cemmap
centre for microdata methods and practice

Valid post-selection and post-regularization inference: An elementary, general approach

Victor Chernozhukov
Christian Hansen
Martin Spindler

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP36/16

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

Valid Post-Selection and Post-Regularization Inference: An Elementary, General Approach

VICTOR CHERNOZHUKOV, CHRISTIAN HANSEN, AND MARTIN SPINDLER

Abstract. Here we present an expository, general analysis of valid post-selection or post-regularization inference about a low-dimensional target parameter in the presence of a very high-dimensional nuisance parameter which is estimated using selection or regularization methods. Our analysis provides a set of high-level conditions under which inference for the low-dimensional parameter based on testing or point estimation methods will be regular despite selection or regularization biases occurring in estimation of the high-dimensional nuisance parameter. The results may be applied to establish uniform validity of post-selection or post-regularization inference procedures for low-dimensional target parameters over large classes of models. The high-level conditions allow one to clearly see the types of structure needed for achieving valid post-regularization inference and encompass many existing results. A key element of the structure we employ and discuss in detail is the use of orthogonal or "immunized" estimating equations that are locally insensitive to small mistakes in estimation of the high-dimensional nuisance parameter. As an illustration, we use the high-level conditions to provide readily verifiable sufficient conditions for a class of affine-quadratic models that include the usual linear model and linear instrumental variables model as special cases. As a further application and illustration, we use these results to provide an analysis of post-selection inference in a linear instrumental variables model with many regressors and many instruments. We conclude with a review of other developments in post-selection inference and note that many of the developments can be viewed as special cases of the general encompassing framework of orthogonal estimating equations provided in this paper.

Key words: Neyman, orthogonalization, $C(\alpha)$ statistics, optimal instrument, optimal score, optimal moment, post-selection and post-regularization inference, efficiency, optimality

1. Introduction

Analysis of high-dimensional models, models in which the number of parameters to be estimated is large relative to the sample size, is becoming increasingly important. Such models arise naturally in readily available high-dimensional data which have many measured characteristics available per individual observation as in, for example, large survey data sets, scanner data, and text data. Such models also arise naturally even in data with a small number of measured characteristics in situations where the exact functional form with which the observed variables enter the model is unknown. Examples of this scenario include semiparametric models with nonparametric nuisance functions.

Date: December, 2014. We thank Denis Chetverikov, Mert Demirer, Anna Mikusheva, seminar participants and the discussant Susan Athey at the AEA Session on Machine Learning in Economics and Econometrics, CEME conference on Non-Standard Problems in Econometrics, Berlin Statistics Seminar, and the students from MIT's 14.387 Applied Econometrics Class for useful comments.

Chernozhukov: Massachussets Institute of Technology, 50 Memorial Drive, E52-361B, Cambridge, MA 02142, vchern@mit.edu. Hansen: University of Chicago Booth School of Business, 5807 S. Woodlawn Ave., Chicago, IL 60637, chansen1@chicagobooth.edu. Spindler: Munich Center for the Economics of Aging, Amalienstr. 33, 80799 Munich, Germany, spindler@mea.mpisoc.mpg.de.

More generally, models with many parameters relative to the sample size often arise when attempting to model complex phenomena.

The key concept underlying the analysis of high-dimensional models is that regularization, such as model selection or shrinkage of model parameters, is necessary if one is to draw meaningful conclusions from the data. For example, the need for regularization is obvious in a linear regression model with number of right-hand-side variables greater than the sample size, but arises far more generally in any setting where the number of parameters is not small relative to the sample size. Given the importance of using regularization in analyzing high-dimensional models, it is then important to explicitly account for the impact of this regularization on the behavior of estimators if one wishes to accurately characterize their finite-sample behavior. The use of such regularization techniques may easily invalidate conventional approaches to inference about model parameters and other interesting target parameters. A major goal of this paper is to provide a general, formal framework which provides guidance about setting up estimating equations and making appropriate use of regularization devices so that inference about parameters of interest will remain valid in the presence of data-dependent model selection or other approaches to regularization.

It is important to note that understanding estimators' behavior in high-dimensional settings is also useful in conventional low-dimensional settings. As noted above, dealing formally with high-dimensional models requires explicitly accounting for model selection or other forms of regularization. Providing results that explicitly account for this regularization then allows us to accommodate and coherently account for the fact that low-dimensional models estimated in practice are often the result of specification searches. As in the high-dimensional setting, failure to account for this variable selection will invalidate usual inference procedures while the approach that we outline will remain valid and can easily be applied in conventional low-dimensional settings.

The chief goal of this overview paper is to offer a general framework that encompasses many existing results regarding inference on model parameters in high-dimensional models. The encompassing framework we present and the key theoretical results are new, though they are clearly heavily influenced and foreshadowed by previous, more specialized results. As application of the framework, we also present new results on inference in a reasonably broad class of models, termed affine-quadratic models, that includes the usual linear model and linear instrumental variables model and then apply these results to provide new results regarding post-regularization inference on the parameters on endogenous variables in a linear instrumental variables model with very many instruments and controls. We also provide a discussion of previous research that aims to highlight that many existing results fall within the general framework.

Formally, we present a series of results for obtaining valid inferential statements about a low-dimensional parameter of interest, α, in the presence of a high-dimensional nuisance parameter η. The general approach we offer relies on two fundamental elements. First, it is important that estimating equations used to draw inferences about α satisfy

a key orthogonality or immunization condition.[1] For example, when estimation and inference for α is based on the empirical analog of a theoretical system of equations

$$\mathrm{M}(\alpha, \eta) = 0,$$

we show that setting up the equations in a manner such that the orthogonality or immunization condition

$$\partial_\eta \mathrm{M}(\alpha, \eta) = 0$$

holds is an important element in providing an inferential procedure for α that remains valid when η is estimated using regularization. We note that this condition can generally be established. For example, we can apply Neyman's classic orthogonalized score in likelihood settings; see, e.g. Neyman (1959) and Neyman (1979). We also describe an extension of this classic approach to the GMM setting. In general, applying this orthogonalization will introduce additional nuisance parameters that will be treated as part of η.

The second key element of our approach is the use of high-quality, structured estimators of η. Crucially, additional structure on η is needed for informative inference to proceed, and it is thus important to use estimation strategies that leverage and perform well under the desired structure. An example of a structure that has been usefully employed in the recent literature is approximate sparsity, e.g. Belloni et al. (2012). Within this framework, η is well approximated by a sparse vector which suggests the use of a sparse estimator such as the Lasso (Frank and Friedman (1993) and Tibshirani (1996)). The Lasso estimator solves the general problem

$$\widehat{\eta}_L = \arg\min_\eta \ \ell(\text{data}, \eta) + \lambda \sum_{j=1}^{p} |\psi_j \eta_j|,$$

where $\ell(\text{data}, \eta)$ is some general loss function that depends on the data and the parameter η, λ is a penalty level, and ψ_j's are penalty loadings. The leading example is the usual linear model where $\ell(\text{data}, \eta) = \sum_{i=1}^{n}(y_i - x_i'\eta)^2$ is the usual least squares loss with y_i denoting the outcome of interest for observation i and x_i denoting predictor variables, and we provide further discussion of this example in the appendix. Other examples of $\ell(\text{data}, \eta)$ include suitable loss functions corresponding to well-known M-estimators, the negative of the log-likelihood, and GMM criterion functions. This estimator and related methods such as those in Candès and Tao (2007), Meinshausen and Yu (2009), Bickel et al. (2009), Belloni and Chernozhukov (2013), and Belloni et al. (2011) are computationally efficient and have been shown to have good estimation properties even when perfect variable selection is not feasible under approximate sparsity. These good estimation properties then translate into providing "good enough" estimates of η to result in valid inference about α when coupled with orthogonal estimating equations as discussed above. Finally, it is important to note that the general results we present do not require or leverage approximate sparsity or sparsity-based estimation strategies. We

[1]We refer to the condition as an orthogonality or immunization condition as orthogonality is a much used term and our usage differs from some other usage in defining orthogonality conditions used in econometrics.

provide this discussion here simply as an example and because the structure offers one concrete setting in which the general results we establish may be applied.

In the remainder of this paper, we present the main results. In Sections 2 and 3, we provide our general set of results that may be used to establish uniform validity of inference about low-dimensional parameters of interest in the presence of high-dimensional nuisance parameters. We provide the framework in Section 2, and then discuss how to achieve the key orthogonality condition in Section 3. In Sections 4 and 5, we provide details about establishing the necessary results for estimation quality of η within the approximately sparse framework. The analysis in Section 4 pertains to a reasonably general class of affine-quadratic models, and the analysis of Section 5 specializes this result to the case of estimating the parameters on a vector of endogenous variables in a linear instrumental variables model with very many potential control variables and very many potential instruments. The analysis in Section 5 thus extends results from Belloni et al. (2012) and Belloni, Chernozhukov and Hansen (2014). We also provide a brief simulation example and an empirical example that looks at logit demand estimation within the linear many instrument and many control setting in Section 5. We conclude with a literature review in Section 6.

Notation. We use "wp $\to$ 1" to abbreviate the phrase "with probability that converges to 1", and we use arrows $\to_{\mathrm{P}_n}$ and $\rightsquigarrow_{\mathrm{P}_n}$ to denote convergence in probability and in distribution under the sequence of probability measures $\{\mathrm{P}_n\}$. The symbol $\sim$ means "distributed as". The notation $a \lesssim b$ means that $a = O(b)$ and $a \lesssim_{\mathrm{P}_n} b$ means $a = O_{\mathrm{P}_n}(b)$. The ℓ_2 and ℓ_1 norms are denoted by $\|\cdot\|$ and $\|\cdot\|_1$, respectively; and the ℓ_0-"norm", $\|\cdot\|_0$, denotes the number of non-zero components of a vector. When applied to a matrix, $\|\cdot\|$ denotes the operator norm. We use the notation $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. Here and below, $\mathbb{E}_n[\cdot]$ abbreviates the average $n^{-1}\sum_{i=1}^n[\cdot]$ over index i. That is, $\mathbb{E}_n[f(w_i)]$ denotes $n^{-1}\sum_{i=1}^n[f(w_i)]$. In what follows, we use the m-sparse norm of a matrix Q defined as

$$\|Q\|_{\mathsf{sp}(m)} = \sup\{|b'Qb|/\|b\|^2 : \|b\|_0 \leq m, \|b\| \neq 0\}.$$

We also consider the pointwise norm of a square matrix matrix Q at a point $x \neq 0$:

$$\|Q\|_{\mathsf{pw}(x)} = |x'Qx|/\|x\|^2.$$

For a differentiable map $x \mapsto f(x)$, mapping $\mathbb{R}^d$ to $\mathbb{R}^k$, we use $\partial_{x'}f$ to abbreviate the partial derivatives $(\partial/\partial x')f$, and we correspondingly use the expression $\partial_{x'}f(x_0)$ to mean $\partial_{x'}f(x)\,|_{x=x_0}$, etc. We use x' to denote the transpose of a column vector x.

2. A Testing and Estimation Approach to Valid Post-Selection and Post-Regularization Inference

2.1. **The Setting.** We assume that estimation is based on the first n elements $(w_{i,n})_{i=1}^n$ of the *stationary* data-stream $(w_{i,n})_{i=1}^\infty$ which lives on the probability space $(\Omega, \mathcal{A}, \mathrm{P}_n)$. The data points $w_{i,n}$ take values in a measurable space $\mathcal{W}$ for each i and n. Here, P_n, the probability law or data-generating process, can change with n. We allow the law to

change with n to claim robustness or uniform validity of results with respect to perturbations of such laws. Thus the data, all parameters, estimators, and other quantities are indexed by n, but we typically suppress this dependence to simplify notation.

The target parameter value $\alpha = \alpha_0$ is assumed to solve the system of theoretical equations

$$\mathrm{M}(\alpha, \eta_0) = 0,$$

where $\mathrm{M} = (\mathrm{M}_l)_{l=1}^k$ is a measurable map from $\mathcal{A} \times \mathcal{H}$ to $\mathbb{R}^k$ and $\mathcal{A} \times \mathcal{H}$ are some convex subsets of $\mathbb{R}^d \times \mathbb{R}^p$. Here the dimension d of the target parameter $\alpha \in \mathcal{A}$ and the number of equations k are assumed to be fixed and the dimension $p = p_n$ of the nuisance parameter $\eta \in \mathcal{H}$ is allowed to be very high, potentially much larger than n. To handle the high-dimensional nuisance parameter η, we employ structured assumptions and selection or regularization methods appropriate for the structure to estimate η_0.

Given an appropriate estimator $\hat{\eta}$, we can construct an estimator $\hat{\alpha}$ as an approximate solution to the estimating equation:

$$\|\hat{\mathrm{M}}(\hat{\alpha}, \hat{\eta})\| \leq \inf_{\alpha \in \mathcal{A}} \|\hat{\mathrm{M}}(\alpha, \hat{\eta})\| + o(n^{-1/2})$$

where $\hat{\mathrm{M}} = (\hat{\mathrm{M}}_l)_{l=1}^k$ is the empirical analog of theoretical equations M, which is a measurable map from $\mathcal{W}^n \times \mathcal{A} \times \mathcal{H}$ to $\mathbb{R}^k$. We can also use $\hat{\mathrm{M}}(\alpha, \hat{\eta})$ to test hypotheses about α_0 and then invert the tests to construct confidence sets.

It is not required in the formulation above, but a typical case is when $\hat{\mathrm{M}}$ and M are formed as theoretical and empirical moment functions:

$$\mathrm{M}(\alpha, \eta) := \mathrm{E}[\psi(w_i, \alpha, \eta)], \quad \hat{\mathrm{M}}(\alpha, \eta) := \mathbb{E}_n[\psi(w_i, \alpha, \eta)],$$

where $\psi = (\psi_l)_{l=1}^k$ is a measurable map from $\mathcal{W} \times \mathcal{A} \times \mathcal{H}$ to $\mathbb{R}^k$. Of course, there are many problems that do not fall in the moment condition framework.

2.2. **Valid Inference via Testing.** A simple introduction to the inferential problem is via the testing problem where we would like to test some hypothesis about the true parameter value α_0. By inverting the test, we create a confidence set for α_0. The key condition for the validity of this confidence region is adaptivity, which can be ensured by using orthogonal estimating equations and using structured assumptions on the high-dimensional nuisance parameter.[2]

The key condition enabling us to perform valid inference on α_0 is the *adaptivity* condition:

$$\sqrt{n}(\hat{\mathrm{M}}(\alpha_0, \hat{\eta}) - \hat{\mathrm{M}}(\alpha_0, \eta_0)) \to_{\mathrm{P}_n} 0. \tag{1}$$

This condition states that using $\sqrt{n}\hat{\mathrm{M}}(\alpha_0, \hat{\eta})$ is as good as using $\sqrt{n}\hat{\mathrm{M}}(\alpha_0, \eta_0)$, at least to the first order. This condition may hold despite using estimators $\hat{\eta}$ that are not asymptotically linear and are non-regular. Verification of adaptivity may involve substantial work as illustrated below. A key requirement which often arises is the *orthogonality* or *immunization* condition:

$$\partial_{\eta'} \mathrm{M}(\alpha_0, \eta_0) = 0. \tag{2}$$

[2]We refer to Bickel (1982) for a definition of and introduction to adaptivity.

This condition states that the equations are locally insensitive to small perturbations of the nuisance parameter around the true parameter values. In several important models, this condition is equivalent to the double-robustness condition (Robins and Rotnitzky (1995)). Additional assumptions regarding the *quality of estimation* of η_0 are also needed and are highlighted below.

The adaptivity condition immediately allows us to use the statistic $\sqrt{n}\hat{\mathrm{M}}(\alpha_0, \hat{\eta})$ to perform inference. Indeed, suppose we have that

$$\Omega^{-1/2}(\alpha_0)\sqrt{n}\hat{\mathrm{M}}(\alpha_0, \eta_0) \rightsquigarrow_{\mathrm{P}_n} \mathcal{N}(0, I_k) \tag{3}$$

for some positive definite $\Omega(\alpha) = \mathrm{Var}(\sqrt{n}\hat{\mathrm{M}}(\alpha, \eta_0))$. This condition can be verified using central limit theorems for triangular arrays. Such theorems are available for both i.i.d. as well as dependent and clustered data. Suppose further that there exists $\hat{\Omega}(\alpha)$ such that

$$\hat{\Omega}^{-1/2}(\alpha_0)\Omega^{1/2}(\alpha_0) \rightarrow_{\mathrm{P}_n} I_k. \tag{4}$$

It is then immediate that the following score statistic, evaluated at $\alpha = \alpha_0$, is asymptotically normal,

$$S(\alpha) := \hat{\Omega}_n^{-1/2}(\alpha)\sqrt{n}\hat{\mathrm{M}}(\alpha, \hat{\eta}) \rightsquigarrow_{\mathrm{P}_n} \mathcal{N}(0, I_k), \tag{5}$$

and that the quadratic form of this score statistic is asymptotically χ^2-square with k degrees of freedom:

$$C(\alpha_0) = \|S(\alpha_0)\|^2 \rightsquigarrow_{\mathrm{P}_n} \chi^2(k). \tag{6}$$

The statistic given in (6) simply corresponds to a quadratic form in appropriately normalized statistics that have the desired immunization or orthogonality condition. We refer to this statistic as a "generalized $C(\alpha)$-statistic" in honor of Neyman's fundamental contributions, e.g. Neyman (1959) and Neyman (1979), because, in likelihood settings, statistic (6) reduces to Neyman's $C(\alpha)$-statistic and the generalized score $S(\alpha_0)$ given in (5) reduces to Neyman's orthogonalized score. We demonstrate these relationships in the special case of likelihood models in Section 3.1 and provide a generalization to GMM models in Section 3.2. Both of these examples serve to illustrate construction of appropriate statistics in different settings, but we note that the framework applies far more generally.

The following elementary result is an immediate consequence of the preceding discussion.

Proposition 1 (Valid Inference After Selection or Regularizaton)**.** *Consider a sequence $\{\mathbf{P}_n\}$ of sets of probability laws such that for each sequence $\{\mathrm{P}_n\} \in \{\mathbf{P}_n\}$ the adaptivity condition (1), the normality condition (3), and the variance consistency condition (4) hold. Then $\mathsf{CR}_{1-a} = \{\alpha \in \mathcal{A} : C(\alpha) \leq c(1-a)\}$, where $c(1-a)$ is the $1-a$-quantile of a $\chi^2(k)$, is a uniformly valid confidence interval for α_0 in the sense that*

$$\lim_{n\to\infty} \sup_{\mathrm{P}\in\mathbf{P}_n} |\mathrm{P}(\alpha_0 \in \mathsf{CR}_{1-a}) - (1-a)| = 0.$$

We remark here that in order to make the uniformity claim interesting we should insist that the sets of probability laws $\mathbf{P}_n$ are non-decreasing in n, i.e. $\mathbf{P}_{\bar{n}} \subseteq \mathbf{P}_n$ whenever $\bar{n} \leq n$.

