Robust Estimation Under Many Instruments Asymptotics

Mikkel Sølvsten*

January 11, 2017

Abstract

This paper considers a new class of robust estimators in a linear instrumental variables (IV) model with many instruments. The estimators are generalized method of moments (GMM) estimators, and the class includes the limited maximum likelihood estimator (LIML) as a special case. Each estimator in the class is consistent and asymptotically normal under many instruments asymptotics, and this paper provides consistent variance estimators that are of the "sandwich" type and can be used to conduct asymptotically correct inference. Furthermore, this paper characterizes an optimal robust estimator among the members of the class. Compared to LIML, the optimal robust estimator is less influenced by outliers and more efficient under thick-tailed error distributions. In an empirical example (Angrist and Krueger, 1991), the optimal robust estimator is approximately 80% more efficient than LIML.

Keywords: robust estimation, many instruments asymptotics, efficiency, generalized method of moments.

1 Introduction

The focus of this paper is robust estimation of the structural coefficient in a linear IV model with many instruments. Models with many instrumental variables have received consider-

* *Department of Economics, UC Berkeley (mikkel@econ.berkeley.edu).* I am grateful to Michael Jansson, Jim Powell, Demian Pouzo, and Noureddine El Karoui for valuable advice, and thank seminar participants at UC Berkeley for helpful comments.

able attention in recent years as estimators using a large number of instruments can be more accurate than estimators using a small number of instruments. It is well understood that inference based on standard asymptotics may lead to incorrect confidence intervals when an estimator uses multiple instruments. Inference based on many instruments asymptotics—where the number of instruments and the sample size grow at the same rate—tends to correct this problem. Under many instruments asymptotics, the two-stage least squares estimator (2SLS) is inconsistent, whereas LIML is consistent and asymptotically normal (see, e.g., Kunitomo, 1980; Morimune, 1983; Bekker, 1994).

LIML is essentially an optimal estimator when the structural and reduced form errors have a joint normal distribution (Chioda and Jansson, 2009), but normality may not be a good approximation to all economic data. To illustrate, this paper takes the model and data from Angrist and Krueger (1991) and documents that the structural error distribution is well-approximated by a normal distribution that has been contaminated with gross outliers. The accuracy of LIML could be negatively affected by the presence of outliers, which makes it important to understand if alternative estimators can be more efficient when the errors are nonnormal.

The first contribution of this paper is to propose a new class of estimators of the structural coefficient in a linear IV model and to show that each member of the class is consistent and asymptotically normal at the usual parametric rate under many instruments asymptotics. Additionally, this paper characterizes both an optimal and an optimal robust estimator within this class. The optimal estimator minimizes asymptotic variance when the shape of the joint error distribution is known (e.g., normal), whereas the optimal robust estimator minimizes the maximal asymptotic variance over a neighborhood of contaminated normal distributions, i.e., a mixture between the standard normal distribution (with high probability) and some unknown contaminating distribution (with low probability). Each estimator of the structural coefficient is an element of a just-identified GMM estimator which corresponds to a vector of sample moments indexed by two functions of the structural residuals. When both functions are linear, the ensuing estimator is LIML. When the first function is proportional to the score of the structural errors and the second is proportional to the conditional mean of the reduced form error given the structural error, either of which may be nonlinear, then the ensuing estimator is optimal. When both functions censor the structural residuals at some data-dependent level, the ensuing estimator is optimal robust.

These contributions add to a growing literature on many instruments asymptotics that started with Kunitomo (1980) and Morimune (1983), who derived asymptotic variances that are larger than the usual IV formulas and depend on the number of instruments. Bekker (1994) provided consistent estimators of these larger variances under normal errors and Hansen, Hausman, and Newey (2008) extended the variance formulas and estimators to allow for nonnormal errors.[1] This paper expands the class of asymptotically normal estimators to include robust alternatives to LIML and provides formulas for their asymptotic variances that are natural extensions of the existing formulas. In addition, this paper provides consistent variance estimators that are of the "sandwich" type, where the outer matrix is the inverse Jacobian of the sample moments and the inner matrix is a sample average of outer products of some moment function.

This paper also adds to the literature on optimal and robust estimation in the linear IV model. Anderson, Kunitomo, and Matsushita (2010) showed optimality of LIML among estimators that are functions of the sufficient statistics from the normal model, and under normality of the errors, Chioda and Jansson (2009) showed optimality of LIML among estimators that are invariant to rotations of the sufficient statistics from the normal model. The optimality results of this paper are complementary to the existing literature, as they imply optimality of LIML under normal errors, but for a different class of estimators than previously considered. However, they also bring a new perspective to these results by presenting estimators that are robust and more efficient than LIML under nonnormal errors and many instruments. In models with a fixed number of instruments, the two-stage least absolute deviations estimator (Amemiya, 1982; Powell, 1983), the resistant esimator of Krasker and Welsch (1985), the two-stage quantiles and two-stage trimmed least squares estimators (Chen and Portnoy, 1996), the IV quantile regression estimator (Chernozhukov and Hansen, 2006), the robust estimators of Honoré and Hu (2004), the nonlinear IV estimators of Hansen, McDonald, and Newey (2010), and the adaptive estimator of Cattaneo, Crump, and Jansson (2012) are also examples of robust estimators or estimators that can be

[1]Some additional papers that consider estimation and inference with many or many weak instruments are Hahn (2002); Hahn and Hausman (2002); Chamberlain and Imbens (2004); Chao and Swanson (2005); Chao, Swanson, Hausman, Newey, and Woutersen (2012); Hausman, Newey, Woutersen, Chao, and Swanson (2012); Hansen and Kozbur (2014); Kolesár (2015); Wang and Kaffo (2016). See also Newey (1990); Belloni, Chen, Chernozhukov, and Hansen (2012) and Kolesár (2013) for estimation in related models.

more efficient than LIML under nonnormal errors. However, I am unaware of papers that establish consistency or asymptotic normality of these estimators under many instruments asymptotics.

