INFERENCE FOR HIGH-DIMENSIONAL SPARSE ECONOMETRIC MODELS

ABSTRACT. This article is about estimation and inference methods for high dimensional sparse (HDS) regression models in econometrics. High dimensional sparse models arise in situations where many regressors (or series terms) are available and the regression function is well-approximated by a parsimonious, yet unknown set of regressors. The latter condition makes it possible to estimate the entire regression function effectively by searching for approximately the right set of regressors. We discuss methods for identifying this set of regressors and estimating their coefficients based on ℓ_1-penalization and describe key theoretical results. In order to capture realistic practical situations, we expressly allow for imperfect selection of regressors and study the impact of this imperfect selection on estimation and inference results. We focus the main part of the article on the use of HDS models and methods in the instrumental variables model and the partially linear model. We present a set of novel inference results for these models and illustrate their use with applications to returns to schooling and growth regression.

Key Words: inference under imperfect model selection, structural effects, high-dimensional econometrics, instrumental regression, partially linear regression, returns-to-schooling, growth regression

1. INTRODUCTION

We consider linear, high dimensional sparse (HDS) regression models in econometrics. The HDS regression model allows for a large number of regressors, p, which is possibly much larger than the sample size, n, but imposes that the model is sparse. That is, we assume only $s \ll n$ of these regressors are important for capturing the main features of the regression function. This assumption makes it possible to estimate HDS models effectively by searching for approximately the right set of regressors. In this article, we review estimation methods for HDS models that make use of ℓ_1-penalization and then provide a set of novel inference results. We also provide empirical examples that illustrate the potential wide applicability of HDS models and methods in econometrics.

The motivation for considering HDS models comes in part from the wide availability of data sets with many regressors. For example, the American Housing Survey records prices as well as a multitude of features of houses sold; and scanner data-sets record prices and numerous characteristics of products sold at a store or on the internet. HDS models are also partly motivated by the use of series methods in econometrics. Series methods use many constructed or series regressors – regressors formed as transformation of elementary regressors – to approximate regression functions. In these applications, it is important to have parsimonious yet accurate approximation of the regression function. One way to achieve this is to use the data to select a small of number of informative terms from among a very large set of control variables or approximating functions. In this article, we formally discuss doing this selection and estimating the regression function.

We organize the article as follows. In the next section, we introduce the concepts of sparse and approximately sparse regression models in the canonical context of modeling a conditional mean function and motivate the use of HDS models via an empirical and analytical examples. In Section 3, we discuss some principal estimation methods and mention extensions of these methods to applications beyond conditional mean models. We discuss some key estimation results for HDS methods and mention various extensions of these results in Section 4. We then develop HDS models and methods in instrumental variables models with many instruments in Section 5 and a partially linear model with many series terms in Section 6, with the main emphasis given to inference. Finally, we present two empirical examples which motivate the use of these methods in Section 7.

Notation. We allow for the models to change with the sample size, i.e. we allow for array asymptotics. In particular we assume that $p = p_n$ grows to infinity as n grows, and $s = s_n$ can also grow with n, although we require that $s \log p = o(n)$. Thus, all parameters are implicitly indexed by the sample size n, but we omit the index to simplify notation. We also use the following empirical process notation, $\mathbb{E}_n[f] = \mathbb{E}_n[f(z_i)] = \sum_{i=1}^n f(z_i)/n$. The l_2-norm is denoted by $\|\cdot\|$, and the l_0-norm, $\|\cdot\|_0$, denotes the number of non-zero components of a vector. We use $\|\cdot\|_\infty$ to denote the maximal element of a vector. Given a vector $\delta \in \mathbb{R}^p$, and a set of indices $T \subset \{1, \ldots, p\}$, we denote by $\delta_T \in \mathbb{R}^p$ the vector in which $\delta_{Tj} = \delta_j$ if $j \in T$, $\delta_{Tj} = 0$ if $j \notin T$. We use the notation $(a)_+ = \max\{a, 0\}$, $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. We also use the notation $a \lesssim b$ to denote $a \leqslant cb$ for some constant $c > 0$ that does not depend on n; and $a \lesssim_P b$ to denote $a = O_P(b)$. For an event E, we say that E wp $\to 1$ when E occurs with probability approaching one as n grows.

2. Sparse and Approximately Sparse Regression Models

In this section we review the modeling foundations for HDS methods and provide motivating examples with emphasis on applications in econometrics. First, let us consider the following

parametric linear regression model:

$$y_i = x_i'\beta_0 + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2), \quad \beta_0 \in \mathbb{R}^p, \quad i = 1, \ldots, n$$

$$T = \text{support}(\beta_0) \text{ has } s \text{ elements where } s < n,$$

where $p > n$ is allowed, T is unknown, and regressors $X = [x_1, \ldots, x_n]'$ are fixed. We assume Gaussian errors to simplify the presentation of the main ideas throughout the article, but note that this assumption can be eliminated without substantially altering the results. It is clear that simply regressing y on all p available x variables is problematic when p is large relative to n which motivates consideration of models that impose some regularization on the estimation problem.

The key assumption that allows effective use of this large set of covariates is sparsity of the model of interest. Sparsity refers to the condition that only $s \ll n$ elements of β_0 are non-zero but allows the identities of these elements to be unknown. Sparsity can be motivated on economic grounds in situations where a researcher believes that the economic outcome could be well-predicted by a small (relative to the sample size) number of factors but is unsure about the identity of the relevant factors. Note that we allow $s = s_n$ to grow with n, as mentioned in the notation section, although $s \log p = o(n)$ will be required for consistency. This simple sparse model substantially generalizes the classical parametric linear model by letting the identities, T, of the relevant regressors be unknown. This generalization is useful in practice since it is problematic to assume that we know the identities of the relevant regressors in many examples.

The previous model is simple and allows us to convey the essential ideas of the sparsity-based approach. However, it is unrealistic in that it presumes *exact* sparsity or that, after accounting for s main regressors, the error in approximating the regression function is zero. We shall make no formal use of the previous model, but instead use a much more general, *approximately sparse* or nonparametric model. In this model, *all* of the regressors potentially have a *non-zero* contribution to the regression function, but no more than s unknown regressors are needed for approximating the regression function with a sufficient degree of accuracy.

We formally define the approximately sparse model as follows.

Condition ASM. *We have data* $\{(y_i, z_i), i = 1, \ldots, n\}$ *that for each* n *obey the regression model:*

$$y_i = f(z_i) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2), \quad i = 1, \ldots, n, \tag{2.1}$$

where y_i *is the outcome variable,* z_i *is a* k_z*-vector of elementary regressors,* $f(z_i)$ *is the regression function, and* ϵ_i *are i.i.d. disturbances. Let* $x_i = P(z_i)$*, where* $P(z_i)$ *is a vector of dimension* $p = p_n$*, that contains a dictionary of possibly technical transformations of* z_i*, including a constant. The values* $x_1, \ldots, x_n$ *are treated fixed, and normalized so that* $\mathbb{E}_n[x_{ij}^2] = 1$ *for* $j = 1, \ldots, p$*. The regression function* $f(z_i)$ *admits the approximately sparse form, namely there exists* β_0 *such that*

$$f(z_i) = x_i'\beta_0 + r_i, \quad \|\beta_0\|_0 \leqslant s, \quad c_s := \{\mathbb{E}_n[r_i^2]\}^{1/2} \leqslant K\sigma\sqrt{s/n}. \tag{2.2}$$

where $s = s_n = o(n/\log p)$ and K is a constant independent of n.

In the set-up we consider the fixed design case, which covers random sampling as a special case where $x_1, \ldots, x_n$ represent a realization of this sample on which we condition throughout. The vector $x_i = P(z_i)$ can include polynomial or spline transformations of the original regressors z_i see, e.g., Newey (1997) and Chen (2007) for various examples of series terms. The approximate sparsity can be motivated similarly to Newey (1997), who assumes that the first $s = s_n$ series terms can approximate the nonparametric regression function well. Condition ASM is more general in that it does not impose that the most important $s = s_n$ terms in the approximating dictionary are the first s terms; in fact, the identity of the most important terms is treated as unknown. We note that in the parametric case, we may naturally choose $x_i'\beta_0 = f(z_i)$ so that $r_i = 0$ for all $i = 1, \ldots, n$. In the nonparametric case, we may think of $x_i'\beta_0$ as any sparse parametric model that yields a good approximation to the true regression function $f(z_i)$ in equation (2.1) so that r_i is "small" relative to the conjectured size of the estimation error. Given (2.2), our target in estimation is the parametric function $x_i'\beta_0$, where we can call

$$T := \text{support}(\beta_0)$$

the "true" model. Here we emphasize that the ultimate target in estimation is, of course, $f(z_i)$. The function $x_i'\beta_0$ is simply a convenient intermediate target introduced so that we can approach the estimation problem as if it were parametric. Indeed, the two targets, $f(z_i)$ and $x_i'\beta_0$, are equal up to the approximation error r_i. Thus, the problem of estimating the parametric target $x_i'\beta_0$ is equivalent to the problem of estimating the nonparametric target $f(z_i)$ modulo approximation errors.