Proof. For any sequence of positive constants ϵ_n approaching 0, let $\mathrm{P}_n \in \mathbf{P}_n$ be any sequence such that

$$|\mathrm{P}_n(\alpha_0 \in \mathsf{CR}_{1-a}) - (1-a)| + \epsilon_n \geq \sup_{\mathrm{P} \in \mathbf{P}_n} |\mathrm{P}(\alpha_0 \in \mathsf{CR}_{1-a}) - (1-a)|.$$

By conditions (3) and (4) we have that

$$\mathrm{P}_n(\alpha_0 \in \mathsf{CR}_{1-a}) = \mathrm{P}_n(C(\alpha_0) \leq c(1-a)) \to \mathbb{P}(\chi^2(k) \leq c(1-a)) = 1-a,$$

which implies the conclusion from the preceding display. ∎

2.3. **Valid Inference via Adaptive Estimation.** Suppose that $\mathrm{M}(\alpha_0, \eta_0) = 0$ holds for $\alpha_0 \in \mathcal{A}$. We consider an estimator $\hat{\alpha} \in \mathcal{A}$ that is an approximate minimizer of the map $\alpha \mapsto \|\hat{\mathrm{M}}(\alpha, \hat{\eta})\|$ in the sense that

$$\|\hat{\mathrm{M}}(\hat{\alpha}, \hat{\eta})\| \leq \inf_{\alpha \in \mathcal{A}} \|\hat{\mathrm{M}}(\alpha, \hat{\eta})\| + o(n^{-1/2}). \tag{7}$$

In order to analyze this estimator, we assume that the derivatives $\Gamma_1 := \partial_{\alpha'} \mathrm{M}(\alpha_0, \eta_0)$ and $\partial_{\eta'} \mathrm{M}(\alpha, \eta_0)$ exist. We assume that α_0 is interior relative to the parameter space $\mathcal{A}$; namely, for some $\ell_n \to \infty$ such that $\ell_n/\sqrt{n} \to 0$,

$$\{\alpha \in \mathbb{R}^d : \|\alpha - \alpha_0\| \leq \ell_n/\sqrt{n}\} \subset \mathcal{A}. \tag{8}$$

We also assume the following local-global identifiability condition holds: For some constant $c > 0$,

$$2\|\mathrm{M}(\alpha, \eta_0)\| \geq \|\Gamma_1(\alpha - \alpha_0)\| \wedge c \quad \forall \alpha \in \mathcal{A}, \quad \text{mineig}(\Gamma_1'\Gamma_1) \geq c. \tag{9}$$

Further, for $\Omega = \mathrm{Var}(\sqrt{n}\hat{\mathrm{M}}(\alpha_0, \eta_0))$, we suppose that the central limit theorem,

$$\Omega^{-1/2}\sqrt{n}\hat{\mathrm{M}}(\alpha_0, \eta_0) \rightsquigarrow_{\mathrm{P}_n} \mathcal{N}(0, I), \tag{10}$$

and the stability condition,

$$\|\Gamma_1'\Gamma_1\| + \|\Omega\| + \|\Omega^{-1}\| \lesssim 1, \tag{11}$$

hold.

Assume that for some sequence of positive numbers $\{r_n\}$ such that $r_n \to 0$ and $r_n n^{1/2} \to \infty$, the following stochastic equicontinuity and continuity conditions hold:

$$\sup_{\alpha \in \mathcal{A}} \frac{\|\hat{\mathrm{M}}(\alpha, \hat{\eta}) - \mathrm{M}(\alpha, \hat{\eta})\| + \|\mathrm{M}(\alpha, \hat{\eta}) - \mathrm{M}(\alpha, \eta_0)\|}{r_n + \|\hat{\mathrm{M}}(\alpha, \hat{\eta})\| + \|\mathrm{M}(\alpha, \eta_0)\|} \to_{\mathrm{P}_n} 0, \tag{12}$$

$$\sup_{\|\alpha - \alpha_0\| \leq r_n} \frac{\|\hat{\mathrm{M}}(\alpha, \hat{\eta}) - \mathrm{M}(\alpha, \hat{\eta}) - \hat{\mathrm{M}}(\alpha_0, \eta_0)\|}{n^{-1/2} + \|\hat{\mathrm{M}}(\alpha, \hat{\eta})\| + \|\mathrm{M}(\alpha, \eta_0)\|} \to_{\mathrm{P}_n} 0. \tag{13}$$

Suppose that uniformly for all $\alpha \neq \alpha_0$ such that $\|\alpha - \alpha_0\| \leq r_n \to 0$, the following conditions on the smoothness of M and the quality of the estimator $\hat{\eta}$ hold, as $n \to \infty$:

$$\begin{array}{l} \|\mathrm{M}(\alpha, \eta_0) - \mathrm{M}(\alpha_0, \eta_0) - \Gamma_1[\alpha - \alpha_0]\| \|\alpha - \alpha_0\|^{-1} \to 0, \\ \sqrt{n}\|\mathrm{M}(\alpha, \hat{\eta}) - \mathrm{M}(\alpha, \eta_0) - \partial_{\eta'}\mathrm{M}(\alpha, \eta_0)[\hat{\eta} - \eta_0]\| \to_{\mathrm{P}_n} 0, \\ \|\{\partial_{\eta'}\mathrm{M}(\alpha, \eta_0) - \partial_{\eta'}\mathrm{M}(\alpha_0, \eta_0)\}[\hat{\eta} - \eta_0]\| \|\alpha - \alpha_0\|^{-1} \to_{\mathrm{P}_n} 0. \end{array} \tag{14}$$

Finally, as before, we assume that the orthogonality condition

$$\partial_{\eta'}\mathrm{M}(\alpha_0, \eta_0) = 0 \tag{15}$$

holds.

The above conditions extend the analysis of Pakes and Pollard (1989) and Chen et al. (2003), which in turn extended Huber's (1964) classical results on Z-estimators. These conditions allow for both smooth and non-smooth systems of estimating equations. The identifiability condition imposed above is mild and holds for broad classes of identifiable models. The equicontinuity and smoothness conditions imposed above require mild smoothness on the function M as well as require that $\hat{\eta}$ is a good-quality estimator of η_0. In particular, these conditions will often require that $\hat{\eta}$ converges to η_0 at a faster rate than $n^{-1/4}$ as demonstrated, for example, in the next section. However, the rate condition alone is not sufficient for adaptivity. We also need the orthogonality condition (15). In addition, we need that $\hat{\eta} \in \mathcal{H}_n$, where $\mathcal{H}_n$ is a set whose complexity does not grow too quickly with the sample size, to verify the stochastic equicontinuity condition; see, e.g., Belloni, Chernozhukov, Fernández-Val and Hansen (2013) and Belloni, Chernozhukov and Kato (2013*b*). In the next section, we use sparsity of $\hat{\eta}$ to control this complexity. Note that conditions (12)-(13) can be simplified by only leaving r_n and $n^{-1/2}$ in the denominator, though this simplification would then require imposing compactness on $\mathcal{A}$ even in linear problems.

Proposition 2 (Valid Inference via Adaptive Estimation after Selection or Regularization)**.** *Consider a sequence $\{\mathbf{P}_n\}$ of sets of probability laws such that for each sequence $\{\mathrm{P}_n\} \in \{\mathbf{P}_n\}$ conditions (7)-(15) hold. Then*

$$\sqrt{n}(\hat{\alpha} - \alpha_0) + [\Gamma_1'\Gamma_1]^{-1}\Gamma_1'\sqrt{n}\hat{\mathrm{M}}(\alpha_0, \eta_0) \to_{\mathrm{P}_n} 0.$$

In addition, for $V_n := (\Gamma_1'\Gamma_1)^{-1}\Gamma_1'\Omega\Gamma_1(\Gamma_1'\Gamma_1)^{-1}$, we have that

$$\lim_{n\to\infty} \sup_{\mathrm{P}\in\mathbf{P}_n} \sup_{R\in\mathcal{R}} |\mathrm{P}(V_n^{-1/2}(\hat{\alpha} - \alpha_0) \in R) - \mathbb{P}(\mathcal{N}(0, I) \in R)| = 0,$$

where $\mathcal{R}$ is a collection of all convex sets. Moreover, the result continues to apply if V_n is replaced by a consistent estimator $\hat{V}_n$ such that $\hat{V}_n - V_n \to_{\mathrm{P}_n} 0$ under each sequence $\{\mathrm{P}_n\}$. Thus, $\mathsf{CR}_{1-a}^l = [l'\hat{\alpha} \pm c(1-a/2)(l'\hat{V}_n l/n)^{1/2}]$ where $c(1-a/2)$ is the $(1-a/2)$-quantile of $\mathcal{N}(0,1)$ is a uniformly valid confidence set for $l'\alpha_0$:

$$\lim_{n\to\infty} \sup_{\mathrm{P}\in\mathbf{P}_n} |\mathrm{P}(l'\alpha_0 \in \mathsf{CR}_{1-a}^l) - (1-a)| = 0.$$

Note that the above formulation implicitly accommodates weighting options. Suppose M^o and $\hat{\mathrm{M}}^o$ are the original theoretical and empirical systems of equations, and let

$\Gamma_1^o = \partial_{\alpha'} \mathrm{M}^o(\alpha_0, \eta_0)$ be the original Jacobian. We could consider $k \times k$ positive-definite weight matrices A and $\hat{\mathrm{A}}$ such that

$$\|\mathrm{A}^2\| + \|(\mathrm{A}^2)^{-1}\| \lesssim 1, \quad \|\hat{\mathrm{A}}^2 - \mathrm{A}^2\| \to_{\mathrm{P}_n} 0. \tag{16}$$

For example, we may wish to use the optimal weighting matrix $\mathrm{A}^2 = \mathrm{Var}(\sqrt{n}\hat{\mathrm{M}}^o(\alpha_0, \eta_0))^{-1}$ which can be estimated by $\hat{\mathrm{A}}^2$ obtained using a preliminary estimator $\hat{\alpha}^o$ resulting from solving the problem with some non-optimal weighting matrix such as I. We can then simply redefine the system of equations and the Jacobian according to

$$\mathrm{M}(\alpha, \eta) = \mathrm{A}\mathrm{M}^o(\alpha, \eta), \quad \hat{\mathrm{M}}(\alpha, \eta) = \hat{\mathrm{A}}\hat{\mathrm{M}}^o(\alpha, \eta), \quad \Gamma_1 = \mathrm{A}\Gamma_1^o. \tag{17}$$

Proposition 3 (Adaptive Estimation via Weighted Equations)**.** *Consider a sequence* $\{\mathbf{P}_n\}$ *of sets of probability laws such that for each sequence* $\{\mathrm{P}_n\} \in \{\mathbf{P}_n\}$ *the conditions of Proposition 2 hold for the original pair of systems of equations* $(\mathrm{M}^o, \hat{\mathrm{M}}^o)$ *and that (16) holds. Then these conditions also hold for the new pair* $(\mathrm{M}, \hat{\mathrm{M}})$ *in (17), so that all the conclusions of Proposition 2 apply to the resulting approximate argmin estimator* $\hat{\alpha}$*. In particular, if we use* $\mathrm{A}^2 = \mathrm{Var}(\sqrt{n}\hat{\mathrm{M}}^o(\alpha_0, \eta_0))^{-1}$ *and* $\hat{\mathrm{A}}^2 - \mathrm{A}^2 \to_{\mathrm{P}_n} 0$*, then the large sample variance* V_n *simplifies to* $V_n = (\Gamma_1'\Gamma_1)^{-1}$*.*

2.4. **Inference via Adaptive "One-Step" Estimation.** We next consider a "one-step" estimator. To define the estimator, we start with an initial estimator $\tilde{\alpha}$ that satisfies, for $r_n = o(n^{-1/4})$,

$$\mathrm{P}_n(\|\tilde{\alpha} - \alpha_0\| \leq r_n) \to 1. \tag{18}$$

The one-step estimator $\check{\alpha}$ then solves a linearized version of (7):

$$\check{\alpha} = \tilde{\alpha} - [\hat{\Gamma}_1'\hat{\Gamma}_1]^{-1}\hat{\Gamma}_1'\hat{\mathrm{M}}(\tilde{\alpha}, \hat{\eta}) \tag{19}$$

where $\hat{\Gamma}_1$ is an estimator of Γ_1 such that

$$\mathrm{P}_n(\|\hat{\Gamma}_1 - \Gamma_1\| \leq r_n) \to 1. \tag{20}$$

Since the one-step estimator is considerably more crude than the argmin estimator, we need to impose additional smoothness conditions. Specifically, we suppose that uniformly for all $\alpha \neq \alpha_0$ such that $\|\alpha - \alpha_0\| \leq r_n \to 0$, the following strengthened conditions on stochastic equicontinuity, smoothness of M and the quality of the estimator $\hat{\eta}$ hold, as $n \to \infty$:

$$\begin{array}{l} n^{1/2}\|\hat{\mathrm{M}}(\alpha, \hat{\eta}) - \mathrm{M}(\alpha, \hat{\eta}) - \hat{\mathrm{M}}(\alpha_0, \eta_0)\| \to_{\mathrm{P}_n} 0, \\ \|\mathrm{M}(\alpha, \eta_0) - \mathrm{M}(\alpha_0, \eta_0) - \Gamma_1[\alpha - \alpha_0]\|\|\alpha - \alpha_0\|^{-2} \lesssim 1, \\ \sqrt{n}\|\mathrm{M}(\alpha, \hat{\eta}) - \mathrm{M}(\alpha, \eta_0) - \partial_{\eta'}\mathrm{M}(\alpha, \eta_0)[\hat{\eta} - \eta_0]\| \to_{\mathrm{P}_n} 0, \\ \sqrt{n}\|\{\partial_{\eta'}\mathrm{M}(\alpha, \eta_0) - \partial_{\eta'}\mathrm{M}(\alpha_0, \eta_0)\}[\hat{\eta} - \eta_0]\| \to_{\mathrm{P}_n} 0. \end{array} \tag{21}$$

Proposition 4 (Valid Inference via Adaptive One-Step Estimators)**.** *Consider a sequence* $\{\mathbf{P}_n\}$ *of sets of probability laws such that for each sequence* $\{\mathrm{P}_n\} \in \{\mathbf{P}_n\}$ *the conditions of Proposition 2 as well as (18), (20), and (21) hold. Then the one-step estimator* $\check{\alpha}$ *defined by (19) is first order equivalent to the argmin estimator* $\hat{\alpha}$*:*

$$\sqrt{n}(\check{\alpha} - \hat{\alpha}) \to_{\mathrm{P}_n} 0.$$

Consequently, all conclusions of Proposition 2 apply to $\check{\alpha}$ in place of $\hat{\alpha}$.

The one-step estimator requires stronger regularity conditions than the argmin estimator. Moreover, there is finite-sample evidence (e.g. Belloni, Chernozhukov and Wei (2013)) that in practical problems the argmin estimator often works much better, since the one-step estimator typically suffers from higher-order biases. This problem could be alleviated somewhat by iterating on the one-step estimator, treating the previous iteration as the "crude" start $\tilde{\alpha}$ for the next iteration.

3. Achieving Orthogonality Using Neyman's Orthogonalization

Here we describe orthogonalization ideas that go back at least to Neyman (1959); see also Neyman (1979). Neyman's idea was to project the score that identifies the parameter of interest onto the ortho-complement of the tangent space for the nuisance parameter. This projection underlies semi-parametric efficiency theory, which is concerned particularly with the case where η is infinite-dimensional, cf. van der Vaart (1998). Here we consider finite-dimensional η of high dimension; for discussion of infinite-dimensional η in an approximately sparse setting, see Belloni, Chernozhukov, Fernández-Val and Hansen (2013) and Belloni, Chernozhukov and Kato (2013*b*).

3.1. **The Classical Likelihood Case.** In likelihood settings, the construction of orthogonal equations was proposed by Neyman (1959) who used them in construction of his celebrated $C(\alpha)$-statistic. The $C(\alpha)$-statistic, or the orthogonal score statistic, was first explicitly used for testing (and also for setting up estimation) in high-dimensional sparse models in Belloni, Chernozhukov and Kato (2013*b*) and Belloni, Chernozhukov and Kato (2013*a*), in the context of quantile regression, and Belloni, Chernozhukov and Wei (2013) in the context of logistic regression and other generalized linear models. More recent uses of $C(\alpha)$-statistics (or close variants) include those in Voorman et al. (2014), Ning and Liu (2014), and Yang et al. (2014).

Suppose that the (possibly conditional, possibly quasi) log-likelihood function associated to observation w_i is $\ell(w_i, \alpha, \beta)$, where $\alpha \in \mathcal{A} \subset \mathbb{R}^d$ is the target parameter and $\beta \in \mathcal{B} \subset \mathbb{R}^{p_0}$ is the nuisance parameter. Under regularity conditions, the true parameter values $\gamma_0 = (\alpha_0', \beta_0)'$ obey

$$\mathrm{E}[\partial_\alpha \ell(w_i, \alpha_0, \beta_0)] = 0, \quad \mathrm{E}[\partial_\beta \ell(w_i, \alpha_0, \beta_0)] = 0. \tag{22}$$

Now consider the moment function

$$\mathrm{M}(\alpha, \eta) = \mathrm{E}[\psi(w_i, \alpha, \eta)], \quad \psi(w_i, \alpha, \eta) = \partial_\alpha \ell(w_i, \alpha, \beta) - \mu \partial_\beta \ell(w_i, \alpha, \beta). \tag{23}$$

Here the nuisance parameter is

$$\eta = (\beta', \mathrm{vec}(\mu)')' \in \mathcal{B} \times \mathcal{D} \subset \mathbb{R}^p, \quad p = p_0 + dp_0,$$

where μ is the $d \times p_0$ *orthogonalization* parameter matrix whose true value μ_0 solves the equation:

$$J_{\alpha\beta} - \mu J_{\beta\beta} = 0 \text{ (i.e., } \mu_0 = J_{\alpha\beta} J_{\beta\beta}^{-1}), \tag{24}$$

where, for $\gamma := (\alpha', \beta')'$ and $\gamma_0 := (\alpha_0', \beta_0')'$,

$$J := \partial_{\gamma'} \mathrm{E}[\partial_\gamma \ell(w_i, \gamma)\ ]|_{\gamma=\gamma_0} \quad =: \quad \left(\begin{array}{cc} J_{\alpha\alpha} & J_{\alpha\beta} \\ J_{\beta\alpha} & J_{\beta\beta} \end{array} \right).$$

Note that μ_0 not only creates the necessary orthogonality but also creates

- the *optimal score* (in statistical language)
- or, equivalently, the *optimal instrument/moment* (in econometric language)[3]

for inference about α_0.

Provided μ_0 is well-defined, we have by (22) that

$$\mathrm{M}(\alpha_0, \eta_0) = 0.$$

Moreover, the function M has the desired orthogonality property:

$$\partial_{\eta'} \mathrm{M}(\alpha_0, \eta_0) = \Big[J_{\alpha\beta} - \mu_0 J_{\beta\beta};\ F\mathrm{E}[\partial_\beta \ell(w_i, \alpha_0, \beta_0)] \Big] = 0, \tag{25}$$

where F is a tensor operator, such that $Fx = \partial \mu x / \partial \mathrm{vec}(\mu)'\ |_{\mu=\mu_0}$ is a $d \times (dp_0)$ matrix for any vector x in $\mathbb{R}^{p_0}$. Note that the orthogonality property holds for Neyman's construction even if the likelihood is misspecified. That is, $\ell(w_i, \gamma_0)$ may be a quasi-likelihood, and the data need not be i.i.d. and may, for example, exhibit complex dependence over i.

