Supplementary Material for Post-Selection and Post-Regularization Inference in Linear Models with Very Many Controls and Instruments

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, α, in the presence of a very high-dimensional nuisance parameter, η, which is estimated using modern selection or regularization methods. Our analysis relies on high-level, easy-to-interpret conditions that allow one to clearly see the structures needed for achieving valid post-regularization inference. Simple, readily verifiable sufficient conditions are provided for a class of affine-quadratic models. We rely on asymptotic statements which dramatically simplifies theoretical statements and helps highlight the structure of the problem. We focus our discussion on estimation and inference procedures based on using the empirical analog of theoretical equations

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

which identify α. Within this structure, we show that setting up such equations in a manner such that the orthogonality/immunization condition

$$\partial_\eta M(\alpha, \eta) = 0$$

at the true parameter values is satisfied, coupled with plausible conditions on the smoothness of M and the quality of the estimator $\hat{\eta}$, guarantees that inference for the main parameter α based on testing or point estimation methods discussed below will be regular despite selection or regularization biases occurring in estimation of η. In particular, the estimator of α will often be uniformly consistent at the root-n rate and uniformly asymptotically normal even though estimators $\hat{\eta}$ will generally not be asymptotically linear and regular. The uniformity holds over large classes of models that do not impose highly implausible "beta-min" conditions. We also show that inference can be carried out by inverting tests formed from Neyman's $C(\alpha)$ (orthogonal score) statistics. As an application and an illustration of these ideas, we provide an analysis of post-selection inference in the linear models with many regressors and many instruments. We conclude with a review of other developments in post-selection inference and argue that many of the developments can be viewed as special cases of the general framework of orthogonalized estimating equations.

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

---

Date: December, 2014. We thank 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.

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 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

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

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 $\partial_{x'} f(x) \mid_{x=x_0}$, etc. We use x' to denote the transpose of a column vector x.

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

1.1. 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 the probability law, sometimes called the data-generating process, P_n, 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 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:

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

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.

1.2. **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 assumption on the high-dimensional nuisance parameter.

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 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:

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

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

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

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

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

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 d degrees of freedom:

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

We refer to this statistic as a generalized $C(\alpha)$-statistic, since in likelihood settings the construction above immediately reduces to the classical Neyman's $C(\alpha)$-statistic and the generalized score $S(\alpha_0)$ reduces to Neyman's orthogonalized score; see, e.g. Neyman (1959) and Neyman (1979). We demonstrate these relationships below. Neyman designed his statistic precisely to deal with inference on α_0 when estimation of the nuisance parameters η_0 is crude, which, for example, arises in panel data and other problems with incidental parameters. Here and elsewhere, we are advancing Neyman's ideas to modern (very) high-dimensional problems.

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) and the normality conditions (3) and (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. ■

1.3. **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

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

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 $c > 0$,

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

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

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

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 $n \to \infty$:

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

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. 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 any convex set R *and*

$$V_n = (\Gamma_1'\Gamma_1)^{-1}\Gamma_1'\Omega\Gamma_1(\Gamma_1'\Gamma_1)^{-1},$$

we have that

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

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,* $\mathsf{CR}^l_{1-a} = [l'\hat{\alpha} \pm c(1 - a/2)(l'\hat{V}_n l/n)^{1/2}]$ *where* $c(1 - a/2)$ *is the* $(1 - a/2)$*-quantile of a* $\mathcal{N}(0, 1)$ *is a uniformly valid confidence set for* $l'\alpha_0$*:*

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

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 $\mathrm{A}^2 = \mathrm{Var}(\sqrt{n}\hat{\mathrm{M}}^o(\alpha_0, \eta_0))$ 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 $\mathrm{A} = I$. 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* $\mathrm{A}^2 = \mathrm{Var}(\sqrt{n}\hat{\mathrm{M}}^o(\alpha_0, \eta_0))$ *and* $\hat{\mathrm{A}}^2 - \mathrm{A}^2 \to_{\mathrm{P}_n} 0$*, then the large sample variance* V_n *simplifies to*

$$V_n = (\Gamma_1'\Gamma_1)^{-1}.$$

1.4. **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 $n \to \infty$:

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

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.

2. 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).

2.1. **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 $\eta \mapsto \hat{\Gamma}_j(\eta)$ 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 shall defer the details to the interested reader.

The estimator in this case is

$$\hat{\alpha} = \arg\min_{\alpha\in\mathbb{R}^d} \|\hat{\mathrm{M}}(\alpha,\hat{\eta})\| = -[\hat{\Gamma}_1(\hat{\eta})'\hat{\Gamma}_1(\hat{\eta})]^{-1}\hat{\Gamma}_1'(\hat{\eta})\hat{\Gamma}_2(\hat{\eta}), \tag{22}$$

provided the inverse is well-defined. It follows that

$$\sqrt{n}(\hat{\alpha} - \alpha_0) = -[\hat{\Gamma}_1(\hat{\eta})'\hat{\Gamma}_1(\hat{\eta})]^{-1}\hat{\Gamma}_1(\hat{\eta})\sqrt{n}\hat{\mathrm{M}}(\alpha_0,\hat{\eta}). \tag{23}$$

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

$$\sqrt{n}(\hat{\mathrm{M}}(\alpha_0,\hat{\eta}_0) - \hat{\mathrm{M}}(\alpha_0,\eta_0)) \to_{\mathrm{P}_n} 0, \quad \hat{\Gamma}_1(\hat{\eta}) - \Gamma_1(\eta_0) \to_{\mathrm{P}_n} 0. \tag{24}$$

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

Lemma 1 (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 (24), the asymptotic normality condition (10), and the stability condition (11) hold. Then all the conditions of Propositions 1 and 2 hold. Moreover, the conclusions of Propositions 1 and 2 hold for the estimator $\hat{\alpha}$ in (22).*

