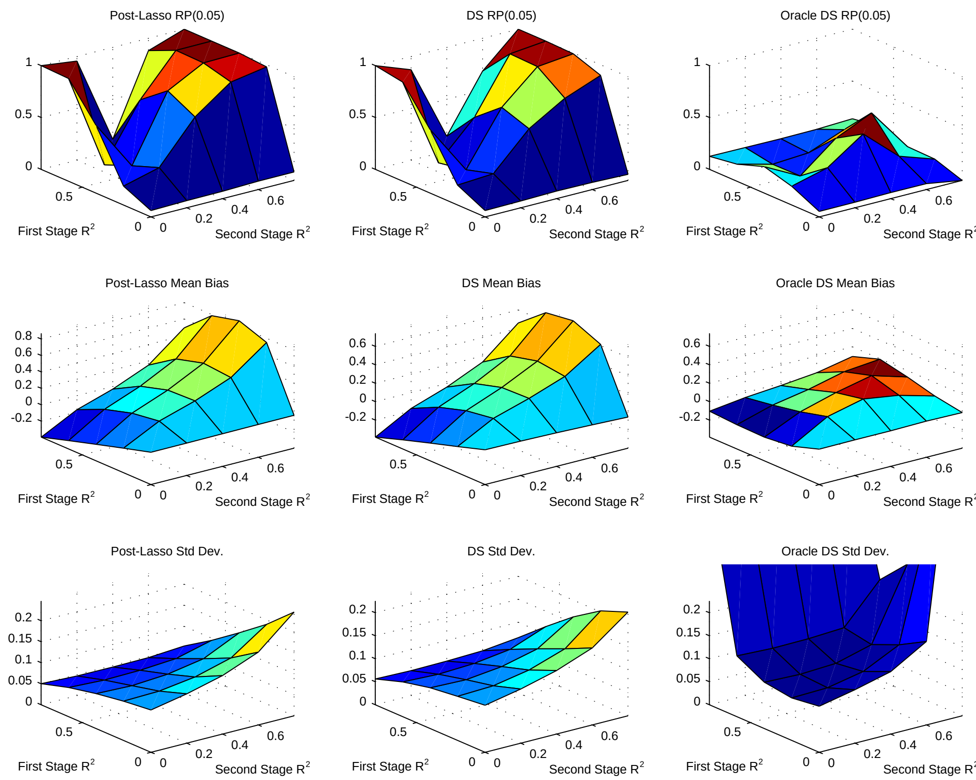

FIGURE 8. Design 6

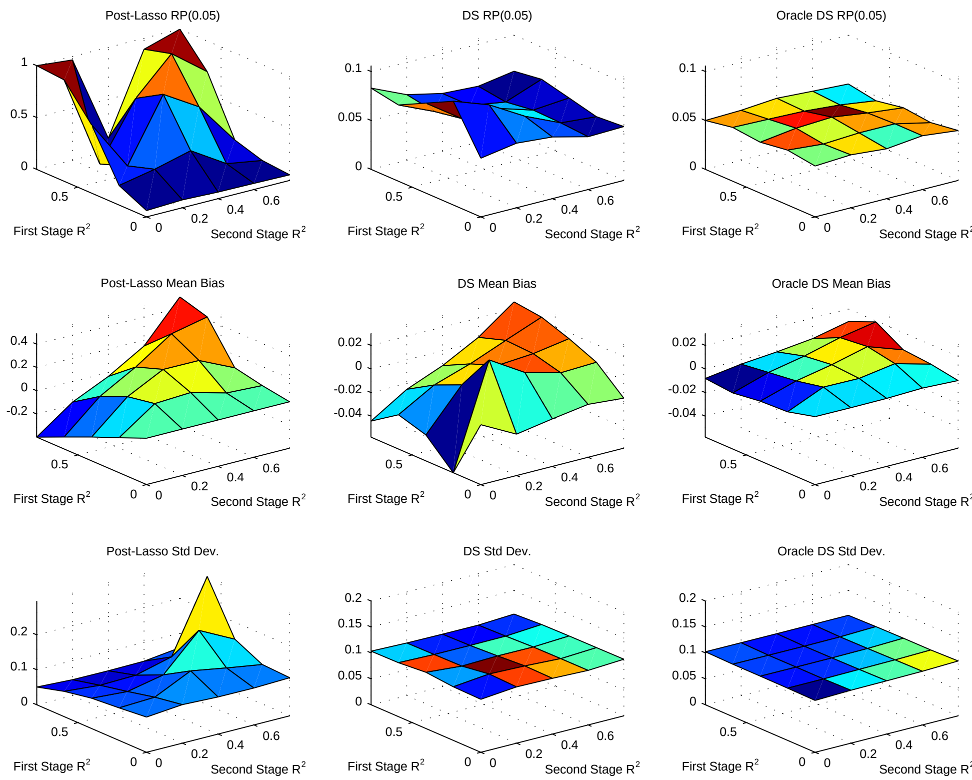

FIGURE 9. Design 7

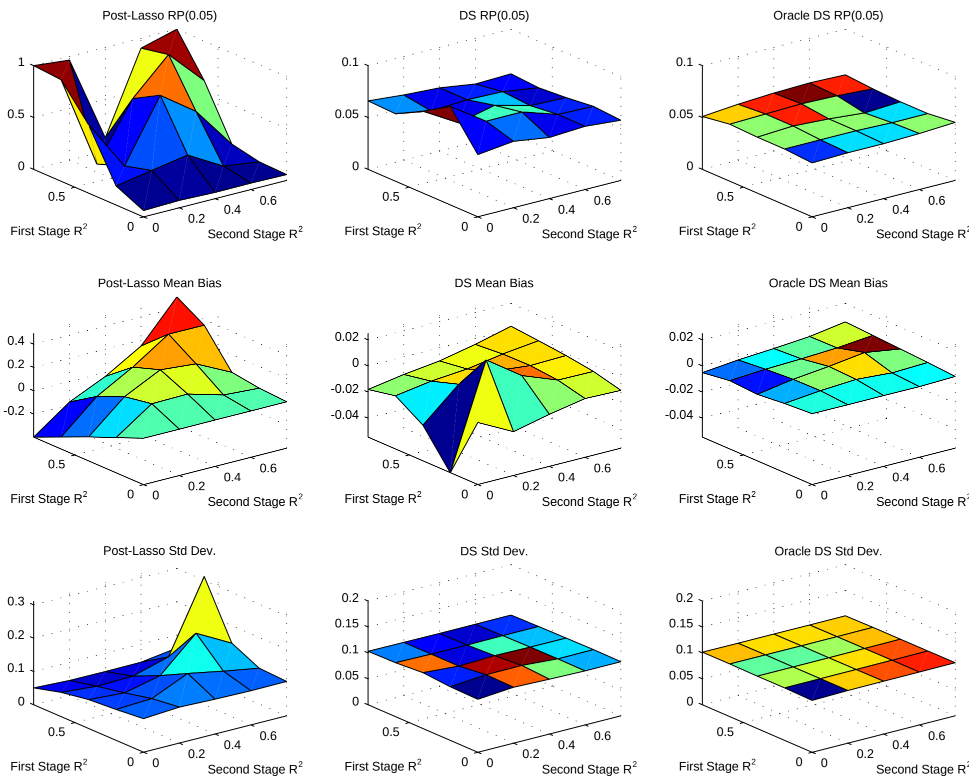

FIGURE 10. Design 72

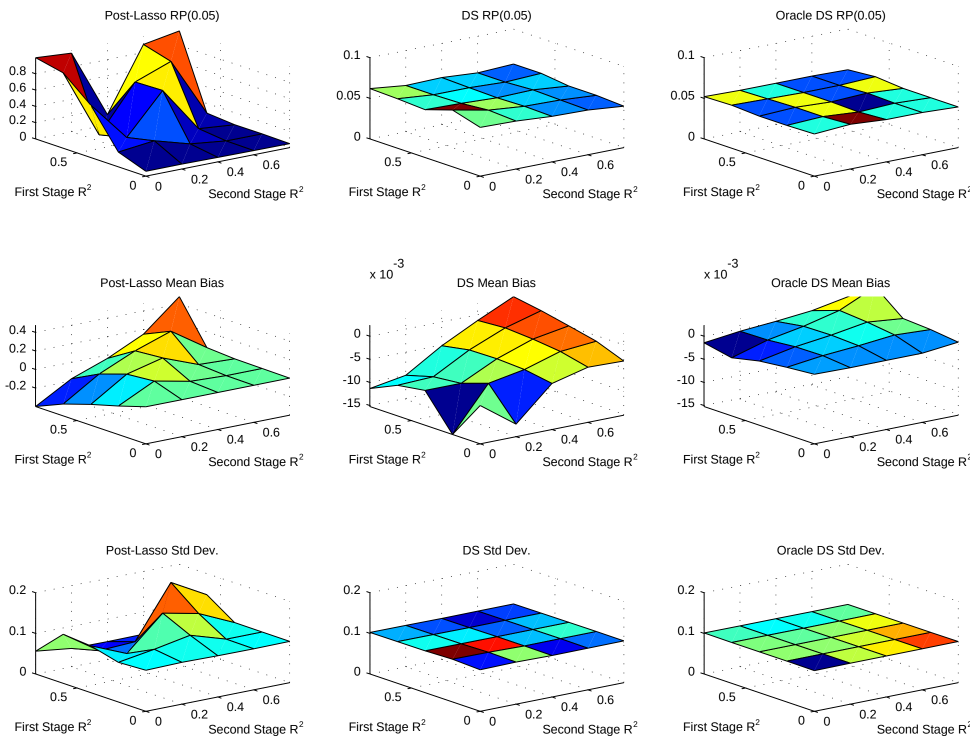

Figure 11. Design 722

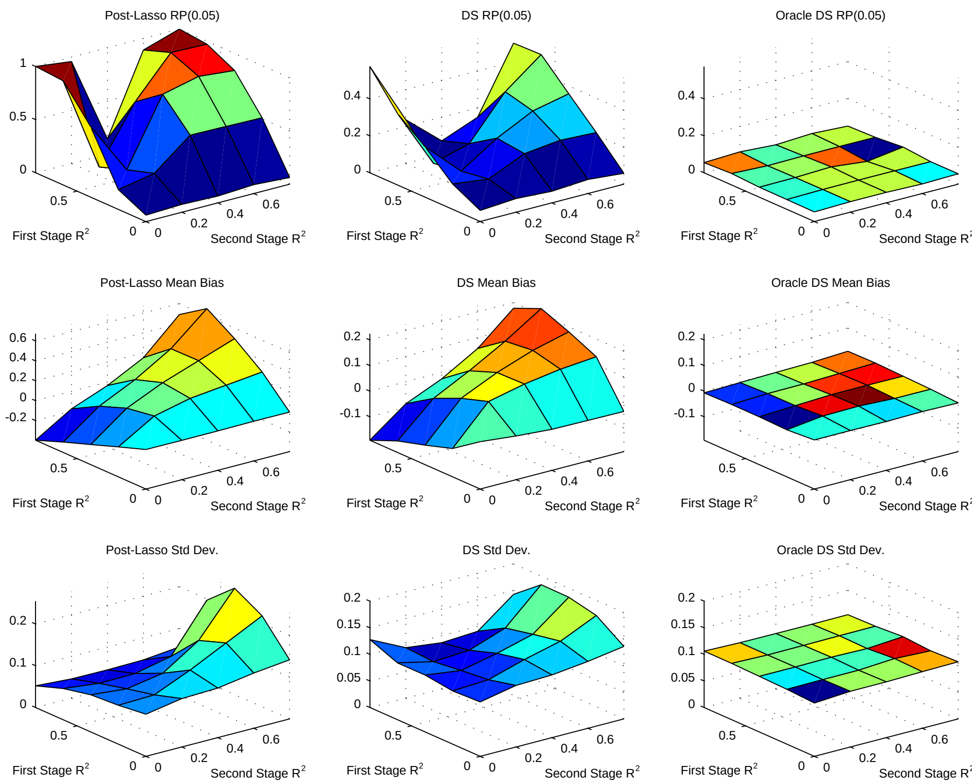

FIGURE 12. Design 8

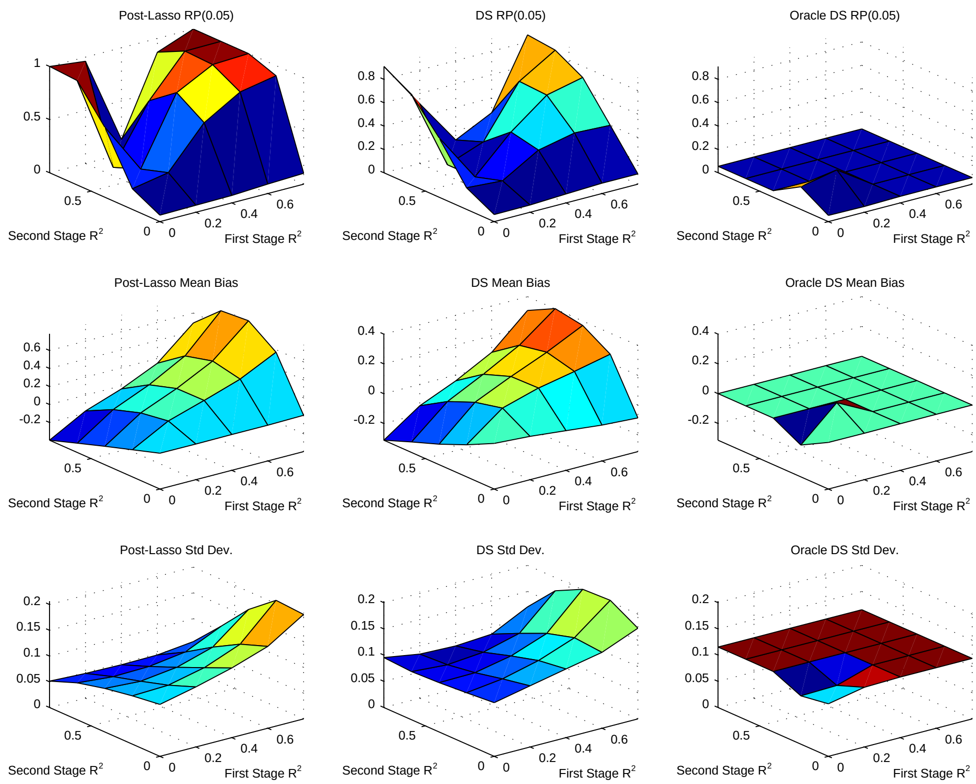

FIGURE 13. Design 1001

