The Fragility of Sparsity*

Michal Kolesár Ulrich K. Müller Sebastian T. Roelsgaard

March 4, 2025

Abstract

We show, using three empirical applications, that linear regression estimates which rely on the assumption of sparsity are fragile in two ways. First, we document that different choices of the regressor matrix that do not impact ordinary least squares (OLS) estimates, such as the choice of baseline category with categorical controls, can move sparsity-based estimates by two standard errors or more. Second, we develop two tests of the sparsity assumption based on comparing sparsity-based estimators with OLS. The tests tend to reject the sparsity assumption in all three applications. Unless the number of regressors is comparable to or exceeds the sample size, OLS yields more robust inference at little efficiency cost.

*Kolesár acknowledges support by the National Science Foundation Grant SES-22049356. All errors are our own. We thank Colin Cameron, Helmut Farbmacher, Helmut Küchenhoff, Bentley MacLeod, Whitney Newey, Alexandre Poirier, and numerous seminar and conference participants for helpful comments, and Max Mongkalakorn for excellent research assistance.

1 Introduction

The linear regression model is the most common workhorse for estimating causal effects. The key assumption allowing for a causal interpretation of the coefficient on the treatment or intervention of interest is that conditional on the set of controls included in the model, the remaining variation in the treatment is as good as random. Since the plausibility of this assumption increases with additional control variables, researchers commonly estimate regression models with many controls.

To fix ideas, consider estimation of the coefficient β on treatment D_i in the regression model

$$Y_i = D_i\beta + W_i'\gamma + U_i, \qquad E[U_i \mid D_i, W_i] = 0, \tag{1}$$

based on n observations indexed by i. Traditional estimation methods like OLS limit the dimension p of the control vector W_i to be smaller than n. This limitation has fueled the development of alternative estimation methods that allow p to exceed n, provided one is willing to make the (approximate) sparsity assumption that the control function $W_i'\gamma$ can be well-approximated, in a mean-squared error sense, by a linear combination of a small number s of the controls, so that, up to log terms, $s^2/n \to 0$. In particular, several papers (see, among others, Javanmard & Montanari, 2014; van de Geer et al., 2014; Zhang & Zhang, 2014) have developed "debiased lasso" estimators of β that are able to purge bias present when eq. (1) is estimated directly using the lasso or related methods. Belloni et al. (2014) propose a corresponding post-double-selection procedure that first estimates the propensity score regression

$$D_i = W_i'\delta + \tilde{D}_i, \qquad E[\tilde{D}_i \mid W_i] = 0, \tag{2}$$

by the lasso, and also runs a lasso regression of the outcome on the controls W_i. In a second step, β is estimated by regressing the outcome on D_i and a union of the controls selected by the two lasso regressions. A number of authors have developed variants of these methods that allow for treatment effect heterogeneity (see, among others, Athey et al., 2018; Belloni et al., 2017; Chernozhukov et al., 2018; Farrell, 2015). These developments have had a large influence on empirical practice in applied economics and related fields, and researchers routinely use these sparsity-based estimators (SBEs) even in the regime $p < n$ as a complement to or replacement of OLS.

This paper argues that this practice may not lead to reliable inference for two reasons. First, SBEs, like the sparsity assumption underlying them, are not invariant to linear reparametrizations of the control matrix W. For instance, while normalization choices such as the choice of baseline category for a set of dummy variables, or how and whether to center

the variables prior to taking powers or interactions, are immaterial when running OLS, they in general affect SBEs. We document that in three empirical applications of SBEs, which include the empirical illustration in Belloni et al. (2014), seemingly innocuous normalization choices of this type induce variation in SBEs of the same magnitude as their standard error. It is thus impossible to interpret the resulting inference without substantively engaging with the question whether the particular normalization choice used is appropriate. Yet, papers using SBEs do not even tell readers what this choice was, let alone argue that it is the right one to obtain a sparse representation. While it is well-known that SBEs are not invariant to linear reparametrizations in the theoretical literature, this practice implies that the extent of their fragility is not currently appreciated, even by researchers in the field.

Second, one might be skeptical of the plausibility of the sparsity assumption more generally. In social science applications, it is often difficult to point to theoretical or institutional arguments for why a small number of controls should be able to soak up most of the confounding. We thus develop two statistical tests of the null hypothesis that (approximate) sparsity holds. Both tests tend to reject in the baseline specification of the three empirical applications. What is more, we are often unable to find any normalization choice for which the tests do not reject, suggesting that this second concern is not just a variation of the first.

Our conclusion is that applied researchers should be wary of using SBEs without a substantive defense of the sparsity assumption for the specific choice of the control matrix used in the analysis. Unless p is close to or exceeds n, a robust default is to simply run OLS.[1]

Our analysis relates to several strands of literature. Many authors mention the SBEs' lack of invariance to linear reparametrization as an undesirable feature. In the specific context of including categorical controls, alternatives to standard lasso regression have been developed that are invariant to the choice of baseline category (e.g. Bondell & Reich, 2009; Gertheiss & Tutz, 2010; Stokell et al., 2021). Similarly, one could perhaps incorporate the choice of a centering constant when taking powers as an additional tuning parameter when fitting SBEs. However, as we discuss further in Section 3.5 below, such solutions come with their own challenges. Using a Bayesian framework, Giannone et al. (2021) present empirical evidence suggesting sparsity may not be a compelling assumption in popular data sets used in economics. Wüthrich and Zhu (2023) give simulation evidence and theoretical arguments showing that even if sparsity arguably holds, SBEs may still display substantial bias in finite samples. Angrist and Frandsen (2022) explore robustness of SBEs to tuning parameter choices.

[1] One only needs to be careful in standard error construction, as OLS no longer consistently estimates the residuals once $p \asymp n$. This renders the usual Eicker-Huber-White standard errors invalid; instead one needs to use a standard error formula that is robust to high-dimensional controls (see, e.g. Cattaneo et al., 2018b; D'Adamo, 2019; Dobriban et al., 2024; Jochmans, 2022; Kline et al., 2020).

The remainder of the paper is organized as follows. In Section 2, we introduce the three empirical examples used to quantify fragility of SBEs: the empirical illustration in Belloni et al. (2014) on the effect of abortion on crime, a Ferrara (2022) study of occupational upgrading by Black southerners, and a study of the effect of moral values on voting behavior by Enke (2020). Our analysis of fragility to linear reparametrizations focuses on two seemingly innocuous normalizations: first, we consider different ways of dropping collinear columns, such as which category to drop when including categorical variables. Second, when W includes powers and interactions of a set of baseline controls, we consider different ways of centering the baseline controls, such as demeaning, subtracting the median or no centering. We find that these normalizations can move the estimates by two standard errors to more. We also consider alternative normalizations that further expand the set of possible specifications of W: expressing a set of categorical controls as indicators for subsets of the categories, rather than just as indicators for each individual category; and continuously varying where a variable is centered prior to taking powers, rather than just considering discrete values such as the mean or the median. These normalizations lead to an even greater variation in the estimates.

In Section 3, we interpret these empirical results through the lens of a thought experiment: suppose a researcher picks one of the possible normalizations at random. How likely would such a random choice lead to an (approximately) sparse representation? We compute that probability in three stylized examples of normalizations and find that it rapidly decreases as function of n. Leaving default choices for how W is constructed to statistical software or happenstance is very unlikely to lead to sparsity.

Section 4 conducts an analysis of the relative efficiency of SBEs and OLS. We focus on the regime where p is smaller, but proportional to n. This is because if $p/n \to 0$, OLS achieves the semiparametric efficiency bound under homoskedastic errors, which limits the arguments in favor of alternative estimators. On the other hand, if p exceeds n, the OLS benchmark is no longer available. We show that under homoskedasticity, the proportional variance reduction of SBEs relative to OLS is capped at $1 - p/n$ (effectively, we can at best avoid a degrees of freedom adjustment). This benchmark forms the basis for our recommendation to simply run OLS unless p is close to n: it yields robust estimates at little efficiency loss.

In Section 5, we develop two tests of the sparsity assumption, again in the regime $p \asymp n$.[2] We circumvent the difficulty that the sparsity assumption only restricts rates, rather than the actual number of non-zero coefficients for a given n, by testing it indirectly. The first

[2]Carpentier and Verzelen (2021) develop a test in the high-dimensional regime $p/n \to \infty$. However, their test requires explicit specification of the sparsity level under the null, as does the test developed by He (2020) in the regime $p < n$. In the context of factor models, Beyhum and Striaukas (2024) consider testing the null of a zero vector against a sparse alternative.

test compares OLS and lasso residuals: under (approximate) sparsity, the residual sum of squares of the lasso should exceed the OLS residual sum of squares only by a small amount. The second test is an application of the Hausman (1978) specification test: any difference between SBE and OLS estimates must be explained by differences in the relative efficiency of the estimators afforded by the sparsity assumption that SBE exploits, but OLS does not. Thus, the common practice of reporting SBE estimates alongside OLS should not be interpreted as a robustness check for the OLS specification; rather, divergence between the estimates indicates a failure of the sparsity assumption.

Section 6 concludes. All proofs are relegated to appendices.

2 Empirical illustrations

In this section we revisit three applications that leveraged SBEs: the application in Belloni et al. (2014, BCH) that probes the investigation by Donohue and Levitt (2001) of the impact of abortion on crime, the Ferrara (2022) study of employment opportunities for Black southerners in the aftermath of World War II (WW2), and the examination of the relationship between moral values and voting behavior by Enke (2020). In each application, we apply the original SBE to eq. (1) after changing the specification of the control matrix W (where rows correspond to the control vectors W_i) in seemingly innocuous ways that do not impact OLS estimates. We show that the impact of these normalizations on SBEs is, on the other hand, substantial. To isolate the effect of the choice of control matrix from other implementation details, our analysis otherwise sticks to software defaults.[3]

In Section 2.1, we consider effects of two normalizations. First, we consider different ways of resolving multicollinearity in the control matrix by changing which columns are dropped to make the matrix W full rank. For example, if multicollinearity arises due to the inclusion of categorical variables, we change which category is dropped. Original implementations of SBEs in each of the three applications remove multicollinearity in W as a data-processing step, similar to standard implementations of least squares regression. Our exercise is therefore equivalent to changing the order of the columns of the control matrix, but otherwise conducting the analysis exactly as in the original. For lasso-based estimators, such a step is typically needed to ensure that so-called "compatibility conditions" or "restricted eigenvalue" assumptions hold; these assumptions ensure fast convergence rates for the lasso (see, e.g. Bickel et al., 2009).[4] Second, when W includes powers and interactions of a baseline set

[3]For estimates using post-double selection, we use the `hdm` package and its defaults, including the choice of the tuning parameter and normalizing the columns of W to have unit variance.

[4]A necessary condition for these assumptions is that submatrices of W with $2s$ columns, where s is the sparsity index, are full rank. It implies that if, say, a categorical variable has fewer than $2s$ categories, we

of controls, we consider different normalizations of the baseline controls (such as demeaning vs subtracting the median) before taking powers and interactions. Appropriately centering the baseline variables puts them on the same scale, and may make the sparsity assumption more plausible and more easily interpretable. Relative to using raw polynomials, it typically also helps with numerical stability.

Section 2.2 considers alternative normalizations of W in the two scenarios above: expressing a set of categorical controls as indicators for subsets of the categories, rather than only as indicators for each individual category; and continuously varying where a variable is centered prior to taking powers and interactions, rather than just considering discrete values such as the mean or the median.

2.1 Normalizations of the control matrix

We begin by considering different ways of resolving multicollinearity issues that arise in each application.

The data in the first application consists of an annual panel of US states over the period 1985–97 originally analyzed by Donohue and Levitt (2001), who ran a two-way fixed effects regression of crime rates on effective abortion rates, state and year fixed effects, and 8 baseline covariates.[5] BCH argue that this set of 8 controls may be insufficient to purge time-varying confounders. They consider a first-differences version of this specification, with 12 time effects and first-differences of the 8 baseline controls, which they augment with 136 further variables obtained by squaring the baseline controls and interacting them with each other and with linear and quadratic trends.[6] In addition, they include 49 variables that are time-invariant within each state, corresponding to initial values and averages of various transformations of the baseline controls, as well as interactions of these 49 variables with a time trend and its square. Since there are only 48 states in the data, these variables span the same column space as state fixed effects. The resulting control matrix has 303 columns, but rank of only 294: because time effects are included, 2 of the time-invariant variables are redundant, as are 2 of the interactions with a time trend and 2 with its square. In addition, one of the baseline controls, a shall-issue dummy, is binary and non-zero only 21 times, but

cannot include all categories as well as the intercept.

[5]These are: lags of the number of prisoners and police per capita, the unemployment rate, per-capita income and beer consumption, poverty rate, AFDC generosity lagged 15 years, and a dummy for a shall-issue concealed carry law.

[6]In particular, BCH add squares of the first differences as well as lags and squared lags of the baseline controls, and they interact the first differences and their squares with a linear and quadratic trend. They also add interactions of the first-differences, and interact them with a linear and a quadratic trend. Since the time series of the shall-carry dummy in each state never changes from 1 to 0, its first difference is also binary. These transformations therefore only yield 136 unique columns.

it is interacted with 24 variables, so that 3 of the interactions are redundant. BCH first drop the collinear columns, and then estimate the model by the post-double lasso estimator that they develop.

Column 2 of Panel A in Table 1 replicates the BCH estimates for each of the three versions of the crime rate outcome variable considered by BCH.[7] Comparing the standard error to the OLS standard error in column 1, we see that the post-double lasso estimator appears to be substantially more precise than the OLS estimate. It is also economically significant: the effective abortion rate is defined as the average legalized abortion rate among arrestee cohorts (the number of abortions per live birth in a cohort weighted by the cohort's share of arrestees, which is outcome-specific). For the violent crime outcome, its standard deviation across states in 1997 is about 1, so that the estimate implies a 16% reduction in crime rate per standard deviation increase in effective abortion rate. However, this result is very sensitive to how we resolve the collinearity. To resolve it, we may drop any 3 of the 24 interactions with the shall-issue dummy, and any 2 of the 49 time-invariant controls; the same holds when we interact these with a time trend and its square. This gives $\binom{24}{3}\binom{49}{2}^3 \approx 3 \times 10^{12}$ possible ways of obtaining a full rank control matrix. Columns 3 and 4 report a range of estimates we obtain by randomly choosing among these possibilities, showing that, depending on the outcome, the estimates move by 1.2 to 1.9 standard errors.

Similar collinearity issues arise in the second application that replicates the analysis in Ferrara (2022), who studies to what extent post WW2 occupational upgrading of Black workers from low-skilled to semi-skilled can be attributed to war casualties among semi-skilled white soldiers. Using a decennial 1920–1960 unbalanced panel of county-level observations in 16 predominantly Southern US states, the study runs a two-way fixed effects specification with county and time fixed effects, interactions between state and time fixed effects, the share of semi-skilled Black workers as the outcome, and treatment given by the white casualty rate interacted with a post-war indicator. To purge time-varying confounders, the study also includes 24 baseline controls (including the county draft rate, WW2 spending, demographic and socioeconomic controls), their squares and interactions, as well as interactions between the 24 baseline controls and state and time effects. In addition, two baseline controls (number of slaves in 1860 and the unemployment rate in 1937) are also included in triple interactions with other controls, time and state effects. After dropping a reference state and a reference year in each interaction, and dropping zero and repeated columns, the resulting matrix has 2270 columns, but rank of only 2252, because the state of Delaware only contains 15 observations, but there are 33 Delaware-specific controls.[8] Ferrara (2022) first

[7]The replication differs slightly from the original (Table 2 in BCH) because the algorithm for dropping collinear columns in BCH had a coding error, yielding a matrix with 296 columns, and a rank of 293.

[8]In particular, since the two baseline controls entering triple interactions are time-invariant and the

Table 1: Fragility of sparsity-based methods to normalizations of the control matrix.