An alternative way to define μ_0 arises by considering that, under correct specification and sufficient regularity, the information matrix equality holds and yields

$$\begin{array}{rcl} J = J^0 & := & \mathrm{E}[\partial_\gamma \ell(w_i, \gamma) \partial_\gamma \ell(w_i, \gamma)']|_{\gamma=\gamma_0} \\ & = & \left(\begin{array}{cc} \mathrm{E}[\partial_\alpha \ell(w_i, \gamma) \partial_\alpha \ell(w_i, \gamma)'] & \mathrm{E}[\partial_\alpha \ell(w_i, \gamma) \partial_\beta \ell(w_i, \gamma)'] \\ \mathrm{E}[\partial_\beta \ell(w_i, \gamma) \partial_\alpha \ell(w_i, \gamma)'] & \mathrm{E}[\partial_\beta \ell(w_i, \gamma) \partial_\beta \ell(w_i, \gamma)'] \end{array} \right) \Bigg|_{\gamma=\gamma_0}, \\ & =: & \left(\begin{array}{cc} J^0_{\alpha\alpha} & J^0_{\alpha\beta} \\ J^0_{\beta\alpha} & J^0_{\beta\beta} \end{array} \right). \end{array}$$

Hence define $\mu_0^* = J^0_{\alpha\beta} J^{0-1}_{\beta\beta}$ as the population *projection coefficient* of the score for the main parameter $\partial_\alpha \ell(w_i, \gamma_0)$ on the score for the nuisance parameter $\partial_\beta \ell(w_i, \gamma_0)$:

$$\partial_\alpha \ell(w_i, \gamma_0) = \mu_0^* \partial_\beta \ell(w_i, \gamma_0) + \varrho, \quad \mathrm{E}[\varrho \partial_\beta \ell(w_i, \gamma_0)'] = 0. \tag{26}$$

We can see this construction as the non-linear version of Frisch-Waugh's "partialling out" from the linear regression model. It is important to note that under misspecification the information matrix equality generally does not hold, and this projection approach does not provide valid orthogonalization.

Lemma 1 (Neyman's orthogonalization for (quasi-) likelihood scores)**.** *Suppose that for each $\gamma = (\alpha, \beta) \in \mathcal{A} \times \mathcal{B}$, the derivative $\partial_\gamma \ell(w_i, \gamma)$ exists and is continuous at γ with probability one, and obeys the dominance condition $\mathrm{E} \sup_{\gamma \in \mathcal{A} \times \mathcal{B}} \|\partial_\gamma \ell(w_i, \gamma)\|^2 < \infty$. Suppose that condition (22) holds for some (quasi-) true value (α_0, β_0). Then, (i) if J exists and is finite and $J_{\beta\beta}$ is invertible, then the orthogonality condition (25) holds; (ii)*

[3]The connection between optimal instruments/moments and likelihood/score has been elucidated by the fundamental work of Chamberlain (1987).

if the information matrix equality holds, namely $J = J^0$, then the orthogonality condition (25) holds for the projection parameter μ_0^* in place of the orthogonalization parameter matrix μ_0.

The claim follows immediately from the computations above.

With the formulations above Neyman's $C(\alpha)$-statistic takes the form

$$C(\alpha) = \|S(\alpha)\|_2^2, \quad S(\alpha) = \hat{\Omega}^{-1/2}(\alpha, \hat{\eta})\sqrt{n}\hat{\mathrm{M}}(\alpha, \hat{\eta}),$$

where $\hat{\mathrm{M}}(\alpha, \hat{\eta}) = \mathbb{E}_n[\psi(w_i, \alpha, \hat{\eta})]$ as before, $\Omega(\alpha, \eta_0) = \mathrm{Var}(\sqrt{n}\hat{\mathrm{M}}(\alpha, \eta_0))$, and $\hat{\Omega}(\alpha, \hat{\eta})$ and $\hat{\eta}$ are suitable estimators based on sparsity or other structured assumptions. The estimator is then

$$\hat{\alpha} = \arg\inf_{\alpha \in \mathcal{A}} C(\alpha) = \arg\inf_{\alpha \in \mathcal{A}} \|\sqrt{n}\hat{\mathrm{M}}(\alpha, \hat{\eta})\|,$$

provided that $\hat{\Omega}(\alpha, \hat{\eta})$ is positive definite for each $\alpha \in \mathcal{A}$. If the conditions of Section 2 hold, we have that

$$C(\alpha) \rightsquigarrow \chi^2(d), \quad V_n^{-1/2}\sqrt{n}(\hat{\alpha} - \alpha_0) \rightsquigarrow \mathcal{N}(0, I), \tag{27}$$

where $V_n = \Gamma_1^{-1}\Omega(\alpha_0, \eta_0)\Gamma_1^{-1}$ and $\Gamma_1 = J_{\alpha\alpha} - \mu_0 J'_{\alpha\beta}$. Under correct specification and i.i.d. sampling, the variance matrix V_n further reduces to the optimal variance

$$\Gamma_1^{-1} = (J_{\alpha\alpha} - J_{\alpha\beta}J_{\beta\beta}^{-1}J'_{\alpha\beta})^{-1},$$

of the first d components of the maximum likelihood estimator in a Gaussian shift experiment with observation $Z \sim \mathcal{N}(h, J_0^{-1})$. Likewise, the result (27) also holds for the one-step estimator $\check{\alpha}$ of Section 2 in place of $\hat{\alpha}$ as long as conditions in Section 2 hold.

Provided that sparsity or its generalizations are plausible assumptions to make regarding η_0, the formulations above naturally lend themselves to sparse estimation. For example, Belloni, Chernozhukov and Wei (2013) used penalized and post-penalized maximum likelihood to estimate β_0, and used the information matrix equality to estimate the orthogonalization parameter matrix μ_0^* by using Lasso or Post-Lasso estimation of the projection equation (26). It is also possible to estimate μ_0 directly by finding approximate sparse solutions to the empirical analog of the system of equations $J_{\alpha\beta} - \mu J_{\beta\beta} = 0$ using ℓ_1-penalized estimation, as, e.g., in van de Geer et al. (2014), or post-ℓ_1-penalized estimation.

3.2. **Achieving Orthogonality in GMM Problems.** Here we consider $\gamma_0 = (\alpha_0', \beta_0')'$ that solve the system of equations:

$$\mathrm{E}[m(w_i, \alpha_0, \beta_0)] = 0,$$

where $m : \mathcal{W} \times \mathcal{A} \times \mathcal{B} \mapsto \mathbb{R}^k$, $\mathcal{A} \times \mathcal{B}$ is a convex subset of $\mathbb{R}^d \times \mathbb{R}^{p_0}$, and $k \geq d + p_0$ is the number of moments. The orthogonal moment equation is

$$\mathrm{M}(\alpha, \eta) = \mathrm{E}[\psi(w_i, \alpha, \eta)], \quad \psi(w_i, \alpha, \eta) = \mu m(w_i, \alpha, \beta). \tag{28}$$

The nuisance parameter is

$$\eta = (\beta', \mathrm{vec}(\mu)')' \in \mathcal{B} \times \mathcal{D} \subset \mathbb{R}^p, \quad p = p_0 + dk,$$

where μ is the $d \times k$ orthogonalization parameter matrix. The "true value" of μ is

$$\mu_0 = (G_\alpha' \Omega_m^{-1} - G_\alpha' \Omega_m^{-1} G_\beta (G_\beta' \Omega_m^{-1} G_\beta)^{-1} G_\beta' \Omega_m^{-1}),$$

where, for $\gamma = (\alpha', \beta')'$ and $\gamma_0 = (\alpha_0', \beta_0')'$,

$$G_\gamma = \partial_{\gamma'} \mathrm{E}[m(w_i, \alpha, \beta)]\Big|_{\gamma=\gamma_0} = \Big[\partial_{\alpha'} \mathrm{E}[m(w_i, \alpha, \beta)], \partial_{\beta'} \mathrm{E}[m(w_i, \alpha, \beta)]\Big]\Big|_{\gamma=\gamma_0} =: \Big[G_\alpha, G_\beta\Big],$$

and

$$\Omega_m = \mathrm{Var}(\sqrt{n}\mathbb{E}_n[m(w_i, \alpha_0, \beta_0)]).$$

As before, we can interpret μ_0 as an operator creating orthogonality while building

- the *optimal instrument/moment* (in econometric language),
- or, equivalently, the *optimal score* function (in statistical language).[4]

The resulting moment function has the required orthogonality property; namely, the first derivative with respect to the nuisance parameter when evaluated at the true parameter values is zero:

$$\partial_{\eta'} \mathrm{M}(\alpha_0, \eta)|_{\eta=\eta_0} = [\mu_0 G_\beta, F\mathrm{E}[m(w_i, \alpha_0, \beta_0)]] = 0, \tag{29}$$

where F is a tensor operator, such that $Fx = \partial \mu x / \partial \mathrm{vec}(\mu)' \,|_{\mu=\mu_0}$ is a $d \times (dk)$ matrix for any vector x in $\mathbb{R}^k$.

Estimation and inference on α_0 can be based on the empirical analog of (28):

$$\hat{\mathrm{M}}(\alpha, \hat{\eta}) = \mathbb{E}_n[\psi(w_i, \alpha, \hat{\eta})],$$

where $\hat{\eta}$ is a post-selection or other regularized estimator of η_0. Note that the previous framework of (quasi)-likelihood is incorporated as a special case with

$$m(w_i, \alpha, \beta) = [\partial_\alpha \ell(w_i, \alpha)', \partial_\beta \ell(w_i, \beta)']'.$$

With the formulations above, Neyman's $C(\alpha)$-statistic takes the form:

$$C(\alpha) = \|S(\alpha)\|_2^2, \quad S(\alpha) = \hat{\Omega}^{-1/2}(\alpha, \hat{\eta}) \sqrt{n} \hat{\mathrm{M}}(\alpha, \hat{\eta}),$$

where $\hat{\mathrm{M}}(\alpha, \hat{\eta}) = \mathbb{E}_n[\psi(w_i, \alpha, \hat{\eta})]$ as before, $\Omega(\alpha, \eta_0) = \mathrm{Var}(\sqrt{n}\hat{\mathrm{M}}(\alpha, \eta_0))$, and $\hat{\Omega}(\alpha, \hat{\eta})$ and $\hat{\eta}$ are suitable estimators based on structured assumptions. The estimator is then

$$\hat{\alpha} = \arg \inf_{\alpha \in \mathcal{A}} C(\alpha) = \arg \inf_{\alpha \in \mathcal{A}} \|\sqrt{n}\hat{\mathrm{M}}(\alpha, \hat{\eta})\|,$$

provided that $\hat{\Omega}(\alpha, \hat{\eta})$ is positive definite for each $\alpha \in \mathcal{A}$. If the high-level conditions of Section 2 hold, we have that

$$C(\alpha) \rightsquigarrow_{\mathrm{P}_n} \chi^2(d), \quad V_n^{-1/2} \sqrt{n}(\hat{\alpha} - \alpha) \rightsquigarrow_{\mathrm{P}_n} \mathcal{N}(0, I), \tag{30}$$

where $V_n = (\Gamma_1')^{-1} \Omega(\alpha_0, \eta_0) (\Gamma_1)^{-1}$ coincides with the optimal variance for GMM; here $\Gamma_1 = \mu_0 G_\alpha$. Likewise, the same result (30) holds for the one-step estimator $\check{\alpha}$ of Section 2 in place of $\hat{\alpha}$ as long as the conditions in Section 2 hold. In particular, the variance V_n corresponds to the variance of the first d components of the maximum likelihood estimator in the normal shift experiment with the observation $Z \sim \mathcal{N}(h, (G_\gamma' \Omega_m^{-1} G_\gamma)^{-1})$.

[4]Cf. previous footnote.

The above is a generic outline of the properties that are expected for inference using orthogonalized GMM equations under structured assumptions. The problem of inference in GMM under sparsity is a very delicate matter due to the complex form of the orthogonalization parameters. One potential approach to the problem is outlined in Chernozhukov et al. (2014).

4. Achieving Adaptivity In Affine-Quadratic Models via Approximate Sparsity

Here we take orthogonality as given and explain how we can use approximate sparsity to achieve the adaptivity property (1).

4.1. The Affine-Quadratic Model.

We analyze the case where $\hat{\mathrm{M}}$ and M are *affine* in α and *affine-quadratic* in η. Specifically, we suppose that for all α

$$\hat{\mathrm{M}}(\alpha, \eta) = \hat{\Gamma}_1(\eta)\alpha + \hat{\Gamma}_2(\eta), \quad \mathrm{M}(\alpha, \eta) = \Gamma_1(\eta)\alpha + \Gamma_2(\eta),$$

where the orthogonality condition holds,

$$\partial_{\eta'} \mathrm{M}(\alpha_0, \eta_0) = 0,$$

and $\eta \mapsto \hat{\Gamma}_j(\eta)$ and $\eta \mapsto \Gamma_j(\eta)$ are affine-quadratic in η for $j = 1$ and $j = 2$. That is, we will have that all second-order derivatives of $\hat{\Gamma}_j(\eta)$ and $\Gamma_j(\eta)$ for $j = 1$ and $j = 2$ are constant over the convex parameter space $\mathcal{H}$ for η.

This setting is both useful, including most widely used linear models as a special case, and pedagogical, permitting simple illustration of the key issues that arise in treating the general problem. The derivations given below easily generalize to more complicated models, but we defer the details to the interested reader.

The estimator in this case is

$$\hat{\alpha} = \arg\min_{\alpha \in \mathbb{R}^d} \|\hat{\mathrm{M}}(\alpha, \hat{\eta})\|^2 = -[\hat{\Gamma}_1(\hat{\eta})'\hat{\Gamma}_1(\hat{\eta})]^{-1}\hat{\Gamma}_1(\hat{\eta})'\hat{\Gamma}_2(\hat{\eta}), \tag{31}$$

provided the inverse is well-defined. It follows that

$$\sqrt{n}(\hat{\alpha} - \alpha_0) = -[\hat{\Gamma}_1(\hat{\eta})'\hat{\Gamma}_1(\hat{\eta})]^{-1}\hat{\Gamma}_1(\hat{\eta})'\sqrt{n}\hat{\mathrm{M}}(\alpha_0, \hat{\eta}). \tag{32}$$

This estimator is adaptive if, for $\Gamma_1 := \Gamma_1(\eta_0)$,

$$\sqrt{n}(\hat{\alpha} - \alpha_0) + [\Gamma_1'\Gamma_1]^{-1}\Gamma_1'\sqrt{n}\hat{\mathrm{M}}(\alpha_0, \eta_0) \to_{\mathrm{P}_n} 0,$$

which occurs under (10) and (11) if

$$\sqrt{n}(\hat{\mathrm{M}}(\alpha_0, \hat{\eta}) - \hat{\mathrm{M}}(\alpha_0, \eta_0)) \to_{\mathrm{P}_n} 0, \quad \hat{\Gamma}_1(\hat{\eta}) - \Gamma_1(\eta_0) \to_{\mathrm{P}_n} 0. \tag{33}$$

Therefore, the problem of adaptivity of the estimator is directly connected to the problem of adaptivity of testing hypotheses about α_0.

Lemma 2 (Adaptive Testing and Estimation in Affine-Quadratic Models)**.** *Consider a sequence $\{\mathbf{P}_n\}$ of sets of probability laws such that for each sequence $\{\mathrm{P}_n\} \in \{\mathbf{P}_n\}$ condition (33), the asymptotic normality condition (10), the stability condition (11), and condition (4) hold. Then all the conditions of Propositions 1 and 2 hold. Moreover, the conclusions of Proposition 1 hold, and the conclusions of Proposition 2 hold for the estimator $\hat{\alpha}$ in (31).*

4.2. **Adaptivity for Testing via Approximate Sparsity.** Assuming the orthogonality condition holds, we follow Belloni et al. (2012) in using approximate sparsity to achieve the adaptivity property (1) for the testing problem in the affine-quadratic models.

We can expand each element $\hat{\mathrm{M}}_j$ of $\hat{\mathrm{M}} = (\hat{\mathrm{M}}_j)_{j=1}^k$ as follows:

$$\sqrt{n}(\hat{\mathrm{M}}_j(\alpha_0, \hat{\eta}) - \hat{\mathrm{M}}_j(\alpha_0, \eta_0)) = T_{1,j} + T_{2,j} + T_{3,j}, \tag{34}$$

where

$$\begin{array}{l} T_{1,j} := \sqrt{n}\partial_\eta \mathrm{M}_j(\alpha_0, \eta_0)'(\hat{\eta} - \eta_0), \\ T_{2,j} := \sqrt{n}(\partial_\eta \hat{\mathrm{M}}_j(\alpha_0, \eta_0) - \partial_\eta \mathrm{M}_j(\alpha_0, \eta_0))'(\hat{\eta} - \eta_0), \\ T_{3,j} := \sqrt{n}2^{-1}(\hat{\eta} - \eta_0)'\partial_\eta \partial_{\eta'} \hat{\mathrm{M}}_j(\alpha_0)(\hat{\eta} - \eta_0). \end{array} \tag{35}$$

The term $T_{1,j}$ vanishes precisely because of orthogonality, i.e.

$$T_{1,j} = 0.$$

However, terms $T_{2,j}$ and $T_{3,j}$ need not vanish. In order to show that they are asymptotically negligible, we need to impose further structure on the problem.

Structure 1: Exact Sparsity. We first consider the case of using an exact sparsity structure where $\|\eta_0\|_0 \leq s$ and $s = s_n \geq 1$ can depend on n. We then use estimators $\hat{\eta}$ that exploit the sparsity structure.

Suppose that the following bounds hold with probability $1 - o(1)$ under P_n:

$$\begin{array}{cc} \|\hat{\eta}\|_0 \lesssim s, & \|\eta_0\|_0 \leq s, \\ \|\hat{\eta} - \eta_0\|_2 \lesssim \sqrt{(s/n)\log(pn)}, & \|\hat{\eta} - \eta_0\|_1 \lesssim \sqrt{(s^2/n)\log(pn)}. \end{array} \tag{36}$$

These conditions are typical performance bounds which hold for many sparsity-based estimators such as Lasso, post-Lasso, and their extensions.

We suppose further that the moderate deviation bound

$$\bar{T}_{2,j} = \|\sqrt{n}(\partial_{\eta'} \hat{\mathrm{M}}_j(\alpha_0, \eta_0) - \partial_{\eta'} \mathrm{M}_j(\alpha_0, \eta_0))\|_\infty \lesssim_{\mathrm{P}_n} \sqrt{\log(pn)} \tag{37}$$

holds and that the sparse norm of the second-derivative matrix is bounded:

$$\bar{T}_{3,j} = \|\partial_\eta \partial_{\eta'} \hat{\mathrm{M}}_j(\alpha_0)\|_{\mathsf{sp}(\ell_n s)} \lesssim_{\mathrm{P}_n} 1 \tag{38}$$

where $\ell_n \to \infty$ but $\ell_n = o(\log n)$.

Following Belloni et al. (2012), we can verify condition (37) using the moderate deviation theory for self-normalized sums (e.g., Jing et al. (2003)), which allows us to avoid making highly restrictive subgaussian or gaussian tail assumptions. Likewise, following Belloni et al. (2012), we can verify the second condition using laws of large numbers

for large matrices acting on sparse vectors as in Rudelson and Vershynin (2008) and Rudelson and Zhou (2011); see Lemma 7. Indeed, condition (38) holds if

$$\|\partial_\eta \partial_{\eta'} \hat{\mathrm{M}}_j(\alpha_0) - \partial_\eta \partial_{\eta'} \mathrm{M}_j(\alpha_0)\|_{\mathsf{sp}(\ell_n s)} \to_{\mathrm{P}_n} 0, \ \|\partial_\eta \partial_{\eta'} \mathrm{M}_j(\alpha_0)\|_{\mathsf{sp}(\ell_n s)} \lesssim 1.$$

The above analysis immediately implies the following elementary result.