2.2. **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 $\hat{\mathrm{M}} = (\hat{\mathrm{M}})_{j=1}^k$ as follows:

$$\sqrt{n}(\hat{\mathrm{M}}_j(\alpha_0,\hat{\eta}) - \hat{\mathrm{M}}_j(\alpha_0,\eta_0)) = T_{1,j} + T_{2,j} + T_{3,j}, \tag{25}$$

where

$$\begin{aligned} T_{1,j} &:= \sqrt{n}\partial_\eta \mathrm{M}_k(\alpha_0,\eta_0)'(\hat{\eta} - \eta_0), \\ T_{2,j} &:= \sqrt{n}(\partial_\eta \hat{\mathrm{M}}_k(\alpha_0,\eta_0) - \partial_\eta \mathrm{M}_k(\alpha_0,\eta_0))'(\hat{\eta} - \eta_0), \\ T_{3,j} &:= \sqrt{n}2^{-1}(\hat{\eta} - \eta_0)'\partial_\eta\partial_{\eta'}\hat{\mathrm{M}}_k(\alpha_0)(\hat{\eta} - \eta_0). \end{aligned} \tag{26}$$

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:

$$\begin{aligned} \|\hat{\eta}\|_0 \lesssim s, \quad & \|\eta_0\|_0 \leq s, \\ \|\hat{\eta} - \eta_0\|_2 \lesssim \sqrt{(s/n)\log(pn)}, \quad & \|\hat{\eta} - \eta_0\|_1 \lesssim \sqrt{(s^2/n)\log(pn)}. \end{aligned} \tag{27}$$

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

$$\bar{T}_{2,j} = \|\sqrt{n}(\partial_{\eta'}\hat{\mathrm{M}}_k(\alpha_0, \eta_0) - \partial_{\eta'}\mathrm{M}_k(\alpha_0, \eta_0))\|_\infty \lesssim_{\mathrm{P}_n} \sqrt{\log(pn)}, \tag{28}$$

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

$$\bar{T}_{3,j} = \|\partial_\eta \partial_{\eta'}\hat{\mathrm{M}}_k(\alpha_0)\|_{\mathsf{sp}(\ell_n s)} \lesssim_{\mathrm{P}_n} 1 \tag{29}$$

where $\ell_n = \log n$.

Following Belloni et al. (2012), we can verify condition (28) 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 (e.g., Rudelson and Zhou (2011)). Indeed, the condition (29) holds if

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

The above analysis immediately implies the following elementary result.

Lemma 2 (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, (ii) that the conditions on sparsity and the quality of estimation (27) hold, and the sparsity index obeys*

$$s^2 \log(pn)^3/n \to 0, \tag{30}$$

(iii) that the moderate deviation bound (28) holds, and (iv) the sparse norm of the second derivatives matrix is bounded as in (29). Then the adaptivity condition (1) holds for the sequence $\{\mathrm{P}_n\}$.

We note that (30) 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 the 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 (27)-(28) if $s^2\log(pn)^3/n \to 0$, then $T_{2,j}$ vanishes in probability, since by Hölder's inequality,

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

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,

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

The conclusion follows from (25). ■

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:

$$\begin{aligned} &\eta_0 = \eta_0^m + \eta_0^r,\ \mathrm{support}(\eta_0^m) \cap \mathrm{support}(\eta_0^r) = \emptyset, \\ &\|\eta_0^m\|_0 \le s,\ \|\eta_0^r\|_2 \le c\sqrt{s/n},\ \|\eta_0^r\|_1 \le c\sqrt{s^2/n}. \end{aligned} \tag{31}$$

This condition allows for much more realistic and richer models than can be accommodated under exact sparsity. For example, under approximate sparsity η_0 need not have *any* zero components at all. In Section 4, we provide an example where (31) arises from a more primitive condition that the absolute values $\{|\eta_{0j}|, j = 1, ..., p\}$, sorted in a decreasing order, decay at a polynomial speed with respect to j.

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

$$\|\hat\eta\|_0 \lesssim s, \|\hat\eta - \eta_0^m\|_2 \lesssim \sqrt{(s/n)\log(pn)}, \quad \|\hat\eta - \eta_0^m\|_1 \lesssim \sqrt{(s^2/n)\log(pn)}. \tag{32}$$

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,

$$\|\hat\eta - \eta_0\|_2 \lesssim \sqrt{(s/n)\log(pn)}, \quad \|\hat\eta - \eta_0\|_1 \lesssim \sqrt{(s^2/n)\log(pn)}. \tag{33}$$

Here 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

$$\|\partial_\eta\partial_{\eta'}\hat{\mathrm{M}}_k(\alpha_0)\|_{\mathsf{pw}(a)} \lesssim_{\mathrm{P}_n} 1. \tag{34}$$

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

Lemma 3 (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*, (ii) that the conditions on approximate sparsity (31) and the quality of estimation (32) hold, and the sparsity index obeys*

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

(iii) that the moderate deviation bound (28) holds, (iv) the sparse norm of the second derivatives matrix is bounded as in (29), and (v) the pointwise norm of the second derivative matrix is bounded as in (34). Then the adaptivity condition (1) holds:

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

2.3. **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,

$$\|\partial_\eta \hat{\Gamma}_{1,ml}(\eta_0) - \partial_\eta \Gamma_{1,ml}(\eta_0)\|_\infty \lesssim_{\mathrm{P}_n} 1, \quad \|\partial_\eta \Gamma_{1,ml}(\eta_0)\|_\infty \lesssim 1. \tag{35}$$

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

$$\|\partial_\eta \partial_{\eta'} \hat{\Gamma}_{1,ml}\|_{\mathsf{sp}(\ell_n s)} + \|\partial_\eta \partial_{\eta'} \hat{\Gamma}_{1,ml}\|_{\mathsf{pw}(a)} \lesssim_{\mathrm{P}_n} 1, \tag{36}$$

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 4 (Elementary Adaptivity for Estimation via Approximate Sparsity)**.** *Consider a sequence of* $\{\mathrm{P}_n\}$ *for which the conditions of the previous lemma hold. In addition assume that the deviation bound (35) holds and the sparse norm and pointwise norms of the second derivatives matrices are stochastically bounded as in (36). Then the adaptivity condition (24) holds for the testing and estimation problem in the affine-quadratic model.*

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 on 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 the 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) among others.