		Sparsity-based estimation				
	OLS	Replication	Range (collinear)		Range (powers)	
Outcome	(1)	(2)	(3)	(4)	(5)	(6)
		A: BCH				
$\Delta \ln$(violent crime/capita)	0.006	−0.160	−0.216	−0.109	−0.160	−0.122
($n = 576$, $p = 294$)	(0.755)	(0.112)	(0.118)	(0.093)	(0.112)	(0.097)
$\Delta \ln$(property crime/capita)	−0.154	−0.110	−0.137	−0.054	−0.127	−0.078
($n = 576$, $p = 294$)	(0.223)	(0.045)	(0.045)	(0.047)	(0.038)	(0.041)
$\Delta \ln$(murder/capita)	2.240	−0.131	−0.225	−0.061	−0.149	−0.066
($n = 576$, $p = 294$)	(2.819)	(0.146)	(0.140)	(0.149)	(0.151)	(0.161)
		B: Ferrara (2022)				
% Semiskilled Black workers	0.118	0.548	0.242	0.657	0.482	0.548
($n = 4,903$, $p = 2,252$)	(0.126)	(0.167)	(0.126)	(0.153)	(0.137)	(0.167)
		C: Enke (2020)				
Trump − avg. GOP	−3.676	−1.921	−2.120	−1.840		
($n = 2,452$, $p = 978$)	(1.416)	(0.941)	(0.950)	(0.934)		
Voted for Trump in 2016	−12.399	−12.335	−12.357	−11.955		
($n = 2,852$, $p = 1,068$)	(1.357)	(1.050)	(1.049)	(1.028)		
Voted for Trump in primaries	−5.318	−7.778	−8.618	−7.716		
($n = 1,544$, $p = 824$)	(2.660)	(1.544)	(1.534)	(1.541)		

Notes: Col. 2 replicates SBEs for the BCH, Ferrara (2022), and Enke (2020) studies discussed in the text. Col. 1 reports the unpenalized OLS estimate for the same specification. Cols. 3 and 4 report the range of estimates obtained under alternative ways of dropping collinear columns of the control matrix. Cols. 5 and 6 report the range of estimates obtained under alternative normalizations of the controls prior to taking powers and interactions: no normalization, demeaning, subtracting the median, and setting the range to $[-1, 1]$ and $[0, 1]$. The outcome variables in Panel C are multiplied by 100. "Trump − avg GOP" refers to the difference between the indicator for voting for Trump in the 2016 general election minus the vote share given to Republican candidates in the previous two presidential elections. Standard errors are given in parentheses. These are robust for Enke (2020), clustered by state for BCH and by county for Ferrara (2022).

forms a full-rank control matrix, and then uses double-t selection to estimate the model.[9]

We replicate this specification in column 2 of Panel B in Table 1.[10] However, similar to panel A, there are many ways of resolving multicollinearity in the control matrix, because there are many ways of specifying a reference state and a reference year in each interaction, and, if we keep 3 Delaware county effects, $\binom{30}{18}$ ways of dropping Delaware-specific controls. As shown in columns 3 and 4, depending on how the multicollinearity is resolved, the estimates vary by over three standard errors depending on how we order the state and time effects.[11]

Our final empirical example uses data from Enke (2020), who examines how voters' moral values affect their voting patterns. Enke uses survey data to construct an index of the relative importance of universalist moral values (individual rights, justice, fairness) vs communal values (loyalty, respect). He then runs a regression of three different measures of voting behavior on this index, controlling for 10 continuous or binary controls, as well as 5 sets of categorical variables.[12] Columns 1 and 2 in panel C of Table 1 replicate the OLS and post-double lasso estimates reported in the paper. Here the inclusion of multiple sets of categorical variables necessitates a specification of a reference category for each set. Columns 3 and 4 report the range of estimates obtained by changing these reference categories, which varies between a third and a half of standard error depending on the outcome. This is a narrower range than in panels A and B, albeit still large enough to affect the economic interpretation of the estimates.

Columns 5 and 6 of Table 1 show the range of estimates we obtain from considering different normalizations when taking powers and interactions between variables in the BCH and Ferrara (2022) applications. In addition to no normalization (the choice in the orig-

specification includes county effects, we need to exclude the main effects of these controls, their squares and state-interactions, which yields a control matrix with 2588 columns. 154 columns are collinear since we need to drop a reference state and year in each interaction, and 18 Delaware-specific controls are also collinear. In addition, 164 columns are repeated or zero, yielding a rank equal to 2252.

[9]Specifically, Ferrara (2022) first runs a selection step that regresses the outcome and the treatment on the control matrix using OLS, and selects controls with a t-statistic that is larger than 2.575 in absolute value in either regression. In a second step, the outcome is regressed on the treatment and selected controls using OLS.

[10]The estimate is slightly higher than in the original because, to keep missing data and normalization issues separate, we restrict the sample to the 4,903 observations with no missing values in both estimation steps described in footnote 9. In addition, we include year fixed effects in both the first and second step, rather than exclude them from the first step.

[11]We also considered sensitivity of the results to ordering of the columns in the control matrix when the double-t estimator is replaced by the post double lasso estimator. This yielded a range of variation in the estimates equal to about two thirds of a standard error.

[12]The continuous or binary controls comprise: political liberalism, log of household income, education, log of population density of respondent's ZIP code, religiosity, gender, employment indicator, altruism, measure of trust, and the absolute value of the moral values index. The categorical variables are: county, year of birth, religious denomination and occupation fixed effects.

inal analyses), we consider demeaning, centering at the median, and setting the range to $[-1, 1]$ and $[0, 1]$. The table shows that such normalizations can again substantially move the estimates, by up to 1.3 standard errors.

2.2 Alternative normalizations of the control matrix

We now consider two alternatives to the construction of W in the presence of categorical variables and power and interactions.

In Section 2.1, a categorical variable with k categories was always specified as a set of $k-1$ dummies for individual categories, with a reference dummy dropped to prevent collinearity. We now consider expressing such variables as $k-1$ indicators for different subsets of the k categories. As discussed in Section 3.3, if the subsets are carefully chosen, such a specification may be more likely to lead to sparsity in certain contexts. Here we pick the subsets at random, subject to the constraint that they span the same column space. For the Enke (2020) application, columns 3 and 4 of Table 1 correspond to a special case of this exercise, when the subsets are constrained to be of cardinality one. Correspondingly, the range of variation in the estimates over all possible subsets, reported in columns 1 and 2 of Table 2, is considerably greater, exceeding 0.9 standard errors in all regressions. For the Ferrara (2022) application, the range exceeds two standard errors.[13]

Similarly, the centering of baseline variables at the mean vs the median we considered in the previous subsection can be thought of as a special case of a more general problem: we could in principle choose the centering constant λ to be any number. At a theoretical level, we are not aware of arguments for why $\lambda = 0$ (no normalization), or setting it to the sample mean should more plausibly lead to sparsity than other choices. To further explore the sensitivity to the choice of λ, we standardize the baseline variables, and then vary the specification of W by picking λ from the range $[-1, 1]$ for each baseline variable. To align the exercise with the analytical calculations in Section 3.4, we use Hermite rather than raw polynomials. Columns 3 and 4 of Table 2 show that the resulting variation in the estimates is again substantial, exceeding 1.7 standard errors in all specifications. The sign of the estimated effect is also sensitive to the choice of λ in two of the BCH specifications.

[13]The range does not exceed that in columns 3 and 4 of Table 1 because we do not vary which Delaware-specific controls are dropped.

Table 2: Fragility of sparsity-based methods to further normalizations of the control matrix.

	Sparsity-based estimation			
	Range (category sums)		Range (offset)	
Outcome	(1)	(2)	(3)	(4)
	A: BCH			
$\Delta \ln$(violent crime/capita)			−0.239	0.107
($n = 576$, $p = 294$)			(0.105)	(0.093)
$\Delta \ln$(property crime/capita)			−0.193	−0.011
($n = 576$, $p = 294$)			(0.047)	(0.053)
$\Delta \ln$(murder/capita)			−0.338	0.128
($n = 576$, $p = 294$)			(0.160)	(0.195)
	B: Ferrara (2022)			
% Semiskilled Black workers	0.401	0.685	0.397	0.646
($n = 4,903$, $p = 2,252$)	(0.134)	(0.162)	(0.148)	(0.164)
	C: Enke (2020)			
Trump − avg. GOP	−2.333	−1.476		
($n = 2,452$, $p = 978$)	(0.918)	(0.906)		
Voted for Trump in 2016	−12.946	−11.202		
($n = 2,852$, $p = 1,068$)	(1.029)	(1.008)		
Voted for Trump in primaries	−9.269	−7.083		
($n = 1,544$, $p = 824$)	(1.526)	(1.525)		

Notes: Cols. 1 and 2 report the range of estimates obtained under alternative ways of expressing a set of categorical variables. Cols. 3 and 4 report the range of estimates obtained when powers are constructed using Hermite polynomials, with an offset λ ranging between −1 and 1. See notes to Table 1 for a description of outcome variables. Standard errors are given in parentheses. These are robust for Enke (2020), clustered by state for BCH and by county for Ferrara (2022).

3 Sparse representations are rare

In this section we consider the thought experiment of picking a particular linear representation of the controls W at random. In particular, we consider (i) a full rotation of the column space; (ii) different ways of controlling for categorical variables; and (iii) an offset for the scalar variable in a polynomial series regression. We find that sparse representations are exceedingly rare in each example. This suggests that in a given application, undiscerning default choices for expressing the column space will typically not satisfy the sparsity assumptions. This helps to explain the findings of the previous section: most normalization choices cannot yield sparse representations. As we discuss in detail in Section 3.5 below, it also implies that the representation of the control matrix is a substantive choice and not just an implementation detail, like, say, the choice of the penalty parameter: it impacts not just the finite-sample estimate, but also the large-sample validity of SBE-based inference.

3.1 Approximate sparsity

For simplicity, the calculations in this section focus on sparsity in the outcome regression (1). Let A be a full-rank $p \times p$ matrix. Then an alternative expression for the column space of W_i is $\tilde{W}_i = AW_i$ with coefficient $\tilde{\gamma} = A'^{-1}\gamma$, so that $W_i'\gamma = \tilde{W}_i'\tilde{\gamma}$. Note that any such transformation leaves the OLS coefficient on D_i numerically unaltered.

In our examples, γ is exactly sparse, with just one non-zero element, $\|\gamma\|_0 = 1$. The researcher uses the transformed regressors $\tilde{W}_i$ with coefficients $\tilde{\gamma}$, which typically is not exactly sparse. However, for sparsity-based inference to be asymptotically valid, it suffices for $\tilde{\gamma}$ to be *approximately sparse*. In particular, under our maintained assumption that $p \asymp n$, it suffices that for a sparsity index

$$s = o(\sqrt{p}/\log p) \tag{3}$$

the mean square approximation error of $W_i'\gamma$ satisfies

$$\min_{\|v\|_0 \le s} E[(W_i'\gamma - \tilde{W}_i'v)^2] = O(s/p) \tag{4}$$

(cf. Belloni et al., 2014, page 614). Here, and throughout the paper, all limits are taken as $p \to \infty$ (or, equivalently, as $n \to \infty$).

3.2 Rotation

By way of establishing a benchmark, we start with an extreme case, and consider the set of transformations

$$\tilde{W}_i = RW_i, \quad R \in \mathbf{R} = \{A \in \mathbb{R}^{p\times p} \colon A'A = I_p,\ \det(A) = 1\},$$

that is, $\tilde{W}_i$ is obtained by rotating W_i by R, and $\tilde{\gamma} = R\gamma$. For the purposes of studying sparsity, this is a very large set of transformations, as for any γ with $\|\gamma\|_2 > 0$, there exists R such that $\tilde{\gamma}$ has only one non-zero element. At the same time, there also exists an R such that all elements of $\tilde{\gamma}$ are equal (and equal to $\|\gamma\|_2/\sqrt{p}$).

Now suppose we take a random draw $\mathcal{R}$ from $\mathbf{R}$, with distribution equal to the Haar measure (so that for any $R \in \mathbf{R}$, $R\mathcal{R} \sim \mathcal{R}$). Recall that the multivariate standard normal distribution is spherically symmetrical, that is, for $\mathcal{Z} \sim \mathcal{N}(0, I_p)$ and any $R \in \mathbf{R}$, $R\mathcal{Z} \sim \mathcal{Z}$. This implies that the induced distribution of $\tilde{\gamma} = \mathcal{R}\gamma$ satisfies

$$\mathcal{R}\gamma \sim \frac{\|\gamma\|_2}{\|\mathcal{Z}\|_2}\mathcal{Z}, \tag{5}$$

so the p elements of $\mathcal{R}\gamma$ are, up to a common rescaling, i.i.d. standard normal. The normal distribution is thin-tailed, making it very unlikely that a few largest absolute values of $\mathcal{Z}$ dominate. This suggests that with very high probability, $\mathcal{R}\gamma$ is not approximately sparse, as formalized in the following result.

Theorem 1. *Suppose that the eigenvalues of $E[W_iW_i']$ are bounded away from zero and infinity, and that $\|\gamma\|_2 \asymp 1$. Then the logarithm of the probability that the model with regressor $\tilde{W}_i = \mathcal{R}W_i$ satisfies eqs. (3) and (4) is of the order $-\frac{p}{4}\log p$.*

Given that a sparsity assumption allows for more informative inference, and that any coefficient vector can be rotated into an (extremely) sparse one, it is not surprising that a random rotation only rarely yields a sparse representation. The contribution of Theorem 1 is to quantify just how exceedingly rare this event is: for $p \geq 50$, say, $p^{-p/4} < 10^{-21}$. Of course, researchers rarely consider all rotations when specifying the regressor matrix. The relevance of this quantification lies in showing that even when the measure of plausible rotations is orders of magnitude smaller than the full set, the probability of arriving at a sparse representation is still minuscule.

3.3 Categorical data

An empirically common form of controls is dummies for categorical variables. If there are p underlying categories, and a constant is included, then one must designate a reference category by dropping one dummy to avoid perfect collinearity. Suppose the model is exactly sparse with sparsity index s when one chooses the right reference category. This means that $p-s$ categories have the same coefficient as the reference category. Thus, if one were to pick a reference category at random, one has a $1-s/p$ chance of inducing sparsity, which is a probability close to one under (3).

However, an assumption that many categories have the same effect is not the only way to induce sparsity for a categorical variable. Suppose, for instance, that we want to nonparametrically control for a variable that measures age. Then a plausible form of sparsity arises under an assumption that the age effect is a step function, with some dividing line between young and old. If this threshold is not close to either very young or very old, then a reference category approach will not yield sparsity, but an appropriate re-expression of the column space will. Or maybe there are three distinct coefficients, one for the young, one for the middle-aged and one for the old, which again requires a different specification of the fixed effects to induce sparsity. Such specifications are less common in applied work. Arguably this is not because they are a priori less implausible, but rather, it is well understood that under OLS, they are all equivalent to the reference category specification. However, when imposing a sparsity assumption on the coefficients, this equivalence breaks down.

A general specification of the column space with categorical data takes the form

$$W_i = A_0 Z_i, \quad A_0 \in \mathbf{A} = \{A \in \mathbb{R}^{p\times p} \colon A_{ij} \in \{0,1\}, \quad A \text{ is full rank}\}, \tag{6}$$

where the Z_i indicates the baseline categories (so each Z_i is a column of I_p). By construction, elements of W_i are all equal to zero or one in this specification, and since $A \in \mathbb{R}^{p\times p}$ is full rank, the constant vector is part of the column space spanned by W_i.

Suppose with $A_0 \in \mathbf{A}$, the coefficient vector on W_i is sparse. Now consider picking $\mathcal{A} \in \mathbf{A}$ at random by repeatedly drawing $p \times p$ matrices with i.i.d. Bernoulli(q) entries, $0 < q \leq 1/2$, until we find one that is full rank. By Theorem A of Tikhomirov (2020), the probability of discarding a matrix in this process is vanishingly small, as it is smaller than $(1-q+\varepsilon)^p$ for all $\varepsilon > 0$ and large enough p. This allows us to essentially ignore rank-deficient matrices, yielding the following result.

Theorem 2. *Suppose a single coefficient on $W_i = A_0 Z_i$ is constant and non-zero, and the number of zeros K in the corresponding row of A_0 satisfies $0 < \lim_{n\to\infty} K/p < 1$. If all baseline categories have population fractions of the same order, then the probability that the*

model with $\tilde{W}_i = \mathcal{A}Z_i$ *satisfies eqs.* (3) *and* (4) *is no larger than*

$$(1 - q + \varepsilon)^K$$

for all $\varepsilon > 0$ *and large enough* p.

3.4 Offset in Hermite polynomial regression

Our last example involves a large number of technical regressors. Specifically, suppose there is a scalar variable $z_i \sim_{\text{i.i.d.}} \mathcal{N}(0,1)$ which we want to control for nonparametrically by a series regression. We consider scaled Hermite polynomials

$$H_j(x) = (j!)^{-1/2}(-1)^j e^{x^2/2}\frac{d^j}{dx^j}e^{-x^2/2}, \quad j = 0, \ldots, p-1$$

which are particularly convenient for an underlying standard Gaussian variable, since we have $E[\tilde{W}_i\tilde{W}_i'] = I_p$ for $\tilde{W}_{i,j} = H_{j-1}(z_i)$.