Lemma 3 (Elementary Adaptivity for Testing via Sparsity)**.** *Let* $\{\mathrm{P}_n\}$ *be a sequence of probability laws. Assume (i)* $\eta \mapsto \hat{\mathrm{M}}(\alpha_0, \eta)$ *and* $\eta \mapsto \mathrm{M}(\alpha_0, \eta)$ *are affine-quadratic in* η *and the orthogonality condition holds, (ii) that the conditions on sparsity and the quality of estimation (36) hold, and the sparsity index obeys*

$$s^2 \log(pn)^3/n \to 0, \tag{39}$$

(iii) that the moderate deviation bound (37) holds, and (iv) the sparse norm of the second derivatives matrix is bounded as in (38). Then the adaptivity condition (1) holds for the sequence $\{\mathrm{P}_n\}$*.*

We note that (39) requires that the true value of the nuisance parameter is sufficiently sparse, which we can relax in some special cases to the requirement $s \log(pn)^c/n \to 0$, for some constant c, by using sample-splitting techniques; see Belloni et al. (2012). However, this requirement seems unavoidable in general.

Proof. We noted that $T_{1,j} = 0$ by orthogonality. Under (36)-(37) if $s^2 \log(pn)^3/n \to 0$, then $T_{2,j}$ vanishes in probability, since by Hölder's inequality,

$$T_{2,j} \leq \bar{T}_{2,j} \|\hat{\eta} - \eta_0\|_1 \lesssim_{\mathrm{P}_n} \sqrt{s^2 \log(pn)^3/n} \to_{\mathrm{P}_n} 0.$$

Also, if $s^2 \log(pn)^2/n \to 0$, then $T_{3,j}$ vanishes in probability, since by Hölder's inequality and for sufficiently large n,

$$T_{3,j} \leq \bar{T}_{3,j} \|\hat{\eta} - \eta_0\|^2 \lesssim_{\mathrm{P}_n} \sqrt{n} s \log(pn)/n \to_{\mathrm{P}_n} 0.$$

The conclusion follows from (34). ∎

Structure 2. Approximate Sparsity. Following Belloni et al. (2012), we next consider an approximate sparsity structure. Approximate sparsity imposes that, given a constant $c > 0$, we can decompose η_0 into a sparse component η_r^m and a "small" non-sparse component η^r:

$$\begin{array}{l} \eta_0 = \eta_0^m + \eta_0^r, \ \mathrm{support}(\eta_0^m) \cap \mathrm{support}(\eta_0^r) = \emptyset, \\ \|\eta_0^m\|_0 \leq s, \ \|\eta_0^r\|_2 \leq c\sqrt{s/n}, \ \|\eta_0^r\|_1 \leq c\sqrt{s^2/n}. \end{array} \tag{40}$$

This condition allows for much more realistic and richer models than can be accommodated under exact sparsity. For example, η_0 needs not have *any* zero components at all under approximate sparsity. In Section 5, we provide an example where (40) arises from a more primitive condition that the absolute values $\{|\eta_{0j}|, j = 1, ..., p\}$, sorted in decreasing order, decay at a polynomial speed with respect to j.

Suppose that we have an estimator $\hat{\eta}$ such that with probability $1 - o(1)$ under P_n the following bounds hold:

$$\|\hat{\eta}\|_0 \lesssim s, \|\hat{\eta} - \eta_0^m\|_2 \lesssim \sqrt{(s/n)\log(pn)}, \quad \|\hat{\eta} - \eta_0^m\|_1 \lesssim \sqrt{(s^2/n)\log(pn)}. \tag{41}$$

This condition is again a standard performance bound expected to hold for sparsity-based estimators under approximate sparsity conditions; see Belloni et al. (2012). Note that by the approximate sparsity condition, we also have that, with probability $1 - o(1)$ under P_n,

$$\|\hat{\eta} - \eta_0\|_2 \lesssim \sqrt{(s/n)\log(pn)}, \quad \|\hat{\eta} - \eta_0\|_1 \lesssim \sqrt{(s^2/n)\log(pn)}. \tag{42}$$

We can employ the same moderate deviation and bounded sparse norm conditions as in the previous subsection. In addition, we require the pointwise norm of the second-derivatives matrix to be bounded. Specifically, for any deterministic vector $a \neq 0$, we require

$$\|\partial_\eta \partial_{\eta'} \hat{\mathrm{M}}_j(\alpha_0)\|_{\mathsf{pw}(a)} \lesssim_{\mathrm{P}_n} 1. \tag{43}$$

This condition can be easily verified using ordinary laws of large numbers.

Lemma 4 (Elementary Adaptivity for Testing via Approximate Sparsity)**.** *Let* $\{\mathrm{P}_n\}$ *be a sequence of probability laws. Assume (i)* $\eta \mapsto \hat{\mathrm{M}}(\alpha_0, \eta)$ *and* $\eta \mapsto \mathrm{M}(\alpha_0, \eta)$ *are affine-quadratic in* η *and the orthogonality condition holds, (ii) that the conditions on approximate sparsity (40) and the quality of estimation (41) hold, and the sparsity index obeys*

$$s^2 \log(pn)^3/n \to 0,$$

(iii) that the moderate deviation bound (37) holds, (iv) the sparse norm of the second derivatives matrix is bounded as in (38), and (v) the pointwise norm of the second derivative matrix is bounded as in (43). Then the adaptivity condition (1) holds:

$$\sqrt{n}(\hat{\mathrm{M}}(\alpha_0, \hat{\eta}) - \hat{\mathrm{M}}(\alpha_0, \eta_0)) \to_{\mathrm{P}_n} 0.$$

4.3. **Adaptivity for Estimation via Approximate Sparsity.** We work with the approximate sparsity setup and the affine-quadratic model introduced in the previous subsections.

In addition to the previous assumptions, we impose the following conditions on the components $\partial_\eta \Gamma_{1,ml}$ of $\partial_\eta \Gamma_1$, where $m = 1, ..., k$ and $l = 1, ..., d$,. First, we need the following deviation and boundedness condition: For each m and l,

$$\|\partial_\eta \hat{\Gamma}_{1,ml}(\eta_0) - \partial_\eta \Gamma_{1,ml}(\eta_0)\|_\infty \lesssim_{\mathrm{P}_n} 1, \quad \|\partial_\eta \Gamma_{1,ml}(\eta_0)\|_\infty \lesssim 1. \tag{44}$$

Second, we require the sparse and pointwise norms of the following second-derivative matrices be stochastically bounded: For each m and l,

$$\|\partial_\eta \partial_{\eta'} \hat{\Gamma}_{1,ml}\|_{\mathsf{sp}(\ell_n s)} + \|\partial_\eta \partial_{\eta'} \hat{\Gamma}_{1,ml}\|_{\mathsf{pw}(a)} \lesssim_{\mathrm{P}_n} 1, \tag{45}$$

where $a \neq 0$ is any deterministic vector. Both of these conditions are mild. They can be verified using self-normalized moderate deviation theorems and by using laws of large numbers for matrices as discussed in the previous subsection.

Lemma 5 (Elementary Adaptivity for Estimation via Approximate Sparsity)**.** *Consider a sequence* $\{\mathrm{P}_n\}$ *for which the conditions of the previous lemma hold. In addition assume that the deviation bound (44) holds and the sparse norm and pointwise norms of the second derivatives matrices are stochastically bounded as in (45). Then the adaptivity condition (33) holds for the testing and estimation problem in the affine-quadratic model.*

5. Analysis of the IV Model with Very Many Control and Instrumental Variables

Note that in the following we write $w \perp v$ to denote $\mathrm{Cov}(w, v) = 0$.

Consider the linear instrumental variable model with response variable:

$$y_i = d_i'\alpha_0 + x_i'\beta_0 + \varepsilon_i, \quad \mathrm{E}[\varepsilon_i] = 0, \quad \varepsilon_i \perp (z_i, x_i), \tag{46}$$

where y_i is the response variable, $d_i = (d_{ik})_{k=1}^{p^d}$ is a p^d-vector of endogenous variables, such that

$$\begin{array}{lll} d_{i1} = x_i'\gamma_{01} + z_i'\delta_{01} + u_{i1}, & \mathrm{E}[u_{i1}] = 0, & u_{i1} \perp (z_i, x_i), \\ \vdots & \vdots & \vdots \\ d_{ip^d} = x_i'\gamma_{0p^d} + z_i'\delta_{0p^d} + u_{ip^d}, & \mathrm{E}[u_{ip^d}] = 0, & u_{ip^d} \perp (z_i, x_i). \end{array} \tag{47}$$

Here $x_i = (x_{ij})_{j=1}^{p^x}$ is a p^x-vector of exogenous control variables, including a constant, and $z_i = (z_i)_{i=1}^{p^z}$ is a p^z-vector of instrumental variables. We will have n *i.i.d.* draws of $w_i = (y_i, d_i', x_i', z_i')'$ obeying this system of equations. We also assume that $\mathrm{Var}(w_i)$ is finite throughout so that the model is well defined.

The parameter value α_0 is our target. We allow $p^x = p_n^x \gg n$ and $p^z = p_n^z \gg n$, but we maintain that p^d is fixed in our analysis. This model includes the many instruments and small number of controls case considered by Belloni et al. (2012) as a special case, and the analysis readily accommodates the many controls and no instruments case – i.e. the linear regression model – considered by Belloni et al. (2010*a*); Belloni, Chernozhukov and Hansen (2014) and Zhang and Zhang (2014). For the latter, we simply set $p_n^z = 0$ and impose the additional condition $\varepsilon_i \perp u_i$ for $u_i = (u_{ij})_{j=1}^{p_d}$, which together with $\varepsilon_i \perp x_i$ implies that $\varepsilon_i \perp d_i$.

We may have that z_i and x_i are correlated so that z_i are valid instruments only after controlling for x_i; specifically, we let $z_i = \Pi x_i + \zeta_i$, for Π a $p_n^z \times p_n^x$ matrix and ζ_i a p_n^z-vector of unobservables with $x_i \perp \zeta_i$. Substituting this expression for z_i as a function of x_i into (46) gives a system for y_i and d_i that depends only on x_i:

$$\begin{array}{lll} y_i = x_i'\theta_0 + \rho_i^y, & \mathrm{E}[\rho_i^y] = 0, & \rho_i^y \perp x_i, \\ \\ d_{i1} = x_i'\vartheta_{01} + \rho_{i1}^d, & \mathrm{E}[\rho_{i1}^d] = 0, & \rho_{i1}^d \perp x_i, \\ \vdots & \vdots & \vdots \\ d_{ip^d} = x_i'\vartheta_{0p^d} + \rho_{ip^d}^d, & \mathrm{E}[\rho_{ip^d}^d] = 0, & \rho_{ip^d}^d \perp x_i. \end{array} \tag{48}$$

Because the dimension $p = p_n$ of

$$\eta_0 = (\theta_0', (\vartheta_{0k}', \gamma_{0k}', \delta_{0k}')_{k=1}^{p^d})'$$

may be larger than n, informative estimation and inference about α_0 is impossible without imposing restrictions on η_0.

In order to state our assumptions, we fix a collection of positive constants $(\mathsf{a}, \mathsf{A}, \mathsf{c}, \mathsf{C})$, where $\mathsf{a} > 1$, and a sequence of constants $\delta_n \searrow 0$ and $\ell_n \nearrow \infty$. These constants will not vary with P, but rather we will work with collections of P defined by these constants.

CONDITION AS.1 *We assume that η_0 is approximately sparse, namely that the decreasing rearrangement $(|\eta_0|_j^*)_{j=1}^p$ of absolute values of coefficients $(|\eta_{0j}|)_{j=1}^p$ obeys*

$$|\eta_0|_j^* \leq \mathsf{A} j^{-\mathsf{a}}, \quad \mathsf{a} > 1, \quad j = 1, ..., p. \tag{49}$$

Given this assumption we can decompose η_0 into a sparse component η_0^m and small non-sparse component η_0^r:

$$\begin{gathered} \eta_0 = \eta_0^m + \eta_0^r, \ \text{support}(\eta_0^m) \cap \text{support}(\eta_0^r) = \emptyset, \\ \|\eta_0^m\|_0 \leq s, \ \|\eta_0^r\|_2 \leq c\sqrt{s/n}, \ \|\eta_0^r\|_1 \leq c\sqrt{s^2/n}, \\ s = cn^{\frac{1}{2\mathsf{a}}}, \end{gathered} \tag{50}$$

where the constant c depends only on (a, A).

CONDITION AS.2 *We assume that*

$$s^2 \log(pn)^3/n \leq o(1). \tag{51}$$

We shall perform inference on α_0 using the empirical analog of theoretical equations:

$$\mathrm{M}(\alpha_0, \eta_0) = 0, \quad \mathrm{M}(\alpha, \eta) := \mathrm{E}\left[\psi(w_i, \alpha, \eta)\right], \tag{52}$$

where $\psi = (\psi_k)_{k=1}^{p^d}$ is defined by

$$\psi_k(w_i, \alpha, \eta) := \left(y_i - x_i'\theta - \sum_{\bar{k}=1}^{p^d} (d_{i\bar{k}} - x_i'\vartheta_{\bar{k}})\alpha_{\bar{k}} \right) (x_i'\gamma_k + z_i'\delta_k - x_i'\vartheta_k).$$

We can verify that the following orthogonality condition holds:

$$\partial_{\eta'} \mathrm{M}(\alpha_0, \eta)\Big|_{\eta=\eta_0} = 0. \tag{53}$$

This means that missing the true value η_0 by a small amount does not invalidate the moment condition. Therefore, the moment condition will be relatively insensitive to non-regular estimation of η_0.

We denote the empirical analog of (52) as

$$\hat{\mathrm{M}}(\alpha, \hat{\eta}) = 0, \quad \hat{\mathrm{M}}(\alpha, \eta) := \mathbb{E}_n\left[\psi_i(\alpha, \eta)\right]. \tag{54}$$

Inference based on this condition can be shown to be immunized against small selection mistakes by virtue of orthogonality.

The above formulation is a special case of the linear-affine model. Indeed, here we have

$$\begin{gathered} \mathrm{M}(\alpha, \eta) = \Gamma_1(\eta)\alpha + \Gamma_2(\eta), \quad \hat{\mathrm{M}}(\alpha, \eta) = \hat{\Gamma}_1(\eta)\alpha + \hat{\Gamma}_2(\eta), \\ \Gamma_1(\eta) = \mathrm{E}[\psi^a(w_i, \eta)], \quad \hat{\Gamma}_1(\eta) = \mathbb{E}_n[\psi^a(w_i, \eta)], \\ \Gamma_2(\eta) = \mathrm{E}[\psi^b(w_i, \eta)], \quad \hat{\Gamma}_2(\eta) = \mathbb{E}_n[\psi^b(w_i, \eta)], \end{gathered}$$

where

$$\psi^a_{k,\bar{k}}(w_i,\eta) = -(d_{i\bar{k}} - x_i'\vartheta_{\bar{k}})(x_i'\gamma_k + z_i'\delta_k - x_i'\vartheta_k),$$
$$\psi^b_k(w_i,\eta) = (y_i - x_i'\theta)(x_i'\gamma_k + z_i'\delta_k - x_i'\vartheta_k).$$

Consequently we can use the results of the previous section. In order to do so we need to provide a suitable estimator for η_0. Here we use the Lasso and Post-Lasso estimators, as defined in Belloni et al. (2012), to deal with non-normal errors and heteroscedasticity.

Algorithm 1 (Estimation of η_0)**.** *(1) For each k, do Lasso or Post-Lasso Regression of d_{ik} on x_i, z_i to obtain $\hat{\gamma}_k$ and $\hat{\delta}_k$. (2) Do Lasso or Post-Lasso Regression of y_i on x_i to get $\hat{\theta}$. (3) Do Lasso or Post-Lasso Regression of $\hat{d}_{ik} = x_i'\hat{\gamma}_k + z_i'\hat{\delta}_k$ on x_i to get $\hat{\vartheta}_k$. The estimator of η_0 is given by $\hat{\eta} = (\hat{\theta}', (\hat{\vartheta}_k', \hat{\gamma}_{0k}', \hat{\delta}_k')_{k=1}^{p^d})'$.*

We then use

$$\hat{\Omega}(\alpha,\hat{\eta}) = \mathbb{E}_n[\psi(w_i,\alpha,\hat{\eta})\psi(w_i,\alpha,\hat{\eta})'].$$

to estimate the variance matrix $\Omega(\alpha,\eta_0) = \mathbb{E}_n[\psi(w_i,\alpha,\eta_0)\psi(w_i,\alpha,\eta_0)']$. We formulate the orthogonal score statistic and the $C(\alpha)$-statistic,

$$S(\alpha) := \hat{\Omega}_n^{-1/2}(\alpha,\hat{\eta})\sqrt{n}\hat{\mathrm{M}}(\alpha,\hat{\eta}), \quad C(\alpha) = \|S(\alpha)\|^2, \tag{55}$$

as well as our estimator $\hat{\alpha}$:

$$\hat{\alpha} = \arg\min_{\alpha\in\mathcal{A}} \|\sqrt{n}\hat{\mathrm{M}}(\alpha,\hat{\eta})\|^2.$$

Note also that $\hat{\alpha} = \arg\min_{\alpha\in\mathcal{A}} C(\alpha)$ under mild conditions, since we work with "exactly identified" systems of equations. We also need to specify a variance estimator $\hat{V}_n$ for the large sample variance V_n of $\hat{\alpha}$. We set $\hat{V}_n = (\hat{\Gamma}_1(\hat{\eta})')^{-1}\hat{\Omega}(\hat{\alpha},\hat{\eta})(\hat{\Gamma}_1(\hat{\eta}))^{-1}$.

To estimate the nuisance parameter we impose the following condition. Let $f_i := (f_{ij})_{j=1}^{p_f} := (x_i', z_i')'$; $h_i := (h_{il})_{l=1}^{p_h} := (y_i, d_i', \bar{d}_i')'$ where $\bar{d}_i = (\bar{d}_{ik})_{k=1}^{p^d}$ and $\bar{d}_{ik} := x_i'\gamma_{0k} + z_i'\delta_{0k}$; $v_i = (v_{il})_{l=1}^{p_h} := (\varepsilon_i, \rho_i^y, \rho_i^{d\prime}, \varrho_i')'$ where $\varrho_i = (\varrho_{ik})_{k=1}^{p^d}$ and $\varrho_{ik} := d_{ik} - \bar{d}_{ik}$. Let $\tilde{h}_i := h_i - \mathrm{E}[h_i]$.

Condition RF. *(i) The eigenvalues of $\mathrm{E}[f_i f_i']$ are bounded from above by C and from below by c. For all j and l, (ii) $\mathrm{E}[h_{il}^2] + \mathrm{E}[|f_{ij}^2\tilde{h}_{il}^2|] + 1/\mathrm{E}[f_{ij}^2 v_{il}^2] \le \mathsf{C}$ and $\mathrm{E}[|f_{ij}^2 v_{il}^2|] \le \mathrm{E}[|f_{ij}^2\tilde{h}_{il}^2|]$, (iii) $\mathrm{E}[|f_{ij}^3 v_{il}^3|]^2 \log^3(pn)/n \le \delta_n$, and (iv) $s\log(pn)/n \le \delta_n$. With probability no less than $1-\delta_n$, we have that (v) $\max_{i\le n,j} f_{ij}^2[s^2\log(pn)]/n \le \delta_n$ and $\max_{l,j} |(\mathbb{E}_n - \mathrm{E})[f_{ij}^2 v_{il}^2]| + |(\mathbb{E}_n - \mathrm{E})[f_{ij}^2\tilde{h}_{il}^2]| \le \delta_n$ and (vi) $\|\mathbb{E}_n[f_i f_i'] - \mathrm{E}[f_i f_i']\|_{\mathsf{sp}(\ell_n s)} \le \delta_n$.*

The conditions are motivated by those given in Belloni et al. (2012). The current conditions are made slightly stronger to account for the fact that we use zero covariance conditions in formulating the moments. Some conditions could be easily relaxed at a cost of more complicated exposition.