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 (α_0, β_0) obey

$$\mathrm{E}[\partial_\alpha \ell(w_i, \alpha_0, \beta_0)] = 0, \quad \mathrm{E}[\partial_\beta \ell(w_i, \alpha_0, \beta_0)] = 0. \tag{37}$$

Now consider the moment function

$$\mathrm{M}(\alpha, \eta) = \mathrm{E}[\psi(w_i, \alpha, \eta)], \quad \psi(w_i, \alpha, \eta) = \partial_\alpha \ell(w_i, \alpha, \beta) - \mu \partial_\beta \ell(w_i, \alpha, \beta). \tag{38}$$

Here the nuisance parameter is

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

where μ is the *orthogonalization* parameter whose true value μ_0 solves the equation:

$$J_{\alpha\beta} - \mu J_{\beta\beta} = 0 \text{ (i.e., } \mu_0 = J_{\alpha\beta} J_{\beta\beta}^{-1}), \tag{39}$$

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

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

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)[1]

for inference about α_0.

Provided μ_0 is well-defined, we have by (37) that

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

Moreover, the function M has the desired orthogonality property:

$$\partial_{\eta'} \mathrm{M}(\alpha_0, \eta_0) = \Big[J_{\alpha\beta} - \mu_0 J_{\beta\beta};\ \partial_\beta \mathrm{E}[\ell(w_i, \alpha_0, \beta_0)] \Big] = 0. \tag{40}$$

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

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

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

where $\mu_0^* = J^0_{\alpha\beta} J^{0-1}_{\beta\beta}$ is 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 $\partial_\beta \ell(w_i,\gamma_0)$:

$$\partial_\alpha \ell(w_i,\gamma_0) = \mu_0^* \partial_\beta \ell(w_i,\gamma_0) + \varrho, \quad \mathrm{E}[\varrho \partial_\beta \ell(w_i,\gamma_0)'] = 0. \tag{41}$$

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 5 (Neyman's orthogonalization for (quasi-) likelihood scores)**.** *Suppose that for each given $\gamma = (\alpha,\beta) \in \mathcal{A} \times \mathcal{B}$, the derivative $\partial_\gamma \ell(w_i,\gamma)$ exists with probability one. Suppose that condition (37) 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 (40) holds; (ii) if also $\mathrm{E}\sup_{\gamma\in\mathcal{N}} \|\partial_\gamma \ell(w_i,\gamma)\|^2 < \infty$, where $\mathcal{N}$ is an open set containing γ_0, then J^0 exists and is finite. If the information matrix equality holds, namely $J = J^0$, then the orthogonality condition (40) holds for the projection parameter μ_0^* in place of the orthogonalization parameter μ_0.*

Proof. The first claim follows from the computations above and definition of μ_0. The second claim follows from the dominated convergence theorem and the computations given 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 1 hold, we have that

$$C(\alpha) \rightsquigarrow \chi^2(d), \quad V_n^{-1/2}\sqrt{n}(\hat{\alpha} - \alpha_0) \rightsquigarrow \mathcal{N}(0,I), \tag{42}$$

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 (42) also holds for the one-step estimator $\check{\alpha}$ of Section 1 in place of $\hat{\alpha}$, provided the conditions in Section 1 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 μ_0^* by using Lasso or Post-Lasso estimation of the projection equation (41). 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

$$\mathrm{M}(\alpha, \eta) = \mathrm{E}[\psi(w_i, \alpha, \eta)], \quad \psi(w_i, \alpha, \eta) = \mu m(w_i, \alpha, \beta). \tag{43}$$

The nuisance parameter is

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

where μ is the orthogonalization parameter. 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 $\gamma_0 = (\alpha'_0, \beta'_0)'$,

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

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).[2]

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

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:

$$\partial_{\eta'} \mathrm{M}(\alpha_0, \eta)|_{\eta=\eta_0} = [\mu_0 G_\beta, \mathrm{E}[m(w_i, \alpha_0, \beta_0)]] = 0. \tag{44}$$

Estimation and inference on α_0 can be based on the empirical analog of (43):

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

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 1 hold, we have that

$$C(\alpha) \rightsquigarrow_{\mathrm{P}_n} \chi^2(d), \quad V_n^{-1/2}\sqrt{n}(\hat{\alpha} - \alpha) \rightsquigarrow_{\mathrm{P}_n} \mathcal{N}(0, I), \tag{45}$$

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 (45) holds for the one-step estimator $\check{\alpha}$ of Section 1 in place of $\hat{\alpha}$, provided the conditions in Section 1 hold. In particular, 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 $Z \sim \mathcal{N}(h, (G_\gamma' \Omega_m^{-1} G_\gamma)^{-1})$.

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. 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:

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

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

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

Here $x_i = (x_{ij})_{j=1}^{p^x}$ is a p^x-vector of exogenous control variables and $z_i = (z_i)_{i=1}^{p_z}$ is a p^z-vector of instrumental variables.

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, Chernozhukov and Hansen (2014) and Zhang and Zhang (2014). For the latter, we set $p_n^z = 0$ and impose the additional condition $\varepsilon_i \perp d_i$.

We will have n *i.i.d.* draws of

$$w_i = (y_i, d_i', x_i', z_i')', \quad i = 1, ..., n,$$

obeying this system of equations. We also assume that $\mathrm{Var}(w_i)$ is finite throughout so that the model is well defined.