The notion of robustness this paper adopts is similar to the one used by Huber (1964), and defines an estimator as robust if its asymptotic variance remains finite when the distribution of the structural errors ranges over a neighborhood of contaminated normal distributions. The optimal robust estimator derived here treats the structural residuals of the IV model the same way as Huber's most robust (or minimax) estimator treats the residuals of the regression model (Huber, 1964, 1973, 1981).

Finally, this paper makes an additional contribution of potential independent interest. The contribution is to give high-level conditions for asymptotic normality of a single element of a just-identified GMM estimator whose dimension grows at the same rate as the sample size. Following Huber (1967), there have been numerous papers giving high-level conditions in GMM setups. See, e.g., Hansen (1982); Pakes and Pollard (1989); Andrews (1994); Newey (1994); Newey and McFadden (1994); Ai and Chen (2003); Chen, Linton, and Van Keilegom (2003); Chen (2007); Newey and Windmeijer (2009). These papers cover cases of smooth and non-smooth objective functions and parametric and semi-parametric estimators, but all of them rely on an intermediate result of consistency of the estimator for some pseudo-true, non-random value. In contrast, this paper allows for the estimator to have a random, sample-dependent "limit." This is a necessary extension, as the reduced form parameters in this setup do not settle down around some non-random value under many instruments asymptotics. This paper presents results for smooth and non-smooth objective functions, and verifies the high-level conditions for examples that are differentiable or Lipschitz continuous.

The remainder of this paper is organized as follows. Section 2 defines the model and describes the class of estimators and the associated asymptotic variance estimators for a simplified version of the model. Section 3 gives high-level conditions for asymptotic normality of a single element of a GMM estimator whose dimension grows at the same rate as the sample size. Section 4 shows that each estimator in the class is asymptotically normal and characterizes optimal and robust estimators. Section 5 describes the class of estimators in the generality of the full model. Section 6 presents simulation results, and section 7 applies some of the estimators to the empirical example provided by Angrist and

Krueger (1991). Section 8 concludes. Proofs are in the appendices.

Notation

For a vector v, let $\|v\| = \sqrt{v'v}$ be the Euclidean norm. For a symmetric matrix A, let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ be the smallest and largest eigenvalues of A. For any matrix A, let $\|A\| = \lambda_{\max}(A'A)^{1/2}$ be the largest singular value of A, and let $\sigma_{\min}(A) = \lambda_{\min}(A'A)^{1/2}$ be the smallest singular value of A. $\|A\|$ is an operator norm (induced by Euclidean norms) and $\sigma_{\min}(A) > 0$ if and only if $A'A$ is invertible. For any absolutely continuous function $f : \mathbb{R} \to \mathbb{R}$, let f' be a derivative of f when it exists and zero otherwise. Let $\{a_{ni}\}_{i,n}$ be shorthand for $\{a_{ni} : i \in \{1, \cdots, n\}, n \in \mathbb{N}\}$, and let $\{a_{ni}\}_i$ be shorthand for $\{a_{ni} : i \in \{1, \cdots, n\}\}$. Let Φ be the distribution function of the standard normal distribution. Limits are considered as $n \to \infty$ unless otherwise noted.

2 Model and Estimators

This paper considers a linear model with two endogenous variables and k_n instrumental variables. The model consists of a structural and a reduced form equation given by

$$y_{in} = x_{in}\beta_0 + w_i'\delta_0 + \varepsilon_i \tag{1a}$$

$$x_{in} = z_{in}'\pi_{0n} + w_i'\eta_0 + u_i \qquad (i = 1, \cdots, n) \tag{1b}$$

where n is the sample size, $y_{in}, x_{in} \in \mathbb{R}$ are endogenous variables, $w_i \in \mathbb{R}^G$ is a vector of included exogenous variables that includes a constant term, $z_{in} \in \mathbb{R}^{k_n}$ is a vector of instruments, $\varepsilon_i, u_i \in \mathbb{R}$ are unobserved stochastic errors, $\beta_0 \in \mathbb{R}$ is the parameter of interest, and $(\pi_{0n}, \delta_0, \eta_0) \in \mathbb{R}^{k_n+2G}$ is a nuisance parameter. The model involves potentially many instruments as

$$\lim_{n\to\infty} \frac{k_n}{n} = \alpha \in [0, 1).$$

The special case $\alpha = 0$ includes standard asymptotics where k_n is fixed as $n \to \infty$. To simplify notation, drop the subscript n on y_i, x_i, z_i, k, and π_0.

Let the matrices Z and W denote the stacked observations of z_i' and w_i', and assume that

(Z, W) has full rank. Residualize and normalize Z so that $n^{-1}Z'Z = I_k$ and $n^{-1}Z'W = 0$.[2] This transformation of (Z, W) simplifies some definitions and proofs and does not affect the asymptotic distribution of the estimators of β_0 considered in this paper. The errors (ε_i, u_i) are assumed *i.i.d.* and independent of (Z, W). The reduced form error, u_i, has finite variance, the intercept in (1b) is normalized so that $\mathbb{E}[u_i] = 0$, and the structural error, ε_i, has a bounded and absolutely continuous density f with nonzero and finite Fisher information, $\mathcal{I}_f = \mathbb{E}[(f'/f)^2(\varepsilon_i)]$. The instruments are jointly strong in the sense that the conditional covariance between x_i and $z_i'\pi_0$ is bounded away from zero, i.e.,

$$\sigma_{xz} = \tfrac{1}{n}\sum_i \mathbb{E}\left[x_i z_i'\pi_0 \mid Z\right] = \tfrac{1}{n}\pi_0' Z'Z\pi_0 = \|\pi_0\|^2 > c + o_p(1) \text{ for some } c > 0.$$

To simplify the exposition, the rest of this section presents the class of estimators when the nuisance parameters (δ_0, η_0) are known and normalized to equal zero, while section 5 treats the case where (δ_0, η_0) are unknown.