Suppose further that if z_i is shifted by a constant $\lambda \in \mathbb{R}$, the expansion is extremely sparse, with only the highest order term being non-zero, that is, $Y_i = D_i\beta + H_{p-1}(z_i + \lambda) + U_i$. If we were to draw λ uniformly on $[0,1]$, what is the probability that a researcher ignoring the shift obtains an approximately sparse representation? The following result shows that the probability will be bounded above by $\varepsilon/\log(p)$, for all $\varepsilon > 0$ and large enough p.

Theorem 3. *Suppose* $\lambda = L/\log p$, $L > 0$. *If* L *is fixed, then for* $1 \le j \le L\sqrt{p}/\log p$ *and* $p \ge \max\{(2L)^2, 6\}$, $\tilde{\gamma}_{p-j}^2 \ge Ce^{j/2}$, *where* C *is an absolute constant, and eqs.* (3) *and* (4) *fail. In contrast, if* $L \to 0$, *eqs.* (3) *and* (4) *hold.*

While the rate of $\log p$ in Theorem 3 is quite slow, note that when sparsity fails, it does so dramatically in the sense that many coefficients diverge. This is not surprising. It is well-known that assuming that the first s (say) terms of the polynomial expansion provide a good approximation to the regression function amounts to imposing a smoothness assumption on it. In this case, the exact basis approximation would matter little, and sparsity-based methods would perform well regardless of the choice of λ. However, under smoothness SBEs are not needed, since a low-order series expansion (i.e. only controlling for the first s polynomial terms, with $s \ll p$) already performs well.

In the current context with technical regressors, the appeal of SBEs is that the assumption of sparsity they leverage is weaker than assuming smoothness, since we allow for the best s-term approximation to include high order polynomials. But if a good approximation to the regression function indeed requires high-order terms, it must be very non-smooth—and

for non-smooth functions even small changes in the approximating basis imply wild swings in the coefficients. In other words, while sparsity is indeed weaker than smoothness, the particular class of non-smooth functions sparsity allows for is tightly linked to the basis used. If one suspects non-smoothness, and one wants to avoid running OLS with many polynomial terms, one therefore needs a substantive argument for why the particular basis chosen is likely to capture it.

3.5 Discussion

The fact that the sparsity assumption depends on normalization choices in the specification of W is well-known in the theoretical literature. The three stylized thought-experiments above quantify the extent of this dependence: the vast majority of normalization choices do not induce an approximately sparse representation. The results in Section 2 show that this lack of invariance induces variation in SBE estimates that is of the same order as sampling uncertainty. This fragility appears underappreciated in the literature. We reviewed 50 most cited papers applying SBEs, and found that, with one exception, none made explicit what the normalization choices were, let alone discussed why they were appropriate to induce sparsity.[14]

A natural reaction to these results is to seek to modify SBEs in a way that they adapt to different forms of sparse representations. For instance, it may be possible to treat the offset λ as another parameter to be estimated. Similarly, a number of proposals have been developed that modify the lasso to make it invariant to the choice of reference category, including the group lasso (Yuan & Lin, 2006), or variants of fused lasso (Tibshirani et al., 2005), such as all-pairs penalties (Bondell & Reich, 2009; Gertheiss & Tutz, 2010) or the SCOPE estimator (Stokell et al., 2021). When combined with the debiasing techniques developed in the literature on debiased lasso (e.g. Belloni et al., 2014; Javanmard & Montanari, 2014; van de Geer et al., 2014; Zhang & Zhang, 2014), these approaches may yield algorithms that are less fragile.

However, it appears non-trivial to extend these modifications to the more complicated designs encountered in practice, such as when interactions are present, or to handle collinearity issues arising due to limited variability in the controls. Both the BCH and the Ferrara (2022) applications of Section 2 exhibit multicollinearity that goes beyond the issue of choosing baseline categories.

[14]We ordered all papers citing one of the "debiased lasso" papers mentioned in the introduction by number of citations on Google Scholar. One paper mentioned that it included all dummy variables and did not pick a reference category; the paper did not discuss the implications of this choice for the restricted eigenvalue condition, or whether other normalization choices were made.

An alternative approach could be to exhaustively consider all plausible normalization choices, analogous to our analysis in Section 2, and report the union of the confidence intervals. However, this is practically difficult, as the number of possible normalizations is often very large. Moreover, such an approach would only yield valid inference if at least one of the considered choices induces a sparse representation.

To ensure that the latter condition holds, one needs a substantive argument. In some instances, such arguments may necessitate further expanding the set of normalization choices. For instance, the grouped lasso approach in the references cited above require that the baseline categories can be partitioned into a small number of types, with all categories of the same type having (nearly) identical effect on the outcome variable. In the context of controlling for profession dummies, say, this amounts to an assumption that there are only a few profession types with heterogeneous effects, which is a strong assumption. It may be more reasonable to instead posit that professions are characterized by a combination of a small number of latent skills, and these skills affect the outcome in an additive manner. This then yields a different set of plausible normalizations that are not captured by grouped lasso-type approaches.

In absence of substantive arguments, the sparsity assumption may well fail to hold, even after considering a large set of potential normalization choices. We present evidence for this in our applications in Section 5 below. Thus, fully data-driven methods to modify SBEs such that they adapt to an appropriate sparse representation regardless of the context does not seem like a feasible solution.

Applications of SBEs require researchers to take a stand on why a particular representation (or a set of representations) admits a sparse representation based on domain-specific knowledge. This is analogous to using problem-specific arguments to defend other substantive assumptions, such as assuming selection on observables. But it is at odds with the purported appeal of SBEs to be fully automatic, with the aim of reducing the researchers' degrees of freedom in model choice.

While the sensitivity of SBEs to the control matrix specification may seem similar to the usual issue that with flexible methods, implementation issues, such as tuning parameter choice, can matter a lot in finite samples, it is fundamentally different. It also differs from the well-known fact that the choice of basis matters in implementing series regression. Under smoothness, many basis choices for series regression are theoretically justified, as well as first-order equivalent in particular settings such as the partially linear model. Large sample validity of flexible methods can be justified under a wide range of tuning parameter choices, including the choice of penalty in implementing SBEs. In contrast, because the specification of the control matrix affects the validity of the sparsity assumption, and thus large-sample

validity of inference, it is a substantive modeling choice, not merely an implementation detail. It needs to be defended as such.

4 Efficiency gains under sparsity

One argument for using SBEs when p is large, but still smaller than the sample size, is that OLS estimates are too noisy. In this section, we quantify the potential efficiency gains of alternative estimators relative to OLS, and we argue that the potential gains are limited unless p is comparable to n.

We consider the model given in eqs. (1) and (2) in the introduction. We maintain the assumption that the residuals U_i and $\tilde{D}_i$ are conditionally mean zero to avoid technical complications; the assumption could be replaced by an assumption that restricts these conditional means to be small, so that the linear specifications in eqs. (1) and (2) could be thought of as approximating non-linear conditional means (see, e.g., Cattaneo et al., 2018b). Using the Frisch-Waugh-Lowell theorem, we can write the OLS estimator of β as

$$\hat{\beta}_{OLS} := \frac{\sum_{i=1}^n \ddot{D}_i Y_i}{\sum_{i=1}^n \ddot{D}_i^2},$$

where $\ddot{D}_i$ denotes the sample propensity score residual from estimating eq. (2) by OLS. Specifically, employing the usual matrix notation, $\ddot{D}_i$ corresponds to the ith element of $\ddot{D} := (I - P)D$, where $P = WW^+$ is the projection matrix onto the space spanned by the controls. Here W^+ denotes the pseudo-inverse (so that if W is full rank, $P = W(W'W)^{-1}W'$).

To analyze $\hat{\beta}_{OLS}$, we need to impose more structure on the model in eqs. (1) and (2).

Assumption 1. For some constants $\eta > 0$ and $K \geq 1$: (i) $\{(U_i, \tilde{D}_i)\}_{i=1}^n$ are independent across i conditional on W; (ii) $E[|U_i|^{2+\eta} \mid D, W] + E[|\tilde{D}_i|^4 \mid W] \leq K$ uniformly over i; (iii) $1/E[\tilde{D}_i^2 \mid W] + 1/E[U_i^2 \mid D, W] \leq K$ uniformly over i; (iv) $\limsup_{n\to\infty} p/n < 1$.

Parts (i)–(iii) of Assumption 1 are standard assumptions on sampling and the regression errors, ensuring that the errors are neither too thick-tailed nor degenerate. Part (iv) ensures that, when OLS estimates of (1) implicitly estimate the propensity score regression, we do not overfit so much that we eliminate any variation in the sample residuals $\ddot{D}_i$. This overfitting precludes using OLS in settings with $p > n$. However, as long as the limit of p/n is strictly smaller than 1, the OLS estimator remains asymptotically normal with the usual sandwich expression for its standard error:

Lemma 1. *Consider the model in eqs. (1) and (2). Under Assumption 1,*

$$\frac{\hat{\beta}_{OLS} - \beta}{s_{OLS}} \xrightarrow{d} \mathcal{N}(0,1), \qquad s^2_{OLS} = \frac{1}{(\ddot{D}'\ddot{D})^2} \sum_{i=1}^{n} \ddot{D}_i^2 U_i^2.$$

The fact that OLS remains asymptotically normal in the regime $p \asymp n$ is not new—results similar to Lemma 1 appeared previously in, for instance, Cattaneo et al. (2018b) who build on earlier work by Mammen (1993). We state it here under a simpler set of assumptions to allow us to consider its implications for potential efficiency gains. To this end, observe that the asymptotic variance is in general larger than the semiparametric efficiency bound unless $p/n \to 0$. In particular, when the errors U_i are homoskedastic, the efficient standard error is given by the square root of

$$s^2_* = \frac{1}{(\tilde{D}'\tilde{D})^2} \sum_{i=1}^{n} \tilde{D}_i^2 U_i^2,$$

in the sense that the semiparametric efficiency bound is given by the probability limit of s^2_*/n (see, e.g., Robinson, 1988). Semiparametrically efficient estimators $\hat{\beta}^*$ of β thus need to satisfy

$$\hat{\beta}^* - \beta = (\tilde{D}'\tilde{D})^{-1}\tilde{D}'U + o_p(n^{-1/2}), \tag{7}$$

This lack of semiparametric efficiency is not a deficiency of OLS: under Gaussian errors, one-sided t-tests based on OLS are uniformly most powerful. Rather, it reflects the fact that when $p \asymp n$, achieving the semiparametric efficiency bound requires some type of restriction on γ. It turns out that restricting γ to be sparse (see Section 3.1) is sufficient; indeed a number of SBEs satisfy eq. (7) (e.g. Belloni et al., 2014; Javanmard & Montanari, 2014; van de Geer et al., 2014; Zhang & Zhang, 2014).

The standard error ratio s_*/s_{OLS} thus represents the potential efficiency gain from imposing sparsity. When U_i is homoskedastic and Assumption 1 holds, the ratio satisfies

$$\frac{s^2_*}{s^2_{OLS}} = (1-p/n)\frac{\ddot{D}'\ddot{D}/(n-p)}{\tilde{D}'\tilde{D}/n}(1+o_p(1)) = (1-p/n)\kappa(1+o_p(1)), \qquad \kappa = \frac{E[\hat{\sigma}^2_{\tilde{D}}]}{\sigma^2_{\tilde{D}}},$$

where $\sigma^2_{\tilde{D}}$ is the variance of the error in the propensity score regression, and $\hat{\sigma}^2_{\tilde{D}} = \ddot{D}'\ddot{D}/(n-p)$, the mean squared error in the propensity score regression, can be thought of as an estimator of $\sigma^2_{\tilde{D}}$. The ratio κ measures the relative bias of this estimator, with $\kappa = 1$ when the estimator is unbiased. This is the case when $\tilde{D}_i$ is homoskedastic, and the standard error ratio s_*/s_{OLS} then simplifies to a degrees-of-freedom correction, $\sqrt{1-p/n}$, as noted previously in Cattaneo et al. (2018a). When $\tilde{D}_i$ is heteroskedastic, $\hat{\sigma}^2_{\tilde{D}}$ may display downward bias if the leverages P_{ii} are positively correlated with the conditional variances $E[\tilde{D}_i^2 \mid W]$; this can be seen from

writing $\kappa = E[\sum_i (1 - P_{ii}) E[\tilde{D}_i^2 \mid W]] / \sum_i (1 - P_{ii}) E[\tilde{D}_i^2]$. However, the correlation would need to be substantial to induce downward bias that is sizable enough to allow for large efficiency gains. When $p/n = 0.2$ and $\kappa = 0.9$, for instance, corresponding to 10% downward bias, the standard error ratio implies a $1 - \sqrt{0.8 \cdot 0.9} = 15.1\%$ reduction in standard error. Consequently, the ratio p/n needs to be much larger than 0.2 to allow for sizable efficiency gains under sparsity: at $\kappa = 0.9$, we need $p/n \geq 6/16 \approx 0.38$ to reduce the standard errors by more than 25%.

While these efficiency calculations rely on homoskedastic errors U_i, similar conclusions are likely to hold under heteroskedasticity. The argument for this is that, in general,

$$\frac{s_*}{s_{OLS}} = \frac{s_*/s_{*,\text{hom}}}{s_{OLS}/s_{OLS,\text{hom}}} \sqrt{(1 - p/n)\kappa} \cdot (1 + o_p(1)),$$

where $s_{*,\text{hom}} = \sigma_U / \sqrt{\tilde{D}'\tilde{D}}$ and $s_{OLS,\text{hom}} = \sigma_U / \sqrt{\ddot{D}'\ddot{D}}$ are homoskedasticity-only standard error formulas. The ratios $s_*/s_{*,\text{hom}}$ and $s_{OLS}/s_{OLS,\text{hom}}$ measure the magnitude of the heteroskedasticity correction on the standard errors. Heteroskedasticity would thus need to have a large *differential* impact on the standard errors for OLS versus $\hat{\beta}_*$ for the standard error ratio to deviate much from $\sqrt{(1 - p/n)\kappa}$. But magnitudes of heteroskedasticity corrections tend to be modest in practice, so that substantive efficiency gains are unlikely when the controls number 20% or less of the sample size. In such cases, simply running OLS offers a robust method of inference at little efficiency loss.

5 Testing sparsity

In this section, we develop two tests of the sparsity assumption under our maintained assumption that p is smaller than, but proportional to the sample size. We then apply these tests to the empirical examples studied in Section 2.

5.1 Hausman test

The first test we develop is a simple application of the idea popularized by Hausman (1978) that if we have two estimators, one of which is more efficient, but requires stronger assumptions for its validity, we can indirectly test these stronger assumptions by checking whether the estimates are statistically significantly different. The next lemma formalizes the result.

Lemma 2. *Consider the model in eqs. (1) and (2). Suppose Assumption 1 holds, and that*

$\mathrm{tr}(P)/n \geq \tilde{K}$ *a.s. for some* $\tilde{K} > 0$*, Then, for any estimator satisfying eq.* (7)*,*

$$\frac{\hat{\beta}_{OLS} - \hat{\beta}_*}{s_H} \xrightarrow{d} \mathcal{N}(0,1), \qquad s_H^2 = \sum_{i=1}^n Z_i^2 U_i^2,$$

where $Z_i = \frac{\ddot{D}_i}{\ddot{D}'\ddot{D}} - \frac{\tilde{D}_i}{\tilde{D}'\tilde{D}}$.

Additionally, suppose the regression functions admit the decomposition $W_i'\gamma = f(W_i) + r_\gamma(W_i)$ *and* $W_i'\delta = g(W_i) + r_\delta(W_i)$*, where the remainder terms satisfy (i)* $\frac{1}{n}\sum_{i=1}^n E[r_\delta(W_i)^2 + r_\gamma(W_i)^2] \to 0$*; (ii)* $\frac{1}{n}\sum_{i=1}^n E[r_\gamma(W_i)^2 r_\delta(W_i)^2]$ *is bounded; and (iii)* $n\sum_{i=1}^n (Z_i^2 U_i^2 - \tilde{z}_i^2 (U_i + r_\gamma(W_i))^2) = o_p(1)$*, with* $\tilde{z}_i = \frac{\ddot{D}_i}{\ddot{D}'\ddot{D}} - \frac{\tilde{D}_i + r_\delta(W_i)}{\sum_{i=1}^n (\tilde{D}_i + r_\delta(W_i))^2}$*. Suppose also that for some estimates* $\hat{U} = \hat{U}(Y, D, W)$ *and* $\hat{D} = \hat{D}(W, D)$ *(iv)* $\max_i |\hat{U}_i - U_i - r_\gamma(W_i)| + \max_i |\hat{D}_i - D_i - r_\delta(W_i)| = o_p(1)$*. Then the same conclusion holds with* s_H^2 *replaced by* $\hat{s}_H^2 = \sum_{i=1}^n \hat{Z}_i^2 \hat{U}_i^2$*, where* $\hat{Z}_i = \frac{\ddot{D}_i}{\ddot{D}'\ddot{D}} - \frac{\hat{D}_i}{\hat{D}'\hat{D}}$.