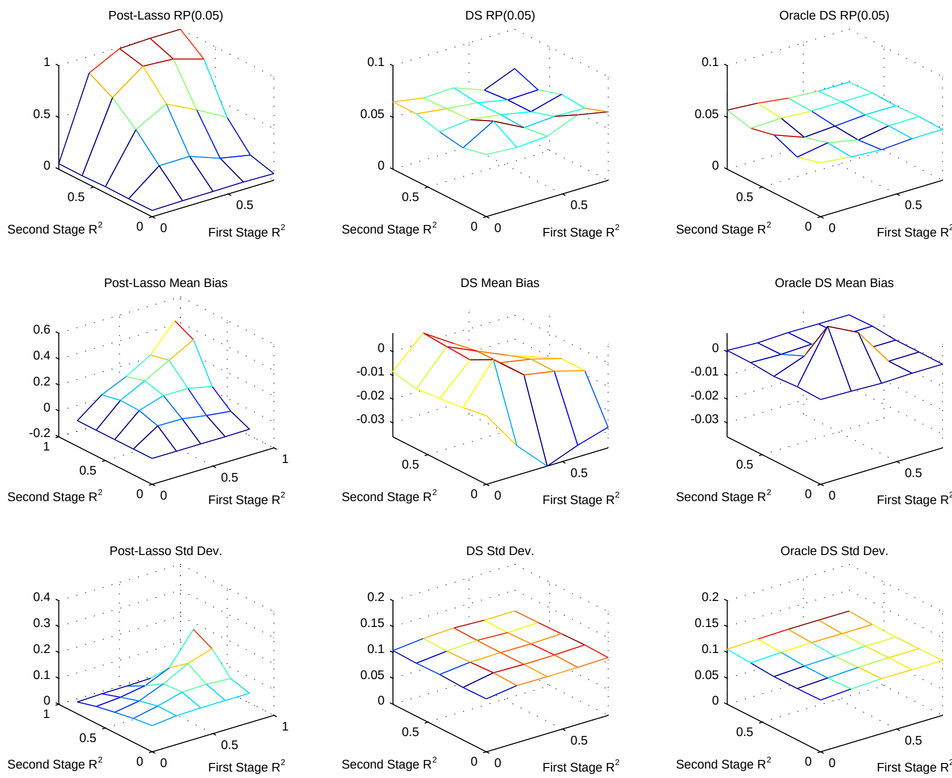

FIGURE 14. Design 1a

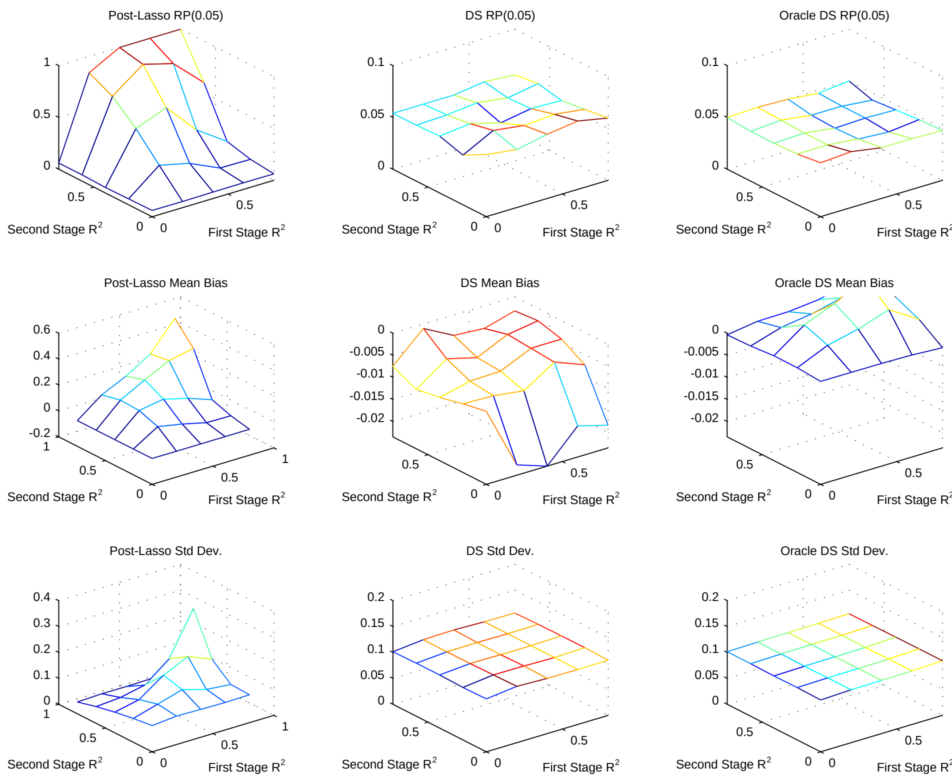

FIGURE 15. Design 2a

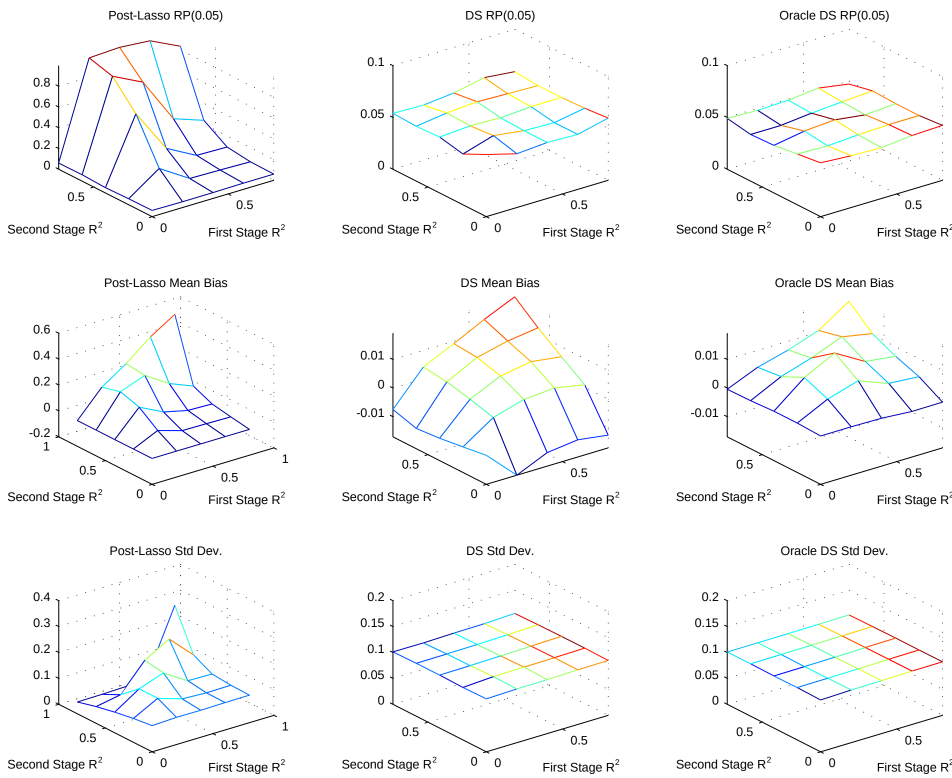

FIGURE 16. Design 22a

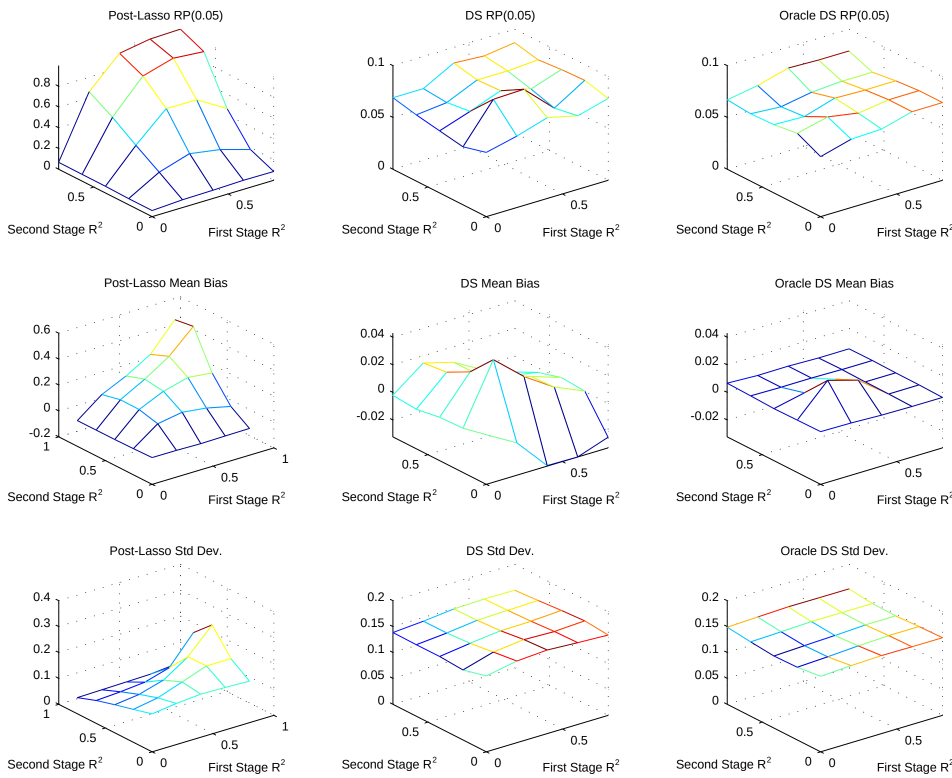

FIGURE 17. Design 3a

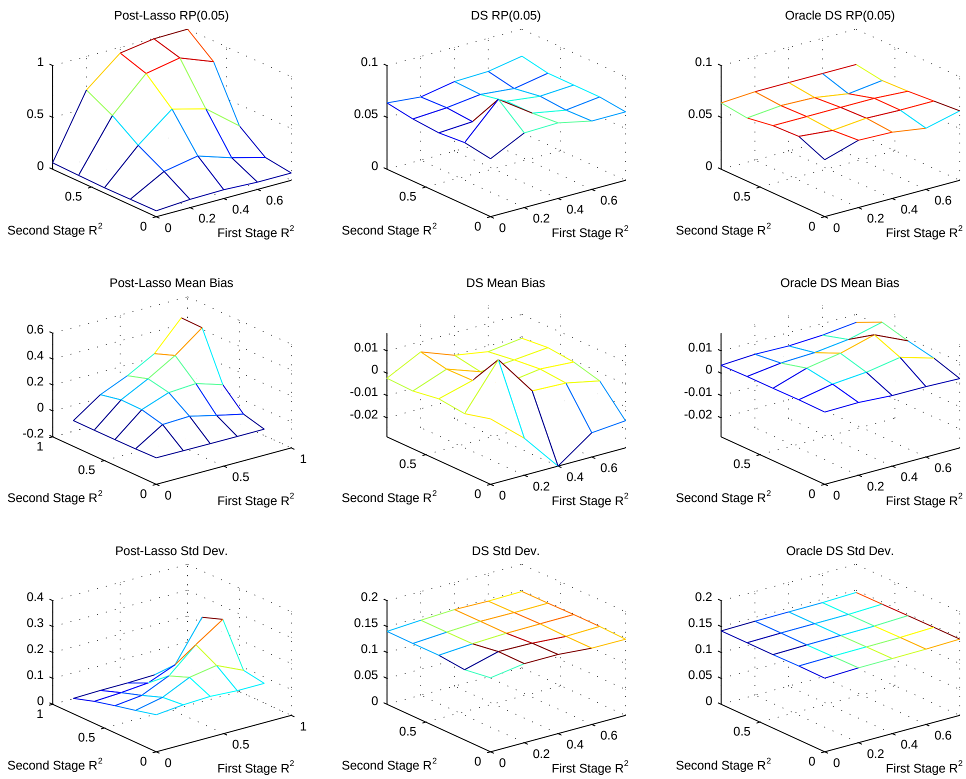

FIGURE 18. Design 4a

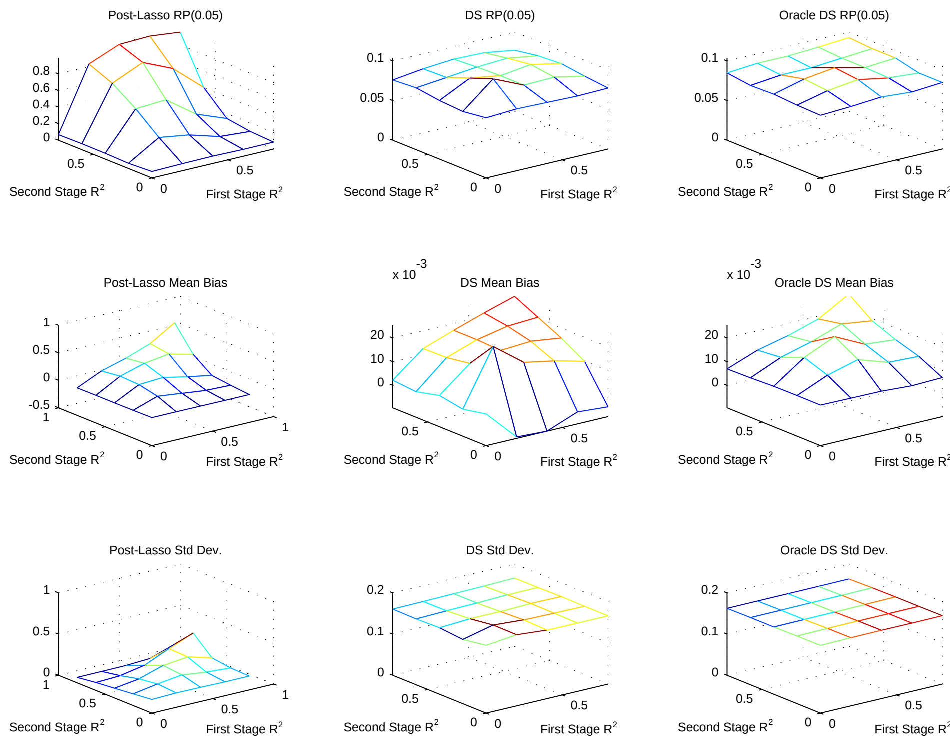