One way to explicitly construct a good approximating model β_0 for (2.2) is by taking β_0 as the solution to

$$\min_{\beta \in \mathbb{R}^p} \mathbb{E}_n[(f(z_i) - x_i'\beta)^2] + \sigma^2 \frac{\|\beta\|_0}{n}. \tag{2.3}$$

We can call (2.3) the oracle problem,[1] and so we can call $T = \text{support}(\beta_0)$ the oracle model. Note that we necessarily have that $s = \|\beta_0\| \leqslant n$. The oracle problem (2.3) balances the approximation error $\mathbb{E}_n[(f(z_i) - x_i'\beta)^2]$ over the design points with the variance term $\sigma^2\|\beta\|_0/n$, where the latter is determined by the number of non-zero coefficients in β. Letting $c_s^2 := \mathbb{E}_n[r_i^2] = \mathbb{E}_n[(f(z_i) - x_i'\beta_0)^2]$ denote the squared error from approximating values $f(z_i)$ by $x_i'\beta_0$, the quantity $c_s^2 + \sigma^2 s/n$ is the optimal value of (2.3). In common nonparametric problems, such as the one described below, the optimal solution in (2.3) would balance the approximation error with the variance term giving that $c_s \leqslant K\sigma\sqrt{s/n}$. Thus, we would have $\sqrt{c_s^2 + \sigma^2 s/n} \lesssim \sigma\sqrt{s/n}$, implying that the quantity $\sigma\sqrt{s/n}$ is the ideal goal for the rate of convergence. If we knew the oracle model T, we would achieve this rate by using the oracle estimator, the least squares estimator based on this model. Of course, we do not generally know T since we do

not observe the $f(z_i)$'s and thus cannot attempt to solve the oracle problem (2.3). Since T is unknown, we will not generally be able to achieve the exact oracle rates of convergence, but we can hope to come close to this rate.

Before considering estimation methods, a natural question is whether exact or approximate HDS models make sense in econometric applications. In order to answer this question, it is helpful to consider the following two examples in which we abstract from estimation completely and only ask whether it is possible to accurately describe some structural econometric function $f(z)$ using a low-dimensional approximation of the form $P(z)'\beta_0$.

Example 1: Sparse Models for Earning Regressions. In this example we consider a model for the conditional expectation of log-wage y_i given education z_i, measured in years of schooling. We can expand the conditional expectation of wage y_i given education z_i:

$$E[y_i|z_i] = \sum_{j=1}^{p} \beta_{0j} P_j(z_i), \tag{2.4}$$

using some dictionary of approximating functions $P(z_i) = (P_1(z_i), \ldots, P_p(z_i))'$, such as polynomial or spline transformations in z_i and/or indicator variables for levels of z_i. In fact, since we can consider an overcomplete dictionary, the representation of the function using $P_1(z_i), \ldots, P_p(z_i)$ may not be unique, but this is not important for our purposes.

A conventional sparse approximation employed in econometrics is, for example,

$$f(z_i) := E[y_i|z_i] = \tilde{\beta}_1 P_1(z_i) + \cdots + \tilde{\beta}_s P_s(z_i) + \tilde{r}_i, \tag{2.5}$$

where the P_j's are low-order polynomials or splines, with typically one or two (linear or linear and quadratic) terms. Of course, there is no guarantee that the approximation error $\tilde{r}_i$ in this case is small or that these particular polynomials form the best possible s-dimensional approximation. Indeed, we might expect the function $E[y_i|z_i]$ to change rapidly near the schooling levels associated with advanced degrees, such as MBAs or MDs. Low-degree polynomials may not be able to capture this behavior very well, resulting in large approximation errors $\tilde{r}_i$.

A sensible question is then, "Can we find a better approximation that uses the same number of parameters?" More formally, can we construct a much better approximation of the sparse form

$$f(z_i) := E[y_i|z_i] = \beta_{k_1} P_{k_1}(z_i) + \cdots + \beta_{k_s} P_{k_s}(z_i) + r_i, \tag{2.6}$$

for some regressor indices $k_1, \ldots, k_s$ selected from $\{1, \ldots, p\}$? Since we can always include (2.5) as a special case, we can in principle do no worse than the conventional approximation; and, in fact, we can construct (2.6) that is much better, if there are some important higher-order terms in (2.4) that are completely missed by the conventional approximation. Thus, the answer to the question depends strongly on the empirical context.

Consider for example the earnings of prime age white males in the 2000 U.S. Census see, e.g., Angrist, Chernozhukov, and Fernandez-Val (2006). Treating this data as the population data,

Sparse Approximation	L_2 error	L_∞ error
Conventional	0.12	0.29
Lasso	0.08	0.12
Post-Lasso	0.04	0.08

TABLE 1. Errors of Conventional and the Lasso-based Sparse Approximations of the Earning Function. The Lasso method minimizes the least squares criterion plus the ℓ_1-norm of the coefficients scaled by a penalty parameter λ. The nature of the penalty forces many coefficients to zero, producing a sparse fit. The Post-Lasso minimizes the least squares criterion over the non-zero components selected by the Lasso estimator. This example deals with a pure approximation problem, in which there is no noise.

we can compute $f(z_i) = E[y_i|z_i]$ without error. Figure 1 plots this function. We then construct two sparse approximations and also plot them in Figure 1. The first is the conventional approximation of the form (2.5) with $P_1, \ldots, P_s$ representing polynomials of degree zero to $s-1$ ($s = 5$ in this example). The second is an approximation of the form (2.6), with $P_{k_1}, \ldots, P_{k_s}$ consisting of a constant, a linear term, and three linear splines terms with knots located at 16, 17, and 19 years of schooling. We find the latter approximation automatically using the ℓ_1-penalization or Lasso methods discussed below,[2] although in this special case we could construct such an approximation just by eye-balling Figure 1 and noting that most of the function is described by a linear function with a few abrupt changes that can be captured by linear spline terms that induce large changes in slope near 17 and 19 years of schooling. Note that an exhaustive search for a low-dimensional approximation in principle requires looking at a very large set of models. Methods for HDS models, such as ℓ_1-penalized least squares (Lasso), which we employed in this example, are designed to avoid this search. □

Example 2: Series approximations and Condition ASM. It is clear from the statement of Condition ASM that this expansion incorporates both substantial generalizations and improvements over the conventional series approximation of regression functions in Newey (1997). In order to explain this consider the set $\{P_j(z), j \geqslant 1\}$ of orthonormal basis functions on $[0,1]^d$, e.g. orthopolynomials, with respect to the Lebesgue measure. Suppose z_i have a uniform distribution on $[0,1]^d$ for simplicity.[3] Assuming $\mathrm{E}[f^2(z_i)] < \infty$, we can represent f via a Fourier expansion, $f(z) = \sum_{j=1}^{\infty} \delta_j P_j(z)$, where $\{\delta_j, j \geqslant 1\}$ are Fourier coefficients that satisfy $\sum_{j=1}^{\infty} \delta_j^2 < \infty$.

Let us consider the case that f is a smooth function so that Fourier coefficients feature a polynomial decay $\delta_j \propto j^{-\nu}$, where ν is a measure of smoothness of f. Consider

[2]The set of functions considered consisted of 12 linear splines with various knots and monomials of degree zero to four. Note that there were only 12 different levels of schooling.

[3]The discussion in this example continues to apply when z_i has a density that is bounded from above and away from zero on $[0,1]^d$.

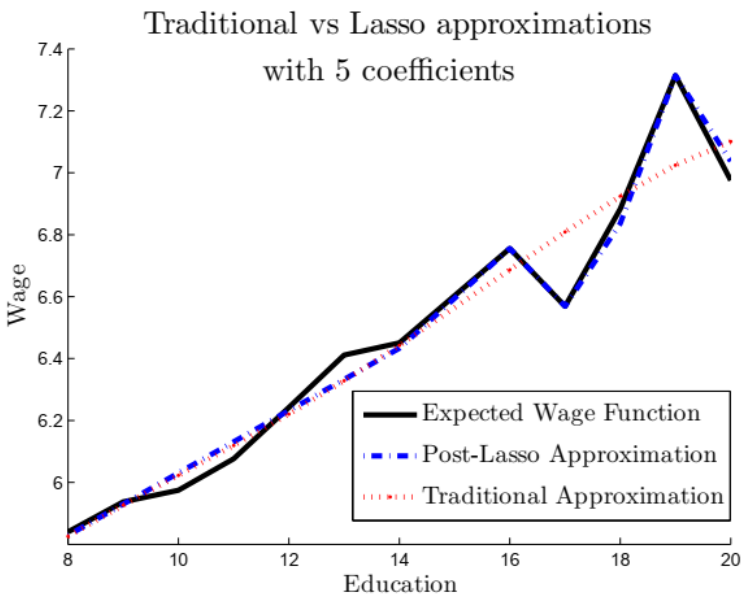

FIGURE 1. The figures illustrates the Post-Lasso sparse approximation and the fourth order polynomial approximation of the wage function.

the conventional series expansion that uses the first K terms for approximation, $f(z) = \sum_{j=1}^{K} \beta_{0j} P_j(z) + a_c(z)$, with $\beta_{0j} = \delta_j$. Here $a_c(z_i)$ is the approximation error which obeys $\sqrt{\mathbb{E}_n[a_c^2(z_i)]} \lesssim_P \sqrt{\mathrm{E}[a_c^2(z_i)]} \lesssim K^{\frac{-2\nu+1}{2}}$. Balancing the order $K^{\frac{-2\nu+1}{2}}$ of approximation error with the order $\sqrt{K/n}$ of the estimation error gives the oracle-rate-optimal number of series terms $s = K \propto n^{1/2\nu}$, and the resulting oracle series estimator, which knows s, will estimate f at the oracle rate of $n^{\frac{1-2\nu}{4\nu}}$. This also gives us the identity of the most important series terms $T = \{1, ..., s\}$, which are simply the first s terms. We conclude that Condition ASM holds for the sparse approximation $f(z) = \sum_{j=1}^{p} \beta_{0j} P_j(z) + a(z)$, with $\beta_{0j} = \delta_j$ for $j \leqslant s$ and $\beta_{0j} = 0$ for $s+1 \leqslant j \leqslant p$, and $a(z_i) = a_c(z_i)$, which coincides with the conventional series approximation above, so that $\sqrt{\mathbb{E}_n[a^2(z_i)]} \lesssim_P \sqrt{s/n}$ and $\|\beta_0\|_0 \leqslant s$.