To estimate the variance matrix and establish asymptotic normality, we also need the following condition. Let $\mathsf{q} > 4$ be a fixed constant.

Condition SM. *For each l and k, (i) $\mathrm{E}[|h_{il}|^{\mathsf{q}}] + \mathrm{E}[|v_{il}|^{\mathsf{q}}] \le \mathsf{C}$, (ii) $\mathsf{c} \le \mathrm{E}[\varepsilon_i^2 \mid x_i, z_i] \le \mathsf{C}$, $\mathsf{c} < \mathrm{E}[\varrho_{ik}^2 \mid x_i, z_i] \le \mathsf{C}$ a.s., (iii) $\sup_{\alpha\in\mathcal{A}} \|\alpha\|_2 \le \mathsf{C}$.*

Under the conditions set forth above, we have the following result on validity of post-selection and post-regularization inference using the $C(\alpha)$-statistic and estimators derived from it.

Proposition 5 (Valid Inference in Large Linear Models using $C(\alpha)$-statistics)**.** *Let* $\mathbf{P}_n$ *be the collection of all* P *such that Conditions AS.1-2, SM, and RF hold for the given* n*. Then uniformly in* $\mathrm{P} \in \mathbf{P}_n$*,* $S(\alpha_0) \rightsquigarrow \mathcal{N}(0, I)$*, and* $C(\alpha_0) \rightsquigarrow \chi^2(p^d)$*. As a consequence, the confidence set* $\mathsf{CR}_{1-a} = \{\alpha \in \mathcal{A} : C(\alpha) \leq c(1-a)\}$*, where* $c(1-a)$ *is the* $1-a$*-quantile of a* $\chi^2(p^d)$ *is uniformly valid for* α_0*, in the sense that*

$$\lim_{n\to\infty} \sup_{\mathrm{P}\in\mathbf{P}_n} |\mathrm{P}(\alpha_0 \in \mathsf{CR}_{1-a}) - (1-a)| = 0.$$

Furthermore, for $V_n = (\Gamma_1')^{-1}\Omega(\Gamma_1)^{-1}$*, we have that*

$$\lim_{n\to\infty} \sup_{\mathrm{P}\in\mathbf{P}_n} \sup_{R\in\mathcal{R}} |\mathrm{P}(V_n^{-1/2}(\hat{\alpha} - \alpha_0) \in R) - \mathbb{P}(\mathcal{N}(0, I) \in R)| = 0,$$

where $\mathcal{R}$ *is the collection of all convex sets. Moreover, the result continues to apply if* V_n *is replaced by* $\hat{V}_n$*. Thus,* $\mathsf{CR}^l_{1-a} = [l'\hat{\alpha} \pm c(1-a/2)(l'\hat{V}_n l/n)^{1/2}]$*, where* $c(1-a/2)$ *is the* $(1-a/2)$*-quantile of a* $\mathcal{N}(0,1)$*, provides a uniformly valid confidence set for* $l'\alpha_0$*:*

$$\lim_{n\to\infty} \sup_{\mathrm{P}\in\mathbf{P}_n} |\mathrm{P}(l'\alpha_0 \in \mathsf{CR}^l_{1-a}) - (1-a)| = 0.$$

5.1. **Simulation Illustration.** In this section, we provide results from a small Monte Carlo simulation to illustrate performance of the estimator resulting from applying Algorithm 1 in a small sample setting. As comparison, we report results from two commonly used "unprincipled" alternatives for which uniformly valid inference over the class of approximately sparse models does not hold. Simulation parameters were chosen so that approximate sparsity holds but exact sparsity is violated in such a way that we expected the unprincipled procedures to perform poorly.

For our simulation, we generate data as n iid draws from the model

$$\begin{array}{l} y_i = \alpha d_i + x_i'\beta + 2\varepsilon_i \\ d_i = x_i'\gamma + z_i'\delta + u_i \\ z_i = \Pi x_i + .125\zeta_i \end{array} \quad \left| \quad \begin{pmatrix} \varepsilon_i \\ u_i \\ \zeta_i \\ x_i \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} 1 & .6 & 0 & 0 \\ .6 & 1 & 0 & 0 \\ 0 & 0 & I_{p_n^z} & 0 \\ 0 & 0 & 0 & \Sigma \end{pmatrix}\right), \right.$$

where Σ is a $p_n^x \times p_n^x$ matrix with $\Sigma_{kj} = (0.5)^{|j-k|}$ and $I_{p_n^z}$ is a $p_n^z \times p_n^z$ identity matrix. We set the number of potential controls variables (p_n^x) to 200, the number of instruments (p_n^z) to 150, and the number of observations (n) to 200. For model coefficients, we set $\alpha = 0$, $\beta = \gamma$ as p_n^x-vectors with entries $\beta_j = \gamma_j = 1/(9\nu_j)$, $\nu_j = 4/9 + \sum_{j=5}^{p_n^x} 1/j^2$ for $j \leq 4$ and $\beta_j = \gamma_j = 1/(j^2\nu_j)$ for $j > 4$, δ as a p_n^z-vector with entries $\delta_j = \frac{3}{j^2}$, and $\Pi = [I_{p_n^z}\ ,\ 0_{p_n^z \times (p_n^x - p_n^z)}]$. We report results based on 1000 simulation replications.

We provide results for four different estimators - an infeasible Oracle estimator that knows the nuisance parameters η (Oracle), two naive estimators, and the proposed "Double-Selection" estimator. The results for the proposed "Double-Selection" procedure are obtained following Algorithm 1 using Post-Lasso at every step. To obtain the Oracle results, we run standard IV regression of $y_i - \mathrm{E}[y_i|x_i]$ on $d_i - \mathrm{E}[d_i|x_i]$ using the

single instrument $\zeta_i'\delta$. The expected values are obtained from the model above and $\zeta_i'\delta$ provides the information in the instruments that is unrelated to the controls.

The two naive alternatives offer unprincipled, though potentially intuitive alternatives. The first naive estimator follows Algorithm 1 but replaces Lasso/Post-Lasso with stepwise regression with p-value for entry of .05 and p-value for removal of .10 (Stepwise). The second naive estimator (Non-orthogonal) corresponds to using a moment condition which does not satisfy the orthogonality condition described previously but will produce valid inference when perfect model selection in the regression of d on x and z is possible or perfect model selection in the regression of y on x is possible and an instrument is selected in the d on x and z regression.[5]

All of the Lasso and Post-Lasso estimates are obtained using the data-dependent penalty level from Belloni and Chernozhukov (2013). This penalty level depends on a standard deviation that is estimated adapting the iterative algorithm described in Belloni et al. (2012) Appendix A using Post-Lasso at each iteration. For inference in all cases, we use standard t-tests based on conventional homoscedastic IV standard errors obtained from the final IV step performed in each strategy.

We display the simulation results in Figure 5.1, and we report the median bias (Bias), median absolute deviation (MAD), and size of 5% level tests (Size) for each procedure in Table 1. For each estimator, we plot the simulation estimate of the sampling distribution of the estimator centered around the true parameter and scaled by the estimated standard error. With this standardization, usual asymptotic approximations would suggest that these curves should line up with a $\mathcal{N}(0,1)$ density function which is displayed as the bold solid line in the figure. We can see that the Oracle estimator and the Double-Selection estimator are centered correctly and line up reasonably well with the $\mathcal{N}(0,1)$, though both estimators exhibit some mild skewness. It is interesting that the sampling distributions of the Oracle and Double-Selection estimators are very similar as predicted by the theory. In contrast, both of the naive estimators are centered far from zero, and it is clear that the asymptotic approximation provides a very poor guide to the finite sample distribution of these estimators in the design considered.

The poor inferential performance of the two naive estimators is driven by different phenomena. The unprincipled use of stepwise regression fails to control spurious inclusion of irrelevant variables which leads to inclusion of many essentially irrelevant variables, resulting in many-instrument-type problems (e.g. Chao et al. (2012)). In addition, the spuriously included variables are those most highly correlated to the noise within sample which adds an additional type of "endogeneity bias". The failure of the

[5]Specifically, for the second naive alternative (Non-orthogonal), we first do Lasso regression of d on x and z to obtain Lasso estimates of the coefficients γ and δ. Denote these estimates as $\hat{\gamma}_L$ and $\hat{\delta}_L$, and denote the indices of the coefficients estimated to be non-zero as $\hat{I}_x^d = \{j : \hat{\gamma}_{Lj} \neq 0\}$ and $\hat{I}_z^d = \{j : \hat{\delta}_{Lj} \neq 0\}$. We then run Lasso regression of y on x to learn the identities of controls that predict the outcome. We denote the Lasso estimates as $\hat{\theta}_L$ and keep track of the indices of the coefficients estimated to be non-zero as $\hat{I}_x^y = \{j : \hat{\theta}_{Lj} \neq 0\}$. We then take the union of the controls selected in either step $\hat{I}_x = \hat{I}_x^y \cup \hat{I}_x^d$. The estimator of α is then obtained as the usual 2SLS estimator of y_i on d_i using all selected elements from x_i, x_{ij} such that $j \in \hat{I}_x$, as controls and the selected elements from z_i, z_{ij} such that $j \in \hat{I}_z^d$, as instruments.

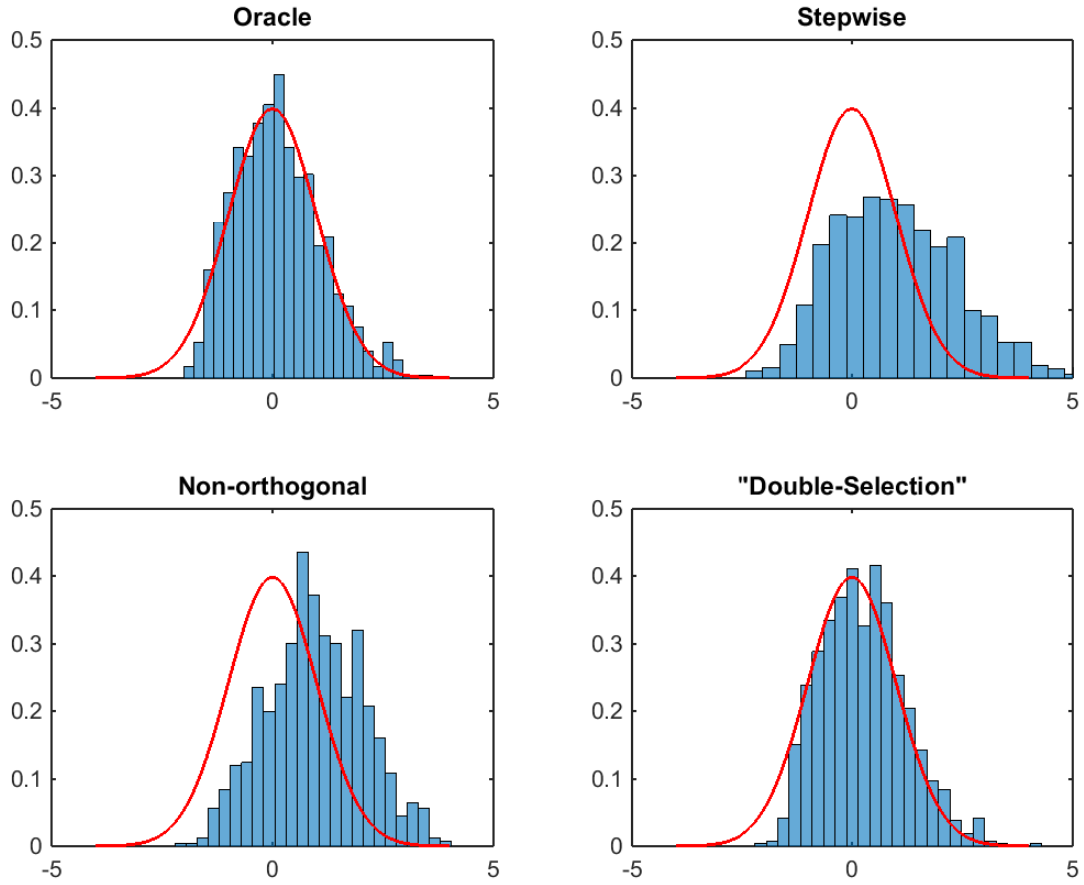

FIGURE 1. The figure presents the histogram of the estimator from each method centered around the true parameters and scaled by the estimated standard error from the simulation experiment. The red curve is the pdf of a standard normal which will correspond to the sampling distribution of the estimator under the asymptotic approximation. Each panel is labeled with the corresponding estimator from the simulation.

"Non-orthogonal" method is driven by the fact that perfect model selection is not possible within the present design: Here we have model selection mistakes in which the control variables that are correlated to the instruments but only moderately correlated to the outcome and endogenous variable are missed. Such exclusions result in standard omitted variables bias in the estimator for the parameter of interest and substantial size distortions. The additional step in the Double-Selection procedure can be viewed as a way to guard against such mistakes. Overall, the results illustrate the uniformity claims made in the preceding section. The feasible Double-Selection procedure following from Algorithm 1 performs similarly to the semi-parametrically efficient infeasible Oracle. We obtain good inferential properties with the asymptotic approximation providing a fairly good guide to the behavior of the estimator despite working in a setting where perfect

TABLE 1. Summary of Simulation Results for the Estimation of α

Method	Bias	MAD	Size
Oracle	0.015	0.247	0.043
Stepwise	0.282	0.368	0.261
Non-orthogonal	0.084	0.112	0.189
Double-Selection	0.069	0.243	0.053

This table summarizes the simulation results from a linear IV model with many instruments and controls. Estimators include an infeasible oracle as a benchmark (Oracle), two naive alternatives (Stepwise and Non-orthogonal) described in the text, and our proposed feasible valid procedure (Double-Selection). Median bias (Bias), median absolute deviation (MAD), and size for 5% level tests (Size) are reported.

model selection is impossible. While simply illustrative of the theory, the results are reassuring and in line with extensive simulations in the linear model with many controls provided in Belloni, Chernozhukov and Hansen (2014), in the instrumental variables model with many instruments and a small number of controls provided in Belloni et al. (2012), and in linear panel data models provided in Belloni, Chernozhukov, Hansen and Kozbur (2014).

5.2. **Empirical Illustration: Logit Demand Estimation.** As further illustration of the approach, we provide a brief empirical example where we estimate the coefficients in a simple logit model of demand for automobiles using market share data. Our example is based on the data and most basic strategy from Berry et al. (1995). Specifically, we estimate the parameters from the model

$$\log(s_{it}) - \log(s_{0t}) = \alpha_0 p_{it} + x_{it}'\beta_0 + \varepsilon_{it},$$
$$p_{it} = z_{it}'\delta_0 + x_{it}'\gamma_0 + u_{it},$$

where s_{it} is the market share of product i in market t with product 0 denoting the outside option, p_{it} is price and treated as endogenous, x_{it} are observed included product characteristics, and z_{it} are instruments. One could also adapt the proposed variable selection procedures to extensions of this model such as the nested logit model or models allowing for random coefficients; see, e.g., Gillen et al. (2014) for an example with a random coefficient.

In our example, we use the same set of product characteristics (x-variables) as used in obtaining the basic results in Berry et al. (1995). Specifically, we use five variables in x_{it}: a constant, an air conditioning dummy, horsepower divided by weight, miles per dollar, and vehicle size. We refer to these five variables as the baseline set of controls.

We also adopt the argument from Berry et al. (1995) to form our potential instruments. Berry et al. (1995) argue that that characteristics of other products will satisfy an exclusion restriction, $\mathrm{E}[\varepsilon_{it}|x_{j\tau}] = 0$ for any τ and $j \neq i$, and thus that any function of characteristics of other products may be used as instrument for price. This condition leaves a very high-dimensional set of potential instruments as any combination of functions of $\{x_{j\tau}\}_{j\neq i,\tau\geq 1}$ may be used to instrument for p_{it}. To reduce the dimensionality,

Berry et al. (1995) use intuition and an exchangeability argument to motivate consideration of a small number of these potential instruments formed by taking sums of product characteristics formed by summing over products excluding product i. Specifically, we form baseline instruments by taking

$$z_{k,it} = \left(\sum_{r \neq i, r \in \mathcal{I}_f} x_{k,rt}, \sum_{r \neq i, r \notin \mathcal{I}_f} x_{k,rt} \right)$$

where $x_{k,it}$ is the k^{th} element of vector x_{it} and $\mathcal{I}_f$ denotes the set of products produced by firm f. This choice yields a vector z_{it} consisting of 10 instruments. We refer to this set of instruments as the baseline instruments.

While the choice of the baseline instruments and controls is motivated by good intuition and economic theory, it should be noted that theory does not clearly state which product characteristics or instruments should be used in the model. Theory also fails to indicate the functional form with which any such variables should enter the model. The high-dimensional methods outlined in this paper offer one strategy to help address these concerns which complements the economic intuition motivating the baseline controls and instruments. As an illustration, we consider an expanded set of controls and instruments. We augment the set of potential controls with all first order interactions of the baseline variables, quadratics and cubics in all continuous baseline variables, and a time trend which yields a total of 24 x-variables. We refer to these as the augmented controls. We then take sums of these characteristics as potential instruments following the original strategy which yields 48 potential instruments.

We report estimation results in Table 2. We report results obtained by applying the method outlined in Algorithm 1 using just the baseline set of five product characterstics and 10 instruments in the row labeled "Baseline 2SLS with Selection" and results obtained by applying the method to the augmented set of 24 controls and 48 instruments in the row labeled "Augmented 2SLS with Selection." In each case, we apply the method outlined in Algorithm 1 using post-Lasso in each step and forcing the intercept to be included in all models. We employ the heteroscedasticity robust version of Post-Lasso of Belloni et al. (2012) following the implementation algorithm provided in Appendix A of Belloni et al. (2012). For comparison, we also report OLS and 2SLS estimates using only the baseline variables in "Baseline OLS" and "Baseline 2SLS," respectively; and we report OLS and 2SLS estimates using the augmented variable set in "Augmented OLS" and "Augmented 2SLS," respectively. All standard errors are conventional heteroscedasticity robust standard errors.

Considering first estimates of the price coefficient, we see that the estimated price coefficient increases in magnitude as we move from OLS to 2SLS and then to the selection based results. After selection using only the original variables, we estimate the price coefficient to be -.185 with an estimated standard error of .014 compared to an OLS estimate of -.089 with estimated standard error of .004 and 2SLS estimate of -.142 with estimated standard error of .012. In this case, all five controls are selected in the log-share on controls regression, all five controls but only four instruments are selected in the price on controls and instruments regression, and four of the controls are selected for the

TABLE 2. Estimates of Price Coefficient

	Price Coefficient	Standard Error	Number Inelastic
	Estimates Without Selection		
Baseline OLS	-0.089	0.004	1502
Baseline 2SLS	-0.142	0.012	670
Augmented OLS	-0.099	0.005	1405
Augmented 2SLS	-0.127	0.014	874
	2SLS Estimates With "Double Selection"		
Baseline 2SLS Selection	-0.185	0.014	139
Augmented 2SLS Selection	-0.221	0.015	12

This table reports estimates of the coefficient on price ("Price Coefficient") along with the estimated standard error ("Standard Error") obtained using different sets of controls and instruments. The rows "Baseline OLS" and "Baseline 2SLS" respectively provide OLS and 2SLS results using the baseline set of variables (5 controls and 10 instruments) described in the text. The rows "Augmented OLS," "Augemented 2SLS "are defined similarly but use the augmented set of variables described in the text (24 controls and 48 instruments). The rows "Baseline 2SLS with Selection" and "Augmented 2SLS with Selection" applies the "double selection" approach developed in this paper to select a set of controls and instruments and perform valid post-selection inference about the estimated price coefficient where selection occurs considering only the baseline variables. For each procedure, we also report the point estimate of the number of products for which demand is estimated to be inelastic in the column "Number Inelastic."

price on controls relationship. The difference between the baseline results is thus largely driven by the difference in instrument sets. The change in the estimated coefficient is consistent with the wisdom from the many-instrument literature that inclusion of irrelevant instruments biases 2SLS toward OLS.