FIGURE 19. Design 44a

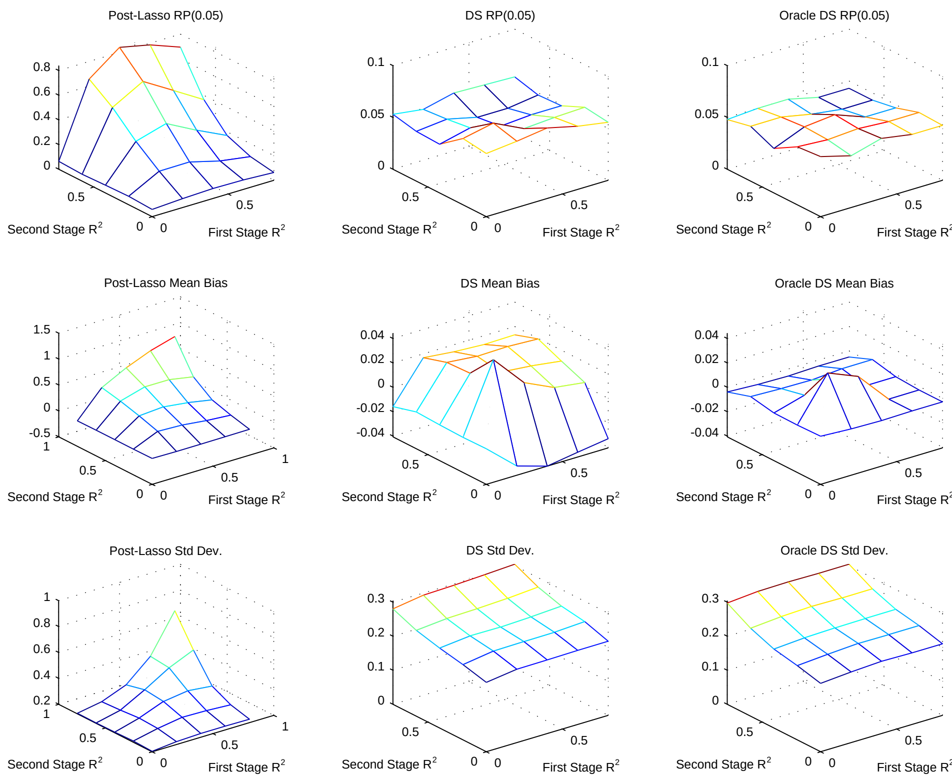

FIGURE 20. Design 5a

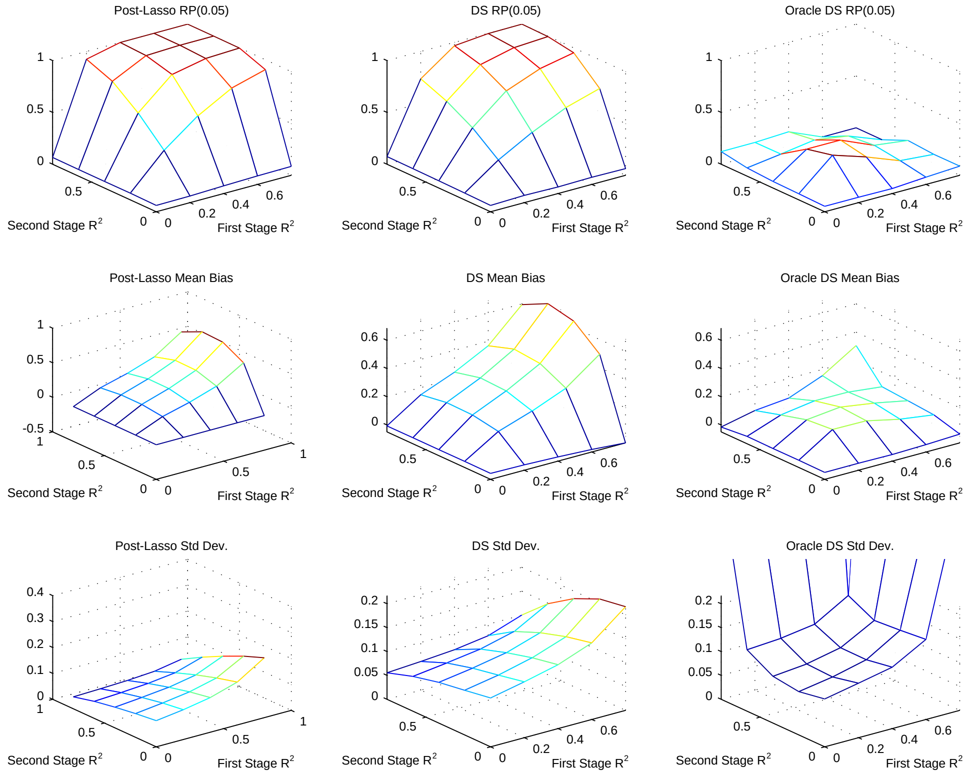

FIGURE 21. Design 6a

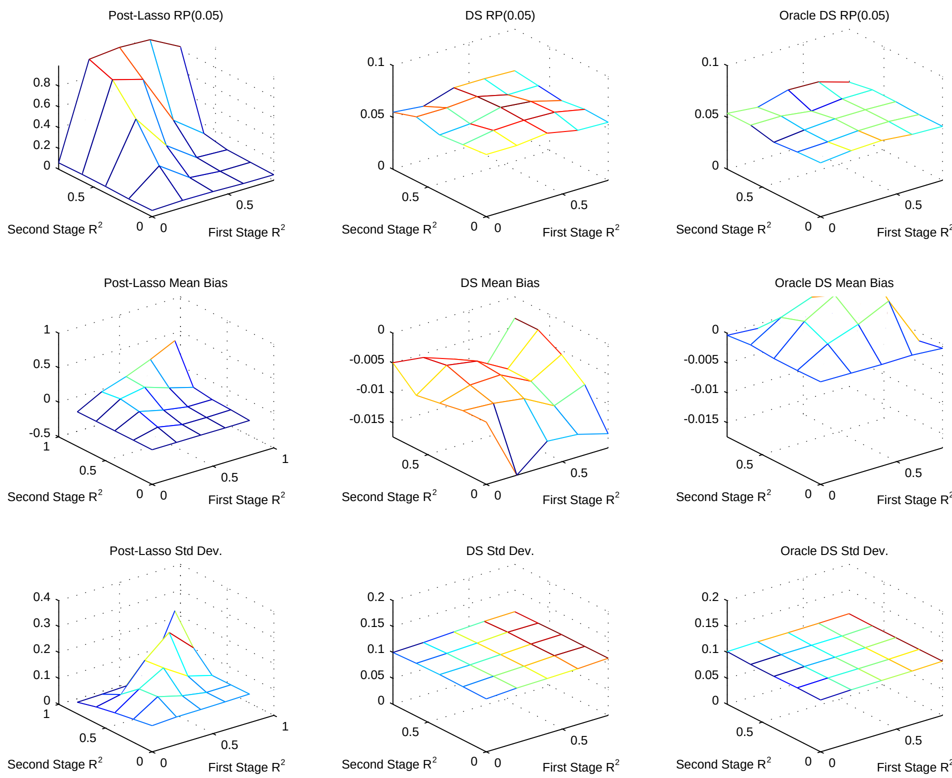

FIGURE 22. Design 7a

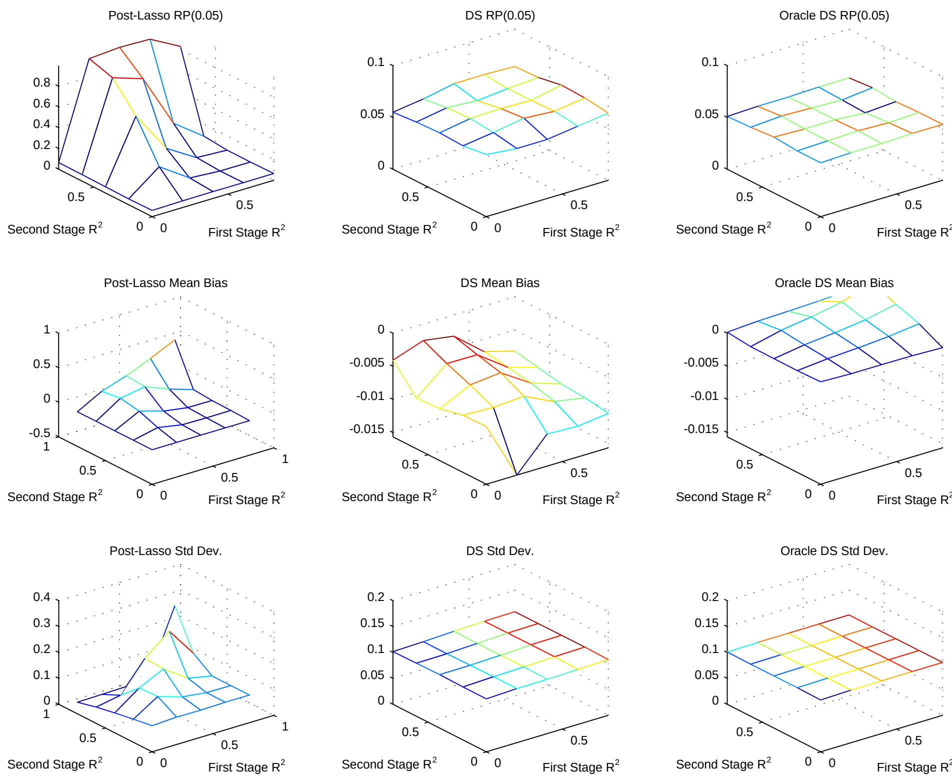

FIGURE 23. Design 72a

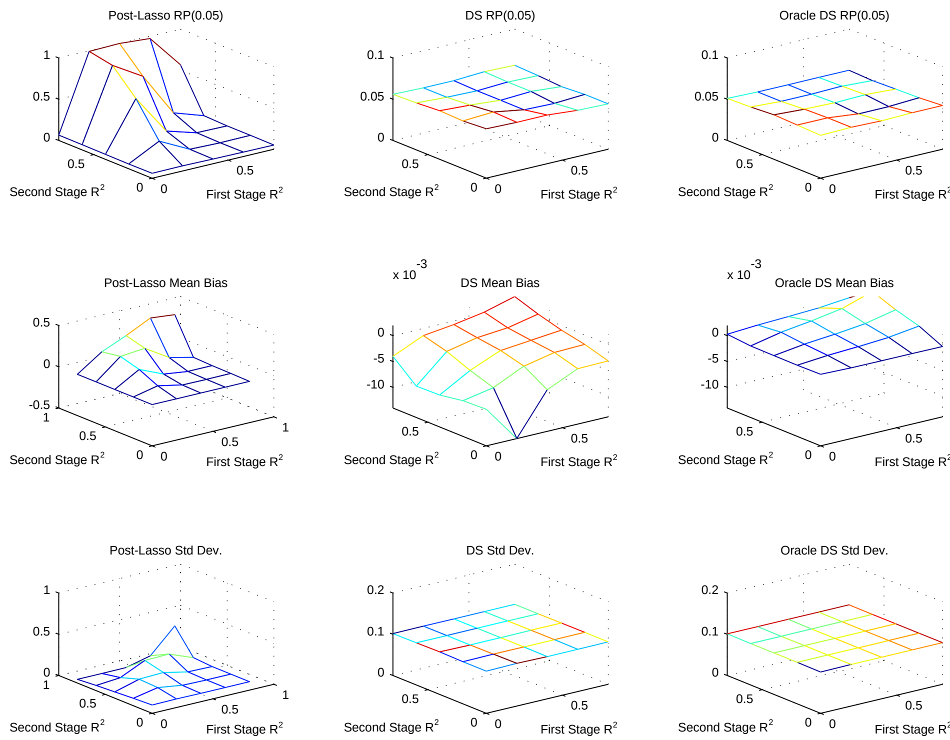

FIGURE 24. Design 722a

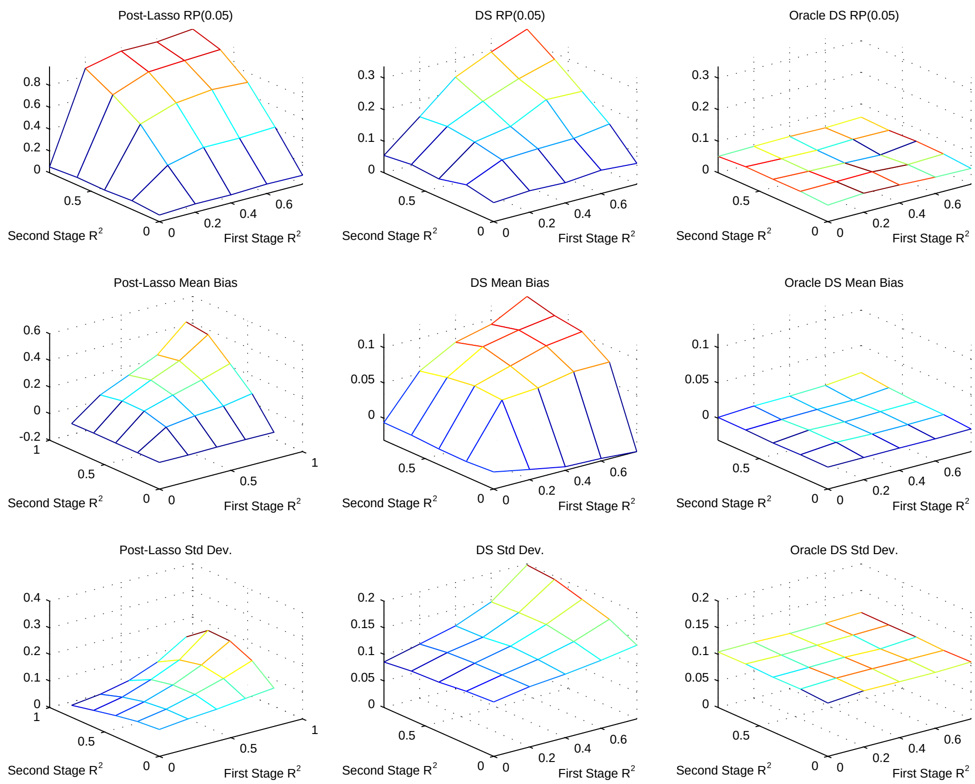