Next suppose that Fourier coefficients feature the following pattern $\delta_j = 0$ for $j \leqslant M$ and $\delta_j \propto (j - M)^{-\nu}$ for $j > M$. Clearly in this case the standard series approximation based on the first $K \leqslant M$ terms, $\sum_{j=1}^{K} \delta_j f_j(z)$, has no predictive power for $f(z)$, and the corresponding standard series estimator based on the first K terms therefore fails completely.[4] In contrast, Condition ASM is easily satisfied in this case, and the Lasso-based estimators will perform at a near-oracle level in this case. Indeed, we can use the first p series terms to form the approximation $f(z) = \sum_{j=1}^{p} \beta_{0j} P_j(z) + a(z)$, where $\beta_{0j} = 0$ for $j \leqslant M$ and $j > M + s$, $\beta_{0j} = \delta_j$ for $M + 1 \leqslant j \leqslant M + s$ with $s \propto n^{1/2\nu}$, and p such that $M + n^{1/2\nu} = o(p)$. Hence $\|\beta_0\|_0 = s$, and we have that $\sqrt{\mathbb{E}_n[a^2(z_i)]} \lesssim_P \sqrt{\mathrm{E}[a^2(z_i)]} \lesssim \sqrt{s/n} \lesssim n^{\frac{1-2\nu}{4\nu}}$. □

[4] This is not merely a finite sample phenomenon but is also accommodated in the asymptotics since we expressly allow for array asymptotics; i.e. the underlying true model could change with n. Recall that we omit the indexing by n for ease of notation.

3. Sparse Estimation Methods

3.1. **ℓ_1-penalized and post ℓ_1-penalized estimation methods.** In order to discuss estimation consider first, as a matter of motivation, the classical AIC/BIC type estimator (Akaike 1974, Schwarz 1978) that solves the empirical (feasible) analog of the oracle problem:

$$\min_{\beta \in \mathbb{R}^p} \mathbb{E}_n[(y_i - x_i'\beta)^2] + \frac{\lambda}{n}\|\beta\|_0,$$

where λ is a penalty level.[5] This estimator has attractive theoretical properties. Unfortunately, it is computationally prohibitive since the solution to the problem may require solving $\sum_{k \leqslant n} \binom{p}{k}$ least squares problems.[6]

One way to overcome the computational difficulty is to consider a convex relaxation of the preceding problem, namely to employ a closest convex penalty – the ℓ_1 penalty – in place of the ℓ_0 penalty. This construction leads to the so called Lasso estimator $\widehat{\beta}$ (Tibshirani 1996), defined as a solution for the following optimization problem:

$$\min_{\beta \in \mathbb{R}^p} \mathbb{E}_n[(y_i - x_i'\beta)^2] + \frac{\lambda}{n}\|\beta\|_1, \tag{3.7}$$

where $\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$. The Lasso estimator is computationally attractive because it minimizes a convex function. A basic choice for penalty level suggested by Bickel, Ritov, and Tsybakov (2009) is

$$\lambda = 2 \cdot c\sigma\sqrt{2n \log(2p/\gamma)}. \tag{3.8}$$

where $c > 1$ and $1-\gamma$ is a confidence level that needs to be set close to 1. The formal motivation for this penalty is that it leads to near-oracle rates of convergence of the estimator.

The penalty level specified above is not feasible since it depends on the unknown σ. Belloni and Chernozhukov (2011c) propose to set

$$\lambda = 2 \cdot c\widehat{\sigma}\Phi^{-1}(1 - \gamma/2p), \tag{3.9}$$

with $\widehat{\sigma} = \sigma + o_P(1)$ obtained via an iteration method defined in Appendix A, where $c > 1$ and $1 - \gamma$ is a confidence level.[7] Belloni and Chernozhukov (2011c) also propose the X-dependent penalty level:

$$\lambda = c \cdot 2\widehat{\sigma}\Lambda(1 - \gamma|X), \tag{3.10}$$

where

$$\Lambda(1 - \gamma|X) = (1 - \gamma) - \text{quantile of } \; n\|\mathbb{E}_n[x_i g_i]\|_\infty \mid X$$

[5] The penalty level λ in the AIC/BIC type estimator needs to account for the noise since it observes y_i instead of $f(z_i)$ unlike the oracle problem (2.3).

[6] Results on the computational intractability of this problem were established in Natarajan (1995), Ge, Jiang, and Ye (2011) and Chen, Ge, Wang, and Ye (2011).

[7] Practical recommendations include the choice $c = 1.1$ and $\gamma = .05$.

where $X = [x_1, \ldots, x_n]'$ and g_i are i.i.d. $N(0,1)$, which can be easily approximated by simulation. We note that

$$\Lambda(1-\gamma|X) \leqslant \sqrt{n}\Phi^{-1}(1-\gamma/2p) \leqslant \sqrt{2n\log(2p/\gamma)}, \tag{3.11}$$

so $\sqrt{2n\log(2p/\gamma)}$ provides a simple upper bound on the penalty level. Note also that Belloni, Chen, Chernozhukov, and Hansen (2010) formulate a feasible Lasso procedure for the case with heteroscedastic, non-Gaussian disturbances. We shall refer to the feasible Lasso method with the feasible penalty levels (3.9) or (3.10) as the *Iterated Lasso*. This estimator has statistical performance that is similar to that of the (infeasible) Lasso described above.

Belloni, Chernozhukov, and Wang (2011) propose a variant called the *Square-root Lasso* estimator $\widehat{\beta}$ defined as a solution to the following program:

$$\min_{\beta\in\mathbb{R}^p} \sqrt{\mathbb{E}_n[(y_i - x_i'\beta)^2]} + \frac{\lambda}{n}\|\beta\|_1, \tag{3.12}$$

with the penalty level

$$\lambda = c\cdot\widetilde{\Lambda}(1-\gamma|X), \tag{3.13}$$

where $c > 1$ and

$$\widetilde{\Lambda}(1-\gamma|X) = (1-\gamma) - \text{quantile of } \; n\|\mathbb{E}_n[x_i g_i]\|_\infty / \sqrt{\mathbb{E}_n[g_i^2]} \mid X,$$

with $g_i \sim N(0,1)$ independent for $i = 1, \ldots, n$. As with Lasso, there is also simple asymptotic option for setting the penalty level:

$$\lambda = c\cdot\Phi^{-1}(1-\gamma/2p). \tag{3.14}$$

The main attractive feature of (3.12) is that the penalty level λ is independent of the value σ, and so it is pivotal with respect to that parameter. Nonetheless, this estimator has statistical performance that is similar to that of the (infeasible) Lasso described above. Moreover, the estimator is a solution to a highly tractable conic programming problem:

$$\min_{t\geqslant 0,\beta\in\mathbb{R}^p} t + \frac{\lambda}{n}\|\beta\|_1 : \quad \sqrt{\mathbb{E}_n[(y_i - x_i'\beta)^2]} \leqslant t, \tag{3.15}$$

where the criterion function is linear in parameters t and positive and negative components of β, while the constraint can be formulated with a second-order cone, informally known also as the "ice-cream cone".

There are several other estimators that make use of penalization by the ℓ_1-norm. An important case includes the Dantzig selector estimator proposed and analyzed by Candes and Tao (2007). It also relies on ℓ_1-regularization but exploits the notion that the residuals should be nearly uncorrelated with the covariates. The estimator is defined as a solution to:

$$\min_{\beta\in\mathbb{R}^p} \quad \|\beta\|_1 \quad : \quad \|\mathbb{E}_n[x_i(y_i - x_i'\beta)]\|_\infty \leqslant \lambda/n \tag{3.16}$$

where $\lambda = \sigma\Lambda(1-\gamma|X)$. In what follows we will focus our discussion on Lasso but virtually all theoretical results carry over to other ℓ_1-regularized estimators including (3.12) and (3.16).

We also refer to Gautier and Tsybakov (2011) for a feasible Dantzig estimator that combines the square-root lasso method (3.15) with the Dantzig method.