With the larger set of variables, our post-model-selection estimator of the price coefficient is -.221 with an estimated standard error .015 compared to OLS estimate of -.099 with estimated standard error of .005 and 2SLS estimate of -.127 with estimated standard error of .014. Here, we see some evidence that the original set of controls may have been overly parsimonious as we select some terms that were not included in the baseline variable set. We also see a closer agreement between the OLS estimate and 2SLS estimate without selection which is likely driven by the larger number of instruments considered and the usual bias towards OLS seen in 2SLS with many weak or irrelevant instruments. In the log-share on controls regression, we have that eight control variables are selected; and we have seven controls and only four instruments selected in the price on controls and instrument regression. We also have that 13 variables are selected for the price on controls relationship. The selection of these additional variables suggests that there is important nonlinearity missed by the baseline set of variables.

The most interesting feature of the results is that estimates of own-price elasticities become more plausible as we move from the baseline results to the results based on variable selection with a large number of controls. Recall that facing inelastic demand is inconsistent with profit maximizing price choice within the present context, so theory would predict that demand should be elastic for all products. However, the baseline point estimates imply inelastic demand for 670 products. When we use the larger set

of instruments without selection, the number of products for which we estimate inelastic demand increases to 874 with the increase being generated by the 2SLS coefficient estimate moving back towards the OLS estimate. Using the variable selection results provides results closer to the theoretical prediction. The point estimates based on selection from only the baseline variables imply inelastic demand for 139 products, and we estimate inelastic demand for only 12 products using the results based on selection from the larger set of variables. Thus, the new methods provide the most reasonable estimates of own-price elasticities.

We conclude by noting that the simple specification above suffers from the usual drawbacks of the logit demand model. However, the example illustrates how the application of the methods we have outlined may be used in estimation of structural parameters in economics and add to the plausibility of the resulting estimates. In this example, we see that we get more sensible estimates of key parameters with at most a modest cost in increased estimation uncertainty after applying the methods in this paper while considering a flexible set of variables.

6. Overview of Related Literature

Inference following model selection or regularization more generally has been an active area of research in econometrics and statistics for the last several years. In this section, we provide a brief overview of this literature highlighting some key developments. This review is necessarily selective due to the large number of papers available and the rapid pace at which new papers are appearing. We choose to focus on papers that deal specifically with high-dimensional nuisance parameter settings, and note that the ideas in these papers apply in low dimensional settings as well.

Early work on inference in high-dimensional settings focused on inference based on the so-called oracle property; see, e.g., Fan and Li (2001) for an early paper, Fan and Lv (2010) for a more recent review, and Bühlmann and van de Geer (2011) for a textbook treatment. A consequence of the oracle property is that model selection does not impact the asymptotic distribution of the parameters estimated in the selected model. This feature allows one to do inference using standard approximate distributions for the parameters of the selected model ignoring that model selection was done. While convenient and fruitful in many applications (e.g. signal processing), such results effectively rely on strong conditions that imply that one will be able to perfectly select the correct model. For example, such results in linear models require the so called "beta-min condition" (Bühlmann and van de Geer (2011)) that all but a small number of coefficients are exactly zero and the remaining non-zero coefficients are bounded away from zero, effectively ruling out variables that have small, non-zero coefficients. Such conditions seem implausible in many applications, especially in econometrics, and relying on such conditions produces asymptotic approximations that may provide very poor approximations to finite-sample distributions of estimators as they are not uniformly valid over sequences of models that include even minor deviations from conditions implying perfect model selection. The concern about the lack of uniform validity of inference based on oracle properties was raised in a series of papers, including Leeb and Pötscher (2008*a*)

and Leeb and Pötscher (2008*b*) among many others, and the more recent work on post-model-selection inference has been focused on offering procedures that provide uniformly valid inference over interesting (large) classes of models that include cases where perfect model selection will not be possible.

To our knowledge, the first work to formally and expressly address the problem of obtaining uniformly valid inference following model selection is Belloni et al. (ArXiv, 2010*b*) which considered inference about parameters on a low-dimensional set of endogenous variables following selection of instruments from among a high-dimensional set of potential instruments in a homoscedastic, Gaussian instrumental variables (IV) model. The approach does not rely on implausible "beta-min" conditions which imply perfect model selection but instead relies on the fact that the moment condition underlying IV estimation satisfies the *orthogonality condition* (2) and the use of high-quality variable selection methods. These ideas were further developed in the context of providing uniformly valid inference about the parameters on endogenous variables in the IV context with many instruments to allow non-Gaussian heteroscedastic disturbances in Belloni et al. (2012). These principles have also been applied in Belloni et al. (2010*a*), which outlines approaches for regression and IV models; Belloni, Chernozhukov and Hansen (2014) (ArXiv 2011), which covers estimation of the parametric components of the partially linear model, estimation of average treatment effects, and provides a formal statement of the orthogonality condition (2); Farrell (2013) which covers average treatment effects with discrete, multi-valued treatments; Kozbur (2014) which covers additive nonparametric models; and Belloni, Chernozhukov, Hansen and Kozbur (2014) which extends the IV and partially linear model results to allow for fixed effects panel data and clustered dependence structures. The most recent, general approach is provided in Belloni, Chernozhukov, Fernández-Val and Hansen (2013) where inference about parameters defined by a continuum of orthogonalized estimating equations with infinite-dimensional nusiance parameters is analyzed and positive results on inference are developed. The framework in Belloni, Chernozhukov, Fernández-Val and Hansen (2013) is general enough to cover the aforementioned papers and many other parametric and semi-parametric models considered in economics.

As noted above, providing uniformly valid inference following model selection is closely related to use of Neyman's $C(\alpha)$-statistic. Valid confidence regions can be obtained by inverting tests based on these statistics, and minimizers of $C(\alpha)$-statistics may be used as point estimators. The use of $C(\alpha)$ statistics for testing and estimation in high-dimensional approximately sparse models was first explored in the context of high-dimensional quantile regression in Belloni, Chernozhukov and Kato (2013*b*) (Oberwolfach, 2012) and Belloni, Chernozhukov and Kato (2013*a*) and in the context of high-dimensional logistic regression and other high-dimensional generalized linear models by Belloni, Chernozhukov and Wei (2013). More recent uses of $C(\alpha)$-statistics (or close variants, under different names) include those in Voorman et al. (2014), Ning and Liu (2014), and Yang et al. (2014) among others.

There have also been parallel developments based upon ex-post "de-biasing" of estimators. This approach is mathematically equivalent to doing classical "one-step" corrections in the general framework of Section 2. Indeed, while at first glance this "de-biasing"

approach may appear distinct from that taken in the papers listed above in this section, it is the same as approximately solving – by doing one Gauss-Newton step – orthogonal estimating equations satisfying (2). The general results of Section 2 suggest that these approaches – the exact solving and "one-step" solving – are generally first-order asymptotically equivalent, though higher-order differences may persist. To the best of our knowledge, the "one-step" correction approach was first employed in high-dimensional sparse models by Zhang and Zhang (2014) (ArXiv 2011) which covers the homoscedastic linear model (as well as in several follow-up works by the authors). This approach has been further used in van de Geer et al. (2014) (ArXiv 2013) which covers homoscedastic linear models and some generalized linear models, and Javanmard and Montanari (2014) (ArXiv 2013) which offers a related, though somewhat different approach. Note that Belloni, Chernozhukov and Kato (2013*b*) and Belloni, Chernozhukov and Wei (2013) also offer results on "one-step" corrections as part of their analysis of estimation and inference based upon the orthogonal estimating equations. We would not expect that the use of orthogonal estimating equations or the use of "one-step" corrections to dominate each other in all cases, though computational evidence in Belloni, Chernozhukov and Wei (2013) suggests that the use of exact solutions to orthogonal estimating equations may be preferable to approximate solutions obtained from "one-step" corrections in the contexts considered in that paper.

Another branch of the recent literature takes a complementary, but logically distinct, approach that aims at doing valid inference for the parameters of a "pseudo-true" model that results from the use of a model selection procedure, see Berk et al. (2013). Specifically, this approach conditions on a model selected by a data-dependent rule and then attempts to do inference – conditional on the selection event – for the parameters of the selected model, which may deviate from the "true" model that generated the data. Related developments within this approach appear in G'Sell et al. (2013), Lee and Taylor (2014), Lee et al. (2013), Lockhart et al. (2014), Loftus and Taylor (2014), Taylor et al. (2014), and Fithian et al. (2014). Some of the developments still explicitly rely on "beta-min" conditions. To remove these conditions, it seems intellectually very interesting to combine the developments of the present paper (and other preceding papers cited above) with developments in this literature.

The previously mentioned work focuses on doing inference for low dimensional parameters in the presence of high dimensional nuisance parameters. There have also been developments on performing inference for high dimensional parameters. Chernozhukov (2009) proposed inverting a Lasso performance bound in order to construct a simultaneous, Scheffé-style confidence band on all parameters. An interesting feature of this approach is that it uses weaker design conditions than many other approaches but requires the data analyst to supply explicit bounds on restricted eigenvalues. Gautier and Tsybakov (2011) (ArXiv 2011) and Chernozhukov et al. (2013) employ similar ideas while also working with various generalizations of restricted eigenvalues. van de Geer and Nickl (2013) construct confidence ellipsoids for the entire parameter vector using sample splitting ideas. Somewhat related to this literature are the results of Belloni, Chernozhukov and Kato (2013*b*) who use the orthogonal estimating equations framework with infinite-dimensional nuisance parameters and construct a simultaneous confidence

rectangle for many target parameters where the number of target parameters could be much larger than the sample size. They relied upon the high-dimensional central limit theorems and bootstrap results established in Chernozhukov et al. (2013).

Most of the aforementioned results rely on (approximate) sparsity and related sparsity-based estimators. Some examples of the use of alternative regularization schemes are available in the many instrument literature in econometrics. For example, Chamberlain and Imbens (2004) use a shrinkage estimator resulting from use of a Gaussian random coefficients structure over first-stage coefficients, and Okui (2010) uses ridge regression for estimating the first-stage regression in a framework where the instruments may be ordered in terms of relevance. Carrasco (2012) employs a different strategy based on directly regularizing the inverse that appears in the definition of the 2SLS estimator allowing for a number of moment conditions that are larger than the sample size; see also Carrasco and Tchuente Nguembu (2012). The theoretical development in Carrasco (2012) relies on restrictions on the covariance structure of the instruments rather than on the coefficients of the instruments. Hansen and Kozbur (2014) considers a combination of ridge-regularization and the jackknife to provide a procedure that is valid allowing for the number of instruments to be greater than the sample size under weak restrictions on the covariance structure of the instruments and the first-stage coefficients. In all cases, the orthogonality condition holds allowing root-n consistent and asymptotically normal estimation of the main parameter α.

Many other interesting procedures beyond those mentioned in this review have been developed for estimating high-dimensional models; see, e.g. Hastie et al. (2009) for a textbook review. Developing new techniques for estimation in high-dimensional settings is also still an active area of research, so the list of methods available to researchers continues to expand. The use of these procedures and the impact of their use on inference about parameters of interest is an interesting research direction to explore. It seems likely that many of these procedures will provide sufficiently high-quality estimates that they may be used for estimating the nuisance parameters η in the present setting.

Appendix A. The Lasso and Post-Lasso Estimators in the Linear Model

Suppose we have data $\{y_i, x_i\}$ for individuals $i = 1, ..., n$ where x_i is a p-vector of predictor variables and y_i is an outcome of interest. Suppose that we are interested in a linear prediction model for y_i, $y_i = x_i'\eta + \varepsilon_i$, and define the usual least squares criterion function:

$$\hat{Q}(\eta) := \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i'\eta)^2.$$

The Lasso estimator is defined as a solution of the following optimization program:

$$\widehat{\eta}_L \in \arg\min_{\eta\in\mathbb{R}^p} \hat{Q}(\eta) + \frac{\lambda}{n}\sum_{j=1}^{p}|\psi_j\eta_j| \tag{56}$$

where λ is the penalty level and $\{\psi_j\}_{j=1}^p$ are covariate specific penalty loadings. The covariate specific penalty loadings are used to accommodate data that may be non-Gaussian, heteroscedastic, and/or dependent and also help ensure basic equivariance of coefficient estimates to rescaling of the covariates.

The Post-Lasso estimator is defined as the ordinary least square regression applied to the model $\widehat{I}$ selected by Lasso:[6]

$$\widehat{I} = \text{support}(\hat{\eta}_L) = \{j \in \{1, \ldots, p\} \ : \ |\hat{\eta}_{Lj}| > 0\}.$$

The Post-Lasso estimator $\hat{\eta}_{PL}$ is then

$$\widehat{\eta}_{PL} \in \arg\min\{\hat{Q}(\eta) : \quad \eta \in \mathbb{R}^p \text{ such that } \eta_j = 0 \ \text{ for all } j \notin \widehat{I}\} \tag{57}$$

In words, this estimator is ordinary least squares (OLS) using only the regressors whose coefficients were estimated to be non-zero by Lasso.

Lasso and Post-Lasso are motivated by the desire to predict the target function well without overfitting. The Lasso estimator is a computationally attractive alternative to some other classic approaches, such as model selection based on information criteria, because it minimizes a convex function. Moreover, under suitable conditions, the Lasso estimator achieves near-optimal rates in estimating the regression function $x_i'\eta$. However, Lasso does suffer from the drawback that the regularization by the ℓ_1-norm employed in (56) naturally shrinks all estimated coefficients towards zero causing a potentially significant shrinkage bias. The Post-Lasso estimator is meant to remove some of this shrinkage bias and achieves the same rate of convergence as Lasso under sensible conditions.

Practical implementation of the Lasso requires setting the penalty parameter and loadings. Verifying good properties of the Lasso typically relies on having these parameters set so that the penalty dominates the score in the sense that

$$\frac{\psi_j \lambda}{n} \geq \max_{j \leq p} 2c \left| \frac{1}{n} \sum_{i=1}^n x_{j,i} \varepsilon_i \right| \text{ or, equivalently } \frac{\lambda}{\sqrt{n}} \geq \max_{j \leq p} 2c \left| \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^n x_{j,i} \varepsilon_i}{\psi_j} \right|$$

for some $c > 1$ with high probability. Heuristically, we would have the term inside the absolute values behaving approximately like a standard normal random variable if we set $\psi_j = \text{Var}\left[\frac{1}{\sqrt{n}} \sum_{i=1}^n x_{j,i} \varepsilon_i\right]$. We could then get the desired domination by setting $\frac{\lambda}{2c\sqrt{n}}$ large enough to dominate the maximum of p standard normal random variables with high probability, for example, by setting $\lambda = 2c\sqrt{n}\Phi^{-1}\left(1 - .1/[2p\log(n)]\right)$ where $\Phi^{-1}(\cdot)$ denotes the inverse of the standard normal cumulative distribution function. Verifying that this heuristic argument holds with large p and data which may not be i.i.d. Gaussian requires careful and delicate arguments as in, for example, Belloni et al. (2012) which covers heteroscedastic non-Gaussian data or Belloni, Chernozhukov, Hansen and Kozbur (2014) which covers panel data with within individual dependence. The choice of the penalty parameter λ can also be refined as in Belloni et al. (2011). Finally, feasible implementation requires that ψ_j be estimated which can be done through the iterative procedures suggested in Belloni et al. (2012) or Belloni, Chernozhukov, Hansen and Kozbur (2014).

Appendix B. Proofs

B.1. **Proof of Proposition 2.** Consider any sequence $\{\mathrm{P}_n\}$ in $\{\mathbf{P}_n\}$.

Step 1 (r_n-rate). Here we show that $\|\hat{\alpha} - \alpha_0\| \leq r_n$ wp $\to 1$. We have by the identifiability condition, in particular the assumption mineig$(\Gamma_1'\Gamma_1) \geq c$, that

$$\mathrm{P}_n(\|\hat{\alpha} - \alpha_0\| > r_n) \leq \mathrm{P}_n(\|\mathrm{M}(\hat{\alpha}, \eta_0)\| \geq \iota(r_n)), \quad \iota(r_n) := 2^{-1}(\{\sqrt{c} r_n\} \wedge c).$$

[6]We note that we can also allow the set $\widehat{I}$ to contain additional variables not selected by Lasso, but we do not consider that here.

Hence it suffices to show that wp $\to 1$, $\|\mathrm{M}(\hat{\alpha}, \eta_0)\| < \iota(r_n)$. By the triangle inequality,

$$\|\mathrm{M}(\hat{\alpha}, \eta_0)\| \leq I_1 + I_2 + I_3, \quad \begin{array}{l} I_1 = \|\mathrm{M}(\hat{\alpha}, \eta_0) - \mathrm{M}(\hat{\alpha}, \hat{\eta})\|, \\ I_2 = \|\mathrm{M}(\hat{\alpha}, \hat{\eta}) - \hat{\mathrm{M}}(\hat{\alpha}, \hat{\eta})\|, \\ I_3 = \|\hat{\mathrm{M}}(\hat{\alpha}, \hat{\eta})\|. \end{array}$$

By assumption (12), wp $\to 1$

$$I_1 + I_2 \leq o(1)\{r_n + I_3 + \|\mathrm{M}(\hat{\alpha}, \eta_0)\|\}.$$

Hence,

$$\|\mathrm{M}(\hat{\alpha}, \eta_0)\|(1 - o(1)) \leq o(1)(r_n + I_3) + I_3.$$

By construction of the estimator,

$$I_3 \leq o(n^{-1/2}) + \inf_{\alpha \in \mathcal{A}} \|\hat{\mathrm{M}}(\alpha, \hat{\eta})\| \lesssim_{\mathrm{P}_n} n^{-1/2},$$

which follows because

$$\inf_{\alpha \in \mathcal{A}} \|\hat{\mathrm{M}}(\alpha, \hat{\eta})\| \leq \|\hat{\mathrm{M}}(\bar{\alpha}, \hat{\eta})\| \lesssim_{\mathrm{P}_n} n^{-1/2}, \tag{58}$$

where $\bar{\alpha}$ is the one-step estimator defined in Step 3, as shown in (59). Hence wp $\to 1$

$$\|\mathrm{M}(\hat{\alpha}, \eta_0)\| \leq o(r_n) < \iota(r_n),$$

where to obtain the last inequality we have used the assumption $\mathrm{mineig}(\Gamma_1'\Gamma_1) \geq c$.