FIGURE 25. Design 8a

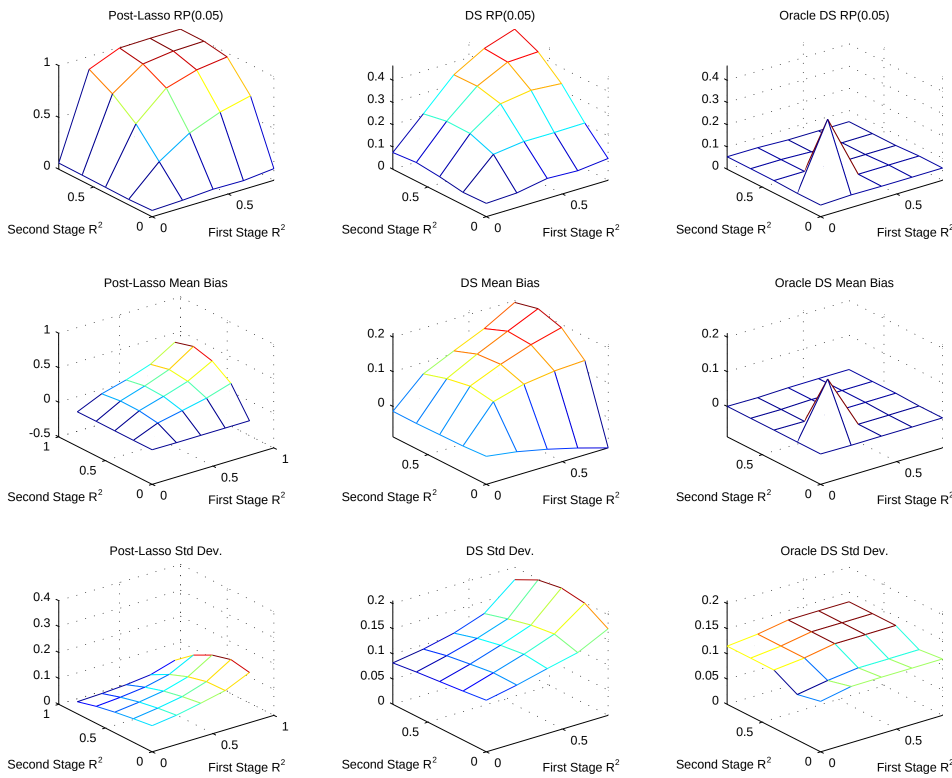

FIGURE 26. Design 1001a

REFERENCES

BACH, F. (2010): "Self-concordant analysis for logistic regression," *Electronic Journal of Statistics*, 4, 384–414.

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.

BELLONI, A., AND V. CHERNOZHUKOV (2011): "ℓ_1-penalized quantile regression in high-dimensional sparse models," *Ann. Statist.*, 39(1), 82–130.