The class of estimators this paper considers is indexed by two Lipschitz continuous functions ϕ and ψ, and the estimators take values in a parameter space $\Theta_n \subset \mathbb{R}^{k+2}$ with a quickly growing dimension. Each estimand is a vector $\theta_0 = (\beta_0, \gamma_0, \pi_0)$ where the additional nuisance parameter γ_0 is defined as

$$\gamma_0 = \frac{\mathbb{E}\left[\phi(\varepsilon_i)u_i\right]}{\mathbb{E}\left[\phi(\varepsilon_i)\psi(\varepsilon_i)\right]},$$

i.e., γ_0 makes $\phi(\varepsilon_i)$ uncorrelated with $u_i - \psi(\varepsilon_i)\gamma_0$. The functions ϕ and ψ, the parameter space, Θ_n, and the nuisance parameter, γ_0, will be discussed further after the definition of the estimators. Each estimator $\hat{\theta} = (\hat{\beta}, \hat{\gamma}, \hat{\pi})$ is an approximate minimizer of an objective function, i.e.,

$$\|m_n(\hat{\theta})\| \leq \inf_{\theta \in \Theta_n} \|m_n(\theta)\| + o_p\left(n^{-1/2}\right),$$

where $m_n(\theta) = \frac{1}{n}\sum_i m_{ni}(\theta)$,

$$m_{ni}(\theta) = \begin{pmatrix} z_i'\pi\phi_i(\beta) \\ \phi_i(\beta)\left(x_i - \psi_i(\beta)\gamma\right) \\ z_i\left(x_i - \psi_i(\beta)\gamma - z_i'\pi\right) \end{pmatrix},$$

[2]The residualization replaces Z with $\bar{Z} = (I - P_W)Z$ where P_W is the projection onto W, and the normalization replaces $\bar{Z}$ with $\tilde{Z} = \bar{Z}(n^{-1}\bar{Z}'\bar{Z})^{-1/2}$. Thus, $x_i = \tilde{z}_i'\tilde{\pi}_0 + w_i'\tilde{\eta}_1 + u_i$ for some $(\tilde{\pi}_0, \tilde{\eta}_1)$, and I remove the tildes to keep the notation simple.

$\phi_i(\beta) = \phi(\varepsilon_i(\beta))$, $\psi_i(\beta) = \psi(\varepsilon_i(\beta))$, and $\varepsilon_i(\beta) = y_i - x_i\beta$. Note that the estimators, $\hat{\theta}$, the additional nuisance parameter, γ_0, the objective functions, m_n, and the asymptotic variances and their estimators (defined below) are all indexed by ϕ and ψ. For simplicity, I do not make that explicit in the notation.

To motivate m_n as a moment function, I relate it to LIML. A standard way to define LIML is as an extremum estimator that minimizes a variance ratio

$$\hat{\beta}_{\text{LIML}} = \arg\min_{\beta\in\mathbb{R}} \frac{\varepsilon(\beta)'P\varepsilon(\beta)}{\varepsilon(\beta)'\varepsilon(\beta)}$$

where $\varepsilon(\beta)$ denotes the stacked observations of $\varepsilon_i(\beta)$ and $P = n^{-1}ZZ'$ denotes the projection onto Z. From this definition it follows that $\hat{\theta}_{\text{LIML}} = (\hat{\beta}_{\text{LIML}}, \hat{\gamma}_{\text{LIML}}, \hat{\pi}_{\text{LIML}})$ solves the nonlinear first order conditions (see appendix D for a derivation)

$$\tfrac{1}{n}\sum_i \begin{pmatrix} z_i'\pi\varepsilon_i(\beta) \\ \varepsilon_i(\beta)\left(x_i - \varepsilon_i(\beta)\gamma\right) \\ z_i'\left(x_i - \varepsilon_i(\beta)\gamma - z_i'\pi\right) \end{pmatrix} = 0, \tag{2}$$

and the left hand side is $m_n(\theta)$ when $\phi(\varepsilon) = \psi(\varepsilon) = \varepsilon$. Thus, it follows that this class of estimators is a generalization of LIML.

A natural interpretation of m_n is that its first entry is the first order condition for a robust IV-regression of y_i on x_i using $z_i'\pi$ as an instrument for x_i and that the remaining entries are the first order conditions for an IV-regression of x_i on $\psi_i(\beta)$ and z_i using $\phi_i(\beta)$ as an instrument for $\psi_i(\beta)$. For this interpretation ones needs to subtract the first entry of m_n from the second, but doing so does not affect the asymptotic distribution of $\hat{\beta}$. The effect of including $\psi_i(\beta)$ in the regression of x_i on z_i is a rotation of the errors which makes $\phi(\varepsilon_i)$ uncorrelated with $u_i - \psi(\varepsilon_i)\gamma_0$. To understand the importance of this rotation, one can consider (2) for $\gamma = 0$ and without the second equation. Solving that for β yields 2SLS which is not consistent under many instruments asymptotics.