The second part of Lemma 2 allows us to construct a simple plug-in estimator of the standard error s_H based on lasso or post-lasso residuals. In particular, when the regression functions admit a sparse approximation in the sense of eqs. (3) and (4), the additional condition (i) in the second part of Lemma 2 will hold with f and g given by the best sparse approximations to $W_i'\gamma$ and $W_i'\delta$, and condition (iv) will hold for the lasso or the post-lasso residuals. Conditions (ii) and (iii) are high-level conditions ensuring that if we include the approximation errors r_δ and r_γ in the definition of the residuals, replacing U_i with $U_i + r_\gamma(W_i)$, and $\tilde{D}_i$ with $\tilde{D}_i + r_\delta(W_i)$, this has negligible impact on the standard error s_H in large samples; it is similar to the condition ASTE (P) (v) in Belloni et al. (2014). We note that Lemma 2 only imposes very weak conditions on the control matrix W, allowing it to be reduced-rank, so long as the rank is proportional to p, and allowing the rows to be dependent, and not identically distributed.

Sometimes, SBEs are used as a "robustness check" alongside a main specification based on OLS. Lemma 2 shows that such practice is in fact the opposite of a robustness check: if the two estimates are not close to one another this indicates failure of the sparsity assumption rather than lack of robustness in the OLS estimates. When U_i is homoskedastic, the Hausman standard error may be written as

$$s_H^2 = (s_{OLS}^2 - s_*^2)(1 + o_p(1)) = s_*^2 \left(\frac{1}{(1 - p/n)\kappa} - 1 \right) (1 + o_p(1)),$$

so that when the efficiency gain $\sqrt{(1 - p/n)\kappa}$ is small, the two estimates need to be tightly coupled, within a fraction of the SBE standard error.

5.2 Residual test

Our second approach to testing sparsity is based on the idea that if the identities $\mathcal{S}^* \subseteq \{1, \dots, p\}$ of the controls that give the best sparse approximation to the regression function are known, testing sparsity is equivalent to testing that the coefficients on the remaining controls are small, which can be gauged using a conventional F-statistic. Although the identities $\mathcal{S}^*$ are unknown in practice, it turns out that under the null hypothesis of sparsity, the residuals from the infeasible short regression that only includes controls in $\mathcal{S}^*$ are sufficiently well approximated by residuals from a lasso regression, so that the hypothesis can be tested by comparing the lasso and OLS residuals.

To make the result precise, consider a linear regression

$$\mathsf{Y}_i = X_i'\alpha + \epsilon_i, \quad E[\epsilon_i \mid X_i] = 0, \qquad i = 1, \dots, n, \tag{8}$$

with ϵ_i independent across i, conditional on the regressors, and $p := \dim(X_i) < n$. In our setup, eq. (8) may correspond to one of three regressions. For testing sparsity of γ in eq. (1), which all SBEs require, we can set $\mathsf{Y}_i = Y_i$ and $X_i = (D_i, W_i')'$. For testing sparsity in the propensity score regression, which is needed, for instance, for the validity of the post-double selection estimator of Belloni et al. (2014), we can set $\mathsf{Y}_i = D_i$ and $X_i = W_i$. Sparsity in both regressions can be tested jointly in the reduced form regression of $\mathsf{Y}_i = Y_i$ onto $X_i = W_i$.

We wish to test the assumption that the regression function $X_i'\alpha$ admits a sparse approximation. To work out the implications of this hypothesis, we introduce some additional notation. For a subset $\mathcal{S} \subseteq \{1, \dots, p\}$ of the regressors, let X_S denote the submatrix of X that drops the columns corresponding to the complement of $\mathcal{S}$, let $P_\mathcal{S} = X_\mathcal{S} X_\mathcal{S}^+$ denote the projection matrix associated with $X_\mathcal{S}$, and let $P = XX^+$ denote the full projection matrix. In a slight departure from Section 3, we gauge the quality of the approximation conditional on X, so that the approximation error from only using the regressors $X_\mathcal{S}$ is given by $(I - P_\mathcal{S})X\alpha$, the residual from projecting $X\alpha$ onto $X_\mathcal{S}$. The assumption that $X\alpha$ is sparse can then be stated as:

Assumption 2. There exists a subset $\mathcal{S}^* \subseteq \{1, \dots, p\}$ with cardinality s, such that $\|(I - P_{\mathcal{S}^*})X\alpha\|_2^2 = O_p(s)$ and $s \log(p)/\sqrt{p} \to 0$.

If we knew the identity of the subset $\mathcal{S}^*$, and we also assumed that the sparsity was exact, so that the $O_p(s)$ term in Assumption 2 was zero, then a natural way of testing Assumption 2 would be to compare the restricted and unrestricted sum of squared residuals,

$$\mathcal{F} = \mathsf{Y}'(I - P_{\mathcal{S}^*})\mathsf{Y} - \mathsf{Y}'(I - P)\mathsf{Y}.$$

The statistic $\mathcal{F}$ is the numerator of the homoskedastic F-statistic. While the F-statistic critical values are only valid under homoskedasticity, we can leverage the fact that when $p \to \infty$, as is the case under our asymptotics, $\mathcal{F}$ is asymptotically normal after centering and scaling to derive a critical value that is robust to heteroskedasticity. Furthermore, if we have an estimator $\hat{\alpha}$ such that $\mathsf{Y} - X\hat{\alpha}$ approximates the residuals $(I - P_{\mathcal{S}^*})\mathsf{Y}$ from the infeasible short regression sufficiently well, we can replace the infeasible sum of squared residuals $\mathsf{Y}'(I - P_{\mathcal{S}^*})\mathsf{Y}$ in $\mathcal{F}$ by $\|\mathsf{Y} - X\hat{\alpha}\|_2^2$ to derive a feasible version of this test. Finally, it turns out that weakening exact sparsity to approximate sparsity in the sense of Assumption 2 does not impact the null rejection probability of the test in large samples. The next lemma formalizes these arguments.

Lemma 3. *Suppose Assumption 2 holds and that, for some $K \geq 1$, (a) $E[\epsilon_i^4 \mid X] \leq K$; (b) $\max_i P_{ii} < 1 - 1/K$; (c) $\mathrm{tr}(P)/n \geq 1/K$; and (d) $E[\epsilon_i^2 \mid X] \geq 1/K$ a.s. Then, as $n \to \infty$,*

$$\frac{\mathcal{F} - \sum_i \epsilon_i^2 P_{ii}}{\sqrt{2\sum_{i \neq j} \epsilon_i^2 \epsilon_j^2 P_{ij}^2}} \xrightarrow{d} \mathcal{N}(0,1). \tag{9}$$

Furthermore, suppose that for some estimator $\hat{\alpha}$, (i) $\|X\hat{\alpha} - X\alpha\|_2 \preceq_p \sqrt{s\log(p)}$; (ii) $\|\hat{\alpha} - (X_{\mathcal{S}^*}'X_{\mathcal{S}^*})^{-1}X_{\mathcal{S}^*}'X\alpha\|_1 \preceq_p s\sqrt{\log(p)/n}$; and, in addition, (iii) $\max_{ij}|X_{ij}|s\sqrt{\log(p)/n} = o_p(1)$; and (iv) $\|\sum_i \epsilon_i P_{ii} X_i\|_\infty + \|\sum_i \epsilon_i X_i\|_\infty \preceq_p \sqrt{n\log(p)}$. Then, letting $\hat{\epsilon}_i = \mathsf{Y}_i - X_i'\hat{\alpha}$,

$$\frac{\|\hat{\epsilon}\|_2^2 - \mathsf{Y}'(I-P)\mathsf{Y} - \sum_i \hat{\epsilon}_i^2 P_{ii}}{\sqrt{2\sum_{i \neq j} \hat{\epsilon}_i^2 \hat{\epsilon}_j^2 P_{ij}^2}} \xrightarrow{d} \mathcal{N}(0,1).$$

Conditions (a) and (d) are standard assumptions on the regression errors, while conditions (b) and (c) impose weak restrictions on the design matrix; condition (c) is analogous to the condition imposed in Lemma 2, and (b) bounds leverage away from one. Conditions (i) and (ii) in the second part of the lemma are standard mean-squared error and ℓ_1 rate conditions that hold for the lasso and post-lasso estimators (Belloni & Chernozhukov, 2013; Bickel et al., 2009). Conditions (iii) and (iv) are tail restrictions on the covariates and residuals analogous to those in Belloni et al. (2014).

Lemma 3 implies that we can test the sparsity assumption by calculating the lasso or post-lasso residuals $\hat{\epsilon}_i$, and then checking whether the residual sum of squares of the lasso is comparable to that of the OLS residual sum of squares. If the difference satifies

$$\|\hat{\epsilon}\|_2^2 - \mathsf{Y}'(I-P)\mathsf{Y} \geq z_{1-\alpha}\sqrt{2\sum_{i \neq j} \hat{\epsilon}_i^2 \hat{\epsilon}_j^2 P_{ij}^2} + \sum_i \hat{\epsilon}_i^2 P_{ii}$$

where $z_{1-\alpha}$ is the $1-\alpha$ quantile of the standard normal distribution (1.645 for $\alpha = 5\%$ level test), we reject Assumption 2.

In general, testing hypotheses about α when $\lim_{n\to\infty} p/n > 0$ can be quite involved, since plugging in regression residuals into asymptotic variance expressions leads to bias (see, e.g., Anatolyev & Sølvsten, 2023; Cattaneo et al., 2018b; Kline et al., 2020). As shown in the proof of Lemma 3 (see eq. (29)), we avoid such difficulties here by virtue of the fact that under the assumptions of the lemma, the lasso residuals $\hat{\epsilon}_i$ are sufficiently accurate so that the variance part of the test statistic, the denominator in eq. (9), can be consistently estimated under the null hypothesis.

5.3 Empirical tests of sparsity

We now apply the tests developed in Sections 5.1 and 5.2 to the empirical illustrations considered in Section 2. For each specification, we consider the Hausman test, which compares OLS with SBE, as well two versions of the residual test. The first version estimates the outcome regression in eq. (1) using the post-lasso, and compares the residuals to those based on OLS. The second version compares post-lasso and OLS residuals from estimating the propensity score regression in eq. (2).

Table 3 reports the results. Column 1 applies the test to the original specification in each paper. We see that for 6 of the 7 outcomes, at least one of the tests rejects the assumption of sparsity.

To check that these results are not driven by finite-sample size distortions of these tests, Appendix C conducts Monte Carlo simulations based on these applications; in these simulations, the size stays close to or below the nominal level.

One response to these findings is to seek alternative normalizations of the control matrix that are consistent with the sparsity assumptions. The remaining columns in Table 3 report the range of p-values under the four different normalizations we considered in Section 2. The table shows that in these applications, we were unable to find a normalization for five of the seven outcomes where at least one test did not reject. This is in spite of the large number of alternative specifications that these normalizations generate.

6 Summary and conclusions

We have argued, using empirical evidence and theoretical arguments, that SBEs display a lack of robustness to the specification of the control matrix. In the three applications we have examined, the range of variation in the SBE estimates under equally plausible

Table 3: p-values (in percentages) for tests of the sparsity assumption under different normalizations of the control matrix.

		Repl.	Collinearity		Powers		Category sums		Offset	
Outcome	Test	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
					A: BCH					
Δ ln(violent	H	81.7	76.0	87.4	81.7	85.8			73.1	100.0
crime/capita)	OR	9.5	9.5	9.5	9.5	9.5			8.2	20.2
	PR	0.3	0.0	0.9	0.0	0.6			0.0	2.2
Δ ln(property	H	82.6	61.8	93.4	70.9	89.7			47.5	100.0
crime/capita)	OR	12.0	12.0	12.0	12.0	12.0			10.4	45.4
	PR	28.8	12.6	35.0	0.0	29.9			0.0	39.9
Δ ln(murder/	H	21.0	19.7	22.5	20.4	22.6			17.3	26.8
capita)	OR	43.3	43.3	43.3	43.3	43.3			43.3	45.4
	PR	0.4	0.2	1.1	0.4	1.2			0.0	26.0
					B: Ferrara (2022)					
% Semiskilled	H	0.0	0.0	5.5	0.0	0.0	0.0	0.0	0.0	0.0
Black workers	OR	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	PR	34.5	19.6	49.7	34.5	53.3	11.0	96.6	34.7	73.3
					C: Enke (2020)					
Trump −	H	6.1	4.9	9.5			1.8	15.2		
avg. GOP	OR	13.0	5.1	13.0			5.0	30.3		
	PR	0.2	0.1	0.2			0.0	4.0		
Voted for Trump	H	94.6	63.1	96.4			19.0	100.0		
in 2016	OR	0.0	0.0	0.5			0.0	14.1		
	PR	0.1	0.0	0.1			0.0	1.4		
Voted for Trump	H	15.5	5.9	16.6			2.6	30.5		
in primaries	OR	9.6	2.8	9.6			0.7	40.6		
	PR	0.0	0.0	0.0			0.0	0.6		

Notes: See notes to Table 1 for description of the specifications. H: Hausman, OR: residual test based on outcome regression, PR: residual test based on propensity score regression. Col. 1 reports p-values under the original specification for the BCH, Ferrara (2022), and Enke (2020) studies. Cols. 2–3 report the range of p-values obtained under alternative ways of dropping collinear columns of the control matrix. Cols. 4–5, 6–7, and 8–9 report, respectively, the p-value range under alternative normalizations of the controls prior to taking powers and interactions, when categorical variables are expressed as indicators for different subsets, and when powers are constructed using Hermite polynomials, with an offset λ ranging between -1 and 1.

alternative specifications is of the same order of magnitude as sampling uncertainty. Two reasons underlie this fragility. First, whether a small number of covariates can account for the bulk of the confounding depends on the particular specification of the control matrix: even if the sparsity assumption holds under a particular way of expressing the column space of the controls, most alternative plausible normalizations do not admit a sparse approximation. Second, it may be the case no control matrix in a large class of normalizations admits a sparse approximation.

What should a practitioner take away from these results? We have argued that unless p is comparable to n, the potential efficiency gains of SBEs over OLS are limited. Simply reporting OLS with standard errors that are robust to the presence of many controls (e.g. Cattaneo et al., 2018a; D'Adamo, 2019; Dobriban et al., 2024; Jochmans, 2022) will deliver credible inference at little efficiency loss. When p is comparable to n, OLS becomes too noisy to be useful, and infeasible when the covariate dimension exceeds the sample size. Sparsity restrictions on the control vector, or other types of restrictions such as limiting the magnitude of the control coefficients (e.g. Armstrong et al., 2023; Li & Müller, 2021), are then necessary for informative inference. However, since these are substantive modeling restrictions, they need to be discussed and defended on substantive grounds, analogous to the discussion of other key assumptions, such as selection on observables. In particular, researchers who opt to leverage SBEs need to explain why sparsity should plausibly hold under the chosen specification of the control matrix, and not leave normalization choices to statistical software. The sparsity tests developed in this paper can serve as a complement to these arguments, provided they serve as a model specification check rather than a pretest.

We have focused on the sparsity assumption and SBEs because these estimators are used frequently, and their theory is well-developed. However, many other modern machine learning methods likewise lack invariance to linear reparametrization of the control matrix. When these methods are used for prediction, this lack of invariance is less important, for two reasons. First, one is typically interested in average performance over many predictions, and the overall prediction performance may be robust even if individual predictions are sensitive to normalizations. Second, one can gauge the performance of a given procedure directly using a test sample. When we incorporate these methods into econometric models, however, we are typically interested in inference on a single causal effect, and test sample benchmarking is unavailable. Understanding more generally when a lack of invariance leads to fragility is an interesting area for future research.

Appendix A Auxiliary results

For the next two results, we consider a quadratic form $\psi' H \psi$, where H is an orthogonal projection matrix with rank bounded by r. Conditional on some σ-algebra $\mathcal{Z}_n$, the elements of ψ are independent and mean zero, and H is non-random. We will prove a law of large numbers and a central limit theorem for $\psi' H \psi$.