ℓ_1-regularized estimators often have a substantial shrinkage bias. In order to remove some of this bias, we consider the post-model-selection estimator that applies ordinary least squares regression to the model $\widehat{T}$ selected by a ℓ_1-regularized estimator $\widehat{\beta}$. Formally, set

$$\widehat{T} = \text{support}(\widehat{\beta}) = \{j \in \{1, \ldots, p\} \ : \ |\widehat{\beta}_j| > 0\},$$

and define the post model selection estimator $\widetilde{\beta}$ as

$$\widetilde{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \mathbb{E}_n[(y_i - x_i'\beta)^2] \quad : \quad \beta_j = 0 \text{ for each } j \in \widehat{T}^c, \tag{3.17}$$

where $\widehat{T}^c = \{1, ..., p\} \setminus \widehat{T}$. In words, the estimator is ordinary least squares applied to the data after removing the regressors that were not selected in $\widehat{T}$. When the ℓ_1-regularized method used to select the model is Lasso (Square-root Lasso), the post-model-selection estimator is called Post-Lasso (Post-Square-root Lasso). If model selection works perfectly – that is, $\widehat{T} = T$ – then the post-model-selection estimator is simply the oracle estimator whose properties are well-known. However, perfect model selection is unlikely in many situations, so we are interested in the properties of the post-model-selection estimator when model selection is imperfect, i.e. when $\widehat{T} \neq T$, and are especially interested in cases where $T \nsubseteq \widehat{T}$. In Section 4 we describe the formal properties of the Post-Lasso estimator.

3.2. **Some Heuristics via Convex Geometry.** Before proceeding to the formal results on estimation, it is useful to consider some heuristics for the ℓ_1-penalized estimators and the choice of the penalty level. For this purpose we consider a parametric model, and a generic ℓ_1-regularized estimator based on a differentiable criterion function $\widehat{Q}$:

$$\widehat{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \widehat{Q}(\beta) + \frac{\lambda}{n}\|\beta\|_1, \tag{3.18}$$

where, e.g., $\widehat{Q}(\beta) = \mathbb{E}_n[(y_i - x_i'\beta)^2]$ for Lasso and $\widehat{Q}(\beta) = \sqrt{\mathbb{E}_n[(y_i - x_i'\beta)^2]}$ for Square-root Lasso. The key quantity in the analysis of (3.18) is the score – the gradient of $\widehat{Q}$ at the true value[8]:

$$S = \nabla \widehat{Q}(\beta_0).$$

The score S is the effective "noise" in the problem that should be dominated by the regularization. However we would like to make the regularization bias as small as possible. This reasoning suggests choosing the smallest penalty level λ that is large enough to dominate the noise with high probability, say $1 - \gamma$, which yields

$$\lambda > c\Lambda, \text{ for } \Lambda := n\|S\|_\infty, \tag{3.19}$$

where Λ is the maximal score scaled by n, and $c > 1$ is a theoretical constant of Bickel, Ritov, and Tsybakov (2009) that guarantees that the score is dominated. We note that the principle

[8]In the case of a nonparametric model the score is similar to the gradient of $\widehat{Q}$ at β_0 but ignores the approximation errors r_i's.

of setting λ to dominate the score of the criterion function is a general principle that carries over to other convex problems with possibly non-differentiable criterion functions and that leads to the optimal – near-oracle – performance of ℓ_1-penalized estimators. See, for instance, Belloni and Chernozhukov (2011a).

It is useful to mention some simple heuristics for the principle (3.19) which arise from considering the simplest case where none of the regressors are significant so that $\beta_0 = 0$. We want our estimator to perform at a near-oracle level in all cases, including this case, but here the oracle estimator β^* sets $\beta^* = \beta_0 = 0$. We also want $\widehat{\beta} = \beta_0 = 0$ in this case, at least with a high probability, say $1 - \gamma$. From the subgradient optimality conditions for (3.18), we must have

$$-S_j + \lambda/n > 0 \text{ and } S_j + \lambda/n > 0 \text{ for all } 1 \leqslant j \leqslant p$$

for this to be true. We can only guarantee this by setting the penalty level λ/n such that $\lambda > n \max_{1 \leqslant j \leqslant p} |S_j| = n\|S\|_\infty$ with probability at least $1 - \gamma$. This is precisely the rule (3.19) appearing above.

Finally, note that in the case of Lasso and Square-root Lasso we have the following expressions for the score:

$$\text{Lasso}: \quad S = 2\mathbb{E}_n[x_i \epsilon_i] =_d 2\sigma \mathbb{E}_n[x_i g_i],$$

$$\text{Square-root Lasso}: \quad S = \frac{\mathbb{E}_n[x_i \epsilon_i]}{\sqrt{\mathbb{E}_n[\epsilon_i^2]}} =_d \frac{\mathbb{E}_n[x_i g_i]}{\sqrt{\mathbb{E}_n[g_i^2]}},$$

where g_i are i.i.d. $N(0,1)$ variables. Note that the score for Square-root Lasso is pivotal, while the score for Lasso is not, as it depends on σ. Thus, the choice of the penalty level for Square-root Lasso need not depend on σ to produce near-oracle performance for this estimator.

3.3. **Beyond Mean Models.** Most of the literature on high dimensional sparse models focuses on the mean regression model discussed above. Here we discuss methods that have been proposed to deal with quantile regression and generalized linear models in high-dimensional sparse settings. We assume i.i.d. sampling for (y_i, x_i) in this subsection.

3.3.1. *Quantile Regression.* We consider a response variable y_i and p-dimensional covariates x_i such that the u-th conditional quantile function of y_i given x_i is given by

$$F^{-1}_{y_i|x_i}(u|x) = x'\beta(u), \quad \beta(u) \in \mathbb{R}^p, \tag{3.20}$$

where $u \in (0,1)$ is quantile index of interest. Recall that the u-th conditional quantile $F^{-1}_{y_i|x_i}(u|x)$ is the inverse of the conditional distribution function $F_{y_i|x_i}(y|x)$ of y_i given $x_i = x$. Suppose that the true model $\beta(u)$ has a sparse support:

$$T_u = \text{support}(\beta(u)) = \{j \in \{1, \ldots, p\} \ : \ |\beta_j(u)| > 0\}$$

has only $s_u \leqslant s \leqslant n/\log(n \vee p)$ non-zero components.

The population coefficient $\beta(u)$ is known to be a minimizer of the criterion function

$$Q_u(\beta) = \mathrm{E}[\rho_u(y_i - x_i'\beta)], \tag{3.21}$$

where $\rho_u(t) = (u - 1\{t \leqslant 0\})t$ is the asymmetric absolute deviation function; see Koenker and Bassett (1978). Given a random sample $(y_1, x_1), \ldots, (y_n, x_n)$, $\widehat{\beta}(u)$, the quantile regression estimator of $\beta(u)$, is defined as a minimizer of the empirical analog of (3.21):

$$\widehat{Q}_u(\beta) = \mathbb{E}_n \left[\rho_u(y_i - x_i'\beta)\right]. \tag{3.22}$$

As before, in high-dimensional settings, ordinary quantile regression is generally not consistent, which motivates the use of penalization in order to remove all, or at least nearly all, regressors whose population coefficients are zero. The ℓ_1-penalized quantile regression estimator $\widehat{\beta}(u)$ is a solution to the following optimization problem:

$$\min_{\beta \in \mathbb{R}^p} \ \widehat{Q}_u(\beta) + \frac{\lambda\sqrt{u(1-u)}}{n}\|\beta\|_1. \tag{3.23}$$

The criterion function in (3.23) is the sum of the criterion function (3.22) and a penalty function given by a scaled ℓ_1-norm of the parameter vector.

In order to describe choice of the penalty level λ, we introduce the random variable

$$\Lambda = n \max_{1 \leqslant j \leqslant p} \left| \mathbb{E}_n \left[\frac{x_{ij}(u - 1\{u_i \leqslant u\})}{\sqrt{u(1-u)}} \right] \right|, \tag{3.24}$$

where $u_1, \ldots, u_n$ are i.i.d. uniform $(0, 1)$ random variables, independently distributed from the regressors, $x_1, \ldots, x_n$. The random variable Λ has a pivotal distribution conditional on $X = [x_1, \ldots, x_n]'$. Then, for $c > 1$, Belloni and Chernozhukov (2011a) propose to set

$$\lambda = c \cdot \Lambda(1 - \gamma|X), \ \text{where} \ \ \Lambda(1 - \gamma|X) := (1 - \gamma)\text{-quantile of } \Lambda \text{ conditional on } X, \tag{3.25}$$

and $1 - \gamma$ is a confidence level that needs to be set close to 1.

The post-penalized estimator (post-ℓ_1-QR) applies ordinary quantile regression to the model $\widehat{T}_u$ selected by the ℓ_1-penalized quantile regression (Belloni and Chernozhukov 2011a). Specifically, set

$$\widehat{T}_u = \mathrm{support}(\widehat{\beta}(u)) = \{j \in \{1, \ldots, p\} \ : \ |\widehat{\beta}_j(u)| > 0\},$$

and define the post-penalized estimator $\widetilde{\beta}(u)$ as

$$\widetilde{\beta}(u) \in \arg\min_{\beta \in \mathbb{R}^p} \widehat{Q}_u(\beta) \ : \ \beta_j = 0, \quad j \in \widehat{T}_u^c \tag{3.26}$$

which is just ordinary quantile regression removing the regressors that were not selected in the first step. Belloni and Chernozhukov (2011a) derive the basic properties of the estimators above; see also Kato (2011) for further important results in nonparametric setting, where group penalization is also studied.