Section 4 shows, under regularity conditions, that an optimal choice of ϕ is proportional to the score function for ε_i, $\frac{f'(\varepsilon)}{f(\varepsilon)}$, and that an optimal choice of ψ is proportional to the conditional mean of u_i given ε_i, $\mathbb{E}[u_i \mid \varepsilon_i = \varepsilon]$. These optimal functions are usually unknown, but a feasible approach would be to let ϕ and ψ be the optimal functions for some fixed distribution, e.g., to let ϕ and ψ be one of the Huber score, $\min\{1, \max\{\varepsilon, -1\}\}$, the Cauchy score, $\varepsilon/(\varepsilon^2 + 1)$, or the Gauss score, ε. See (Hansen et al., 2010) for a

version of this approach under standard asymptotics and for a different class of estimators. An alternative approach, introduced by Huber (1964) in the regression model, is to fix some error distribution (e.g. normal) and use an estimator that minimizes the maximal asymptotic variance over a neighborhood of the fixed error distribution. Section 4 applies this approach to the current model and derives that the ensuing optimal robust (minimax) estimator is indexed by ϕ and ψ that are equal to $\phi_{\nu_0}(\varepsilon) = \min\{\nu_0, \max\{\varepsilon, -\nu_0\}\}$ for some value of ν_0, i.e., ϕ_{ν_0} is linear around zero and censors extreme values of ε_i at $\pm\nu_0$. Section 5 proposes a data driven censoring level $\hat{\nu}$ which also serves to make the estimator scale invariant. For this censoring level the optimal robust estimator is approximately 5% less efficient than LIML when (ε_i, u_i) has a joint normal distribution. On the other hand, the optimal robust estimator is approximately 80% more efficient than LIML for certain thick-tailed error distributions (and in the empirical example), in the sense that for such error distributions, LIML needs an 80% larger sample to achieve the same level of precision as the optimal robust estimator.

The main conditions on the functions ϕ and ψ are (i) that for each ϕ there exist a unique $\delta^* \in \mathbb{R}$ such that $\mathbb{E}[\phi(\varepsilon_i + \delta^*)] = 0$, and (ii) that

$$\mathbb{E}[\phi'(\varepsilon_i + \delta^*)] \neq 0 \quad \text{and} \quad \mathbb{E}\left[\phi(\varepsilon_i + \delta^*)\psi(\varepsilon_i + \delta^*)\right] \neq 0. \tag{3}$$

These conditions will, in general, be satisfied when ϕ and ψ equals one of the Huber, Cauchy, or Gauss scores. For example, if ϕ and ψ equal the Gauss score, then $\delta^* = -\mathbb{E}[\varepsilon_i]$ and (3) is satisfied when the variance of ε_i is nonzero (and finite). Similarly, if ϕ and ψ equal the Huber score, then there exist a unique $\delta^* \in \mathbb{R}$ such that $\mathbb{E}[\phi(\varepsilon_i + \delta^*)] = 0$ and (3) is satisfied when $\varepsilon_i + \delta^*$ puts positive mass on $[-1, 1]$. To simplify the notation, I normalize the intercept in (1a) so that $\delta^* = 0$. Furthermore, I normalize the function ψ such that $\mathbb{E}[\psi(\varepsilon_i)] = 0$. The latter normalization implies that the asymptotic results depends on $\psi(\varepsilon_i)$ instead of $\psi(\varepsilon_i) - \mathbb{E}[\psi(\varepsilon_i)]$.

A necessary condition for consistency of $\hat{\beta}$ is that m_n uniquely identifies β_0 over the parameter space Θ_n, but this is not the case if $\Theta_n = \mathbb{R}^{k+2}$. To see this, consider the special case where $\phi(\varepsilon) = \psi(\varepsilon) = \varepsilon$. In this case, it follows that the first order conditions in (2) have a second solution where $\tilde{\beta} = \arg\max_{\beta \in \mathbb{R}} \varepsilon(\beta)' P \varepsilon(\beta) / \varepsilon(\beta)' \varepsilon(\beta)$, and this solution is not consistent for β_0 in general (see appendix D for a derivation). To ensure consistency of $\hat{\beta}$, I let the parameter space $\Theta_n = \left[\hat{\beta}_{\text{init}} \pm b_n\right] \times \mathbb{R}^{k+1}$ where $\hat{\beta}_{\text{init}}$ is an initial consistent

estimator of β_0 and b_n is a bandwidth that slowly shrinks to zero. This can be thought of as using an initial estimator to select the consistent root of m_n, and section 4 shows that the asymptotic distribution of $\hat{\beta}$ does not depend on the initial estimator or the bandwidth. A natural choice of initial estimator is LIML, which is consistent when ε_i has finite variance.[3]

Section 4 shows, under regularity conditions, that $\sqrt{n}\hat{\Sigma}_n^{-1/2}(\hat{\beta} - \beta_0) \xrightarrow{d} \mathcal{N}(0,1)$ for some sequence of asymptotic variance estimators $\hat{\Sigma}_n$. This implies that confidence intervals or hypothesis tests for β_0 can be constructed using standard methods. In order to describe $\hat{\Sigma}_n$, let J_n be a Jacobian of m_n and suppose that $J_n(\theta)$ is invertible at $\hat{\theta}$. Then, $\hat{\Sigma}_n$ is the upper left entry of a sandwich estimator, i.e.,

$$\hat{\Sigma}_n = \left(J_n^{-1}(\hat{\theta}) \frac{1}{n} \sum_i m_{ni}^s(\hat{\theta}) m_{ni}^s(\hat{\theta})' J_n^{-1}(\hat{\theta})' \right)_{11} \tag{4}$$

where

$$m_{ni}^s(\theta) = \begin{pmatrix} z_i' \pi \phi_i(\beta) \\ \phi_i(\beta) \left(x_i - \psi_i(\beta)\gamma \right) \\ 0_k \end{pmatrix}.$$

The estimator $\hat{\Sigma}_n$ differs from a straightforward application of GMM formulas, as it uses $m_{ni}^s(\hat{\theta})$ rather than $m_{ni}(\hat{\theta})$ to form the inner matrix. This difference is related to the quickly growing number of parameters and sample moments in $m_n(\theta)$ and section 4 provides a further discussion of this difference.