Lemma 4. *Suppose that uniformly over i, $E[|\psi_i|^{2+\eta} \mid \mathcal{Z}_n] \leq K$ for some $K > 0$ and some $\eta \in [0, 2]$. Then*

$$\psi' H \psi = E[\psi' H \psi \mid \mathcal{Z}_n] + O_p(r^{1/2} + r^{2/(2+\eta)}).$$

Proof. Write

$$\psi' H \psi - E[\psi' H \psi \mid \mathcal{Z}_n] = \sum_i (\psi_i^2 - E[\psi_i^2 \mid \mathcal{Z}_n]) H_{ii} + 2 \sum_{i<j} \psi_i \psi_j H_{ij} =: T_1 + T_2.$$

By iterated expectations, and the inequality of von Bahr and Esseen (1965),

$$E|T_1|^{1+\eta/2} \leq 2E \sum_i |\psi_i^2 - E[\psi_i^2 \mid \mathcal{Z}_n]|^{1+\eta/2} |H_{ii}|^{1+\eta/2} \leq 2Kr,$$

so by Markov's inequality, $T_1 = O_p(r^{2/(2+\eta)})$. The term T_2 is mean zero with variance

$$4 \sum_{i<j} E[\psi_i^2 \psi_j^2 H_{ij}^2] \preceq r,$$

so by Markov's inequality, $T_2 = O_p(r^{1/2})$. □

Lemma 5. *Let $\omega^2 = 2 \sum_{i \neq j} H_{ij}^2 E[\psi_i^2 \psi_j^2 \mid \mathcal{Z}_n]$. Suppose that for some constant $K \geq 1$, (a) $r/\omega^2 \leq K$, and (b) $E[\psi_i^4 \mid \mathcal{Z}_n] \leq K$ a.s. Then, as $r \to \infty$,*

$$\frac{\psi' H \psi - \sum_i H_{ii} \psi_i^2}{\omega} \xrightarrow{d} \mathcal{N}(0, 1).$$

Proof. Write $(\psi' H \psi - \sum_i H_{ii} \psi_i^2)/\omega = \sum_i \mathcal{Y}_i$, where $\mathcal{Y}_i = \frac{2}{\omega} \psi_i \sum_{j=1}^{i-1} \psi_j H_{ij}$ is a martingale difference array with respect to the filtration $\mathcal{F}_i = \sigma(\psi_1, \ldots, \psi_i, \mathcal{Z}_n)$. By the martingale central limit theorem, it suffices to verify the Lyapunov condition $\sum_i E[\mathcal{Y}_i^4] \to 0$, and convergence of the conditional variance, $\sum_i E[\mathcal{Y}_i^2 \mid \mathcal{F}_{i-1}] = 1 + o_p(1)$.

The Lyapunov condition follows from the bound

$$\sum_i E[\mathcal{Y}_i^4] \le \frac{2^4}{\omega^4} E\left[\psi_i^4 \cdot 3 \sum_{j=1}^{i-1} \psi_j^2 H_{ij}^2 \sum_{k=1}^{i-1} \psi_k^2 H_{ik}^2\right]$$
$$\le \sum_i E \frac{48K^3}{\omega^4} \sum_{j=1}^{i-1} \sum_{k=1}^{i-1} H_{ij}^2 H_{ik}^2 \le \frac{48K^3}{\omega^4} r \le \frac{48K^5}{r} \to 0,$$

where the last inequality uses condition (a).

To show convergence of the conditional variance, decompose

$$\sum_i E[\mathcal{Y}_i^2 \mid \mathcal{F}_{i-1}] - 1 = \frac{4}{\omega^2} \sum_{i=1}^{n} \sum_{j=1}^{i-1} E[\psi_i^2 \mid \mathcal{Z}_n](\psi_j^2 - E[\psi_j^2 \mid \mathcal{Z}_n]) H_{ij}^2$$
$$+ \frac{8}{\omega^2} \sum_{i=1}^{n} E[\psi_i^2 \mid \mathcal{Z}_n] \sum_{j=1}^{i-1} \psi_j H_{ij} \sum_{k=1}^{j-1} \psi_k H_{ik} =: T_1 + T_2.$$

We have

$$E[T_1^2 \mid \mathcal{Z}_n] \le \frac{16}{\omega^4} \sum_{j=1}^{n} K^3 \left(\sum_{i=j+1}^{n} H_{ij}^2\right)^2 \le \frac{16K^5}{r},$$

so that $T_1 = o_p(1)$ as $r \to \infty$ by Markov's inequality. It remains to show that $T_2 = o_p(1)$. Let $\|A\|_F$ denote the Frobenius norm of a matrix, and let $L_{ij} = \mathrm{I}\{i > j\} H_{ij}$. The second moment of T_2 can be bounded as

$$E[T_2^2 \mid \mathcal{Z}_n] = \frac{64}{\omega^4} \sum_{k>j} E[\psi_j^2 \psi_k^2 \mid \mathcal{Z}_n] \left(\sum_{i=1}^{n} E[\psi_i^2 \mid \mathcal{Z}_n] L_{ij} H_{ik}\right)^2$$
$$\le \frac{64K^4}{r^2} \sum_{j,k} \left(\sum_{i=1}^{n} E[\psi_i^2 \mid \mathcal{Z}_n] L_{ij} H_{ik}\right)^2$$
$$= \frac{64K^4}{r^2} \|H \operatorname{diag}(E[\psi_i^2 \mid \mathcal{Z}_n]) L\|_F^2 \preceq \frac{\|L\|_F^2}{r^2} \le \frac{\|H\|_F^2}{r^2} = \frac{1}{r},$$

where the first inequality uses conditions (a) and (b), and the second inequality applies the inequality $\|AL\|_F^2 \le \lambda_{\max}(A'A)\|L\|_F^2$ twice. The claim then follows by Markov's inequality. □

Lemma 6. *Let $\mathcal{B}_p^K$ be a $K \times p$ matrix with i.i.d. Bernoulli(q) entries, $0 < q \le 1/2$, and assume $0 < \lim_{p\to\infty} K/p \le 1$. For any $\varepsilon > 0$, there exist $C_1, C_2 > 0$ only depending on ε and*

q such that

$$P\left(\min_{\|v\|_2\le\lfloor C_1 p\rfloor,\|v\|_2=1}\|\mathcal{B}_p^K v\|_2^2 \le C_2 p\right) < (1-q+\varepsilon)^K$$

for all large enough p.

Proof. The proof follows from the same arguments as Proposition 3.6 of Tikhomirov (2020). There are only two instances in the proof where $p - K \neq 1$ matters: first, in the second to last displayed inequality on Tikhomirov's page 601, the second inequality now invokes Lemma 3.5 with $m = K$, so in his notation, the two $n-1$ terms are replaced by K (where his n corresponds to our p). Second, in the last displayed inequality on page 601, the $n-1$ term is again replaced by K. Since $\lim_{n\to\infty} K/p = c_1 > 0$, the second inequality now still holds for all $|\sum x_i|$ that are larger by a factor of $2/\sqrt{c_1}$, at least for all large enough n. This only affects the constant C in Tikhomirov's first displayed inequality on page 602, and the result follows as in Tikhomirov's proof. □

Lemma 7. *Suppose that Assumption 2 holds, and that condition (a), and conditions (i), (ii), and (iv) of Lemma 3 hold. Then*

$$\|\mathsf{Y} - X\hat{\alpha}\|_2^2 - \mathsf{Y}'(I - P_{\mathcal{S}^*})\mathsf{Y} = O_p(s\log(p)).$$

Proof of Lemma 7. Letting $r = (I - P_{\mathcal{S}^*})X\alpha$, and $\tilde{\alpha} = \hat{\alpha} - (X'_{\mathcal{S}^*}X_{\mathcal{S}^*})^{-1}X'_{\mathcal{S}^*}X\alpha$, we may write,

$$\begin{aligned}\|\mathsf{Y} - X\hat{\alpha}\|_2^2 - \mathsf{Y}'(I - P_{\mathcal{S}^*})\mathsf{Y} &= \|X\alpha - X\hat{\alpha}\|_2^2 + 2\epsilon'(X\alpha - X\hat{\alpha}) + \epsilon'\epsilon - \|(I - P_{\mathcal{S}^*})\epsilon + r\|_2^2 \\ &= \|X\alpha - X\hat{\alpha}\|_2^2 - 2\epsilon' X\tilde{\alpha} + \epsilon' P_{\mathcal{S}^*}\epsilon - \|r\|_2^2.\end{aligned}$$

Hence, by eq. (26), Assumption 2, and condition (i) of Lemma 3,

$$\left|\|\mathsf{Y} - X\hat{\alpha}\|_2^2 - \mathsf{Y}'(I - P_{\mathcal{S}^*})\mathsf{Y}\right| \preceq_p s\log(p) + \|\epsilon' X\|_\infty\|\tilde{\alpha}\|_1 \preceq_p s\log(p),$$

where the second inequality uses conditions (ii) and (iv) of Assumption 2. □

Appendix B Proofs

Proof of Theorem 1. Let A_p be the probability of the event that the model with regressor $\mathcal{R}W_i$ satisfies eqs. (3) and (4), and let

$$r_{s,\mathcal{R}} = \min_{\|v\|_0\le s} E[(W_i'\mathcal{R}'\mathcal{R}\gamma - W_i'\mathcal{R}'v)^2 \mid \mathcal{R}] = \min_{\|v\|_0\le s}(\mathcal{R}\gamma - v)'\mathcal{R}E[W_iW_i']\mathcal{R}'(\mathcal{R}\gamma - v)$$

denote the mean square approximation error under the best s-sparse approximation in this model. Using the assumption that $E[W_i W_i']$ has eigenvalues bounded from below, eq. (5), and the assumption that $\|\gamma\|_2$ is bounded from below, it follows that for some $K > 0$ large enough,

$$r_{s,\mathcal{R}} \geq \frac{1}{K} \min_{\|v\|_0 \leq s} \|\mathcal{R}\gamma - v\|_2^2 \geq \frac{1}{K^2} \frac{\sum_{j=1}^{p-s} \mathcal{Z}_{j:p}^2}{\sum_{j=1}^{p} \mathcal{Z}_j^2}$$

where $\mathcal{Z}_{j:p}^2$ are the order statistics $\mathcal{Z}_{1:p}^2 \leq \mathcal{Z}_{2:p}^2 \leq \cdots \leq \mathcal{Z}_{p:p}^2$ of the sample $\{\mathcal{Z}_j^2\}_{j=1}^p$, and $\mathcal{Z}_j$ are i.i.d. standard normal.

If eqs. (3) and (4) hold, then for p large enough, $r_{s,\mathcal{R}} \leq \eta := K^{-2} s/p$ for $s = \lfloor \sqrt{p}/\log(p) \rfloor$. Hence, by the union bound, for p large enough,

$$\begin{aligned}
A_p &\leq P\left(r_{s,\mathcal{R}} \leq \eta\right) \\
&\leq \binom{p}{s} P\left(\frac{1}{K^2} \frac{\sum_{j=1}^{p-s} \mathcal{Z}_j^2}{\sum_{j=1}^{p} \mathcal{Z}_j^2} \leq \eta \right) \\
&= \binom{p}{s} P\left(\frac{\sum_{j=1}^{p-s} \mathcal{Z}_j^2/(p-s)}{\sum_{j=p-s}^{p} \mathcal{Z}_j^2/s} \leq \frac{s}{p-s} \frac{K^2\eta}{1-K^2\eta} \right) \\
&= \binom{p}{s} I_{K^2\eta}(p/2 - s/2, s/2),
\end{aligned}$$

where the last line uses the fact that an F-distribution with d_1 and d_2 degrees of freedom has c.d.f. $I_{d_1 x/(d_1 x + d_2)}(d_1/2, d_2/2)$. Here $I_x(a;b) = \int_0^x t^{a-1}(1-t)^{b-1} dt / B(a,b)$ is the regularized incomplete beta function, and $B(a,b)$ is the beta function. Since $\int_0^x t^{a-1}(1-t)^{b-1} dt \leq \int_0^x t^{a-1} dt = x^a/a$,

$$\begin{aligned}
A_p &\leq \binom{p}{s} \frac{1}{B(p/2 - s/2, s/2)} \frac{(K^2\eta)^{p/2-s/2}}{p/2 - s/2} \\
&\leq e^{3s/2} (p/s)^{3s/2-1} (K^2\eta)^{p/2-s/2} = e^{3s/2} (p/s)^{2s-1-p/2} \leq e^{3s/2} (\sqrt{p})^{2s-1-p/2}, \quad (10)
\end{aligned}$$

where the second line uses the identity $\binom{p}{s} = \frac{p}{s(p-s)} \frac{1}{B(s,p-s)}$ and a double application of the beta function bound $\frac{p}{s(p-s)} \frac{1}{B(s,p-s)} \leq (ep/s)^s$ for all $s \geq 1$, and the last inequality uses the definition of s and holds for p large enough so that $2s - 1 - p/2$ is negative. The beta function bound follows from the generalization of Stirling's inequality to gamma functions due to Gordon (1994, Theorem 5),

$$\Gamma(t) = \sqrt{2\pi} t^{t-1/2} e^{-t} R_t, \qquad R_t \in [1, \sqrt{2}] \quad (11)$$

for all $t \geq 1/2$.[15] Taking logs of eq. (10) then yields the upper bound

$$\log(A_p) \leq \tfrac{3}{2}s - (p/4 - s + 1/2)\log(p) \asymp -\frac{p}{4}\log(p).$$

To derive a lower bound for A_p, observe that $r_{s,\mathcal{R}} \leq K^2 \sum_{j=1}^{p-s} \mathcal{Z}_{j:p}^2 / \sum_{j=1}^{p} \mathcal{Z}_j^2$ for a large enough K. Since the probability of an approximately sparse representation is lower bounded by the probability that $r_{s,\mathcal{R}} \leq K^2 s/p$ for $s = \lfloor \sqrt{p}/\log(p)^2 \rfloor$, we have

$$\begin{aligned} A_p &\geq P\left(\frac{\sum_{j=1}^{p-s} \mathcal{Z}_j^2}{\sum_{j=1}^{p} \mathcal{Z}_j^2} \leq s/p \right) = I_{s/p}(p/2 - s/2, s/2) \\ &\geq \frac{(1-s/p)^{s/2-1}(s/p)^{p/2-s/2}}{(p/2-s/2)B(p/2-s/2, s/2)} \\ &\geq \frac{1}{2\sqrt{2\pi}} \frac{(s/p)^{p/2-s+1}}{(s/2)^{1/2}(1-s/p)^{p/2-s+3/2}} \geq \frac{1}{2\sqrt{\pi}} \frac{(s/p)^{p/2-s+1}}{\sqrt{s}}, \end{aligned}$$

where the second line uses the bound $\int_0^x t^{a-1}(1-t)^{b-1}dt \geq (1-x)^{b-1}\int_0^x t^{a-1}dt = (1-x)^{b-1}x^a/a$, and the third line uses the inequality $\frac{p}{s(p-s)B(s,p-s)} \geq \frac{1}{2\sqrt{2\pi}} \frac{(p/s)^s}{s^{1/2}(1-s/p)^{p-s+1/2}}$ that follows from eq. (11). Hence, $\log(A_p) \succeq -\frac{p}{4}\log(p)$, as claimed. □

Proof of Theorem 2. Let $s = \lfloor \sqrt{p}/\log(p) \rfloor$. If eqs. (3) and (4) hold, then for p large enough, $\min_{\|v\|_0 \leq s} E[(Z_i' A_0' \gamma - Z_i' \mathcal{A}' v)^2 \mid \mathcal{A}] \leq s/p$. It hence suffices to show that given $\varepsilon > 0$, for all large enough p,

$$P\left(\min_{\|v\|_0 \leq s} E[(Z_i' A_0' \gamma - Z_i' \mathcal{A}' v)^2 \mid \mathcal{A}] \leq \frac{s}{p} \right) < (1 - q + \varepsilon)^K.$$

Let $\psi = A_0'\gamma$, a $p \times 1$ vector with K elements equal to zero and $p - K$ elements equal to $\|\gamma\|_2$ (which is equal to the single non-zero element of γ). By assumption, $pE[Z_i Z_i']$ is a diagonal matrix with all diagonal elements bounded below by some constant $C_0 > 0$. Thus, it suffices to show that

$$P\left(\min_{\|v\|_0 \leq s} \|\psi - \mathcal{A}' v\|_2^2 \leq s/C_0 \right) < (1 - q + \varepsilon)^K \tag{12}$$

[15] Specifically,

$$\begin{aligned} \frac{p}{s(p-s)} \frac{1}{B(s,p-s)} = \frac{p\Gamma(p)}{s\Gamma(s)(p-s)\Gamma(p-s)} &= \frac{R_p}{\sqrt{2\pi} R_s R_{p-s}} \frac{(p/s)^s}{s^{1/2}} \left(1 + \frac{s}{p-s}\right)^{p-s+1/2} \\ &\leq \frac{1}{\sqrt{\pi}} \sqrt{\frac{p}{s(p-s)}} (ep/s)^s \leq (ep/s)^s, \end{aligned}$$

where the second equality uses eq. (11), the first inequality uses $1 + x \leq e^x$ and the bounds on R_t in eq. (11), and the last inequality uses $p/(s(p-s)) \leq 2$.

for all p large enough.