Step 2 ($n^{-1/2}$-rate). Here we show that $\|\hat{\alpha} - \alpha_0\| \lesssim_{\mathrm{P}_n} n^{-1/2}$. By condition (14) and the triangle inequality, wp $\to 1$

$$\|\mathrm{M}(\hat{\alpha}, \eta_0)\| \geq \|\Gamma_1(\hat{\alpha} - \alpha_0)\| - o(1)\|\hat{\alpha} - \alpha_0\| \geq (\sqrt{c} - o(1))\|(\hat{\alpha} - \alpha_0)\| \geq \sqrt{c}/2\|(\hat{\alpha} - \alpha_0)\|.$$

Therefore, it suffices to show that $\|\mathrm{M}(\hat{\alpha}, \eta_0)\| \lesssim_{\mathrm{P}_n} n^{-1/2}$. We have that

$$\|\mathrm{M}(\hat{\alpha}, \eta_0)\| \leq II_1 + II_2 + II_3, \quad \begin{array}{l} II_1 = \|\mathrm{M}(\hat{\alpha}, \eta_0) - \mathrm{M}(\hat{\alpha}, \hat{\eta})\|, \\ II_2 = \|\mathrm{M}(\hat{\alpha}, \hat{\eta}) - \hat{\mathrm{M}}(\hat{\alpha}, \hat{\eta}) - \hat{\mathrm{M}}(\alpha_0, \eta_0)\|, \\ II_3 = \|\hat{\mathrm{M}}(\hat{\alpha}, \hat{\eta})\| + \|\hat{\mathrm{M}}(\alpha_0, \eta_0)\|. \end{array}$$

Then, by the orthogonality $\partial_{\eta'}\mathrm{M}(\alpha_0, \eta_0) = 0$ and condition (14), wp $\to 1$,

$$\begin{array}{rcl} II_1 & \leq & \|\mathrm{M}(\hat{\alpha}, \hat{\eta}) - \mathrm{M}(\hat{\alpha}, \eta_0) - \partial_{\eta'}\mathrm{M}(\hat{\alpha}, \eta_0)[\hat{\eta} - \eta_0]\| + \|\partial_{\eta'}\mathrm{M}(\hat{\alpha}, \eta_0)[\hat{\eta} - \eta_0]\| \\ & \leq & o(1)n^{-1/2} + o(1)\|\hat{\alpha} - \alpha_0\| \\ & \leq & o(1)n^{-1/2} + o(1)(2/\sqrt{c})\|\mathrm{M}(\hat{\alpha}, \eta_0)\|. \end{array}$$

Then, by condition (13) and by $I_3 \lesssim_{\mathrm{P}_n} n^{-1/2}$,

$$\begin{array}{rcl} II_2 & \leq & o(1)\{n^{-1/2} + \|\hat{\mathrm{M}}(\hat{\alpha}, \hat{\eta})\| + \|\mathrm{M}(\hat{\alpha}, \eta_0)\|\} \\ & \lesssim_{\mathrm{P}_n} & o(1)\{n^{-1/2} + n^{-1/2} + \|\mathrm{M}(\hat{\alpha}, \eta_0)\|\}. \end{array}$$

Since $II_3 \lesssim_{\mathrm{P}_n} n^{-1/2}$ by (58) and $\|\hat{\mathrm{M}}(\alpha_0, \eta_0)\| \lesssim_{\mathrm{P}_n} n^{-1/2}$, it follows that wp $\to 1$, $(1 - o(1))\|\mathrm{M}(\hat{\alpha}, \eta_0)\| \lesssim_{\mathrm{P}_n} n^{-1/2}$.

Step 3 (Linearization). Define the linearization map $\alpha \mapsto \widehat{\mathrm{L}}(\alpha)$ by $\widehat{\mathrm{L}}(\alpha) := \hat{\mathrm{M}}(\alpha_0, \eta_0) + \Gamma_1(\alpha - \alpha_0)$. Then

$$\|\hat{\mathrm{M}}(\hat{\alpha}, \hat{\eta}) - \widehat{\mathrm{L}}(\hat{\alpha})\| \leq III_1 + III_2 + III_3, \quad \begin{array}{l} III_1 = \|\mathrm{M}(\hat{\alpha}, \hat{\eta}) - \mathrm{M}(\hat{\alpha}, \eta_0)\|, \\ III_2 = \|\mathrm{M}(\hat{\alpha}, \eta_0) - \Gamma_1(\hat{\alpha} - \alpha_0)\|, \\ III_3 = \|\hat{\mathrm{M}}(\hat{\alpha}, \hat{\eta}) - \mathrm{M}(\hat{\alpha}, \hat{\eta}) - \hat{\mathrm{M}}(\alpha_0, \eta_0)\|. \end{array}$$

Then, using the assumptions (14) and (13), conclude

$$\begin{aligned} III_1 &\leq \|\mathrm{M}(\hat{\alpha},\hat{\eta}) - \mathrm{M}(\hat{\alpha},\eta_0) - \partial_{\eta'}\mathrm{M}(\hat{\alpha},\eta_0)[\hat{\eta}-\eta_0]\| + \|\partial_{\eta'}\mathrm{M}(\hat{\alpha},\eta_0)[\hat{\eta}-\eta_0]\| \\ &\leq o(1)n^{-1/2} + o(1)\|\hat{\alpha}-\alpha_0\|, \\ III_2 &\leq o(1)\|\hat{\alpha}-\alpha_0\|, \\ III_3 &\leq o(1)(n^{-1/2} + \|\hat{\mathrm{M}}(\hat{\alpha},\hat{\eta})\| + \|\mathrm{M}(\hat{\alpha},\eta_0)\|) \\ &\leq o(1)(n^{-1/2} + n^{-1/2} + III_2 + \|\Gamma_1(\hat{\alpha}-\alpha_0)\|). \end{aligned}$$

Conclude that wp $\to 1$, since $\|\Gamma_1'\Gamma_1\| \lesssim 1$ by assumption (11),

$$\|\hat{\mathrm{M}}(\hat{\alpha},\hat{\eta}) - \widehat{\mathrm{L}}(\hat{\alpha})\| \lesssim_{\mathrm{P}_n} o(1)(n^{-1/2} + \|\hat{\alpha}-\alpha_0\|) = o(n^{-1/2}).$$

Also consider the minimizer of the map $\alpha \mapsto \|\widehat{\mathrm{L}}(\alpha)\|$, namely,

$$\bar{\alpha} = \alpha_0 - (\Gamma_1'\Gamma_1)^{-1}\Gamma_1'\hat{\mathrm{M}}(\alpha_0,\eta_0)$$

which obeys $\|\sqrt{n}(\bar{\alpha}-\alpha_0)\| \lesssim_{\mathrm{P}_n} n^{-1/2}$ under the conditions of the proposition. We can repeat the argument above to conclude that wp $\to 1$, $\|\hat{\mathrm{M}}(\bar{\alpha},\hat{\eta}) - \widehat{\mathrm{L}}(\bar{\alpha})\| \lesssim_{\mathrm{P}_n} o(n^{-1/2})$. This implies, since $\|\widehat{\mathrm{L}}(\bar{\alpha})\| \lesssim_{\mathrm{P}_n} n^{-1/2}$,

$$\|\hat{\mathrm{M}}(\bar{\alpha},\hat{\eta})\| \lesssim_{\mathrm{P}_n} n^{-1/2}. \tag{59}$$

This also implies that $\|\widehat{\mathrm{L}}(\hat{\alpha})\| = \|\widehat{\mathrm{L}}(\bar{\alpha})\| + o_{\mathrm{P}_n}(n^{-1/2})$, since $\|\widehat{\mathrm{L}}(\bar{\alpha})\| \leq \|\widehat{\mathrm{L}}(\hat{\alpha})\|$ and

$$\|\widehat{\mathrm{L}}(\hat{\alpha})\| - o_{\mathrm{P}_n}(n^{-1/2}) \leq \|\hat{\mathrm{M}}(\hat{\alpha},\hat{\eta})\| \leq \|\hat{\mathrm{M}}(\bar{\alpha},\hat{\eta})\| + o(n^{-1/2}) = \|\widehat{\mathrm{L}}(\bar{\alpha})\| + o_{\mathrm{P}_n}(n^{-1/2}).$$

The former assertion implies that $\|\widehat{\mathrm{L}}(\hat{\alpha})\|^2 = \|\widehat{\mathrm{L}}(\bar{\alpha})\|^2 + o_{\mathrm{P}_n}(n^{-1})$, so that

$$\|\widehat{\mathrm{L}}(\hat{\alpha})\|^2 - \|\widehat{\mathrm{L}}(\bar{\alpha})\|^2 = \|\Gamma_1(\hat{\alpha}-\bar{\alpha})\|^2 = o_{\mathrm{P}_n}(n^{-1}),$$

from which we can conclude that $\sqrt{n}\|\hat{\alpha}-\bar{\alpha}\| \to_{\mathrm{P}_n} 0$.

Step 4. (Conclusion). Given the conclusion of the previous step, the remaining claims are standard and follow from the Continuous Mapping Theorem and Lemma 8. ∎

B.2. **Proof of Proposition 3.** We have wp $\to 1$ that, for some constants $0 < u < l < 0$, $l\|x\| \leq \|\mathrm{A}x\| \leq u\|x\|$ and $l\|x\| \leq \|\hat{\mathrm{A}}x\| \leq u\|x\|$. Hence

$$\begin{aligned} &\sup_{\alpha\in\mathcal{A}} \frac{\|\hat{\mathrm{A}}\hat{\mathrm{M}}^o(\alpha,\hat{\eta}) - \mathrm{A}\mathrm{M}^o(\alpha,\hat{\eta})\| + \|\mathrm{A}\mathrm{M}^o(\alpha,\hat{\eta}) - \mathrm{A}\mathrm{M}^o(\alpha,\eta_0)\|}{r_n + \|\hat{\mathrm{A}}\hat{\mathrm{M}}^o(\alpha,\hat{\eta})\| + \|\mathrm{A}\mathrm{M}^o(\alpha,\eta_0)\|} \\ &\leq \sup_{\alpha\in\mathcal{A}} \frac{u}{l} \frac{\|\hat{\mathrm{M}}^o(\alpha,\hat{\eta}) - \mathrm{M}^o(\alpha,\hat{\eta})\| + \|\mathrm{M}^o(\alpha,\hat{\eta}) - \mathrm{M}^o(\alpha,\eta_0)\|}{(r_n/l) + \|\hat{\mathrm{M}}^o(\alpha,\hat{\eta})\| + \|\mathrm{M}^o(\alpha,\eta_0)\|} \\ &\quad + \sup_{\alpha\in\mathcal{A}} \frac{\|\hat{\mathrm{A}}-\mathrm{A}\|\|\hat{\mathrm{M}}^o(\alpha,\hat{\eta})\|}{r_n + l\|\hat{\mathrm{M}}^o(\alpha,\hat{\eta})\|} \lesssim_{\mathrm{P}_n} o(1) + \|\hat{\mathrm{A}}-\mathrm{A}\|/l \to_{\mathrm{P}_n} 0. \end{aligned}$$

The proof that the rest of the conditions hold is analogous and is therefore omitted. ∎

B.3. **Proof of Proposition 4.** Step 1. We define the feasible and infeasible "one-steps"

$$\begin{aligned} \check{\alpha} &= \tilde{\alpha} - \hat{F}\hat{\mathrm{M}}(\tilde{\alpha},\hat{\eta}), \qquad \hat{F} = (\hat{\Gamma}_1'\hat{\Gamma}_1)^{-1}\hat{\Gamma}_1', \\ \bar{\alpha} &= \alpha_0 - F\hat{\mathrm{M}}(\alpha_0,\eta_0), \quad F = (\Gamma_1'\Gamma_1)^{-1}\Gamma_1'. \end{aligned}$$

We deduce by (20) and (11) that

$$\|\hat{F}\| \lesssim_{\mathrm{P}_n} 1, \quad \|\hat{F}\Gamma_1 - I\| \lesssim_{\mathrm{P}_n} r_n, \quad \|\hat{F} - F\| \lesssim_{\mathrm{P}_n} r_n.$$

Step 2. By Step 1 and by condition (21), we have that

$$\begin{aligned}\mathsf{D} &= \|\hat{F}\hat{\mathrm{M}}(\tilde{\alpha},\hat{\eta}) - \hat{F}\hat{\mathrm{M}}(\alpha_0,\eta_0) - \hat{F}\Gamma_1(\tilde{\alpha}-\alpha_0)\| \\ &\leq \|\hat{F}\|\|\hat{\mathrm{M}}(\tilde{\alpha},\hat{\eta}) - \hat{\mathrm{M}}(\alpha_0,\eta_0) - \Gamma_1(\tilde{\alpha}-\alpha_0)\| \\ &\lesssim_{\mathrm{P}_n} \|\hat{\mathrm{M}}(\tilde{\alpha},\hat{\eta}) - \mathrm{M}(\tilde{\alpha},\hat{\eta}) - \hat{\mathrm{M}}(\alpha_0,\eta_0)\| + \mathsf{D}_1 \lesssim_{\mathrm{P}_n} o(n^{-1/2}) + \mathsf{D}_1,\end{aligned}$$

where $\mathsf{D}_1 := \|\mathrm{M}(\tilde{\alpha},\hat{\eta}) - \Gamma_1(\hat{\alpha}-\alpha_0)\|$.

Moreover, $\mathsf{D}_1 \leq IV_1 + IV_2 + IV_3$, where wp $\to 1$ by condition (21) and $r_n^2 = o(n^{-1/2})$

$$\begin{aligned}IV_1 &:= \|\mathrm{M}(\tilde{\alpha},\eta_0) - \Gamma_1(\tilde{\alpha}-\alpha_0)\| \lesssim \|\tilde{\alpha}-\alpha_0\|^2 \lesssim r_n^2 = o(n^{-1/2}), \\ IV_2 &:= \|\mathrm{M}(\tilde{\alpha},\hat{\eta}) - \mathrm{M}(\tilde{\alpha},\eta_0) - \partial_{\eta'}\mathrm{M}(\tilde{\alpha},\eta_0)[\hat{\eta}-\eta_0]\| \lesssim o(n^{-1/2}), \\ IV_3 &:= \|\partial_{\eta'}\mathrm{M}(\tilde{\alpha},\eta_0)[\hat{\eta}-\eta_0]\| \lesssim o(n^{-1/2}).\end{aligned}$$

Conclude that $n^{1/2}\mathsf{D} \to_{\mathrm{P}_n} 0$.

Step 3. We have by the triangle inequality and Steps 1 and 2 that

$$\begin{aligned}\sqrt{n}\|\breve{\alpha}-\bar{\alpha}\| &\leq \sqrt{n}\|(I-\hat{F}\Gamma_1)(\tilde{\alpha}-\alpha_0)\| + \sqrt{n}\|(\hat{F}-F)\hat{\mathrm{M}}(\alpha_0,\eta_0)\| + \sqrt{n}\mathsf{D} \\ &\leq \sqrt{n}\|(I-\hat{F}\Gamma_1)\|\|\tilde{\alpha}-\alpha_0\| + \|\hat{F}-F\|\|\sqrt{n}\hat{\mathrm{M}}(\alpha_0,\eta_0)\| + \sqrt{n}\mathsf{D} \\ &\lesssim_{\mathrm{P}_n} \sqrt{n}r_n^2 + o(1) = o(1).\end{aligned}$$

Thus, $\sqrt{n}\|\breve{\alpha}-\bar{\alpha}\| \to_{\mathrm{P}_n} 0$, and $\sqrt{n}\|\breve{\alpha}-\hat{\alpha}\| \to_{\mathrm{P}_n} 0$ follows from the triangle inequality and the fact that $\sqrt{n}\|\hat{\alpha}-\bar{\alpha}\| \to_{\mathrm{P}_n} 0$. ■

B.4. **Proof of Lemma 2.** The conditions of Proposition 1 are clearly satisfied, and thus the conclusions of Proposition 1 immediately follow. We also have that, for $\hat{\Gamma}_1 = \hat{\Gamma}_1(\hat{\eta})$,

$$\begin{aligned}\sqrt{n}(\hat{\alpha}-\alpha_0) &= -\hat{F}\sqrt{n}\hat{\mathrm{M}}(\alpha_0,\hat{\eta}), \quad \hat{F} = (\hat{\Gamma}_1'\hat{\Gamma}_1)^{-1}\hat{\Gamma}_1, \\ \sqrt{n}(\bar{\alpha}-\alpha_0) &:= -F\sqrt{n}\hat{\mathrm{M}}(\alpha_0,\eta_0), \quad F = (\Gamma_1'\Gamma_1)^{-1}\Gamma_1.\end{aligned}$$

We deduce by (33) and (11) that $\|\hat{F}\| \lesssim_{\mathrm{P}_n} 1$ and $\|\hat{F}-F\| \to_{\mathrm{P}_n} 0$. Hence we have by triangle and Hölder inequalities and condition (33) that

$$\sqrt{n}\|\hat{\alpha}-\bar{\alpha}\| \leq \|\hat{F}\|\sqrt{n}\|\hat{\mathrm{M}}(\alpha_0,\hat{\eta}) - \hat{\mathrm{M}}(\alpha_0,\eta_0))\| + \|\hat{F}-F\|\sqrt{n}\|\hat{\mathrm{M}}(\alpha_0,\eta_0)\| \to_{\mathrm{P}_n} 0.$$

The conclusions regarding the uniform validity of inference using $\hat{\alpha}$, of the form stated in conclusions of Proposition 2, follow from the conclusions regarding the uniform validity of inference using $\bar{\alpha}$, which follow from the Continuous Mapping Theorem, Lemma 8, and the assumed stability conditions (11). This establishes the second claim of the Lemma. Verification of the conditions of Proposition 2 is omitted. ■

B.5. **Proof of Lemma 3 and 4.** The proof of Lemma 3 is given in the main text. As in the proof of Lemma 3, we can expand:

$$\sqrt{n}(\hat{\mathrm{M}}_j(\alpha_0,\hat{\eta}) - \hat{\mathrm{M}}_j(\alpha_0,\eta_0)) = T_{1,j} + T_{2,j} + T_{3,j}, \tag{60}$$

where the terms $(T_{l,j})_{l=1}^3$ are as defined in the main text. We can further bound $T_{3,j}$ as follows:

$$T_{3,j} \leq T_{3,j}^m + T_{4,j}, \qquad \begin{aligned}T_{3,j}^m &:= \sqrt{n}|(\hat{\eta}-\eta_0^m)'\partial_\eta\partial_{\eta'}\hat{\mathrm{M}}_j(\alpha_0)(\hat{\eta}-\eta_0^m)|, \\ T_{4,j} &:= \sqrt{n}|\eta_0^{r\prime}\partial_\eta\partial_{\eta'}\hat{\mathrm{M}}_j(\alpha_0)\eta_0^r|.\end{aligned} \tag{61}$$

Then $T_{1,j} = 0$ by orthogonality, $T_{2,j} \to_{\mathrm{P}_n} 0$ as in the proof of Lemma 3. Since $s^2\log(pn)^2/n \to 0$, $T_{3,j}^m$ vanishes in probability because, by Hölder's inequality and for sufficiently large n,

$$T_{3,j}^m \leq \bar{T}_{3,j}\|\hat{\eta}-\eta_0^m\|^2 \lesssim_{\mathrm{P}_n} \sqrt{n}s\log(pn)/n \to_{\mathrm{P}_n} 0.$$

Also, if $s^2 \log(pn)^2/n \to 0$, $T_{4,j}$ vanishes in probability because, by Hölder's inequality and (43),

$$T_{4,j} \leq \sqrt{n}\|\partial_\eta \partial_{\eta'} \hat{\mathrm{M}}_j(\alpha_0)\|_{\mathsf{pw}(\eta_0^r)}\|\eta_0^r\|^2 \lesssim_{\mathrm{P}_n} \sqrt{n} s \log(pn)/n \to_{\mathrm{P}_n} 0.$$

The conclusion follows from (60). ∎

B.6. **Proof of Lemma 5.** For $m = 1, ..., k$ and $l = 1, ..., d$, we can bound each element $\hat{\Gamma}_{1,ml}(\eta)$ of matrix $\hat{\Gamma}_1(\eta)$ as follows:

$$|\hat{\Gamma}_{1,ml}(\hat{\eta}) - \hat{\Gamma}_{1,ml}(\eta_0)| \leq \sum_{k=1}^4 T_{k,ml}, \qquad \begin{array}{l} T_{1,ml} := |\partial_\eta \Gamma_{1,ml}(\eta_0)'(\hat{\eta} - \eta_0)|, \\ T_{2,ml} := |(\partial_\eta \hat{\Gamma}_{1,ml}(\eta_0) - \partial_\eta \Gamma_{1,ml}(\eta_0))'(\hat{\eta} - \eta_0)|, \\ T_{3,ml} := |(\hat{\eta} - \eta_0^m)' \partial_\eta \partial_{\eta'} \hat{\Gamma}_{1,ml}(\hat{\eta} - \eta_0^m)|, \\ T_{4,ml} := |\eta_0^{r\prime} \partial_\eta \partial_{\eta'} \hat{\Gamma}_{1,ml} \eta_0^r|. \end{array}$$

Under conditions (44) and (45) we have that wp $\to 1$

$$\begin{aligned} T_{1,ml} &\leq \|\partial_\eta \Gamma_{1,ml}(\eta_0)\|_\infty \|\hat{\eta} - \eta_0\|_1 \lesssim_{\mathrm{P}_n} \sqrt{s^2 \log(pn)/n} \to 0, \\ T_{2,ml} &\leq \|\partial_\eta \hat{\Gamma}_{1,ml}(\eta_0) - \partial_\eta \Gamma_{1,ml}(\eta_0)\|_\infty \|\hat{\eta} - \eta_0\|_1 \lesssim_{\mathrm{P}_n} \sqrt{s^2 \log(pn)/n} \to 0, \\ T_{3,ml} &\leq \|\partial_\eta \partial_{\eta'} \hat{\Gamma}_{1,ml}\|_{\mathsf{sp}(\ell_n s)} \|\hat{\eta} - \eta_0^m\|^2 \lesssim_{\mathrm{P}_n} s \log(pn)/n \to 0, \\ T_{4,ml} &\leq \|\partial_\eta \partial_{\eta'} \hat{\Gamma}_{1,ml}\|_{\mathsf{pw}(\eta_0^r)} \|\eta_0^r\|^2 \lesssim_{\mathrm{P}_n} s \log(pn)/n \to 0. \end{aligned}$$

The claim follows from the assumed growth conditions, since d and k are bounded. ∎

Appendix C. Key Tools

Let Φ and Φ^{-1} denote the distribution and quantile function of $\mathcal{N}(0,1)$. Note that in particular $\Phi^{-1}(1-a) \leq \sqrt{2\log(a)}$ for all $a \in (0, 1/2)$.