When $\hat{\beta}$ is LIML, it is instructive to compare $\hat{\Sigma}_n$ to the variance estimator adopted by Hansen et al. (2008). That paper proposes an asymptotic variance estimator that separately estimates four different terms, and these terms are the standard variance estimator, the correction term proposed by Bekker (1994), and two additional terms that are present under some forms of nonnormality. The variance estimator considered here is numerically different from the Hansen et al. (2008) estimator, but asymptotically equivalent, i.e., $\hat{\Sigma}_n$ automatically accounts for all four terms.

[3] Without an initial consistent estimator for β_0, one could still use $m_n(\theta)$ to create a test statistic for hypotheses about β_0. Such an approach would be closely related to Kleibergen (2002)

3 High-level Conditions for Asymptotic Normality

This section gives high-level conditions for asymptotic normality of a single element of a GMM estimator whose dimension grows at the same rate as the sample size, and section 4 verifies these high-level conditions for the estimators of the IV model. I discuss some of the high-level conditions in the context of the IV model, but the results apply to just-identified GMM estimators $\hat{\theta}$ of $\theta_0 \in \mathbb{R}^p$, where $p/n \to \alpha \in [0, 1)$,

$$\|m_n(\hat{\theta})\| \leq \inf_{\theta \in \Theta_n \subset \mathbb{R}^p} \|m_n(\theta)\| + o_p\left(n^{-1/2}\right), \tag{5}$$

and the first entry of θ_0, say β_0, is the object of interest. The results can be seen as extensions of Pakes and Pollard (1989, theorem 3.1) and Newey and McFadden (1994, theorems 3.2 and 7.2) to a setup where the dimension of the parameter space grows as quickly as n, and the conditions are similar to the conditions of the theorems they generalize.

The proofs rely on expansions of $m_n(\hat{\theta})$ around some "limit" of $\hat{\theta}$ which I denote $\bar{\theta}$. This is quite standard, but differs from the classical theorems since $\bar{\theta}$ is different from θ_0 and random. This approach is similar to the one taken by El Karoui, Bean, Bickel, Lim, and Yu (2013) who study robust regression when the number of covariates is of the same order as the sample size (as in the first stage equation). Furthermore, the approach has some similarities with the analysis of two-step estimators in Newey and McFadden (1994, section 6), although in their setup the first step estimator has a non-random limit.

3.1 The "limit" of $\hat{\theta}$

The definition of $\bar{\theta}$ is constructed to satisfy two requirements. First, $\|\hat{\theta} - \bar{\theta}\|$ should be small enough that a Taylor expansion of m_n yields good approximations to the asymptotic behavior of the object of interest, $\hat{\beta} - \beta_0$. Second, $\bar{\theta}$ should be simple enough that one can characterize the limiting behavior of $m_n(\bar{\theta})$. To accommodate this, assume that the first entry of $\bar{\theta}$ is β_0, that $\bar{\theta}$ sets all but a fixed number of entries in $m_n(\bar{\theta})$ equal to zero, and that $\bar{\theta}$ is unique.

In the context of the IV model, define $\bar{\theta} = (\beta_0, \gamma_0, \bar{\pi})$ where $\bar{\pi}$ is the unique solution to

$$0 = \tfrac{1}{n} \sum_i z_i \left(x_i - \psi(\varepsilon_i)\gamma_0 - z_i'\bar{\pi}\right).$$

This definition of $\bar{\theta}$ sets the last k entries of $m_n(\bar{\theta})$ equal to zero. Furthermore, the two nonzero entries of $m_n(\bar{\theta})$ have mean zero, and this is (almost) a necessary condition for some of the high-level conditions of this section. One can think of $\bar{\pi}$ as an infeasible estimator of π_0, since it depends on the unknown parameters (β_0, γ_0). Note also that $\|\bar{\theta} - \theta_0\|$ is, in general, bounded away from zero under many instruments asymptotics, i.e., $\|\bar{\theta} - \theta_0\| > c + o_p(1)$ for some $c > 0$ when $\alpha > 0$.

Lemma 3.1. *Suppose that* $\mathbb{P}\left(\bar{\theta} \in \Theta_n\right) \to 1$. *If*

(i) $\|m_n(\bar{\theta})\| = o_p(1)$;

(ii) for every $\delta > 0$, *there exists a* $c > 0$, *such that*

$$\inf_{\substack{\theta \in \Theta_n, \\ \|\theta - \bar{\theta}\| > \delta}} \|m_n(\theta)\| > c + o_p(1);$$

then $\|\hat{\theta} - \bar{\theta}\| \xrightarrow{p} 0$.