Let $\mathcal{B}$ be a $p \times p$ matrix with i.i.d. Bernoulli(q) entries, and without loss of generality, assume $\mathcal{A} = \mathcal{B}$ if $\mathcal{B}$ is non-singular. By Theorem A of Tikhomirov (2020), given any $\varepsilon_0 > 0$, the probability of $\mathcal{B}$ being singular is smaller than $(1-q+\varepsilon_0)^p$ for all large enough p. Applying this result with $\varepsilon_0 = \varepsilon/2$, and using the law of total probability, we can bound the left-hand side by

$$P\left(\min_{\|v\|_0 \le s} \|\psi - \mathcal{A}'v\|_2^2 \le s/C_0\right) \le P\left(\mathcal{B} \text{ is singular}\right) + P\left(\min_{\|v\|_0 \le s} \|\psi - \mathcal{B}'v\|_2^2 \le s/C_0\right)$$
$$\le (1 - q + \varepsilon/2)^p + P\left(\min_{\|v\|_0 \le s} \|\psi - \mathcal{B}'v\|_2^2 \le s/C_0\right).$$

Since $2(1-q+\varepsilon/2)^p \le (1-q+\varepsilon)^p$ for $p \ge \log(2)/\log(\frac{1-q+\varepsilon}{1-q+\varepsilon/2})$, to show eq. (12), suffices to show that the second term in the above display is bounded by $(1-q+\varepsilon/2)^p$.

With $\eta = \|\gamma\|_2/(2\sqrt{s})$, by the union bound,

$$P\left(\min_{\|v\|_0 \le s} \|\psi - \mathcal{B}v\|_2^2 \le s/C_0\right) \le P\left(\min_{\|v\|_0 \le s, \|v\|_2 < \eta} \|\psi - \mathcal{B}v\|_2^2 \le s/C_0\right) +$$
$$P\left(\min_{\|v\|_0 \le s, \|v\|_2 \ge \eta} \|\psi - \mathcal{B}v\|_2^2 \le s/C_0\right). \quad (13)$$

Note that all elements of $\mathcal{B}$ are either zero or one. By the Cauchy-Schwarz inequality, if $\|v\|_0 \le s$ and $\|v\|_2 \le \eta$, all elements of $\mathcal{B}v$ are bounded above by $\frac{1}{2}\|\gamma\|_2$, and hence $\|\psi - \mathcal{B}v\|_2^2 \ge \frac{1}{4}\|\gamma\|_2^2(p-K)$ almost surely. The first probability on the right-hand side of eq. (13) is thus equal to zero for all large enough p.

Let $\mathcal{B}_p^K$ be the $K \times p$ matrix that collects the K rows of $\mathcal{B}$ where the corresponding element in ψ is zero. Then for any $C_1 > 0$,

$$\min_{\|v\|_0 \le s, \|v\|_2 > \eta} \|\psi - \mathcal{B}v\|_2^2 \ge \min_{\|v\|_0 \le s, \|v\|_2 \ge \eta} \|\mathcal{B}_p^K v\|_2^2 \ge \eta^2 \min_{\|v\|_0 \le \lfloor C_1 p \rfloor, \|v\|_2 = 1} \|\mathcal{B}_p^K v\|_2^2$$

almost surely, where the last inequality holds for all p large enough so that $\lfloor C_1 p \rfloor \ge s$. Thus, for all $C_2 > 0$ and large enough p,

$$P\left(\min_{\|v\|_0 \le s, \|v\|_2 \ge \eta} \|\psi - \mathcal{B}v\|_2^2 \le s/C_0\right) \le P\left(\min_{\|v\|_2 \le \lfloor C_1 p \rfloor, \|v\|_2 = 1} \|\mathcal{B}_p^K v\|_2^2 \le s/\eta^2 \cdot 1/C_0\right)$$
$$\le P\left(\min_{\|v\|_2 \le \lfloor C_1 p \rfloor, \|v\|_2 = 1} \|\mathcal{B}_p^K v\|_2^2 \le C_2 p\right)$$

since $s/\eta^2 \asymp p/(\ln p)^2$. The result now follows from choosing C_1 and C_2 from Lemma 6, with

$\varepsilon/2$ playing the role of ε. □

Proof of Theorem 3. Let $p_0 = p - 1$. Using the identity

$$\sqrt{j!}H_j(x+y) = \sum_{k=0}^{j} \binom{j}{k} x^{j-k}\sqrt{k!}H_k(y),$$

we find

$$H_{p_0}(z_i + \lambda) = \sum_{j=0}^{p_0} \binom{p_0}{j} \sqrt{j!/p_0!}\lambda^{p_0-j}\tilde{W}_{i,j+1} = \sum_{k=0}^{p_0} \tilde{\gamma}_{p-k}\tilde{W}_{i,p-k},$$

where $\tilde{\gamma}_{p-k} = \binom{p_0}{p_0-k}\sqrt{(p_0-k)!/p_0!}\lambda^k$.

Applying the Stirling formula (Robbins, 1955)

$$\log(n!) = \frac{1}{2}\log(2\pi) + (n+1/2)\log(n) - n + R_n, \qquad R_n \in [1/(12n+1), 1/12n]$$

we obtain, for $k \geq 1$,

$$\begin{aligned} \log(\tilde{\gamma}_{p-k}^2) &= \log\left(\frac{p_0!\lambda^{2k}}{k!^2(p_0-k)!}\right) \\ &= (p_0 - k + 1/2)\log\left(\frac{p_0}{p_0-k}\right) + k - \log(k) + 2k\log\left(\frac{\sqrt{p_0}\lambda}{k}\right) + R'_{p_0,k}, \end{aligned} \tag{14}$$

where $R'_{p_0,k} = R_{p_0} - 2R_k - R_{p_0-k} - \log(2\pi)$ is bounded above and below by a constant, since $R_n \in [0, 1/12]$. Now, $\log(p_0/(p_0-k)) = \log(1+k/(p_0-k)) \geq k/p_0$ since $\log(1+x) \geq x/(1+x)$ for $x \geq -1$. Hence, for some constant C, and for $1 \leq k \leq L\sqrt{p_0}/\log(p_0)$

$$\begin{aligned} \log(\tilde{\gamma}_{p-k}^2) &\geq k\left[2 + \frac{1/2-k}{p_0} - \log(k)/k\right] + 2k\log\left(\frac{\sqrt{p_0}L}{k\log(p_0)}\right) + C \\ &\geq k\left[2 + \frac{1/2 - L\sqrt{p_0}/\log(p_0)}{p_0} - \log(k)/k\right] + 2k\log(1) + C \\ &\geq k\left[1 - \frac{L}{\log(p_0)\sqrt{p_0}}\right] + C, \end{aligned}$$

where the last inequality uses $\log(k)/k \leq 1$. The expression in brackets is bounded below by $1/2$ for $p \geq \max\{6, (2L)^2\}$, which yields the result for L fixed.

To prove the result when $L \to 0$, eq. (14) implies, for some constant C,

$$\begin{aligned}\log(\tilde{\gamma}_{p-k}^2) &\leq (p_0 - k + 1/2)\log\left(1 + \frac{k}{p_0 - k}\right) + k - \log(k) + 2k\log\left(\frac{\sqrt{p_0}\lambda}{k}\right) + C \\ &\leq 2k + \frac{1}{2}\frac{k}{p_0 - k} + 2k\log\left(\frac{\sqrt{p_0}\lambda}{k}\right) + C \leq k\left[3 + 2\log\left(\frac{\sqrt{p_0}\lambda}{k}\right)\right] + C,\end{aligned}$$

where the second line uses $\log(1 + x) \leq x$. Since for $k \geq \sqrt{p_0}\lambda e^2$ the expression in square brackets is smaller than -1, if we approximate the regression function using the last $s = \lceil e^2 \max\{L/\log(p_0), 1/\log(p_0)^2\}\sqrt{p_0}\rceil$ regressors, we make an approximation error that is bounded by

$$\sum_{k \geq s} \tilde{\gamma}_{p-k}^2 \leq C \sum_{k \geq s} e^{-k} \preceq e^{-s} \preceq e^{-e^2\sqrt{p}/\log(p)^2},$$

which is of lower order than s/p. □

Proof of Lemma 1. Write

$$\hat{\beta}_{OLS} - \beta = \sum_{i=1}^n \eta_i, \qquad \eta_i = \frac{1}{\ddot{D}'\ddot{D}}\ddot{D}_i U_i.$$

Conditional on (D, W), η_i are independent, with variance $\omega_{OLS}^2 := \text{var}(\sum_i \eta_i \mid D, W) = \frac{\sum_i \ddot{D}_i^2 E[U_i^2|D,W]}{(\ddot{D}'\ddot{D})^2}$. By a conditional version of the Lindeberg-Feller theorem (Bulinski, 2017, Theorem 1), it therefore suffices to verify a conditional version of the Lyapunov condition

$$\sum_{i=1}^n \frac{E\left[|\eta_i|^{2+\eta} \mid D, W\right]}{\omega_{OLS}^{2+\eta}} = o_p(1),$$

and that

$$s_{OLS}^2/\omega_{OLS}^2 = 1 + o_p(1). \tag{15}$$

To show the Lyapunov condition, substitute in the definition of η_i and ω_{OLS} to write the left-hand side as

$$\begin{aligned}&\sum_{i=1}^n \frac{E[|U_i|^{2+\eta} \mid D, W]|\ddot{D}_i|^{2+\eta}}{(\sum_{j=1}^n \ddot{D}_j^2 E[U_j^2 \mid D, W])^{1+\eta/2}} \\ &\quad \leq K^{2+\eta/2}\frac{\sum_{i=1}^n |\ddot{D}_i|^{2+\eta}}{(\sum_{j=1}^n \ddot{D}_j^2)^{1+\eta/2}} \leq K^{2+\eta/2}\frac{\max_{i'}|\ddot{D}_{i'}|^\eta \sum_{i=1}^n \ddot{D}_i^2}{(\sum_{j=1}^n \ddot{D}_j^2)^{1+\eta/2}} = K^{2+\eta/2}\left(\frac{\max_i \ddot{D}_i^2/n}{\frac{1}{n}\sum_{j=1}^n \ddot{D}_j^2}\right)^{\eta/2},\end{aligned}$$

where the first inequality uses Assumption 1 (ii) and (iii). Therefore, to verify the Lyapunov

condition, it suffices to show that

$$\max_i |\ddot{D}_i| = O_p(n^{1/4}), \tag{16}$$

and that

$$\frac{1}{n}\sum_{i=1}^n \ddot{D}_i^2 \asymp_p 1. \tag{17}$$

Now, by Lemma 4 and Assumption 1 (ii) and (iii), $\frac{1}{n}\sum_{i=1}^n \ddot{D}_i^2 = \frac{1}{n}\sum_i M_{ii} E[\tilde{D}_i^2 \mid W] + o_p(1)$, $M_{ii} = (I-P)_{ii}$, with the first term bounded between $(1-p/n)/K$ and K. Therefore, eq. (17) follows by Assumption 1 (iv). Furthermore, by the union bound and Markov's inequality

$$P(\max_i |\ddot{D}_i|/n^{1/4} \geq \epsilon) \leq \sum_{i=1}^n P(|\ddot{D}_i| \geq n^{1/4}\epsilon) \leq \sum_{i=1}^n \frac{E[\ddot{D}_i^4]}{n\epsilon^4} \leq \frac{1}{\epsilon^4} 4K,$$

which implies eq. (16). Here the last inequality follows from the bound

$$E[\ddot{D}_i^4 \mid W] = \sum_j M_{ij}^4 E[\tilde{D}_j^4 \mid W] + 3\sum_{j\neq k} M_{ij}^2 M_{ik}^2 E[\tilde{D}_j^2 \tilde{D}_k^2 \mid W] \leq 4K \sum_{j,k} M_{ij}^2 M_{ik}^2 = 4K M_{ii}^2. \tag{18}$$

It remains to verify eq. (15). Write

$$\frac{s_{OLS}^2}{\omega_{OLS}^2} - 1 = \frac{\frac{1}{n}\sum_{i=1}^n \ddot{D}_i^2 (U_i^2 - E[U_i^2 \mid D, W])}{\frac{1}{n}\sum_{i=1}^n \ddot{D}_i^2 E[U_i^2 \mid D, W]}$$

The denominator satisfies $\frac{1}{n}\sum_{i=1}^n \ddot{D}_i^2 E[U_i^2 \mid D, W] \geq \frac{1}{n}\sum_{i=1}^n \ddot{D}_i^2/K \asymp_p 1$ by eq. (17). The result therefore follows if we can show that the numerator is of the order $o_p(1)$. This follows from the fact that for any variable $\mathcal{U}_i$, that, conditional on (D, W) has mean zero and uniformly bounded $1 + \eta/2$ moments,

$$\frac{1}{n}\sum_i \ddot{D}_i^2 \mathcal{U}_i = o_p(1). \tag{19}$$

In turn, eq. (19) follows by Markov's inequality, and the bound

$$E\left|\frac{1}{n}\sum_{i=1}^n \ddot{D}_i^2 \mathcal{U}_i\right|^{1+\eta/2} \preceq \frac{1}{n^{1+\eta/2}}\sum_{i=1}^n E|\ddot{D}_i|^{2+\eta} \leq \frac{1}{n^{1+\eta/2}}\sum_{i=1}^n E[\ddot{D}_i^4]^{(2+\eta)/4}$$
$$\leq \frac{1}{n^{1+\eta/2}} E\sum_{i=1}^n \left(4K M_{ii}^2\right)^{(2+\eta)/4} \to 0$$

Here the first inequality uses iterated expectations, and the inequality of von Bahr and Esseen (1965), the second inequality uses Jensen's inequality, the third uses eq. (18), and the final limit uses $M_{ii}^2 \leq 1$. □

Proof of Lemma 2. Observe first that

$$n\sum_{i=1}^{n} Z_i^2 = \frac{\tilde{D}'P\tilde{D}/n}{\tilde{D}'\tilde{D}/n \cdot \ddot{D}'\ddot{D}/n} = \frac{\frac{1}{n}\sum_i P_{ii}E[\tilde{D}_i^2 \mid W] + O_p(p^{1/2}/n)}{\tilde{D}'\tilde{D}/n \cdot \ddot{D}'\ddot{D}/n} \asymp_p \frac{\operatorname{tr}(W)}{n}, \tag{20}$$

where the second equality follows by Lemma 4, and the last step follows by the law of large numbers, eq. (17), and Assumption 1 (ii) and (iii). Hence, letting $\omega_H^2 = \sum_{i=1}^n Z_i^2 E[U_i^2 \mid D, W]$, it follows from Assumption 1 (iii), and the fact that $s_*^2 = O_p(1/n)$ that

$$\frac{s_*^2}{\omega_H^2} = \frac{O_p(1)}{n\sum_{i=1}^n Z_i^2 E[U_i^2 \mid D, W]} = O_p(n/\operatorname{tr}(P)).$$

Therefore, under eq. (7), since $\liminf_{n\to\infty} n/\operatorname{tr}(P)$ is bounded,

$$\frac{\hat{\beta}_{OLS} - \hat{\beta}^*}{\omega_H} = \frac{1}{\omega_H}\sum_{i=1}^{n} Z_i U_i + o_p(s_*/\omega_H) = \frac{1}{\omega_H}\sum_{i=1}^{n} Z_i U_i + o_p(1).$$

By a conditional version of the Lindeberg-Feller theorem (Bulinski, 2017, Theorem 1), the first term is asymptotically standard normal if a conditional version of the Lyapunov condition holds,

$$\sum_{i=1}^{n} \frac{|Z_i|^{2+\eta} E[|U_i|^{2+\eta} \mid D, W]}{\omega_H^{2+\eta}} = o_p(1). \tag{21}$$

By Assumption 1 (ii) and (iii) and eq. (20), the left-hand side is bounded above by

$$K^{2+\eta/2}\left(\frac{n\max_i Z_i^2}{n\sum_i Z_i^2}\right)^{\eta/2} \preceq_p (n^2/\operatorname{tr}(P) \cdot \max_i Z_i^2)^{\eta/2}.$$

The claim in eq. (21) then follows from the bound

$$n\max_i Z_i^2 \leq 2\frac{\max_i \ddot{D}_i^2/n}{(\ddot{D}'\ddot{D}/n)^2} + 2\frac{\max_i \tilde{D}_i^2/n}{(\tilde{D}'\tilde{D}/n)^2}.$$

The first term is $O_p(n^{-1/2})$ by eqs. (16) and (17). The second term is also $O_p(n^{-1/2})$, since $\tilde{D}_i'\tilde{D}_i/n \asymp_p 1$ by the law or large numbers and Assumption 1 (ii) and (iii), and since

$$\max_i |\tilde{D}_i| = O_p(n^{1/4}). \tag{22}$$

Specifically, eq. (22) holds since by the union bound and Markov's inequality,

$$P(\max_i |\tilde{D}|_i/n^{1/4} > \epsilon) \leq \sum_{i=1}^n P(|\tilde{D}|_i/n^{1/4} > \epsilon) \leq \sum_{i=1}^n \frac{E[\tilde{D}_i^4]}{n\epsilon^4} \leq \frac{K}{\epsilon^4}.$$