3.3.2. *Generalized Linear Models.* From the discussion above, it is clear that ℓ_1-regularized methods can be extended to other criterion functions $\widehat{Q}$ beyond least squares and quantile regression. ℓ_1-regularized generalized linear models were considered in van de Geer (2008). Let $y \in \mathbb{R}$ denote the response variable and $x \in \mathbb{R}^p$ the covariates. The criterion function of interest is defined as

$$\widehat{Q}(\beta) = \frac{1}{n}\sum_{i=1}^{n} h(y_i, x_i'\beta)$$

where h is convex and 1-Lipschitz with respect the second argument, $|h(y,t)-h(y,t')| \leqslant |t-t'|$. We assume h is differentiable in the second argument with derivative denoted ∇h to simplify exposition. Let the true model parameter be defined by $\beta_0 \in \arg\min_{\beta\in\mathbb{R}^p} \mathrm{E}[h(y_i, x_i'\beta)]$, and consequently we have $\mathrm{E}[x_i \nabla h(y_i, x_i'\beta_0)] = 0$. The ℓ_1-regularized estimator is given by the solution of

$$\min_{\beta\in\mathbb{R}^p} \widehat{Q}(\beta) + \frac{\lambda}{n}\|\beta\|_1.$$

Under high level conditions van de Geer (2008) derived bounds on the excess forecasting loss, $\mathrm{E}[h(y_i, x_i'\widehat{\beta})] - \mathrm{E}[h(y_i, x_i'\beta_0)]$, under sparsity-related assumptions, and also specialized the results to logistic regression, density estimation, and other problems.[9] The choice of penalty parameter λ derived in van de Geer (2008) relies on using the contraction inequalities of Ledoux and Talagrand (1991) in order to bound the score:

$$n\|\nabla\widehat{Q}(\beta_0)\|_\infty = \left\|\sum_{i=1}^{n} x_i \nabla h(y_i, x_i'\beta_0)\right\|_\infty \lesssim_P \left\|\sum_{i=1}^{n} x_i \xi_i\right\|_\infty, \tag{3.27}$$

where ξ_i are independent Rademacher random variables, $P(\xi_i = 1) = P(\xi_i = -1) = 1/2$. Then van de Geer (2008) suggests further bounds on the right side of (3.27). For efficiency reasons, we suggest simulating the $1-\gamma$ quantiles of the right side of (3.27) conditional on regressors. In either way one can achieve the domination of "noise" $\lambda/n \geqslant c\|\nabla\widehat{Q}(\beta_0)\|_\infty$ with high probability. Note that since h is 1-Lipschitz, this choice of the penalty level is pivotal.

4. Estimation Results for High Dimensional Sparse Models

4.1. **Convergence Rates for Lasso and Post-Lasso.** Having introduced Condition ASM and the target parameter defined via (2.3), our task becomes to estimate β_0. We will focus on convergence results in the *prediction norm* for $\delta = \widehat{\beta} - \beta_0$, which measures the accuracy of predicting $x_i'\beta_0$ over the design points $x_1, \ldots, x_n$,

$$\|\delta\|_{2,n} := \sqrt{\mathbb{E}_n[(x_i'\delta)^2]} = \sqrt{\delta'\mathbb{E}_n[x_i x_i']\delta}.$$

The prediction norm directly depends on the the Gram matrix $\mathbb{E}_n[x_i x_i']$. Whenever $p > n$, the empirical Gram matrix $\mathbb{E}_n[x_i x_i']$ does not have full rank and in principle is not well-behaved.

[9] Results in other norms of interest could also be derived, and the behavior of the post-ℓ_1-regularized estimators would also be interesting to consider. This is an interesting venue for future work.

However, we only need good behavior of certain moduli of continuity of the Gram matrix called sparse eigenvalues. We define the minimal m-sparse eigenvalue of a semi-definite matrix M as

$$\phi_{\min}(m)[M] := \min_{\|\delta\|_0 \leqslant m, \delta \neq 0} \frac{\delta' M \delta}{\|\delta\|^2}, \tag{4.28}$$

and the maximal m-sparse eigenvalue as

$$\phi_{\max}(m)[M] := \max_{\|\delta\|_0 \leqslant m, \delta \neq 0} \frac{\delta' M \delta}{\|\delta\|^2}, \tag{4.29}$$

To assume that $\phi_{\min}(m)[\mathbb{E}_n[x_i x_i']] > 0$ requires that all empirical Gram submatrices formed by any m components of x_i are positive definite. To simplify asymptotic statements for Lasso and Post-Lasso, we use the following condition:

Condition SE. *There is $\ell_n \to \infty$ such that*

$$\kappa' \leqslant \phi_{\min}(\ell_n s)[\mathbb{E}_n[x_i x_i']] \leqslant \phi_{\max}(\ell_n s)[\mathbb{E}_n[x_i x_i']] \leqslant \kappa'',$$

where $0 < \kappa' < \kappa'' < \infty$ are constants that do not depend on n.

Comment 4.1. It is well-known that Condition SE is quite plausible for many designs of interest. For instance, Condition SE holds with probability approaching one as $n \to \infty$ if x_i is a normalized form of $\tilde{x}_i$, namely $x_{ij} = \tilde{x}_{ij} / \sqrt{\mathbb{E}_n[\tilde{x}_{ij}^2]}$, and

- $\tilde{x}_i$, $i = 1, \ldots, n$, are i.i.d. zero-mean Gaussian random vectors that have population Gram matrix $\mathrm{E}[\tilde{x}_i \tilde{x}_i']$ with ones on the diagonal and its minimal and maximal $s \log n$-sparse eigenvalues bounded away from zero and from above, where $s \log n = o(n/\log p)$;
- $\tilde{x}_i$, $i = 1, \ldots, n$, are i.i.d. bounded zero-mean random vectors with $\|\tilde{x}_i\|_\infty \leqslant K_n$ a.s. that have population Gram matrix $\mathrm{E}[\tilde{x}_i \tilde{x}_i']$ with ones on the diagonal and its minimal and maximal $s \log n$-sparse eigenvalues bounded from above and away from zero, where $K_n^2 s \log^5(p \vee n) = o(n)$.

Recall that a standard assumption in econometric research is to assume that the population Gram matrix $\mathrm{E}[x_i x_i']$ has eigenvalues bounded from above and below, see e.g. Newey (1997). The conditions above allow for this and more general behavior, requiring only that the $s \log n$ sparse eigenvalues of the population Gram matrix $\mathrm{E}[x_i x_i']$ are bounded from below and from above. The latter is important for allowing functions x_i to be formed as a combination of elements from different bases, e.g. a combination of B-splines with polynomials. □

The following theorem describes the rate of convergence for feasible Lasso in the Gaussian model under Conditions ASM and SE. We formally define the *feasible Lasso* estimator $\widehat{\beta}$ as either the Iterated Lasso with penalty level given by X-independent rule (3.9) or X-dependent rule (3.10) or Square-root Lasso with penalty level given by X-dependent rule (3.13) or X-independent rule (3.14), with the confidence level $1 - \gamma$ such that

$$\gamma = o(1) \text{ and } \log(1/\gamma) \lesssim \log(p \vee n). \tag{4.30}$$

Theorem 1 (Rates for Feasible Lasso). *Suppose that conditions ASM and SE hold. Then for n large enough the following bounds hold with probability at least $1-\gamma$:*

$$C'\|\widehat{\beta}-\beta_0\| \leqslant \|\widehat{\beta}-\beta_0\|_{2,n} \leqslant C\sigma\sqrt{\frac{s\log(2p/\gamma)}{n}},$$

where $C>0$ and $C'>0$ are constants, $C' \gtrsim \sqrt{\kappa'}$ and $C \lesssim 1/\sqrt{\kappa'}$, and $\log(p/\gamma) \lesssim \log(p\vee n)$.

Comment 4.2. Thus the rate for estimating β_0 is $\sqrt{s/n}$, i.e. the root of the number of parameters s in the "true" model divided by the sample size n, times a logarithmic factor $\sqrt{\log(p\vee n)}$. The latter factor can be thought of as the price of not knowing the "true" model. Note that the rate for estimating the regression function f over design points follows from the triangle inequality and Condition ASM:

$$\sqrt{\mathbb{E}_n[(f(z_i)-x_i'\widehat{\beta})^2]} \leqslant \|\widehat{\beta}-\beta_0\|_{2,n} + c_s \lesssim_P \sigma\sqrt{\frac{s\log(p\vee n)}{n}}. \tag{4.31}$$

Comment 4.3. The result of Theorem 1 is an extension of the results in the fundamental work of Bickel, Ritov, and Tsybakov (2009) and Meinshausen and Yu (2009) on infeasible Lasso and Candes and Tao (2007) on the Dantzig estimator. The result of Theorem 1 is derived in Belloni and Chernozhukov (2011c) for Iterated Lasso, and in Belloni, Chernozhukov, and Wang (2011) and Belloni, Chernozhukov, and Wang (2010) for Square-root Lasso (with constants C given explicitly). Similar results also hold for ℓ_1-QR (Belloni and Chernozhukov 2011a) and other M-estimation problems (van de Geer 2008). The bounds of Theorem 1 allow the constructions of confidence sets for β_0, as noted in Chernozhukov (2009); see also Gautier and Tsybakov (2011). Such confidence sets rely on efficiently bounding C. Computing bounds for C requires computation of combinatorial quantities depending on the unknown model T which makes the approach difficult in practice. In the subsequent sections, we will present completely different approaches to inference which have provable confidence properties for parameters of interest and which are computationally tractable. □

As mentioned before, ℓ_1-regularized estimators have an inherent bias towards zero and Post-Lasso was proposed to remove this bias, at least in part. It turns out that we can bound the performance of Post-Lasso as a function of Lasso's rate of convergence and Lasso's model selection ability. For common designs, this bound implies that Post-Lasso performs at least as well as Lasso, and it can be strictly better in some cases. Post-Lasso also has a smaller shrinkage bias than Lasso by construction.