The two details of this lemma that sets it apart from Pakes and Pollard (1989, theorem 3.1, hereafter PP3.1) are that the dimensions of $\bar{\theta}$ and m_n grow quickly with n and that the random vector $\bar{\theta}$ plays the role the parameter θ_0 does in PP3.1. These details have a limited impact on the proof of lemma 3.1, which is along the same lines as the proof of PP3.1, but they have a larger impact on the methods that can be used to verify the high-level conditions (i) and (ii). In the context of PP3.1, the usual way to verify (i) and (ii) is to establish a uniform law of large numbers for m_n, i.e., to show that

$$\sup_{\theta \in \Theta_n} \|m_n(\theta) - M_n(\theta)\| = o_p(1)$$

for some non-random M_n which satisfies (i) and (ii) with θ_0 in the place of $\bar{\theta}$ (see, e.g., Pakes and Pollard, 1989, corollary 3.2). In the current context, where the dimension of $\bar{\theta}$ and m_n grows quickly with n, it appears that the usual approach is infeasible. However, the definition of $\bar{\theta}$ implies that only a fixed number of entries in $m_n(\bar{\theta})$ are nonzero, so (i) will be satisfied if a law of large numbers holds for each of the remaining entries in $m_n(\bar{\theta})$. Furthermore, if m_n is sufficiently smooth, (ii) can be verified directly.

For the IV model, the nonzero entries of $m_n(\bar{\theta})$ are quadratic forms with mean zero and I use the Efron-Stein inequality (see appendix A for a discussion and references) to verify (i). Furthermore, m_n is linear in π which simplifies the verification of (ii).

3.2 Rate of convergence of $\hat{\theta}$ and asymptotic normality of $\hat{\beta}$

This subsection gives conditions that lead to $\|\hat{\theta}-\bar{\theta}\| = O_p(n^{-1/2})$ and asymptotic normality of $\hat{\beta}$. Theorem 3.2 below treats the case where m_n is continuously differentiable, and proposition 3.3 relaxes this smoothness condition. Both results use the following notation. Let J_n be a Jacobian (of m_n), and let the vector θ^* solve

$$0 = m_n(\bar{\theta}) + J_n(\bar{\theta})\left(\theta^* - \bar{\theta}\right).$$

When $J_n(\bar{\theta})$ is invertible it follows that

$$\beta^* - \beta_0 = -J_n^1(\bar{\theta})m_n(\bar{\theta}),$$

where $J_n^1(\bar{\theta})$ is the first row of $J_n^{-1}(\bar{\theta})$. The conditions of theorem 3.2 imply that β^* is asymptotically normal and that the same conclusion can be transferred to $\hat{\beta}$.

Theorem 3.2. *Suppose that* $\|\hat{\theta} - \bar{\theta}\| \xrightarrow{p} 0$, $\mathbb{P}\left(\bar{\theta} \in \Theta_n\right) \to 1$ *and* $\sqrt{n}\inf_{\theta\in\partial\Theta_n}\|\theta - \bar{\theta}\| \xrightarrow{p} \infty$ *where* $\partial\Theta_n$ *is the boundary of* Θ_n*. If*

(i) m_n *is continuously differentiable with Jacobian* J_n*;*

(ii) there exists a $c > 0$ *such that* $\sigma_{\min}\left(J_n(\bar{\theta})\right) > c + o_p(1)$*;*

(iii) for any sequence $\{\delta_n\}$ *of positive numbers converging to zero,*

$$\sup_{\|\theta-\bar{\theta}\|<\delta_n} \|J_n(\theta) - J_n(\bar{\theta})\| = o_p(1);$$

(iv) $\|m_n(\bar{\theta})\| = O_p(n^{-1/2})$ *and* $\sqrt{n}\Sigma_n^{-1/2}J_n^1(\bar{\theta})m_n(\bar{\theta}) \xrightarrow{d} \mathcal{N}(0,1)$ *for some* Σ_n *with* $\Sigma_n^{-1} = O_p(1)$*;*

then $\|\hat{\theta} - \bar{\theta}\| = O_p\left(n^{-1/2}\right)$ *and*

$$\sqrt{n}\Sigma_n^{-1/2}\left(\hat{\beta} - \beta_0\right) \xrightarrow{d} \mathcal{N}(0,1).$$

If, additionally, $\Sigma_n \xrightarrow{p} \Sigma$ *then* $\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} \mathcal{N}(0,\Sigma)$*.*

Condition (i) states that m_n is continuously differentiable, so it is natural to compare the conditions of theorem 3.2 to those of Newey and McFadden (1994, theorem 3.2, hereafter NM3.2). The main differences are that in theorem 3.2 the dimensions of θ, m_n, and J_n

can grow as quickly as n, that the "limiting" object, $\bar{\theta}$ is a random vector, and that the conclusion only gives a limiting distribution for a single element of $\hat{\theta}$. In NM3.2 these dimensions are fixed, the limiting object, θ_0, is nonrandom, and the conclusion is a limiting distribution of $\hat{\theta}$. A further difference is that NM3.2 allows for overidentification in the sense that the dimension of m_n can be larger than that of θ.

NM3.2 requires that the Jacobian converges in probability to a matrix of full rank, which in turn implies the bound on the singular values in (ii). Here, the dimension of $J_n(\bar{\theta})$ grows quickly with n so it will, generally, not converge in probability. However, (ii) is sufficient for the desired result. Similarly, NM3.2 imposes continuity on the limit of the Jacobian, but J_n may not have a limit. Instead, (iii) imposes a stochastic equicontinuity condition on J_n.

NM3.2 assumes that $m_n(\theta_0)$ satisfies a central limit theorem (CLT), whereas the essence of (iv) is that the (fixed number of) nonzero entries of $m_n(\bar{\theta})$ satisfy a CLT. The presence of the random $\bar{\theta}$ makes $m_n(\bar{\theta})$ a sample average of dependent observations, and there are multiple specialized tools to deal with such averages. Appendix A gives a CLT inspired by Chatterjee (2008) which can be used to establish (iv), and section 4 applies the CLT to the IV model. The final condition of NM3.2 is that the limiting object, θ_0, is an interior point of the parameter space, and theorem 3.2 places a similar condition on $\bar{\theta}$, but accommodates that $\bar{\theta}$ and Θ_n are random.