To establish the first claim of Lemma 2, it now suffices to show that $\omega_H/s_H = 1 + o_p(1)$. This follows from writing

$$\frac{s_H^2}{\omega_H^2} - 1 = \frac{n\sum_i Z_i^2(U_i^2 - E[U_i^2 \mid D, W])}{n\sum_i Z_i^2 E[U_i^2 \mid D, W]}.$$

The denominator is of the order p/n by eq. (20) and Assumption 1 (ii) and (iii). To show that the numerator is $o_p(1)$, let $\mathcal{U}_i = U_i^2 - E[U_i^2 \mid D, W]$, and decompose it as

$$\begin{aligned} n\sum_i Z_i^2 \mathcal{U}_i &= \frac{\frac{1}{n}\sum_i \ddot{D}_i^2 \mathcal{U}_i}{(\ddot{D}'\ddot{D}/n)^2} + \frac{\frac{1}{n}\sum_i \tilde{D}_i^2 \mathcal{U}_i}{(\tilde{D}'\tilde{D}/n)^2} - 2\frac{\frac{1}{n}\sum_i \ddot{D}_i\tilde{D}_i\mathcal{U}_i}{(\ddot{D}'\ddot{D}/n)(\tilde{D}'\tilde{D}/n)} \\ &= O_p(1)\frac{1}{n}\sum_i \ddot{D}_i^2\mathcal{U}_i + O_p(1)\frac{1}{n}\sum_i \tilde{D}_i^2\mathcal{U}_i + O_p(1)\frac{1}{n}\sum_i \ddot{D}_i\tilde{D}_i\mathcal{U}_i = o_p(1). \end{aligned}$$

Here the second equality uses eq. (17), and the last inequality uses eq. (19), the law of large numbers, and the result that

$$\frac{1}{n}\sum_i \ddot{D}_i\tilde{D}_i\mathcal{U}_i = o_p(1), \tag{23}$$

which follows by Markov's inequality and the bound

$$\begin{aligned} E\left|\frac{1}{n}\sum_i \ddot{D}_i\tilde{D}_i\mathcal{U}_i\right|^{1+\eta/2} &\preceq \frac{1}{n^{1+\eta/2}}\sum_i E|\ddot{D}_i\tilde{D}_i|^{1+\eta/2} \leq \frac{1}{n^{1+\eta/2}}\sum_i (E\ddot{D}_i^2\tilde{D}_i^2)^{(2+\eta)/4} \\ &= \frac{1}{n^{1+\eta/2}}\sum_i (E\sum_j M_{ij}^2\tilde{D}_j^2\tilde{D}_i^2)^{(2+\eta)/4} \leq \frac{1}{n^{1+\eta/2}}\sum_i (M_{ii}K)^{(2+\eta)/4} \to 0. \end{aligned}$$

Here the first inequality uses iterated expectations, and the inequality of von Bahr and Esseen (1965), the second inequality uses Jensen's inequality, the third uses Assumption 1 (ii), and the final limit uses $M_{ii} \leq 1$.

To prove the second claim in Lemma 2, let $r_{\gamma,i} = r_\gamma(W_i)$, $r_{\delta,i} = r_\delta(W_i)$, $\tilde{u}_i = r_{\gamma,i} + U_i$ and $\tilde{d}_i = r_{\delta,i} + \tilde{D}_i$ so that we may write

$$\left|\frac{\hat{s}_H^2}{\omega_H^2} - 1\right| \preceq_p \left|n\sum_i (\hat{Z}_i^2\hat{U}_i^2 - \tilde{z}_i^2\tilde{u}_i^2)\right| + \left|n\sum_i (Z_i^2U_i^2 - \tilde{z}_i^2\tilde{u}_i^2)\right| + o_p(1). \tag{24}$$

The second term is of the order $o_p(1)$ by condition (iii), so it suffices to show that the first term is of the order $o_p(1)$.

Let $\check{U}_i = \hat{U}_i - \tilde{u}_i$ and $\check{D}_i = \hat{D}_i - \tilde{d}_i$, so we may write $n\hat{Z}_i\hat{U}_i = \frac{\ddot{D}_i\check{U}_i}{\ddot{D}'\ddot{D}/n} + \frac{\ddot{D}_i\tilde{u}_i}{\ddot{D}'\ddot{D}/n} - \frac{\check{D}_i\check{U}_i + \check{D}_i\tilde{u}_i + \tilde{d}_i\check{U}_i}{\hat{D}'\hat{D}/n} - \frac{\tilde{d}_i\tilde{u}_i}{\hat{D}'\hat{D}/n}$ and $n\tilde{z}_i\tilde{u}_i = \frac{\tilde{u}_i\ddot{D}_i}{\ddot{D}'\ddot{D}/n} - \frac{\tilde{u}_i\tilde{d}_i}{\tilde{d}'\tilde{d}/n}$. Plugging these expressions into the first term in eq. (24) and expanding the term yields

$$
\begin{aligned}
n\sum_i(\hat{Z}_i^2\hat{U}_i^2 - \tilde{z}_i^2\tilde{u}_i^2) &= \frac{\frac{1}{n}\sum_i[(\check{D}_i\check{U}_i + \check{D}_i\tilde{u}_i + \tilde{d}_i\check{U}_i)^2 + 2(\check{D}_i\check{U}_i + \check{D}_i\tilde{u}_i + \tilde{d}_i\check{U}_i)\tilde{d}_i\tilde{u}_i]}{(\hat{D}'\hat{D}/n)^2} \\
&+ \frac{\frac{1}{n}\sum_i(\ddot{D}_i^2\check{U}_i^2 + 2\ddot{D}_i^2\check{U}_i\tilde{u}_i)}{(\ddot{D}'\ddot{D}/n)^2} - \frac{2}{n}\sum_i\frac{(\ddot{D}_i\check{U}_i + \ddot{D}_i\tilde{u}_i)(\check{D}_i\check{U}_i + \check{D}_i\tilde{u}_i + \tilde{d}_i\check{U}_i) + \ddot{D}_i\check{U}_i\tilde{d}_i\tilde{u}_i}{\ddot{D}'\ddot{D}/n\cdot\hat{D}'\hat{D}/n} \\
&+ \frac{1}{n}\sum_i\tilde{d}_i^2\tilde{u}_i^2\left[\frac{1}{(\hat{D}'\hat{D}/n)^2} - \frac{1}{(\tilde{d}'\tilde{d}/n)^2}\right] + 2\frac{\frac{1}{n}\sum_i\tilde{d}_i\ddot{D}_i\tilde{u}_i^2}{\ddot{D}'\ddot{D}/n}\left[\frac{1}{\tilde{d}'\tilde{d}/n} - \frac{1}{\hat{D}'\hat{D}/n}\right].
\end{aligned}
$$

Using the inequality $|ab| \leq (a^2+b^2)/2$, eq. (17), and the fact that by condition (iv), $\hat{D}'\hat{D}/n = \tilde{d}'\tilde{d}/n + o_p(1)$ and that $\tilde{d}'\tilde{d}/n \asymp_p 1$ by condition (i) and Assumption 1 (ii) and (iii), it follows that the right-hand side is smaller than

$$
\begin{aligned}
\left|n\sum_i(\hat{Z}_i^2\hat{U}_i^2 - \tilde{z}_i^2\tilde{u}_i^2)\right| &\leq O_p(1)\frac{1}{n}\sum_i[\check{D}_i^2\check{U}_i^2 + \check{D}_i^2\tilde{u}_i^2 + \tilde{d}_i^2\check{U}_i^2 + |\check{D}_i\tilde{d}_i|\tilde{u}_i^2 + \tilde{d}_i^2|\check{U}_i\tilde{u}_i|] \\
&+ O_p(1)\frac{1}{n}\sum_i[\ddot{D}_i^2\check{U}_i^2 + (\ddot{D}_i^4 + \tilde{u}_i^2)|\check{U}_i|] \\
&+ O_p(1)\frac{1}{n}\sum_i[\check{U}_i^2\check{D}_i^2 + \ddot{D}_i^2\check{U}_i^2 + \check{D}_i^2\tilde{u}_i^2 + \tilde{d}_i^2\check{U}_i^2 + \tilde{u}_i^2|\ddot{D}_i\check{D}_i| + |\check{U}_i|(\ddot{D}_i^2 + \tilde{d}_i^2\tilde{u}_i^2)] \\
&+ o_p(1)\frac{1}{n}\sum_i\tilde{d}_i^2\tilde{u}_i^2 + o_p(1)\frac{1}{n}\sum_i(\ddot{D}_i^2 + \tilde{d}_i^2)\tilde{u}_i^2 \\
&\leq o_p(1)\frac{1}{n}\sum_i[1 + \tilde{u}_i^2 + \tilde{d}_i^2 + \tilde{d}_i^2\tilde{u}_i^2 + \ddot{D}_i^4 + \ddot{D}_i^2\tilde{u}_i^2],
\end{aligned}
$$

where the second inequality uses condition (iv) and the inequality $2|a| \leq 1 + a^2$. We now show that each summand is bounded in probability. By condition (i), and Assumption 1 (ii), $\frac{1}{n}\sum_i(\tilde{u}_i^2 + \tilde{d}_i^2) = \frac{1}{n}\sum_i(U_i^2 + D_i^2) + o_p(1) = O_p(1)$. Next,

$$
\frac{1}{n}\sum_i\tilde{d}_i^2\tilde{u}_i^2 \leq \frac{4}{n}\sum_i(\tilde{D}_i^2 + r_{\delta,i}^2)(r_{\gamma,i}^2 + U_i^2)
$$

which by Assumption 1 (ii) and conditions (i) and (ii) has expectation bounded by a constant times $\frac{1}{n}\sum_{i=1}^n E[1 + r_{\gamma,i}^2 + r_{\delta,i}^2 + r_{\delta,i}^2 r_{\gamma,i}^2] = O(1)$, so $\frac{1}{n}\sum_i\tilde{d}_i^2\tilde{u}_i^2 = O_p(1)$ by Markov's inequality.

Likewise, it follows from eq. (18) that $\frac{1}{n}\sum_i \ddot{D}_i^4 = O_p(1)$. Finally, by Assumption 1 (ii), and the bound $E[\ddot{D}_i^2 \mid W] \leq KM_{ii} \leq K$, we have $\frac{1}{n}\sum_i E[\ddot{D}_i^2 \tilde{u}_i^2] \leq 2\frac{1}{n}\sum_i E[\ddot{D}_i^2(K + r_\gamma^2)] \leq 2E[K(K + r_\gamma^2)] = O(1)$, so that by Markov's inequality, the last term is also bounded in probability. □

Proof of Lemma 3. Letting $r = (I - P_{\mathcal{S}^*})X\alpha$, we may write

$$\mathcal{F} = \epsilon' P\epsilon - \epsilon' P_{\mathcal{S}^*}\epsilon + 2\epsilon' r + r'r = \epsilon' P\epsilon + O_p(s + \|r\|_2 + \|r\|_2^2), \tag{25}$$

where the second equality follows since the second term satisfies

$$\epsilon' P_{\mathcal{S}^*}\epsilon = E[\epsilon' P_{\mathcal{S}^*}\epsilon] + O_p(s^{1/2}) \leq Ks + O_p(s^{1/2}) \tag{26}$$

by Lemma 4 and condition (a), and the third term is mean zero with variance bounded by $K\|r\|^2$ by condition (a). Furthermore, by conditions (b) and (d),

$$\omega^2 := 2\sum_{i\neq j} E[\epsilon_i^2\epsilon_j^2 \mid X]P_{ij}^2 \geq 2\sum_{i\neq j} P_{ij}^2/K^2 \geq \frac{2}{K^2}\operatorname{tr}(P)(1 - \max_i P_{ii}) \geq 2\operatorname{tr}(P)/K^3.$$

Hence, by condition (c), condition (a) of Lemma 5 holds. Since condition (b) of Lemma 5 holds by condition (a), we can apply Lemma 5 to $\epsilon' P\epsilon$ to yield $(\epsilon' P\epsilon - \sum_{i=1}^n \epsilon_i^2 P_{ii})/\omega \xrightarrow{d} \mathcal{N}(0,1)$. Combining this result with eq. (25) and Assumption 2 yields

$$\frac{\mathcal{F} - \sum_{i=1}^n \epsilon_i^2 P_{ii}}{\omega} = \frac{\epsilon' P\epsilon - \sum_{i=1}^n \epsilon_i^2 P_{ii}}{\omega} + o_p(1) \xrightarrow{d} \mathcal{N}(0,1). \tag{27}$$

Furthermore, letting $\psi_i = \epsilon_i^2 - E[\epsilon_i^2 \mid X]$, and using the identity $\epsilon_i^2\epsilon_j^2 - E[\epsilon_i^2\epsilon_j^2 \mid X] = \psi_i\psi_j + \psi_i E[\epsilon_j^2 \mid X] + E[\epsilon_i^2 \mid X]\psi_j$, we have

$$\begin{aligned}\left|\frac{2\sum_{i\neq j}\epsilon_i^2\epsilon_j^2 P_{ij}^2 - \omega^2}{\omega^2}\right| &\leq \frac{K^3}{\operatorname{tr}(P)}\left|\sum_{i\neq j}(\epsilon_i^2\epsilon_j^2 - E[\epsilon_i^2\epsilon_j^2 \mid X])P_{ij}^2\right| \\ &\leq \frac{2K^3}{\operatorname{tr}(P)}\left|\sum_{i<j}\psi_i\psi_j P_{ij}^2\right| + \frac{2K^3}{\operatorname{tr}(P)}\left|\sum_{i\neq j}\psi_i E[\epsilon_j^2 \mid X]P_{ij}^2\right|.\end{aligned}$$

By condition (a), conditional on X, the term inside the first absolute value function is mean zero with variance bounded by a constant times $\sum_{i,j} P_{ij}^4 \leq \operatorname{tr}(P)$, and the term inside the second absolute value function is also mean zero with variance bounded by a constant times $\sum_{i=1}^n(\sum_{j=1}^n P_{ij}^2)^2 \leq \operatorname{tr}(P)$, so that the above display is $O_p(\operatorname{tr}(P)^{-1/2}) = o_p(1)$ by Markov's inequality. Combining this result with eq. (27) then yields the first claim.

To prove the second claim in Lemma 3, it suffices to show that

$$\|\mathsf{Y} - X\hat{\alpha}\|_2^2 - \mathsf{Y}'(I - P_{\mathcal{S}^*})\mathsf{Y} = O_p(s\log(p)),$$

that

$$\sum_i P_{ii}(\hat{\epsilon}_i^2 - \epsilon_i^2) = o_p(p^{1/2}) \tag{28}$$

and that

$$\sum_{i\neq j}(\hat{\epsilon}_i^2\hat{\epsilon}_j^2 - \epsilon_i^2\epsilon_j^2)P_{ij}^2 = o_p(p). \tag{29}$$

The first assertion follows by Lemma 7. To show eq. (28), decompose $\hat{\epsilon}_i = \epsilon_i - \tilde{f}_i = \epsilon_i + r_i - X_i'\tilde{\alpha}$, where $\tilde{f}_i = X_i'(\hat{\alpha} - \alpha)$, and $\tilde{\alpha} = \hat{\alpha} - (X_{\mathcal{S}^*}'X_{\mathcal{S}^*})^{-1}X_{\mathcal{S}^*}'X\alpha$, so that we may write

$$\begin{aligned}\left|\sum_i P_{ii}(\hat{\epsilon}_i^2 - \epsilon_i^2)\right| &= \left|\sum_i P_{ii}\tilde{f}_i^2 + 2\sum_i P_{ii}\epsilon_i r_i - 2\sum_i P_{ii}\epsilon_i X_i'\tilde{\alpha}\right| \\ &\preceq_p s\log(p) + \|r\|_2 + 2\left\|\sum_i P_{ii}\epsilon_i X_i\right\|_\infty \|\tilde{\alpha}\|_1 \preceq_p s\log(p) + \|r\|_2,\end{aligned}$$

where the first inequality follows by condition (i) and Markov's inequality, and since conditional on X, the second term is mean zero with standard deviation bounded by a constant times $\|r\|_2$. The second inequality applies conditions (ii) and (iv). Equation (28) then follows by Assumption 2.