The following theorem applies to any Post-Lasso estimator $\widetilde{\beta}$ computed using the model $\widehat{T}=\text{support}(\widehat{\beta})$ selected by a Feasible Lasso estimator $\widehat{\beta}$ defined before Theorem 1.

Theorem 2 (Rates for Feasible Post-Lasso). *Suppose the conditions of Theorem 1 hold and let $\varepsilon>0$. Then there are constants C' and C_ε such that with probability $1-\gamma$*

$$\widehat{s}=|\widehat{T}| \leqslant C's,$$

and with probability $1-\gamma-\varepsilon$

$$\sqrt{\kappa'}\|\widetilde{\beta}-\beta_0\| \leqslant \|\widetilde{\beta}-\beta_0\|_{2,n} \leqslant \quad C_\varepsilon \sigma \sqrt{\frac{s\log(p\vee n)}{n}}. \tag{4.32}$$

If further $|\|\widehat{\beta}\|_0 - s| = o(s)$ *and* $T \subseteq \widehat{T}$ *with probability approaching one, then*

$$\|\widetilde{\beta}-\beta_0\|_{2,n} \lesssim_P \quad \sigma\left[\sqrt{\frac{o(s)\log(p\vee n)}{n}} + \sqrt{\frac{s}{n}}\right]. \tag{4.33}$$

If $\widehat{T} = T$ *with probability approaching one, then Post-Lasso achieves the oracle performance*

$$\|\widetilde{\beta}-\beta_0\|_{2,n} \lesssim_P \quad \sigma\sqrt{s/n}. \tag{4.34}$$

Comment 4.4. The theorem above shows that Feasible Post-Lasso achieves the same near-oracle rate as Feasible Lasso. Notably, this occurs despite the fact that Feasible Lasso may in general fail to correctly select the oracle model T as a subset, that is $T \not\subseteq \widehat{T}$. The intuition for this result is that any components of T that Feasible Lasso misses are very unlikely to be important. Theorem 2 was derived in Belloni and Chernozhukov (2011c) and Belloni, Chernozhukov, and Wang (2010). Similar results have been shown before for ℓ_1-QR (Belloni and Chernozhukov 2011a), and can be derived for other methods that yield sparse estimators. □

4.2. **Monte Carlo Example.** In this section we compare the performance of various estimators relative to the ideal oracle linear regression estimator. The oracle estimator applies ordinary least square to the true model by regressing the outcome on only the control variables with non-zero coefficients. Of course, the oracle estimator is not available outside Monte Carlo experiments.

We considered the following regression model:

$$y = x'\beta_0 + \epsilon, \quad \beta_0 = (1, 1, 1/2, 1/3, 1/4, 1/5, 0, \ldots, 0)',$$

where $x = (1, z')'$ consists of an intercept and covariates $z \sim N(0, \Sigma)$, and the errors ϵ are independently and identically distributed $\epsilon \sim N(0, \sigma^2)$. The dimension p of the covariates x is 500, and the dimension s of the true model is 6. The sample size n is 100. The regressors are correlated with $\Sigma_{ij} = \rho^{|i-j|}$ and $\rho = .5$. We consider the levels of noise to be $\sigma = 1$ and $\sigma = 0.1$. For each repetition we draw new x's and ϵ's.

We consider infeasible Lasso and Post-Lasso estimators, feasible Lasso and Post-Lasso estimators described in the previous section, all with X-dependent penalty levels, as well as (5-fold) cross-validated (CV) Lasso and Post-Lasso. We summarize results on estimation performance in Table 2 which records for each estimator $\bar{\beta}$ the norm of the bias $\|\mathrm{E}[\bar{\beta} - \beta_0]\|$ and also the empirical risk $\{\mathrm{E}[(x_i'(\bar{\beta} - \beta_0))^2]\}^{1/2}$ for recovering the regression function. In this design, infeasible Lasso, Square-root Lasso, and Iterated Lasso exhibit substantial bias toward zero. This bias is somewhat alleviated by choosing the penalty-level via cross-validation, though the

remaining bias is still substantial. It is also apparent that, as intuition and theory would suggest, the post-penalized estimators remove a large portion of this shrinkage bias. We see that among the feasible estimators, the best performing methods are the Post-Square-root Lasso and Post-Iterated Lasso. Interestingly, cross-validation also produces a Post-Lasso estimator that performs nearly as well, although the procedure is much more expensive computationally. The Post-Lasso estimators perform better than Lasso estimators primarily due to a much lower shrinkage bias which is beneficial in the design considered.

	High Noise ($\sigma = 1$)		**Low Noise** ($\sigma = 0.1$)	
Estimator	Bias	Prediction Error	Bias	Prediction Error
Lasso	0.444	0.654	0.0487	0.0700
Post-Lasso	0.129	0.347	0.0054	0.0300
Square-root Lasso	0.526	0.770	0.0615	0.0870
Post-Square-root Lasso	0.187	0.364	0.0035	0.0238
Iterated Lasso	0.437	0.644	0.0477	0.0687
Post-Iterated Lasso	0.133	0.360	0.0056	0.0297
CV Lasso	0.265	0.516	0.0233	0.0987
CV Post-Lasso	0.148	0.415	0.0035	0.0237
Oracle	0.035	0.238	0.0035	0.0237

TABLE 2. The table displays the mean bias and the mean prediction error. The average number of components selected by Lasso was 5.18 in the high noise case and 6.44 in the low noise case. In the case of CV Lasso, the average size of the model was 29.6 in the high noise case and 10.0 in the low noise case. Finally, the CV Post-Lasso selected models with average size of 7.1 in the high noise case and 6.0 in the low noise case.

5. Inference on Structural Effects with High-Dimensional Instruments

5.1. **Methods and Theoretical Results.** In this section, we consider the linear instrumental variable (IV) model with many instruments. Consider the Gaussian simultaneous equation model:

$$y_{1i} = y_{2i}\alpha_1 + w_i'\alpha_2 + \zeta_i, \tag{5.35}$$

$$y_{2i} = f(z_i) + v_i, \tag{5.36}$$

$$\begin{pmatrix} \zeta_i \\ v_i \end{pmatrix} \mid z_i \sim N\left(0, \begin{pmatrix} \sigma_\zeta^2 & \sigma_{\zeta v} \\ \sigma_{\zeta v} & \sigma_v^2 \end{pmatrix}\right). \tag{5.37}$$

Here y_{1i} is the response variable, y_{2i} is the endogenous variable, w_i is a k_w-vector of control variables, $z_i = (u_i', w_i')'$ is a vector of instrumental variables (IV), and (ζ_i, v_i) are disturbances that are independent of z_i. The function $f(z_i) = \mathrm{E}[y_{2i}|z_i]$, the optimal instrument, is an unknown, potentially complicated function of the elementary instruments z_i. The main parameter of interest is the coefficient on y_{2i}, whose true value is α_1. We treat $\{z_i\}$ as fixed throughout.

Based on these elementary instruments, we create a high-dimensional vector of technical instruments, $x_i = P(z_i)$, with dimension p possibly much larger than the sample size though restricted via conditions stated below. We then estimate the the optimal instrument $f(z_i)$ by

$$\widehat{f}(z_i) = x_i'\widehat{\beta}, \tag{5.38}$$

where $\widehat{\beta}$ is a feasible Lasso or Post-Lasso estimator as formally defined in the previous section.

Sparse-methods take advantage of approximate sparsity and ensure that many elements of $\widehat{\beta}$ are zero when p is large. In other words, sparse-methods will select a small subset of the available technical instruments. Let $A_i = (f(z_i), w_i')'$ be the ideal instrument vector, and let

$$\widehat{A}_i = (\widehat{f}(z_i), w_i')' \tag{5.39}$$

be the estimated instrument vector. Denoting $d_i = (y_{2i}, w_i')'$, we form the feasible IV estimator using the estimated instrument vector as

$$\widehat{\alpha}^* = \Big(\mathbb{E}_n[\widehat{A}_i d_i']\Big)^{-1}\Big(\mathbb{E}_n[\widehat{A}_i y_{1i}]\Big). \tag{5.40}$$

The main regularity condition is recorded as follows.

Condition ASIV. *In the linear IV model (5.35)-(5.37) with technical instruments $x_i = P(z_i)$, the following assumptions hold: (i) the parameter values σ_v, σ_ζ and the eigenvalues of $Q_n = \mathbb{E}_n[A_i A_i']$ are bounded away from zero and from above uniformly in n, (ii) condition ASM holds for (5.36), namely for each $i = 1, ..., n$, there exists $\beta_0 \in \mathbb{R}^p$, such that $f(z_i) = x_i'\beta_0 + r_i$, $\|\beta_0\| \leqslant s$, $\{\mathbb{E}_n[r_i^2]\}^{1/2} \leqslant K\sigma_v\sqrt{s/n}$, where constant K does not depend on n, (iii) condition SE holds for $\mathbb{E}_n[x_i x_i']$, and (iv) $s^2\log^2(p \vee n) = o(n)$.*

The main inference result is as follows.