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:

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

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 and a constant $c > 0$, we can decompose η_0 into a sparse component η_0^m and small non-sparse component η_0^r:

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

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

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

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

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

where

$$\psi^a(w_i,\eta) = \Big\{ \sum_{\bar k=1}^{p_d} (d_i - x_i'\vartheta_{\bar k})(x_i'\gamma_k + z_i'\delta_k - x_i'\vartheta_k) \Big\}_{k=1}^{p_d},$$

$$\psi^b(w_i,\eta) = \Big\{ (y_i - x_i'\theta)(x_i'\gamma_k + z_i'\delta_k - x_i'\vartheta_k) \Big\}_{k=1}^{p_d}.$$

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 algorithms, 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_i 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 $\hat\alpha$:

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

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 $\bar d_{ik} := x_i'\gamma_{0k} + z_i'\delta_{0k}$; $v_i = (v_{il})_{l=1}^{p_h} := (\rho_i^y, \rho_i^{d'}, \varrho_i')'$ 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 $\mathrm{E}[f_if_i']$ are bounded from above by C and from below by c. For all j and l, (ii) $\mathrm{E}[h_{il}^2] + \mathrm{E}[|f_{ij}^2\tilde h_{il}^2|] + 1/\mathrm{E}[f_{ij}^2v_{il}^2] \le \mathsf{C}$ and $\mathrm{E}[|f_{ij}^2v_{il}^2|] \le \mathrm{E}[|f_{ij}^2\tilde h_{il}^2|]$, (iii) $\mathrm{E}[|f_{ij}^3v_{il}^3|]^2\log^3(pn)/n \le \delta_n$, 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 $\max_{l,j}|(\mathbb{E}_n - \mathrm{E})[f_{ij}^2v_{il}^2]| + |(\mathbb{E}_n - \mathrm{E})[f_{ij}^2\tilde h_{il}^2]| \le \delta_n$ and (vi) $\|\mathbb{E}_n[f_if_i'] - \mathrm{E}[f_if_i']\|_{\mathsf{sp}(\ell_ns)} \le \delta_n$.*

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)* $\mathrm{E}[|h_{il}|^q] + \mathrm{E}[|v_{il}|^q] \leq \mathsf{C}$, *(ii)* $\mathsf{c} \leq \mathrm{E}[\epsilon_i^2 \mid x_i, z_i] \leq \mathsf{C}$, $\mathsf{c} < \mathrm{E}[\varrho_{ik}^2 \mid x_i, z_i] \leq \mathsf{C}$ *a.s., (iii)* $\sup_{\alpha \in \mathcal{A}} \|\alpha\|_2 \leq \mathsf{C}$.

Under the conditions set forth above, we have the following result on validity of post-selection and post-regularization inference using $C(\alpha)$-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 $C(\alpha_0) \rightsquigarrow \chi^2(d)$. 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 $\chi^2(d)$ 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 any convex set R and

$$V_n = (\Gamma_1')^{-1}\Omega(\Gamma_1)^{-1},$$

we have that

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

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 $\mathcal{N}(0, 1)$, is 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. Implementation Details for Simulation and Empirical Results

In this section, we provide details for the simulation and empirical results reported in the main paper, Chernozhukov et al. (2015).

5.1. **Simulation.**

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

$$\left.\begin{array}{ll} y_i & = \alpha d_i + x_i'\beta + \varepsilon_i \\ d_i & = x_i'\gamma + z_i'\delta + u_i \\ z_i & = \Pi x_i + \zeta_i \end{array}\right| \quad \begin{pmatrix} \varepsilon_i \\ u_i \\ \zeta_i \\ x_i \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} 1 & .6 & 0 & 0 \\ .6 & 1 & 0 & 0 \\ 0 & 0 & .25 I_{p_n^z} & 0 \\ 0 & 0 & 0 & \Sigma \end{pmatrix}\right),$$

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

$\alpha = 1$, $\beta = \gamma$ as p_n^x-vectors with entries $\beta_j = \gamma_j = \frac{1}{j^2}$, δ as a p_n^z-vector with entries $\delta_j = \frac{1}{j^2}$, and

$$\Pi = [I_{p_n^z}\ ,\ 0_{p_n^z \times (p_n^x - p_n^z)}].$$

We report results from four estimators. The results for our proposed 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 on d_i using the single instrument $z_i'\delta$ and the single control $x_i'\beta$. For the results based on stepwise regression, we follow Algorithm 1 from the main text but use stepwise regression with p-value for entry of .05 and p-value for removal of .10 which corresponds to the default decision rule for stepwise regression in many software packages. For the second naive alternative based on non-orthogonal moment conditions, we first do Post-Lasso regression of d on x and z which provides Post-Lasso fitted values for d, $\hat{d}$. We then do Post-Lasso regression of y on x and take the residuals as $\hat{\rho}^y$. We also do Post-Lasso regression of d on x and take these residuals as $\hat{\rho}^d$. The estimator of α is then obtained from the usual IV estimator of $\hat{\rho}^y$ on $\hat{\rho}^d$ using $\hat{d}$ as the instrument. All of the Post-Lasso estimates in the simulation 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. 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.

5.2. **Empirical Example.** To obtain the results in the empirical example, we employ the strategy outlined in Algorithm 1 in Chernozhukov et al. (2015) using Post-Lasso at every step. 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). All standard errors reported in the empirical example are conventional heteroscedasticity robust standard errors for IV estimators.

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 conditions implying 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 (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 1. 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 1 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 and some generalized linear models, and Javanmard and Montanari (2014) (ArXiv 2013) which offers a related, though somewhat different approach. Note also 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.

There is also a complementary, but logically distinct, branch of the recent literature that aims at doing valid inference for the parameters of a "pseudo-true" model that results from the use of a model selection procedure. Specifically, this approach conditions on a model selected by a data-dependent rule and then attempts to do valid inference – conditional on the selection event – for the parameters of the selected model, which may deviate from the "true" model that generated the data. See, for example, Berk et al. (2013), 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) for some recent examples. It seems intellectually very interesting to combine the developments of the present paper (and other preceding papers cited above) with developments in this literature.

There have also been developments on building confidence intervals for high (as opposed to low) 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. Further expanding the set of results for uniform high-dimensional joint inference over a high-dimensional set of target parameters is an interesting topic for future work.