BELLONI, A., V. CHERNOZHUKOV, AND K. KATO (2013): "Uniform Post Selection Inference for LAD Regression Models," *arXiv preprint arXiv:1304.0282*.

BELLONI, A., V. CHERNOZHUKOV, AND Y. WEI (2013): "Honest Confidence Regions for Logistic Regression with a Large Number of Controls," *arXiv preprint arXiv:1304.3969*.

HAHN, J. (1998): "On the role of the propensity score in efficient semiparametric estimation of average treatment effects," *Econometrica*, pp. 315–331.

Leeb, H., and B. M. Pötscher (2008): "Recent developments in model selection and related areas," *Econometric Theory*, 24(2), 319–322.

MacKinnon, J. G., and H. White (1985): "Some heteroskedasticity consistent covariance matrix estimators with improved finite sample properties," *Journal of Econometrics*, 29, 305–325.

Neyman, J. (1979): "$C(\alpha)$ tests and their use," *Sankhya*, 41, 1–21.

Rudelson, M., and R. Vershynin (2008): "On sparse reconstruction from Fourier and Gaussian measurements," *Communications on Pure and Applied Mathematics*, 61, 10251045.

Rudelson, M., and S. Zhou (2011): "Reconstruction from anisotropic random measurements," *ArXiv:1106.1151*.

van de Geer, S. A. (2008): "High-dimensional generalized linear models and the lasso," *Annals of Statistics*, 36(2), 614–645.

von Bahr, B., and C.-G. Esseen (1965): "Inequalities for the rth absolute moment of a sum of random variables, $1 \leqslant r \leqslant 2$," *Ann. Math. Statist*, 36, 299–303.