Theorem 3 (Asymptotic Normality for IV Estimator Based on Lasso and Post-Lasso)**.** *Suppose Condition ASIV holds. The IV estimator constructed in (5.40) is $\sqrt{n}$-consistent and is asymptotically efficient, namely as n grows:*

$$(\sigma_\zeta^2 Q_n^{-1})^{-1/2}\sqrt{n}(\widehat{\alpha}^* - \alpha) = N(0, I) + o_P(1),$$

and the result also holds with Q_n replaced by $\widehat{Q}_n = \mathbb{E}_n[\widehat{A}_i\widehat{A}_i']$ and σ_ζ^2 by $\widehat{\sigma}_\zeta^2 = \mathbb{E}_n[(y_{1i} - \widehat{A}_i'\widehat{\alpha}^*)^2]$.

Comment 5.1. The theorem shows that the IV estimator based on estimating the first-stage with Lasso or Post-Lasso is asymptotically as efficient as the infeasible optimal IV estimator that uses A_i and thus achieves the semi-parametric efficiency bound of Chamberlain (1987). Belloni, Chernozhukov, and Hansen (2010) show that the result continues to hold when other sparse methods are used to estimate the optimal instruments. The sufficient conditions for showing the IV estimator obtained using sparse-methods to estimate the optimal instruments is asymptotically efficient include a set of technical conditions and the following key growth condition: $s^2\log^2(p \vee n) = o(n)$. This rate condition requires the optimal instruments to be sufficiently smooth so that a relatively small number of series terms can be used to approximate

them well. This smoothness ensures that the impact of instrument estimation on the IV estimator is asymptotically negligible. The rate condition $s^2 \log^2(p \vee n) = o(n)$ can be substantive and cannot be substantially weakened for the full-sample IV estimator considered above. However, we can replace this condition with the weaker condition that $s \log(p \vee n) = o(n)$ by employing a sample splitting method from the many instruments literature (Angrist and Krueger 1995) as established in Belloni, Chernozhukov, and Hansen (2010) and Belloni, Chen, Chernozhukov, and Hansen (2010). Moreover, Belloni, Chen, Chernozhukov, and Hansen (2010) show that the result of the theorem, with some appropriate modifications, continues to apply under heteroscedasticity though the estimator does not necessarily attain the semi-parametric efficiency bound. In order to achieve full efficiency allowing for heteroscedasticity, we would need to estimate the conditional variance of the structural disturbances in the second stage equation. In principle, this estimation could be done using sparse methods. □

5.2. **Weak Identification Robust Inference with Very Many Instruments.** Consider the simultaneous equation model:

$$y_{1i} = y_{2i}\alpha_1 + w_i'\alpha_2 + \zeta_i, \quad \zeta_i \mid z_i \sim N\left(0, \sigma_\zeta^2\right), \tag{5.41}$$

where y_{1i} is the response variable, y_{2i} is the endogenous variable, w_i is a k_w-vector of control variables, $z_i = (u_i', w_i')'$ is a vector of instrumental variables (IV), and ζ_i is a disturbance that is independent of z_i. We treat $\{z_i\}$ as fixed throughout.

We would like to use a high-dimensional vector $x_i = P(z_i)$ of technical instruments for inference that is robust to weak identification. We propose a method for inference based on inverting pointwise tests performed using a sup-score statistic defined below. The procedure is similar in spirit to Anderson and Rubin (1949) and Staiger and Stock (1997) but uses a very different statistics that is well-suited to cases with very many instruments.

In order to formulate the sup-score statistic, we first partial-out the effect of controls w_i on the key variables. For an n-vector $\{u_i, i = 1, ..., n\}$, define $\tilde{u}_i = u_i - w_i'\mathbb{E}_n[w_i w_i']^{-1}\mathbb{E}_n[w_i u_i]$, i.e. the residuals left after regressing this vector on $\{w_i, i = 1, ..., n\}$. Hence $\tilde{y}_{1i}$, $\tilde{y}_{2i}$, and $\tilde{x}_{ij}$ are residuals obtained by partialling out controls. Also, let $\tilde{x}_i = (\tilde{x}_{i1}, ..., \tilde{x}_{ip})'$. In this formulation, we omit elements of w_i from $\tilde{x}_{ij}$ since they are eliminated by partialling out. We then normalize without loss of generality

$$\mathbb{E}_n[\tilde{x}_{ij}^2] = 1, \quad j = 1, ..., p. \tag{5.42}$$

The sup-score statistic for testing the hypothesis $\alpha_1 = a$ takes the form:

$$\Lambda_a = \max_{1 \leqslant j \leqslant p} \frac{|n\mathbb{E}_n[(\tilde{y}_{1i} - \tilde{y}_{2i}a)\tilde{x}_{ij}]|}{\sqrt{\mathbb{E}_n[(\tilde{y}_{1i} - \tilde{y}_{2i}a)^2\tilde{x}_{ij}^2]}}.$$

If the hypothesis $\alpha_1 = a$ is true, then the critical value for achieving level γ is

$$\Lambda(1 - \gamma | W, X) = 1 - \gamma - \text{quantile of} \quad \max_{1 \leqslant j \leqslant p} \frac{|n\mathbb{E}_n[\tilde{g}_i\tilde{x}_{ij}]|}{\sqrt{\mathbb{E}_n[\tilde{g}_i^2\tilde{x}_{ij}^2]}} \mid W, X \tag{5.43}$$

where $W = [w_1, ..., w_n]'$, $X = [x_1, ..., x_n]'$, and $g_1, ..., g_n$ are i.i.d. $N(0,1)$ variables independent of W and X; $\tilde{g}_i$ denotes the residuals left after projecting $\{g_i\}$ on $\{w_i\}$ as defined above. We can approximate the critical value $\Lambda(1-\gamma|W,X)$ by simulation conditional on X and W. It is also possible to use a simple asymptotic bound on this critical value of the form

$$\Lambda(1-\gamma) := c\sqrt{n}\Phi^{-1}(1-\gamma/2p) \leq c\sqrt{2n\log(2p/\gamma)}, \tag{5.44}$$

for $c > 1$. The finite-sample $(1-\gamma)$ – confidence region for α_1 is then given by

$$\mathcal{C} := \{a \in \mathbb{R} : \Lambda_a \leqslant \Lambda(1-\gamma|W,X)\},$$

while a large sample $(1-\gamma)$ – confidence region is given by $\mathcal{C}' := \{a \in \mathbb{R} : \Lambda_a \leqslant \Lambda(1-\gamma)\}$.

The main regularity condition is recorded as follows.

Condition HDIV. *Suppose the linear IV model (5.41) holds. Consider the p-vector of instruments $x_i = P(z_i)$, $i = 1, ..., n$, such that $(\log p)/n \to 0$. Suppose further that the following assumptions hold uniformly in n: (i) the parameter value σ_ζ is bounded away from zero and from above, (ii) the dimension of w_i is bounded and the eigenvalues of the Gram matrix $\mathbb{E}_n[w_i w_i']$ are bounded away from zero, (iii) $\|w_i\| \leqslant K$ and $|\tilde{x}_{ij}| \leqslant K$ for all $1 \leqslant i \leqslant n$ and all $1 \leqslant j \leqslant p$, where K is a constant, independent of n.*

The main inference result is as follows.

Theorem 4 (Valid Inference based on the Sup-Score Statistic)**.** *(1) Suppose the linear IV model (5.41) holds. Then $\mathrm{P}(\alpha_1 \in \mathcal{C}) = 1-\gamma$. (2) Suppose further that condition HDIV holds, then $\mathrm{P}(\alpha_1 \in \mathcal{C}') \geqslant 1-\gamma-o(1)$. (3) Moreover, if a is such that that*

$$\max_{1\leqslant j\leqslant p} \frac{|a-\alpha_1|\sqrt{n}|\mathbb{E}_n[\tilde{y}_{2i}\tilde{x}_{ij}]|/\sqrt{\log p}}{\sigma_\zeta + |a-\alpha_1|\sqrt{\mathbb{E}_n[\tilde{y}_{2i}^2\tilde{x}_{ij}^2]}} \to \infty,$$

then $\mathrm{P}(a \in \mathcal{C}) = o(1)$ and $\mathrm{P}(a \in \mathcal{C}') = o(1)$.