Appendix A. Proofs

A.1. **Proof of Proposition 1.** This is given in the text. ■

A.2. **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 $\text{mineig}(\Gamma_1'\Gamma_1) \geq c$, that

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

Hence it suffices to show that wp $\to$ 1

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

By the triangle inequality, $\|\mathrm{M}(\hat\alpha, \eta_0)\| \leq I_1 + I_2 + I_3$, where

$$I_1 = \|\mathrm{M}(\hat\alpha, \eta_0) - \mathrm{M}(\hat\alpha, \hat\eta)\|, \;\; I_2 = \|\mathrm{M}(\hat\alpha, \hat\eta) - \hat{\mathrm{M}}(\hat\alpha, \hat\eta)\|, \;\; I_3 = \|\hat{\mathrm{M}}(\hat\alpha, \hat\eta)\|.$$

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

$$\inf_{\alpha \in \mathcal{A}} \|\hat{\mathrm{M}}(\alpha, \hat{\eta})\| \leq \|\hat{\mathrm{M}}(\bar{\alpha}, \hat{\eta})\| \lesssim_{\mathrm{P}_n} n^{-1/2}, \tag{56}$$

where $\bar{\alpha}$ is the one-step estimator defined in Step 3, as shown in (57). 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

$$\|\mathrm{M}(\hat{\alpha}, \eta_0)\| \leq II_1 + II_2 + II_3,$$

where $II_1 = \|\mathrm{M}(\hat{\alpha}, \eta_0) - \mathrm{M}(\hat{\alpha}, \hat{\eta})\|$, $II_2 = \|\mathrm{M}(\hat{\alpha}, \hat{\eta}) - \hat{\mathrm{M}}(\hat{\alpha}, \hat{\eta}) - \hat{\mathrm{M}}(\alpha_0, \eta_0)\|$, $II_3 = \|\hat{\mathrm{M}}(\hat{\alpha}, \hat{\eta})\| + \|\hat{\mathrm{M}}(\alpha_0, \eta_0)\|$. Then, by the orthogonality $\partial_{\eta'}\mathrm{M}(\alpha_0, \eta_0) = 0$ and condition (14), wp $\to 1$,

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

Then, by condition (13),

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

Since $II_3 \lesssim_{\mathrm{P}_n} n^{-1/2}$ by (56), 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

$$\|\hat{\mathrm{M}}(\hat{\alpha}, \hat{\eta}) - \widehat{\mathrm{L}}(\hat{\alpha})\| \leq III_1 + III_2 + III_3,$$

where $III_1 = \|\mathrm{M}(\hat{\alpha}, \hat{\eta}) - \mathrm{M}(\hat{\alpha}, \eta_0)\|$, $III_2 = \|\mathrm{M}(\hat{\alpha}, \eta_0) - \Gamma_1(\hat{\alpha} - \alpha_0)\|$, $III_3 = \|\hat{\mathrm{M}}(\hat{\alpha}, \hat{\eta}) - \mathrm{M}(\hat{\alpha}, \hat{\eta}) - \hat{\mathrm{M}}(\alpha_0, \eta_0)\|$. 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 $\|\widehat{\mathrm{L}}(\bar{\alpha})\| \lesssim_{\mathrm{P}_n} n^{-1/2}$,

$$\|\hat{\mathrm{M}}(\bar{\alpha}, \hat{\eta})\| \lesssim_{\mathrm{P}_n} n^{-1/2}. \tag{57}$$

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 the Portmanteau Theorem. ■

A.3. **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

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

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

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

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

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

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

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

A.5. **Proof of Lemma 1.** We 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 (24) 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 (24) 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 1 and 2, follow from the conclusions regarding the uniform validity of inference using $\bar{\alpha}$, which follow from the Continuous Mapping Theorem, the Portmanteau Theorem, and the assumed stability conditions (11). This establishes the second claim of the Lemma.

Regarding the first claim, verification of the conditions of Propositions 1 and 2 is left for the interested reader. Having proven the first claim, we could have derived the second claim as a corollary of Propositions 1 and 2, but instead we proved it above directly. ■

A.6. **Proof of Lemma 2.** This is given in the main text. ■

A.7. **Proof of Lemma 3.** As in the proof of Lemma 2, we can expand:

$$\sqrt{n}(\hat{\mathrm{M}}_j(\alpha_0,\hat{\eta}) - \hat{\mathrm{M}}_j(\alpha_0,\eta_0)) = T_{1,j} + T_{2,j} + T_{3,j}, \tag{58}$$

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:

$$\begin{aligned} T_{3,j} &\leq T_{3,j}^m + T_{4,j}, \\ T_{3,j}^m &:= \sqrt{n}|(\hat{\eta}-\eta_0^m)'\partial_\eta\partial_{\eta'}\hat{\mathrm{M}}_k(\alpha_0)(\hat{\eta}-\eta_0^m)|, \\ T_{4,j} &:= \sqrt{n}|\eta_0^{r\prime}\partial_\eta\partial_{\eta'}\hat{\mathrm{M}}_k(\alpha_0)\eta_0^r|. \end{aligned} \tag{59}$$

Then $T_{1,j} = 0$ by orthogonality, $T_{2,j} \to_{\mathrm{P}_n} 0$ as in the proof of Lemma 2. 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 (34),

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

The conclusion follows from (58). ■

A.8. **Proof of Lemma 4.** 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:

$$|\hat{\Gamma}_{1,ml}(\hat{\eta}) - \hat{\Gamma}_{1,ml}(\eta_0)| \leq T_{1,ml} + T_{2,ml} + T_{3,ml},$$

where

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

Under conditions (35) and (36) 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. ■

A.9. **Proof of Propositions 5.** We present the proofs for the case of $p_d = 1$; the general case follows similarly.

We proceed to verify the assumptions of Lemma 3 and 4, from which the desired result follows from Propositions 1 and 2 and Lemma 1. In what follows, we consider an arbitrary sequence $\{\mathrm{P}_n\}$ in $\{\mathbf{P}_n\}$.