Lemma 6 (Moderate Deviation Inequality for Maximum of a Vector)**.** *Suppose that* $\mathcal{S}_j := \sum_{i=1}^n U_{ij}/\sqrt{\sum_{i=1}^n U_{ij}^2}$, *where* U_{ij} *are independent random variables across* i *with mean zero and finite third-order moments. Then*

$$\mathrm{P}\left(\max_{1 \leq j \leq p} |\mathcal{S}_j| > \Phi^{-1}(1 - \gamma/2p)\right) \leq \gamma\left(1 + \frac{A}{\ell_n^3}\right),$$

where A *is an absolute constant, provided for* $\ell_n > 0$

$$0 \leq \Phi^{-1}(1 - \gamma/(2p)) \leq \frac{n^{1/6}}{\ell_n} \min_{1 \leq j \leq p} M_j^2 - 1, \quad M_j := \frac{\left(\frac{1}{n}\sum_{i=1}^n \mathrm{E}[U_{ij}^2]\right)^{1/2}}{\left(\frac{1}{n}\sum_{i=1}^n \mathrm{E}[|U_{ij}|^3]\right)^{1/3}}.$$

This result is essentially due to Jing et al. (2003). The proof of this result, given in Belloni et al. (2012), follows from a simple combination of union bounds with their result.

Lemma 7 (Laws of Large Numbers for Large Matrices in Sparse Norms)**.** *Let* s_n, p_n, k_n *and* ℓ_n *be sequences of positive constants such that* $\ell_n \to \infty$ *but* $\ell_n/\log n \to 0$ *and* c_1 *and* c_2 *be fixed positive constants. Let* $(x_i)_{i=1}^n$ *be i.i.d. vectors such that* $\|\mathrm{E}[x_i x_i']\|_{\mathsf{sp}(s_n \log n)} \leq c_1$, *and either one of the following holds: (a)* x_i *is a sub-Gaussian random vector with* $\sup_{\|u\| \leq 1} \|x_i' u\|_{\psi_2, \mathrm{P}} \leq c_2$, *where* $\|\cdot\|_{\psi_2, \mathrm{P}}$ *denotes the* ψ_2*-Orlizs norm of a random variable, and* $s_n(\log n)(\log(p_n \vee n))/n \to 0$*; or (b)* $\|x_i\|_\infty \leq k_n$ *a.s. and* $k_n^2 s_n(\log^4 n)\log(p_n \vee n)/n \to 0$. *Then there is* $o(1)$ *term such that with probability* $1 - o(1)$*:*

$$\|\mathbb{E}_n[x_i x_i'] - \mathrm{E}[x_i x_i']\|_{\mathsf{sp}(s_n \ell_n)} \leq o(1), \quad \|\mathbb{E}_n[x_i x_i']\|_{\mathsf{sp}(s_n \ell_n)} \leq c_1 + o(1).$$

Under (a) the result follows from Theorem 3.2 in Rudelson and Zhou (2011) and under (b) the result follows from Rudelson and Vershynin (2008), as shown in the Supplemental Material of Belloni and Chernozhukov (2013).

Lemma 8 (Useful implications of CLT in $\mathbb{R}^m$)**.** *Consider a sequence of random vectors Z_n in $\mathbb{R}^m$ such that $Z_n \rightsquigarrow Z = \mathcal{N}(0, I_m)$. The elements of the sequence and the limit variable need not be defined on the same probability space. Then*

$$\lim_{n\to\infty} \sup_{R\in\mathcal{R}} |\mathbb{P}(Z_n \in R) - \mathbb{P}(Z \in R)| = 0,$$

where $\mathcal{R}$ is the collection of all convex sets in $\mathbb{R}^m$.

Proof. Let R denote a generic convex set in $\mathbb{R}^m$. Let $R^\epsilon = \{z \in \mathbb{R}^m : d(z, R) \leq \epsilon\}$ and $R^{-\epsilon} = \{z \in R : B(z, \epsilon) \subset R\}$, where d is the Euclidean distance and $B(z, \epsilon) = \{y \in \mathbb{R}^m : d(y, z) \leq \epsilon\}$. The set R^ϵ may be empty. By Theorem 11.3.3 in Dudley (2002), $\epsilon_n := \rho(Z_n, Z) \to 0$, where ρ is the Prohorov metric. The definition of the metric implies that $\mathbb{P}(Z_n \in R) \leq \mathbb{P}(Z \in R^{\epsilon_n}) + \epsilon_n$. By the reverse isoperimetric inequality [Prop 2.5. Chen and Fang (2011)] $|\mathbb{P}(Z \in R^{\epsilon_n}) - \mathbb{P}(Z \in R)| \leq m^{1/2}\epsilon_n$. Hence $\mathbb{P}(Z_n \in R) \leq \mathbb{P}(Z \in R) + \epsilon_n(1 + m^{1/2})$. Furthermore, for any convex set R, $(R^{-\epsilon_n})^{\epsilon_n} \subset R$ (interpreting the expansion of an empty set as an empty set). Hence for any convex R we have $\mathbb{P}(Z \in R^{-\epsilon_n}) \leq \mathbb{P}(Z_n \in R) + \epsilon_n$ by definition of Prohorov's metric. By the reverse isoperimetric inequality $|\mathbb{P}(Z \in R^{-\epsilon_n}) - \mathbb{P}(Z \in R)| \leq m^{1/2}\epsilon_n$. Conclude that $\mathbb{P}(Z_n \in R) \geq \mathbb{P}(Z \in R) - \epsilon_n(1 + m^{1/2})$. ■

References

Belloni, A. and Chernozhukov, V. (2013). 'Least Squares After Model Selection in High-dimensional Sparse Models', *Bernoulli* 19(2), 521–547. ArXiv, 2009.

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

Belloni, Alexandre, Chen, Daniel, Chernozhukov, Victor and Hansen, Christian. (2012). 'Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain', *Econometrica* 80, 2369–2429. Arxiv, 2010.

Belloni, Alexandre, Chernozhukov, Victor, Fernández-Val, Ivan and Hansen, Christian. (2013). 'Program Evaluation with High-Dimensional Data', *arXiv:1311.2645* . ArXiv, 2013.

Belloni, Alexandre, Chernozhukov, Victor and Hansen, Christian. (2010*a*). 'Inference for High-Dimensional Sparse Econometric Models', *Advances in Economics and Econometrics. 10th World Congress of Econometric Society. August 2010* III, 245–295. ArXiv, 2011.

Belloni, Alexandre, Chernozhukov, Victor and Hansen, Christian. (2014). 'Inference on Treatment Effects After Selection Amongst High-Dimensional Controls', *Review of Economic Studies* 81, 608–650. ArXiv, 2011.

Belloni, Alexandre, Chernozhukov, Victor and Hansen, Christian. (ArXiv, 2010*b*), LASSO Methods for Gaussian Instrumental Variables Models. arXiv:1012.1297.

Belloni, Alexandre, Chernozhukov, Victor, Hansen, Christian and Kozbur, Damian. (2014). 'Inference in High Dimensional Panel Models with an Application to Gun Control', *arXiv:1411.6507* . ArXiv, 2014.

Belloni, Alexandre, Chernozhukov, Victor and Kato, Kengo. (2013*a*). 'Robust Inference in Approximately Sparse Quantile Regression Models (with an Application to Malnutrition)', *arXiv preprint arXiv:1312.7186* . ArXiv, 2013.

Belloni, Alexandre, Chernozhukov, Victor and Kato, Kengo. (2013*b*). 'Uniform Post Selection Inference for LAD Regression Models and Other Z-estimation Problems', *arXiv preprint arXiv:1304.0282* . ArXiv, 2013; Oberwolfach, 2012, Luminy, 2012.

Belloni, Alexandre, Chernozhukov, Victor and Wei, Ying. (2013). 'Honest Confidence Regions for Logistic Regression with a Large Number of Controls', *arXiv preprint arXiv:1304.3969* . ArXiv, 2013.

Berk, Richard, Brown, Lawrence, Buja, Andreas, Zhang, Kai and Zhao, Linda. (2013). 'Valid post-selection inference', *Annals of Statistics* 41, 802–837.

Berry, Steven, Levinsohn, James and Pakes, Ariel. (1995). 'Automobile Prices in Market Equilibrium', *Econometrica* 63, 841–890.

Bickel, P. J. (1982). 'On Adaptive Estimation', *Ann. Statist.* 10(3), 647–671.
URL: *http://dx.doi.org/10.1214/aos/1176345863*

Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009). 'Simultaneous analysis of Lasso and Dantzig selector', *Annals of Statistics* 37(4), 1705–1732.

Bühlmann, Peter and van de Geer, Sara. (2011), *Statistics for high-dimensional data: methods, theory and applications*, Springer.

Candès, E. and Tao, T. (2007). 'The Dantzig selector: statistical estimation when p is much larger than n', *Ann. Statist.* 35(6), 2313–2351.

Carrasco, Marine. (2012). 'A Regularization Approach to the Many Instruments Problem', *Journal of Econometrics* 170(2), 383–398.

Carrasco, Marine and Tchuente Nguembu, Guy. (2012), Regularized LIML with Many Instruments, Technical report, University of Montreal Working paper.

Chamberlain, G. (1987). 'Asymptotic Efficiency in Estimation with Conditional Moment Restrictions', *Journal of Econometrics* 34, 305–334.

Chamberlain, G. and Imbens, G. (2004). 'Random Effects Estimators with Many Instrumental Variables', *Econometrica* 72, 295–306.

Chao, John C., Swanson, Norman R., Hausman, Jerry A., Newey, Whitney K. and Woutersen, Tiemen. (2012). 'Asymptotic Distribution of JIVE in a Heteroskedastic IV Regression with Many Instruments', *Econometric Theory* 28(1), 42–86.

Chen, Louis HY and Fang, Xiao. (2011). 'Multivariate Normal Approximation by Stein's Method: The Concentration Inequality Approach', *arXiv preprint arXiv:1111.4073* .

Chen, Xiaohong, Linton, Oliver and Keilegom, Ingred Van. (2003). 'Estimation of semiparametric models when the criterion function is not smooth', *Econometrica* 71, 1591–1608.

Chernozhukov, V. (2009), High-Dimensional Sparse Econometric Models. (Lecture notes) Stats in the Château, https://studies2.hec.fr/jahia/Jahia/statsinthechateau.

Chernozhukov, V., Chetverikov, D. and Kato, K. (2013). 'Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors', *Annals of Statistics* 41, 2786–2819.

Chernozhukov, Victor, Liu, Han, Lu, Junwei and Ning, Yan. (2014), "Statistical Inference in High-Dimensional Sparse Models using Generalized Method of Moments". mimeo, MIT and Princeton.

Dudley, Richard M. (2002), *Real analysis and probability*, Vol. 74, Cambridge University Press.

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

Fan, J. and Lv, J. (2010). 'A selective overview of variable selection in high dimensional feature space', *Statistica Sinica* 20, 101–148.

Farrell, Max. (2013). 'Robust Inference on Average Treatment Effects with Possibly More Covariates than Observations'.

Fithian, William, Sun, Dennis and Taylor, Jonathan. (2014). 'Optimal Inference After Model Selection', *arXiv preprint arXiv:1410.2597v1* . ArXiv, 2014.

Frank, Ildiko E. and Friedman, Jerome H. (1993). 'A Statistical View of Some Chemometrics Regression Tools', *Technometrics* 35(2), 109–135.

Gautier, Eric and Tsybakov, Alexander B. (2011). 'High-Dimensional Instrumental Variables Regression and Confidence Sets', *ArXiv:1105.2454v4* .

Gillen, Benjamin J., Shum, Matthew and Moon, Hyungsik Roger. (2014). 'Demand Estimation with High-Dimensional Product Charateristics', *Advances in Econometrics* . forthcoming.

G'Sell, Max Grazier, Taylor, Jonathan and Tibshirani, Robert. (2013). 'Adaptive testing for the graphical lasso', *arXiv preprint arXiv:1307.4765* .

Hansen, Christian and Kozbur, Damian. (2014). 'Instrumental variables estimation with many weak instruments using regularized JIVE', *Journal of Econometrics* 182, 290–308.

Hastie, Trevor, Tibshirani, Robert and Friedman, Jerome. (2009), *Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York, NY.

Huber, P.J. (1964), The Behavior of Maximum Likelihood Estimates under Nonstandard Conditions. in J. Neyman, ed., Proceedings of the Fifth Berkeley Symposium 1:221-223. Berkeley: University of California Press.

Javanmard, Adel and Montanari, Andrea. (2014). 'Confidence Intervals and Hypothesis Testing for High-Dimensional Regression', *arXiv:1306.3171v2* . ArXiv, 2013.

Jing, Bing-Yi, Shao, Qi-Man and Wang, Qiying. (2003). 'Self-normalized Cramer-type large deviations for independent random variables', *Ann. Probab.* 31(4), 2167–2215.

Kozbur, Damian. (2014), Inference in Nonparametric Models with a High-Dimensional Component. working paper.

Leeb, Hannes and Pötscher, Benedikt M. (2008*a*). 'Recent developments in model selection and related areas', *Econometric Theory* 24(2), 319–322.
URL: *http://dx.doi.org/10.1017/S0266466608080134*

Leeb, Hannes and Pötscher, Benedikt M. (2008*b*). 'Sparse estimators and the oracle property, or the return of Hodges' estimator', *J. Econometrics* 142(1), 201–211.
URL: *http://dx.doi.org/10.1016/j.jeconom.2007.05.017*

Lee, Jason D, Sun, Dennis L, Sun, Yuekai and Taylor, Jonathan E. (2013). 'Exact post-selection inference with the lasso', *arXiv preprint arXiv:1311.6238* .

Lee, Jason D and Taylor, Jonathan E. (2014). 'Exact post model selection inference for marginal screening', *rXiv preprint arXiv:1402.5596* .

Lockhart, Richard, Taylor, Jonathan, Tibshirani, Ryan J and Tibshirani, Robert. (2014). 'A significance test for the lasso (with discussion)', *Annals of Statistics* 42, 413–468.

Loftus, Joshua R and Taylor, Jonathan E. (2014). 'A significance test for forward stepwise model selection', *arXiv preprint arXiv:1405.3920* .

Meinshausen, N. and Yu, B. (2009). 'Lasso-type recovery of sparse representations for high-dimensional data', *Annals of Statistics* 37(1), 2246–2270.

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

Neyman, Jerzy. (1959), Optimal asymptotic tests of composite statistical hypotheses, *in* **U. Grenander**., ed., 'Probability and Statistics, the Harald Cramer Volume', New York, Wiley.

Ning, Yang and Liu, Han. (2014). 'SPARC: Optimal Estimation and Asymptotic Inference under Semiparametric Sparsity', *arXiv preprint arXiv:1412.2295* .

Okui, R. (2010). 'Instrumental Variable Estimation in the Presence of Many Moment Conditions', *forthcoming Journal of Econometrics* .

Pakes, A. and Pollard, David. (1989). 'Simulation and Asymptotics of Optimization Estimators', *Econometrica* 57, 1027–1057.

Robins, James M. and Rotnitzky, Andrea. (1995). 'Semiparametric efficiency in multivariate regression models with missing data', *J. Amer. Statist. Assoc.* 90(429), 122–129.

Rudelson, Mark and Vershynin, Roman. (2008). 'On sparse reconstruction from Fourier and Gaussian measurements', *Communications on Pure and Applied Mathematics* 61(8), 1025–1045.

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

Taylor, Jonathan, Lockhart, Richard, Tibshirani, Ryan J and Tibshirani, Robert. (2014). 'Post-selection adaptive inference for least angle regression and the lasso', *arXiv preprint arXiv:1401.3889* .

Tibshirani, R. (1996). 'Regression shrinkage and selection via the Lasso', *J. Roy. Statist. Soc. Ser. B* 58, 267–288.

van de Geer, Sara, Bühlmann, Peter, Ritov, Ya'acov and Dezeure, Ruben. (2014). 'On Asymptotically Optimal Confidence Regions and Tests for High-Dimensional Models', *Annals of Statistics* 42, 1166–1202. ArXiv, 2013.

van de Geer, S. and Nickl, R. (2013). 'Confidence sets in sparse regression', *Annals of Statistics*

41, 28522876.

van der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge University Press.

Voorman, Arend, Shojaie, Ali and Witten, Daniela. (2014). 'Inference in High Dimensions with the Penalized Score Test', *arXiv preprint arXiv:1401.2678* .

Yang, Zhuoran, Ning, Yang and Liu, Han. (2014). 'On Semiparametric Exponential Family Graphical Models', *arXiv preprint arXiv:1412.8697* .

Zhang, Cun-Hui and Zhang, Stephanie S. (2014). 'Confidence Intervals for Low Dimensional Parameters in High Dimensional Linear Models', *Journal of the Royal Statistical Society: Series B* 76, 217–242. ArXiv, 2011.