Comment 5.2. The theorem shows that the confidence regions $\mathcal{C}$ and $\mathcal{C}'$ constructed above have finite-sample and large sample validity, respectively. Moreover, the probability of including a false point a in either $\mathcal{C}$ or $\mathcal{C}'$ tends to zero as long as a is sufficiently distant from α_1 and instruments are not too weak. In particular, if there is a strong instrument, the confidence regions will eventually exclude points a that are further than $\sqrt{(\log p)/n}$ away from α_1. Moreover, if there are instruments whose correlation with the endogenous variable is of greater order than $\sqrt{(\log p)/n}$, then the confidence regions will asymptotically be bounded. Finally, note that a nice feature of the construction is that it provides provably valid confidence regions and does not require computation of some combinatorial quantities, in sharp contrast to other recent proposals for inference, e.g. Gautier and Tsybakov (2011). Lastly, we note that it is not difficult to generalize the results to allow for an increasing number of controls w_i under suitable technical conditions that restrict the number of controls and their envelope in relation to the sample size. Here we did not consider this possibility in order to highlight the impact of

very many instruments more clearly. The result (2) extends to non-Gaussian, heteroscedastic cases; we refer to Belloni, Chen, Chernozhukov, and Hansen (2010) for relevant details. □

Comment 5.3 (Inverse Lasso Interpretation)**.** The construction of confidence regions above can be given the following *Inverse Lasso* interpretation. Let

$$\widehat{\beta}_a = \arg\min_{\beta\in\mathbb{R}^p} \mathbb{E}_n[(\tilde{y}_{1i} - a\tilde{y}_{2i}) - \tilde{x}_{ij}'\beta]^2 + \frac{\lambda}{n}\sum_{j=1}^{p}|\beta_j|\gamma_{aj}, \quad \gamma_{aj} = \sqrt{\mathbb{E}_n[(\tilde{y}_{1i} - \tilde{y}_{2i}a)^2\tilde{x}_{ij}^2]}.$$

If $\lambda = 2\Lambda(1-\gamma|W,X)$, then $\mathcal{C}$ is equivalent to the region $\{a \in \mathbb{R} : \widehat{\beta}_a = 0\}$. If $\lambda = 2\Lambda(1-\gamma)$, then $\mathcal{C}'$ is equivalent to the region $\{a \in \mathbb{R} : \widehat{\beta}_a = 0\}$. In words, to construct these confidence regions, we collect all potential values of the structural parameter, where the Lasso regression of the potential structural disturbance on the instruments yields zero coefficients on the instruments. This idea is akin to the Inverse Quantile Regression and Inverse Least Squares ideas in Chernozhukov and Hansen (2008a) and Chernozhukov and Hansen (2008b). □

5.3. **Monte Carlo Example: Instrumental Variable Model.** The theoretical results presented in the previous sections suggest that using Lasso to aid in fitting the first-stage regression should result in IV estimators with good estimation and inference properties. In this section, we provide simulation evidence on these properties of IV estimators using iterated Lasso to select instrumental variables for a second-stage estimator. We also considered Square-root Lasso for variable selection. The results were similar to those for iterated Lasso, so we report only the iterated Lasso results.

Our simulations are based on a simple instrumental variables model of the form

$$\begin{array}{ll} y_{1i} & = \alpha y_{2i} + \zeta_i \\ y_{2i} & = x_i'\Pi + v_i \end{array} \qquad \begin{pmatrix} \zeta_i \\ v_i \end{pmatrix} \mid x_i \sim N\left(0, \begin{pmatrix} \sigma_\zeta^2 & \sigma_{\zeta v} \\ \sigma_{\zeta v} & \sigma_v^2 \end{pmatrix}\right) \text{ i.i.d.},$$

where $\alpha = 1$ is the parameter of interest, and $x_i = (x_{i1}, ..., x_{i100})' \sim N(0, \Sigma_X)$ is the instrument vector with $E[x_{ih}^2] = \sigma_x^2$ and $\text{Corr}(x_{ih}, x_{ij}) = .5^{|j-h|}$. In all simulations, we set $\sigma_\zeta^2 = 1$ and $\sigma_x^2 = 1$. We also use $\text{Corr}(\zeta, v) = .3$.

We consider several different settings for the other parameters. We provide simulation results for sample sizes, n, of 100 and 500. In one simulation design, we set $\Pi = 0$ and $\sigma_v^2 = 1$. In this case, the instruments have no information about the endogenous variable, so α is unidentified. We refer to this as the "No Signal" design. In the remaining cases, we use an "exponential" design for the first stage coefficients, Π, that sets the coefficient on $x_{ih} = .7^{h-1}$ for $h = 1, ..., 100$ to provide an example of Lasso's performance in settings where the instruments are informative. This model is approximately sparse, since the majority of explanatory power is contained in the first few instruments, and obeys the regularity conditions put forward above. We consider values of σ_v^2 which are chosen to benchmark three different strengths of instruments. The three values of σ_v^2 are found as $\sigma_v^2 = \frac{n\Pi'\Sigma_Z\Pi}{F^*\Pi'\Pi}$ for F^* of 10, 40, or 160.

For each setting of the simulation parameter values, we report results from several estimation procedures. A simple possibility when presented with $p < n$ instrumental variables is to just estimate the model using 2SLS and all of the available instruments. It is well-known that this will result in poor-finite sample properties unless there are many more observations than instruments; see, for example, Bekker (1994). Fuller's (1977) estimator (FULL)[10] is robust to many instruments as long as the presence of many instruments is accounted for when constructing standard errors and $p < n$; see Bekker (1994) and Hansen, Hausman, and Newey (2008) for example. We report results for these estimators in rows labeled 2SLS(All) and FULL(All) respectively.[11] In addition, we report Fuller and IV estimates based on the set of instruments selected by Lasso with two different penalty selection methods. IV-Lasso and FULL-Lasso are respectively 2SLS and Fuller using instruments selected by Lasso with penalty obtained using the iterated method outlined in Appendix A. We use an initial estimate of the noise level obtained using the regression of y_2 on the instrument that has the highest simple correlation with y_2. IV-Lasso-CV and FULL-Lasso-CV are respectively 2SLS and Fuller using instruments selected by Lasso using 10-fold cross-validation to choose the penalty level. We also report inference results based on the Sup-Score test developed in Section 5.2.

In Table 3, we report root-mean-squared-error (RMSE), median bias (Med. Bias), rejection frequencies for 5% level tests (rp(.05)), and the number of times the Lasso-based procedures select no instruments ($\|\widehat{\Pi}\|_0 = 0$). For computing rejection frequencies, we estimate conventional 2SLS standard errors for all 2SLS estimators, and the many instrument robust standard errors of Hansen, Hausman, and Newey (2008) for the Fuller estimators. In cases where Lasso selects no instruments, the reported Lasso point estimation properties are based on the feasible procedure that enforces identification by lowering the penalty until one variable is selected. Rejection frequencies in cases where no instruments are selected are based on the feasible procedure that uses conventional IV inference using the selected instruments when this set is non-empty and otherwise uses the Sup-Score test.

The simulation results show that Lasso-based IV estimators is useful in situations with many instruments. As expected, 2SLS(All) does extremely poorly along all dimensions. FULL(All) also performs worse than the Lasso-based estimators in terms of estimator risk (RMSE) in all cases. The Lasso-based procedures do not dominate FULL(All) in terms of median bias, though all of the Lasso-based procedures have smaller median bias than FULL(All) when $n = 100$ and there is some signal in the instruments and are very similar with $n = 500$. In terms of size of 5% level tests, we see that the Sup-Score test uniformly controls size as indicated by the theory. IV-Lasso and FULL-Lasso using the iterated penalty selection method also do a very good job controlling size across all of the simulation settings with a worst-case rejection frequency of .064 (with simulation standard error of .01) and the majority of rejection

[10] The Fuller estimator requires a user-specified parameter. We set this parameter equal to one which produces a higher-order unbiased estimator. See Hahn, Hausman, and Kuersteiner (2004) for additional discussion.

[11] All models include an intercept. With $n = 100$, we randomly select 98 instruments to use for 2SLS(All) and FULL(All).

frequencies below .05. Interestingly, when there is no signal in the instrument, the Lasso-based estimators using penalty selected by CV have substantial size-distortions when $n = 100$ which is due to the CV penalty being small enough that instruments are still selected despite there being no signal. The iterated penalty is such that, at least approximately, only instruments whose coefficients are outside of a $\sqrt{n}$ neighborhood of 0 are selected and thus overselection in cases with little signal is guarded against. Despite the problem with using CV when there is no signal, it is worth noting that the Lasso-based procedures with CV penalty produce tests with approximately correct size in all other parameter settings.

To further examine the properties of the inference procedures that appear to give small size distortions, we plot the power curves of 5% level tests using the Sup-Score test and IV-Lasso with the iterated and CV penalty choices with $n = 100$ in Figure 2.[12] We see that both the Sup-Score test and IV-Lasso using the iterated procedure augmented with Sup-Score test when no instruments are selected appear to uniformly control size and have some power against alternatives when the model is identified. It is also clear that of these two procedures, the IV-Lasso has substantially more power than the Sup-Score test. The figures also show that IV-Lasso with iterated penalty has almost as much power as IV-Lasso using the CV penalty while avoiding the substantial size distortion and spurious power produced by using CV when there is no signal.

Overall, the simulation results are favorable to the Lasso-based IV methods. The Lasso-based estimators dominate the other estimators considered based on RMSE and have relatively small finite sample biases. The Lasso-based procedures also do a good job in producing tests with size close to the nominal level. There is some evidence that the Fuller-Lasso may do better than 2SLS-Lasso in terms of testing performance though these procedures are very similar in the designs considered. It also seems that tests based on IV-Lasso using the iterated penalty selection rule may perform better than tests based on IV-Lasso using cross-validation to choose the Lasso penalty levels, especially when there is little explanatory power in the instruments.