Step 1. (Performance bounds for $\hat{\eta}$). We noted that Condition AS.1 implies the decomposition (50). A straightforward modification of the proofs of Belloni et al. (2012) yields the following performance bounds for estimator $\hat{\eta}$ of η_0: wp $\to 1$,

$$\|\hat{\eta}\|_0 \lesssim s, \|\hat{\eta} - \eta_0^m\|_2 \lesssim \sqrt{(s/n)\log(pn)}, \quad \|\hat{\eta} - \eta_0^m\|_1 \lesssim \sqrt{(s^2/n)\log(pn)}. \tag{60}$$

Note that modification concerns the use of the assumption that errors are not correlated with predictive regressors instead of mean independence. Also the third step of the algorithm requires regressing an estimated response variable on the predictive regressors; the estimation error in the response variable can be treated as approximation errors in Belloni et al. (2012)'s proofs; we omit the details for brevity.

Step 2. (Preparation). It will also be convenient to lift the nuisance parameter η into a higher dimension and redefine the signs of its components as follows:

$$\eta := (\eta_1', \eta_2', \eta_3', \eta_4', \eta_5')' := [-\theta', -\vartheta', \gamma', \delta', -\vartheta']'.$$

With this re-definition, we have

$$\psi(w_i, \alpha, \eta) = \{(y_i + x_i'\eta_1) + (d_i + x_i'\eta_2)\alpha\}\{x_i'\eta_3 + z_i'\eta_4 + x_i'\eta_5\}.$$

Note also that

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

$$\Gamma_1(\eta) = \mathrm{E}[\partial_\alpha \psi(w_i, \alpha, \eta)], \quad \hat{\Gamma}_1(\eta) = \mathbb{E}_n[\partial_\alpha \psi(w_i, \alpha, \eta)].$$

We compute the following partial derivatives:

$$\begin{aligned}
&\partial_\eta \psi(w_i) := \partial_\eta \psi(w_i, \alpha_0, \eta_0) = [x_i'\varrho_i, \alpha_0 x_i'\varrho_i, x_i'\rho_i^y, z_i'\rho_i^y, x_i'\rho_i^y],\\
&\partial_\alpha \psi(w_i, \alpha, \eta) = \{d_i + x_i'\eta_2\}\{x_i'\eta_3 + z_i'\eta_4 + x_i'\eta_5\},\\
&\partial_\alpha \psi(w_i) := \partial_\alpha \psi(w_i, \alpha_0, \eta_0) = \rho_i^d \varrho_i,\\
&\partial_{\eta'}\partial_\alpha \psi(w_i) := \partial_{\eta'}\partial_\alpha \psi(w_i, \alpha_0, \eta_0) = [0, x_i'\varrho_i, x_i'\rho_i^d, z_i'\rho_i^d, x_i'\rho_i^d]',
\end{aligned}$$

$$\partial_\eta \partial_{\eta'} \psi(w_i, \alpha, \eta) = \begin{bmatrix} 0 & 0 & x_i x_i' & x_i z_i' & x_i x_i' \\ 0 & 0 & \alpha x_i x_i' & \alpha x_i z_i' & \alpha x_i x_i' \\ x_i x_i' & \alpha x_i x_i' & 0 & 0 & 0 \\ z_i x_i' & \alpha z_i x_i' & 0 & 0 & 0 \\ x_i x_i' & \alpha x_i x_i' & 0 & 0 & 0 \end{bmatrix},$$

$$\partial_\eta \partial_{\eta'} \partial_\alpha \psi(w_i, \alpha, \eta) = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & x_i x_i' & x_i z_i' & x_i x_i' \\ 0 & x_i x_i' & 0 & 0 & 0 \\ 0 & z_i x_i' & 0 & 0 & 0 \\ 0 & x_i x_i' & 0 & 0 & 0 \end{bmatrix}.$$

Step 3. (Verification of Conditions of Lemma 3).

Application of Lemma 6, condition $\|\alpha_0\| \leq \mathsf{C}$ holding by Condition SM, and Condition RF, yields that wp $\to 1$,

$$\sqrt{n}\|\partial_\eta \mathrm{M}(\alpha_0, \eta_0) - \partial_\eta \hat{\mathrm{M}}(\alpha_0, \eta_0)\|_\infty = \|\sqrt{n}(\mathbb{E}_n - \mathrm{E})\partial_\eta \psi(w_i)\|_\infty \lesssim$$
$$\lesssim \max_j \sqrt{[\mathbb{E}_n[(\partial_\eta \psi(w_i))_j^2]}\sqrt{\log(pn)} \lesssim_{\mathrm{P}_n} \sqrt{\log(pn)}.$$

Application of the triangle inequality, of condition $\|\alpha_0\| \leq \mathsf{C}$, and Condition RF yields:

$$\|\partial_\eta \partial_{\eta'} \hat{\mathrm{M}}(\alpha_0, \eta_0)\|_{\mathsf{sp}(\ell_n s)} \leq C \|\partial_\eta \partial_{\eta'} \mathbb{E}_n[f_i f_i]\|_{\mathsf{sp}(\ell_n s)} \lesssim_{\mathrm{P}_n} 1,$$

where C depends on C.

Moreover, application of the triangle inequality and the Markov inequality yields, for any deterministic $a \neq 0$,

$$\|\partial_\eta \partial_{\eta'} \hat{\mathrm{M}}(\alpha_0, \eta_0)\|_{\mathsf{pw}(a)} \leq C \|\partial_\eta \partial_{\eta'} \mathbb{E}_n[f_i f_i]\|_{\mathsf{pw}(a)} \lesssim_{\mathrm{P}_n} 1,$$

where C depends on C.