It is possible to get rid of the assumption that m_n is continuously differentiable, provided that m_n is well-approximated by a continuously differentiable random function M_n. The following proposition outlines what is meant by well-approximated.

Proposition 3.3. *Suppose that* $\|\hat{\theta}-\bar{\theta}\| \xrightarrow{p} 0$, $\mathbb{P}\left(\bar{\theta} \in \Theta_n\right) \to 1$ *and* $\sqrt{n}\inf_{\theta\in\partial\Theta_n}\|\theta-\bar{\theta}\| \xrightarrow{p} \infty$ *where* $\partial\Theta_n$ *is the boundary of* Θ_n*. If*

(i) for any sequence $\{\delta_n\}$ *of positive numbers converging to zero,*

$$\sup_{\substack{\theta\in\Theta_n,\\ \|\theta-\bar{\theta}\|<\delta_n}} \frac{\sqrt{n}\|m_n(\theta) - m_n(\bar{\theta}) - (M_n(\theta) - M_n(\bar{\theta}))\|}{1+\sqrt{n}\|\theta-\bar{\theta}\|} = o_p(1);$$

(ii) M_n *is continuously differentiable with Jacobian* J_n*;*

(iii) J_n *satisfies theorem* 3.2*(ii) and theorem* 3.2*(iii) and* m_n *satisfies theorem* 3.2*(iv);*

then $\sqrt{n}\Sigma_n^{-1/2}\left(\hat{\beta}-\beta_0\right) \xrightarrow{d} \mathcal{N}(0,1)$.

The conditions of this proposition are similar to the conditions of Newey and McFadden (1994, theorem 7.2), and the discussion following theorem 3.2 about differences and similarities applies here as well. I have verified the conditions of proposition 3.3 for m_n with discontinuous derivatives (e.g., when ϕ or ψ is the Huber score), but not when m_n is discontinuous (e.g., when ϕ or ψ are the sign function). The conditions of Newey and McFadden (1994, theorem 7.2) have been verified in cases with a discontinuous m_n (see, e.g., Andrews, 1994), but I leave it to future work to establish whether the conditions of proposition 3.3 can be verified for a discontinuous m_n.

4 Primitive Conditions for Asymptotic Normality

This section presents three results. First, it gives primitive conditions on the model and estimators of section 2 that are sufficient for the high-level conditions of section 3 and therefore sufficient for asymptotic normality. Second, it presents a consistency result for the asymptotic variance estimators. Third, it characterizes the functions ϕ and ψ that lead to an optimal estimator or to an optimal robust estimator.

4.1 Asymptotic Normality of $\hat{\beta}$

In order to show asymptotic normality of $\hat{\beta}$, this section imposes the following regularity conditions in addition to the assumptions stated together with the model.

Assumption 1. (i) $\{(z_i'\pi_0)^2\}_{i,n}$ is uniformly integrable and one of the following is satisfied:

(a) ϕ and ψ are bounded and $\mathbb{E}[u_i^4] < \infty$; or

(b) $\mathbb{E}[\phi^6(\varepsilon_i) + \psi^6(\varepsilon_i) + u_i^6] < \infty$.

(ii) $\Theta_n = \left[\hat{\beta}_{\text{init}} \pm b_n\right] \times \mathbb{R}^{k+1}$, where $b_n = o_p(1)$ and $n^{-1/2} + |\hat{\beta}_{\text{init}} - \beta_0| = o_p(b_n)$.

(iii) There exists a finite set $A \subseteq \mathbb{R}$ such that ϕ' and ψ' are Lipschitz continuous on each connected set in $\mathbb{R} \setminus A$.

The uniform integrability assumption on $z_i'\pi_0$ and existence of various moments are sufficient for the the law of large numbers and central limit theorem in appendix A. The conditions on the bandwidth, b_n, satisfy two requirements. First, b_n approaches zero, which ensures consistency of $\hat{\beta}$. Second, b_n approaches zero slowly, which implies that the asymptotic distribution of $\hat{\beta}$ neither depends on the bandwidth nor on the initial estimator, $\hat{\beta}_{\text{init}}$. When $\hat{\beta}_{\text{init}}$ is consistent for β_0 there will always exist a b_n that satisfies (ii), and if the initial estimator is LIML and $b_n = \sqrt{\hat{\Sigma}_{n,LIML}}/n^{1/4}$, then (ii) is satisfied when (i) is satisfied for $\phi(\varepsilon) = \psi(\varepsilon) = \varepsilon$. Here, $\hat{\Sigma}_{n,LIML}$ is as defined in (4) for $\phi(\varepsilon) = \psi(\varepsilon) = \varepsilon$. The smoothness condition in (iii) implies that m_n can be approximated by a continuously differentiable M_n as in proposition 3.3, and that the stochastic equicontinuity condition on the Jacobian of M_n, theorem 3.2(iii), is satisfied. The smoothness condition is satisfied by the Huber, Cauchy and Gauss scores, where the Cauchy and Gauss scores satisfy that the derivative is continuous.

The following result verifies that assumption 1 is sufficient for the high-level conditions of lemma 3.1.