To show eq. (29), write the left-hand side as

$$\begin{aligned}&\sum_{i\neq j}(\hat{\epsilon}_i^2\hat{\epsilon}_j{}^2 - \epsilon_i^2\epsilon_j^2)P_{ij}^2 \\ &\quad= \sum_{i\neq j}\tilde{f}_i^2\tilde{f}_j^2P_{ij}^2 + 2\sum_{i\neq j}\epsilon_i^2\tilde{f}_j^2P_{ij}^2 - 4\sum_{i\neq j}\epsilon_j\tilde{f}_j\tilde{f}_i^2P_{ij}^2 - 4\sum_{i\neq j}\epsilon_i\tilde{f}_i\epsilon_j^2P_{ij}^2 + 4\sum_{i\neq j}\epsilon_i\tilde{f}_i\epsilon_j\tilde{f}_jP_{ij}^2 \\ &\qquad=: T_1 + 2T_2 - 4T_3 - 4T_4 + 4T_5 = O_p(s^2\log(p)^2 + \|r\|_2^2) + o_p(p),\end{aligned}$$

where the bounds follow by bounding each term, as derived next. First, by condition (i), T_1 is bounded by a constant times $\|X(\hat{\alpha} - \alpha)\|_2^4 \preceq_p s^2\log(p)^2$. Second,

$$T_2 \leq 2\sum_{i\neq j}\epsilon_i^2r_j^2P_{ij}^2 + 2\sum_{i\neq j}\epsilon_i^2(X_j'\tilde{\alpha})^2P_{ij}^2 = O_p(\|r\|_2^2) + o_p(p).$$

where the second equality follows by Markov's inequality, since the first term, conditional on X, has expectation bounded by a constant times $\sum_j r_j^2P_{jj} \leq \|r\|_2^2$, and the second term

is bounded by $\max_j (X_j'\tilde{\alpha})^2 \cdot \sum_{i \neq j} \epsilon_i^2 P_{ij}^2 = o_p(1) \cdot O_p(p)$, since $\sum_{i \neq j} E[\epsilon_i^2 P_{ij}^2]$ is bounded by a constant times $\sum_{i \neq j} P_{ij}^2 \leq p$, and by conditions (ii) and (iii),

$$\max_j |X_j'\tilde{\alpha}| \leq \max_j \|X_j\|_\infty \|\tilde{\alpha}\|_1 = o_p(1). \tag{30}$$

Third,

$$|T_3| \leq 2 \sum_{i \neq j} \epsilon_j^2 \tilde{f}_i^2 P_{ij}^2 + 2 \sum_{i \neq j} \tilde{f}_j^2 \tilde{f}_i^2 P_{ij}^2 = 2T_1 + 2T_2.$$

Fourth,

$$|T_4| \leq \left| \sum_{i \neq j} \epsilon_i X_i'\tilde{\alpha} \epsilon_j^2 P_{ij}^2 \right| + 2 \left| \sum_{i \neq j} \epsilon_i r_i (\epsilon_j^2 - E[\epsilon_j^2 \mid X]) P_{ij}^2 \right| + \left| \sum_{i \neq j} \epsilon_i r_i E[\epsilon_j^2 \mid X] P_{ij}^2 \right|$$
$$= o_p(p) + O_p(\|r\|_2).$$

Here the conclusion follows since the first term is bounded by $\max_i |X_i'\tilde{\alpha}| \sum_{i \neq j} |\epsilon_i| \epsilon_j^2 P_{ij}^2 = o_p(p)$ by eq. (30) and the fact that $E \sum_{i \neq j} |\epsilon_i| \epsilon_j^2 P_{ij}^2 \preceq \sum_{i,j} P_{ij}^2 \leq p$, the second term is mean zero with variance bounded by a constant times $\sum_{i,j} r_i^2 P_{ij}^4 \leq \|r\|_2^2$, and the third term is mean zero with variance bounded by a constant times $\sum_i r_i^2 P_{ii}^2 \leq \|r\|_2^2$. Finally,

$$T_5 = 2 \sum_{i<j} \epsilon_i \epsilon_j r_j r_i P_{ij}^2 + 2 \sum_{i \neq j} \epsilon_i r_i \epsilon_j X_j'\tilde{\alpha} P_{ij}^2 + \sum_{i \neq j} \epsilon_i \epsilon_j (X_j'\tilde{\alpha})(X_i'\tilde{\alpha}) P_{ij}^2.$$

The first term is mean zero with variance bounded by a constant times $\sum_{i<j} r_j^2 r_i^2 P_{ij}^4 \leq \|r\|_2^4$. The second term is bounded by a constant times $\max_j |X_j'\tilde{\alpha}| \cdot \sum_{i \neq j} r_i^2 |\epsilon_j| P_{ij}^2 + \max_j |X_j'\tilde{\alpha}| \cdot \sum_{i \neq j} \epsilon_i^2 |\epsilon_j| P_{ij}^2 = o_p(\|r\|_2^2 + p)$, and the last term is bounded by $\max_j |X_j'\tilde{\alpha}|^2 \cdot \sum_{i \neq j} |\epsilon_i \epsilon_j| P_{ij}^2 = o_p(p)$. Hence, $T_5 = O_p(\|r\|^2) + o_p(p)$, concluding the proof. □

Appendix C Finite-sample size properties of sparsity tests

We now conduct a simple Monte Carlo simulation to assess the finite-sample performance of the Hausman and residual tests considered in Section 5.

The control matrix W in these simulations corresponds to one of seven designs, each corresponding to one of the seven specifications reported in the first two columns of Table 1. The treatment and outcome vectors D and Y are drawn from independent normal distributions with zero means and homoskedastic variances σ_Y^2 and σ_D^2. For each control matrix, we

consider two values for these variances: either the variance of the outcome and treatment, or else the variances of OLS residuals in each empirical specification. The first method yields standard deviation values (σ_Y, σ_D) given by $\{(0.082, 0.083), (0.052, 0.095), (0.264, 0.083)\}$ in the three BCH specifications, $(14.2, 2.1)$ in the Ferrara (2022) application, and $\{(37.5, 1.1), (49.71.0), (50.00.8)\}$ in the three Enke (2020) specifications. The second method gives standard deviation values $\{(0.045, 0.004), (0.030, 0.009), (0.197, 0.004)\}$ in the three BCH specifications, $(3.1, 0.7)$ in the Ferrara (2022) application, and $\{(27.2, 0.5), (29.8, 0.6), (31.4, 0.5)\}$ in the three Enke (2020) specifications.

In other words, each design corresponds exactly to the empirical specifications we studied in Section 5.3, except we replace the outcome and treatment with mean zero Gaussian variables. The parameter values β, γ and δ in the outcome and propensity score regressions in eqs. (1) and (2) therefore all equal 0, so that the sparsity assumption holds trivially. For each design, we apply the same three tests as in Table 3, with the same implementation. In particular, the tests allow for heteroskedasticity in the Enke (2020) specification, and for clustering in the other two specifications. To benchmark the performance of the tests, we also compare the test to an oracle which replaces the SBE estimates with a constrained OLS estimate that sets all penalized coefficients to zero.[16]

We report the empirical null rejection rates of the sparsity assumption for each test in Table C.1. The rejection rates do not exceed the nominal 5% level by more than one percentage point, and do not deviate from the oracle rejection rates by more than two percentage points in any specification. Both the feasible and the oracle version of the residual test is conservative in several specifications: this appears to be a consequence of the fact that the mean of the infeasible test statistic $\mathcal{F}$ is given by $E[\sum_i \epsilon_i^2 P_{ii} - \sum_i \epsilon_i^2 P_{\mathcal{S}^*,ii}]$, but the recentering in eq. (9) only accounts for the first term, since the second term is asymptotically negligible. This may render the test conservative in finite samples when the true sparsity level is not zero, or when unpenalized covariates (that enter estimate of the sparsity index $\mathcal{S}^*$) are present.

References

Anatolyev, S., & Sølvsten, M. (2023). Testing many restrictions under heteroskedasticity. *Journal of Econometrics*, *236*(1), Article 105473. https://doi.org/10.1016/j.jeconom.2023.03.011

[16]The unpenalized controls comprise 12 year fixed effects in BCH, county fixed effects Ferrara (2022), and the intercept in Enke (2020).

Table C.1: Empirical null rejection rates (in percentages) of sparsity tests with nominal level 5% in simulations.

		Large variance		Small variance	
		SBE	Oracle	SBE	Oracle
Outcome	Test	(1)	(2)	(3)	(4)
		A: BCH			
Δ ln(violent	H	5.5	5.6	5.1	5.1
crime/capita)	PR	1.8	1.8	1.5	1.5
	OR	1.7	1.7	1.6	1.6
Δ ln(property	H	4.9	4.9	5.6	5.5
crime/capita)	PR	1.7	1.7	1.6	1.6
	OR	1.7	1.7	1.5	1.5
Δ ln(murder/	H	5.4	5.4	4.9	5.0
capita)	PR	1.6	1.6	1.6	1.6
	OR	1.8	1.8	1.7	1.7
		B: Ferrara (2022)			
% Semiskilled	H	4.9	4.8	5.1	5.1
Black workers	PR	3.9	4.2	3.9	4.2
	OR	4.1	4.4	4.1	4.4
		C: Enke (2020)			
Trump −	H	5.1	5.0	5.2	5.2
avg. GOP	PR	3.0	4.6	3.3	5.0
	OR	3.2	4.8	3.3	4.9
Voted for Trump	H	5.1	5.1	4.9	4.8
in 2016	PR	3.1	4.8	3.2	4.9
	OR	3.3	4.8	3.3	4.8
Voted for Trump	H	5.1	5.0	5.0	5.1
in primaries	PR	3.1	4.5	3.3	5.0
	OR	3.2	4.7	3.2	4.8

Notes: H: Hausman, OR: residual test based on outcome regression, PR: residual test based on propensity score regression. Oracle corresponds to a constrained OLS estimator that sets all penalized coefficients to zero. The control matrix W corresponds to the control matrix in Cols. 1–2 of Table 1. In Cols. 1–2, the outcome and treatment variance correspond to the variance of the outcome and treatment in the corresponding specification in Cols. 1–2 of Table 1. In Cols. 3–4, the outcome and treatment variances correspond to variances of OLS residuals in these specifications. 10,000 simulation draws.

Angrist, J. D., & Frandsen, B. (2022). Machine labor. *Journal of Labor Economics*, *40*(S1), S97–S140. https://doi.org/10.1086/717933

Armstrong, T., Kolesár, M., & Kwon, S. (2023). *Bias-aware inference in regularized regression models*. arXiv: 2012.14823.

Athey, S., Imbens, G. W., & Wager, S. (2018). Approximate residual balancing: Debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society*, *80*(4), 597–623. https://doi.org/10.1111/rssb.12268

von Bahr, B., & Esseen, C.-G. (1965). Inequalities for the rth absolute moment of a sum of random variables, $1 \leq r \leq 2$. *The Annals of Mathematical Statistics*, *36*(1), 299–303. https://doi.org/10.1214/aoms/1177700291

Belloni, A., & Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, *19*(2), 521–547. https://doi.org/10.3150/11-BEJ410

Belloni, A., Chernozhukov, V., & Hansen, C. B. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, *81*(2), 608–650. https://doi.org/10.1093/restud/rdt044

Belloni, A., Chernozhukov, V., Fernández-Val, I., & Hansen, C. B. (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica*, *85*(1), 233–298. https://doi.org/10.3982/ECTA12723

Beyhum, J., & Striaukas, J. (2024). Testing for sparse idiosyncratic components in factor-augmented regression models. *Journal of Econometrics*, *244*(1), Article 105845. https://doi.org/10.1016/j.jeconom.2024.105845

Bickel, P. J., Ritov, Y., & Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, *37*(4), 1705–1732. https://doi.org/10.1214/08-AOS620

Bondell, H. D., & Reich, B. J. (2009). Simultaneous factor selection and collapsing levels in ANOVA. *Biometrics*, *65*(1), 169–177. https://doi.org/10.1111/j.1541-0420.2008.01061.x

Bulinski, A. V. (2017). Conditional central limit theorem. *Theory of Probability & Its Applications*, *61*(4), 613–631. https://doi.org/10.1137/S0040585X97T98837X

Carpentier, A., & Verzelen, N. (2021). Optimal sparsity testing in linear regression model. *Bernoulli*, *27*(2), 727–750. https://doi.org/10.3150/20-BEJ1224

Cattaneo, M. D., Jansson, M., & Newey, W. K. (2018a). Alternative asymptotics and the partially linear model with many regressors. *Econometric Theory*, *34*(2), 277–301. https://doi.org/10.1017/S026646661600013X

Cattaneo, M. D., Jansson, M., & Newey, W. K. (2018b). Inference in linear regression models with many covariates and heteroscedasticity. *Journal of the American Statistical Association*, *113*(523), 1350–1361. https://doi.org/10.1080/01621459.2017.1328360

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. M. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*(1), C1–C68. https://doi.org/10.1111/ectj.12097

D'Adamo, R. (2019). *Cluster-robust standard errors for linear regression models with many controls*. arXiv: 1806.07314.

Dobriban, E., Su, W. J., Yang, Y., & Zhang, Z. (2024). *Robust inference under heteroskedasticity via the Hadamard estimator*. arXiv: 1807.00347.

Donohue, J. J., III, & Levitt, S. D. (2001). The impact of legalized abortion on crime. *The Quarterly Journal of Economics*, *116*(2), 379–420. https://doi.org/10.1162/00335530151144050

Enke, B. (2020). Moral values and voting. *Journal of Political Economy*, *128*(10), 3679–3729. https://doi.org/10.1086/708857

Farrell, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, *189*(1), 1–23. https://doi.org/10.1016/j.jeconom.2015.06.017

Ferrara, A. (2022). World War II and black economic progress. *Journal of Labor Economics*, *40*(4), 1053–1091. https://doi.org/10.1086/716921

van de Geer, S., Bühlmann, P., Ritov, Y., & Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, *42*(3), 1166–1202. https://doi.org/10.1214/14-AOS1221

Gertheiss, J., & Tutz, G. (2010). Sparse modeling of categorial explanatory variables. *The Annals of Applied Statistics*, *4*(4), 2150–2180. https://doi.org/10.1214/10-AOAS355

Giannone, D., Lenza, M., & Primiceri, G. E. (2021). Economic predictions with big data: The illusion of sparsity. *Econometrica*, *89*(5), 2409–2437. https://doi.org/10.3982/ECTA17842

Gordon, L. (1994). A stochastic approach to the gamma function. *The American Mathematical Monthly*, *101*(9), 858–865. https://doi.org/10.2307/2975134

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica*, *46*(6), 1251–1271. https://doi.org/10.2307/1913827

He, J. (2020). *A test for sparsity* [Unpublished manuscript, Sciences Po].

Javanmard, A., & Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, *15*(82), 2869–2909. http://jmlr.org/papers/v15/javanmard14a.html

Jochmans, K. (2022). Heteroscedasticity-robust inference in linear regression models with many covariates. *Journal of the American Statistical Association*, *117*(538), 887–896. https://doi.org/10.1080/01621459.2020.1831924

Kline, P., Saggio, R., & Sølvsten, M. (2020). Leave-out estimation of variance components. *Econometrica*, *88*(5), 1859–1898. https://doi.org/10.3982/ECTA16410

Li, C. (, & Müller, U. K. (2021). Linear regression with many controls of limited explanatory power. *Quantitative Economics*, *12*(2), 405–442. https://doi.org/10.3982/QE1577

Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *The Annals of Statistics*, *21*(1), 255–285. https://doi.org/10.1214/aos/1176349025

Robbins, H. (1955). A remark on Stirling's formula. *The American Mathematical Monthly*, *62*(1), 26–29. https://doi.org/10.2307/2308012

Robinson, P. M. (1988). Root-N-consistent semiparametric regression. *Econometrica*, *56*(4), 931–954. https://doi.org/10.2307/1912705

Stokell, B. G., Shah, R. D., & Tibshirani, R. J. (2021). Modelling high-dimensional categorical data using nonconvex fusion penalties. *Journal of the Royal Statistical Society*, *83*(3), 579–611. https://doi.org/10.1111/rssb.12432

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., & Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society*, *67*(1), 91–108. https://doi.org/10.1111/j.1467-9868.2005.00490.x

Tikhomirov, K. (2020). Singularity of random Bernoulli matrices. *Annals of Mathematics*, *191*(2), 593–634. https://doi.org/10.4007/annals.2020.191.2.6

Wüthrich, K., & Zhu, Y. (2023). Omitted variable bias of lasso-based inference methods: A finite sample analysis. *The Review of Economics and Statistics*, *105*(4), 982–997. https://doi.org/10.1162/rest_a_01128

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society*, *68*(1), 49–67. https://doi.org/10.1111/j.1467-9868.2005.00532.x

Zhang, C.-H., & Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society*, *76*(1), 217–242. https://doi.org/10.1111/rssb.12026