We have by Condition SM and the law of iterated expectations:

$$\Omega = \mathrm{E}[\psi^2(w_i, \alpha_0, \eta_0)] = \mathrm{E}[\epsilon_i^2 \varrho_i^2] \in \mathrm{E}[\varrho_i^2] \cdot [\mathsf{c}, \mathsf{C}] \in [\mathsf{c}^2, \mathsf{C}^2],$$
$$\mathrm{E}[\psi^{\mathsf{q}/2}(w_i, \alpha_0, \eta_0)] \leq \mathrm{E}[|\epsilon_i \varrho_i|^{\mathsf{q}/2}] \leq \sqrt{\mathrm{E}[|\epsilon_i|^{\mathsf{q}}]}\sqrt{\mathrm{E}[|\varrho_i|^{\mathsf{q}}]} \leq \mathsf{C}.$$

Application of Lyapunov's Central Limit Theorem yields,

$$\Omega^{-1/2}\hat{\mathrm{M}}(\alpha_0, \eta_0) \rightsquigarrow \mathcal{N}(0, 1).$$

Next, $\hat{\Omega}(\alpha_0) = \mathbb{E}_n[\psi^2(w_i, \alpha_0, \hat{\eta})]$ is consistent for Ω. The proof of this result follows similarly to the (slightly more difficult) proof of consistency of $\hat{\Omega} = \mathbb{E}_n[\psi^2(w_i, \hat{\alpha}, \hat{\eta})]$ for Ω, which is given below.

All conditions of Lemma 3 are now verified.

Step 4. (Verification of Conditions of Lemma 4).

Application of Lemma 6 and Conditions RF yields that with probability $1 - o(1)$,

$$\sqrt{n}\|\partial_\eta \hat{\Gamma}_1(\alpha_0, \eta_0) - \partial_\eta \Gamma_1(\alpha_0, \eta_0)\|_\infty = \|\sqrt{n}(\mathbb{E}_n - \mathrm{E})\partial_\eta \partial_\alpha \psi(w_i)\|_\infty \lesssim$$
$$\lesssim \max_j \sqrt{[\mathbb{E}_n[(\partial_\eta \partial_\alpha \psi(w_i))_j^2]}\sqrt{\log(pn)} \lesssim_{\mathrm{P}_n} \sqrt{\log(pn)}.$$

Application of the triangle inequalities and Condition RF yields:

$$\|\partial_\eta \partial_{\eta'} \hat{\Gamma}_1(\alpha_0, \eta_0)\|_{\mathsf{sp}(\ell_n s)} \lesssim \|\partial_\eta \partial_{\eta'} \mathbb{E}_n[f_i f_i\|_{\mathsf{sp}(\ell_n s)} \lesssim_{\mathrm{P}_n} 1.$$

Moreover, application of the triangle inequalities and the Markov inequality yields, for any deterministic $a \neq 0$,

$$\|\partial_\eta \partial_{\eta'} \hat{\Gamma}_1(\alpha_0, \eta_0)\|_{\mathsf{pw}(a)} \lesssim \|\partial_\eta \partial_{\eta'} \mathbb{E}_n[f_i f_i\|_{\mathsf{pw}(a)} \lesssim_{\mathrm{P}_n} 1.$$

Next, by Condition SM we have

$$\Gamma_1 = \mathrm{E}[\rho_i^d \varrho_i] = \mathrm{E}[\varrho_i^2] \in [\mathsf{c}, \mathsf{C}].$$

By Conditon SM we have

$$\begin{aligned}
\|\partial_\eta \Gamma_1(\alpha_0, \eta_0)\|_\infty = \|\mathrm{E}[\partial_{\eta'} \partial_\alpha \psi(w_i)]\|_\infty &\leq \max_j \left(\mathrm{E}[|f_{ij} \varrho_i|] \vee \mathrm{E}[|f_{ij} \rho_i^d|] \right) \\
&\leq \max_j \left(\sqrt{[\mathrm{E}[|f_{ij} \varrho_i|^2]} \vee \sqrt{|\mathrm{E}[|f_{ij} \rho_i^d|^2]} \right) \leq \sqrt{\mathsf{C}}.
\end{aligned}$$

This, as well as previous steps verify conditions of the Lemma 4, which are sufficient to establish that $|\hat{\alpha} - \alpha_0| \lesssim_{\mathrm{P}_n} n^{-1/2}$, which is needed in the last step below.

Next, we show consistency $\hat{V}_n - V_n \to_{\mathrm{P}_n} 0$. Given the stability conditions established above, this follows from $\hat{\Gamma}_1(\hat{\eta}) - \Gamma_1 \to_{\mathrm{P}_n} 0$, which follows from Lemma 4, and from the consistency: $\hat{\Omega} - \Omega \to_{\mathrm{P}_n} 0$. Recall that $\hat{\Omega} = \mathbb{E}_n[\psi^2(w_i, \hat{\alpha}, \hat{\eta})]$ and let $\hat{\Omega}_0 = \mathbb{E}_n[\psi^2(w_i, \alpha_0, \eta_0)]$. Since $\hat{\Omega}_0 - \Omega \to_{\mathrm{P}_n} 0$ by Markov inequality, it suffices to show that $\hat{\Omega} - \Omega_0 \to_{\mathrm{P}_n} 0$. Since $\hat{\Omega} - \Omega_0 = (\sqrt{\hat{\Omega}} - \sqrt{\hat{\Omega}_0})(\sqrt{\hat{\Omega}} + \sqrt{\hat{\Omega}_0})$, it suffices to show that $(\sqrt{\hat{\Omega}} - \sqrt{\hat{\Omega}_0}) \to_{\mathrm{P}_n} 0$. By the triangle inequality and some simple calculations, we have

$$|\sqrt{\hat{\Omega}} - \sqrt{\hat{\Omega}_0}| \lesssim \mathsf{D} := I_2 I_\infty \sqrt{\mathbb{E}_n[\varrho_i^4]} + I_2 I_\infty II_2 II_\infty + II_2 II_\infty \sqrt{\mathbb{E}_n[\epsilon_i^4]},$$

where the terms are defined below. Let

$$\begin{aligned}
&\hat{\epsilon}_i = \hat{\rho}_i^y - \hat{\rho}_i^d \hat{\alpha}, \quad \epsilon_i = \rho_i^y - \rho_i^d \alpha, \\
&\hat{\varrho}_i = x_i' \hat{\gamma} + z_i' \hat{\delta} - x_i' \hat{\vartheta}, \quad \varrho_i = x_i' \gamma_0 + z_i' \delta_0 - x_i' \vartheta_0.
\end{aligned}$$

Then

$$\begin{aligned}
&|\hat{\epsilon}_i - \epsilon_i| \leq |x_i'(\theta_0 - \hat{\theta})| + |\rho_i^d(\alpha_0 - \hat{\alpha})| + |x_i'(\hat{\vartheta} - \vartheta_0)\alpha_0| + |x_i'(\hat{\vartheta} - \vartheta_0)(\hat{\alpha} - \alpha_0)|, \\
&|\hat{\varrho}_i - \varrho_i| \leq |z_i'(\hat{\delta} - \delta_0)| + |x_i'(\hat{\gamma} - \gamma_0)| + |x_i'(\hat{\vartheta} - \vartheta_0)|.
\end{aligned}$$

Then the terms $I_2, I_\infty, II_2, II_\infty$ are defined and bounded, using elementary inequalities and Condition RF, as follows:

$$\begin{aligned}
I_2 &:= \sqrt{\mathbb{E}_n[(\hat{\epsilon}_i - \epsilon_i)^2]} \lesssim_{\mathrm{P}_n} \sqrt{s \log(pn)/n} + \sqrt{\mathbb{E}_n[\rho_i^{d2}]} |\hat{\alpha} - \alpha_0| \\
&\quad + \sqrt{s \log(pn)/n} |\hat{\alpha} - \alpha_0| \to_{\mathrm{P}_n} 0, \\
I_\infty &:= \max_{i \leq n} |\hat{\epsilon}_i - \epsilon_i| \lesssim \max_{ij} |f_{ij}| \sqrt{s^2 \log(pn)/n} + \max_{i \leq n} |\rho_i^d| |\hat{\alpha} - \alpha_0| \\
&\quad + \max_{ij} |f_{ij}| \sqrt{s^2 \log(pn)/n} |\hat{\alpha} - \alpha_0| \to_{\mathrm{P}_n} 0, \\
II_2 &:= \sqrt{\mathbb{E}_n[(\hat{\varrho}_i - \varrho_i)^2]} \lesssim_{\mathrm{P}_n} \sqrt{s \log(pn)/n} \to 0, \\
II_\infty &:= \max_{i \leq n} |\hat{\varrho}_i - \varrho_i| \lesssim \max_{ij} |f_{ij}| \sqrt{s^2 \log(pn)/n} \to_{\mathrm{P}_n} 0,
\end{aligned}$$

where we have used the relations, $|\hat{\alpha}-\alpha_0| \lesssim_{\mathrm{P}_n} n^{-1/2}$, and $\mathbb{E}_n[|\rho_i^d|^2] \lesssim_{\mathrm{P}_n} 1$, $\max_{i\leq n} \rho_i^d \lesssim_{\mathrm{P}_n} n^{1/\mathsf{q}}$, for $\mathsf{q} > 4$, holding by Condition SM, and we have used the fact that

$$\mathbb{E}_n[(x_i'\{\hat{\theta}-\theta_0\})^2 + (x_i'\{\hat{\vartheta}-\vartheta_0\})^2 + (x_i'\{\hat{\gamma}-\gamma_0\})^2 + (z_i'\{\hat{\delta}-\delta_0\})^2] \lesssim_{\mathrm{P}_n} \frac{s\log(pn)}{n}.$$

The latter follows from the following argument, for example, wp $\to 1$,

$$\begin{aligned}
&\mathbb{E}_n[(x_i'\{\hat{\theta}-\theta_0\})^2] \leq 2\mathbb{E}_n[(x_i'\{\hat{\theta}-\theta_0^m\})^2] + 2\mathbb{E}_n[(x_i'\theta_0^r)^2]\\
&\lesssim \|\mathbb{E}_n[f_if_i']\|_{\mathsf{sp}(\ell_n s)}\|\hat{\theta}-\theta_0^m\|^2 + \|\mathbb{E}_n[f_if_i']\|_{\mathsf{pw}(\theta_0^r)}\|\theta_0^r\|^2\\
&\lesssim_{\mathrm{P}_n} s\log(pn)/n,
\end{aligned}$$

since wp $\to 1$ $\|\hat{\theta}-\theta_0^m\|_0 \leq \ell_n s$, $\|\hat{\theta}-\theta_0^m\|^2 \lesssim s\log(pn)/n$, $\|\theta_0^r\|^2 \lesssim s/n$, by Step 1 and Condition AS.1 (see decomposition (50)), and $\|\mathbb{E}_n[f_if_i']\|_{\mathsf{sp}(\ell_n s)} \lesssim_{\mathrm{P}_n} 1$ holding by Condition RF and $\|\mathbb{E}_n[f_if_i']\|_{\mathsf{pw}(\theta_0^r)} \lesssim_{\mathrm{P}_n} 1$ holding by Markov inequality and Condition RF.

Since $\mathbb{E}_n[\varrho_i^4] + \mathbb{E}_n[\epsilon_i^4] \lesssim_{\mathrm{P}_n} 1$ by Condition SM, we conclude that $\mathsf{D} \to_{\mathrm{P}_n} 0$. ■

Appendix B. Some Tools

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

Lemma 6 (Moderate Deviation Inequality for Maximum of a Vector)**.** *Suppose that*

$$\mathcal{S}_j = \frac{\sum_{i=1}^n U_{ij}}{\sqrt{\sum_{i=1}^n U_{ij}^2}},$$

where U_{ij} are independent variables across i with mean zero. We have that

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

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.

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, 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.

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

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

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, Hansen, Christian and Spindler, Martin. (2015). 'Valid Post-Selection and Post-Regularization Inference In Linear Models with Many Controls and Instruments', *American Economic Review: Papers and Proceedings* .

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.

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.

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

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

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* .

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* .

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, 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* .

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.