Lemma 4.1 (Consistency)**.** *If $\hat{\theta}$ is indexed by ϕ and ψ that satisfy assumption 1, then $\|\hat{\theta} - \bar{\theta}\| \xrightarrow{p} 0$.*

I now turn to the main result of the paper, which shows that assumption 1 is sufficient for the conditions of theorem 3.3 and therefore for asymptotic normality of $\hat{\beta}$. The asymptotic variance, Σ_n, of $\hat{\beta}$ takes on a sandwich form, $\Sigma_n = D_n^{-1}\Omega_n D_n^{-1}$, where

$$\begin{aligned}\Omega_n = {} & (1-\alpha)^2\sigma_{xz}\mathbb{E}\left[\phi(\varepsilon_i)^2\right] + \alpha(1-\alpha)\mathbb{E}\left[\phi(\varepsilon_i)^2\right]\mathbb{E}\left[(u_i - \psi(\varepsilon_i)\gamma_0)^2\right] \\ & + 2(1-\alpha)\tfrac{1}{n}\sum_i (P_{ii}-\alpha) z_i'\pi_0 \mathbb{E}\left[\phi(\varepsilon_i)^2(u_i - \psi(\varepsilon_i)\gamma_0)\right] \\ & + \tfrac{1}{n}\sum_i (P_{ii}-\alpha)^2 cov\left(\phi(\varepsilon_i)^2, (u_i - \psi(\varepsilon_i)\gamma_0)^2\right),\end{aligned}$$

and P_{ii} is the i'th diagonal element of the projection matrix P. Furthermore,

$$\begin{aligned}D_n = {} & (1-\alpha)\sigma_{xz}\mathbb{E}\left[\phi'(\varepsilon_i)\right] \\ & + \tfrac{1}{n}\sum_i (P_{ii}-\alpha) z_i'\pi_0 \mathbb{E}\left[\phi'(\varepsilon_i)(u_i - \psi(\varepsilon_i)\gamma_0) - \gamma_0\phi(\varepsilon_i)\psi'(\varepsilon_i)\right].\end{aligned}$$

Theorem 4.2 (Asymptotic Normality)**.** *If $\hat{\theta}$ is indexed by ϕ and ψ that satisfy assumption 1 and there exists a $c > 0$ such that $|D_n| > c + o_p(1)$ and $\Omega_n > c + o_p(1)$, then*

$$\sqrt{n}\Sigma_n^{-1/2}\left(\hat{\beta} - \beta_0\right) \xrightarrow{d} \mathcal{N}(0, 1).$$

In the case where $\hat{\beta}$ is LIML, this theorem is a special case of Hansen et al. (2008, theorem 3). In this case, the first term of Σ_n is the variance of LIML under standard asymptotics, the second term of Σ_n is a "many instruments penalty" characterized by Bekker (1994, eq. (4.7)) under joint normality of (ε_i, u_i), and the third and fourth terms of Ω_n are the terms named $A + A'$ and B in Hansen et al. (2008). In the cases where the number of instruments is fixed or grows slowly ($\alpha = 0$), the asymptotic variances in theorem 4.2 are the same as the asymptotic variances for the class of nonlinear IV (NLIV) estimators studied in Hansen et al. (2010). For the remaining cases, i.e., when ϕ or ψ are nonlinear and $\alpha > 0$, theorem 4.2 has no antecedents (to the best of my knowledge).

In order to make further comparisons between the NLIV estimators and the robust estimators of this paper, one could consider the large sample properties of the NLIV estimators under many instruments asymptotics. However, a heuristic application of the argument in Hausman et al. (2012, section 3, see also Han and Phillips (2006)) suggest that the NLIV estimators are inconsistent under many instruments asymptotics. A potential conclusion from this is that the relationship between the NLIV estimators and the robust estimators of this paper is similar to the relationship between 2SLS and LIML, i.e., under standard asymptotics, the estimators of the two classes are asymptotically equivalent, and under many instruments asymptotics, the NLIV estimators are inconsistent whereas the robust estimators of this paper are consistent and asymptotically normal.

4.2 Consistency of the Asymptotic Variance Estimator

The assumptions that are sufficient for asymptotic normality are also sufficient for consistency of the asymptotic variance estimator.

Lemma 4.3 (Variance Estimation)**.** *If $\hat{\theta}$ is indexed by ϕ and ψ that satisfy assumption 1 and there exists a $c > 0$ such that $|D_n| > c + o_p(1)$ and $\Omega_n > c + o_p(1)$, then*

$$\sqrt{n}\hat{\Sigma}_n^{-1/2}\left(\hat{\beta} - \beta_0\right) \xrightarrow{d} \mathcal{N}(0, 1).$$

The underlying observations leading to lemma 4.3 are that the last k elements of $m_n(\bar{\theta})$ are zero, and that the first two elements of $m_{ni}(\bar{\theta})$ are uncorrelated across observations. This implies that

$$\begin{aligned}\Sigma_n &= \left(J_n^{-1}(\bar{\theta}) n\mathbb{E}[m_n(\bar{\theta})m_n(\bar{\theta})' \mid Z]J_n^{-1}(\bar{\theta})'\right)_{11} \\ &= \left(J_n^{-1}(\bar{\theta})\tfrac{1}{n}\sum_i \mathbb{E}[m_{ni}^s(\bar{\theta})m_{ni}^s(\bar{\theta})' \mid Z]J_n^{-1}(\bar{\theta})'\right)_{11}, \end{aligned} \tag{6}$$

where

$$m_{ni}^s(\theta) = \begin{pmatrix} z_i'\pi\phi_i(\beta) \\ \phi_i(\beta)\left(x_i - \psi_i(\beta)\gamma\right) \\ 0_k \end{pmatrix}.$$

The variance estimator is $\hat{\Sigma}_n = (J_n^{-1}(\hat{\theta})\frac{1}{n}\sum_i m_{ni}^s(\hat{\theta})m_{ni}^s(\hat{\theta})'J_n^{-1}(\hat{\theta})')_{11}$, and in the light of (6), $\hat{\Sigma}_n$ is an analog estimator.

Under standard asymptotics, one natural variance estimator that would emerge from GMM (see, e.g., Newey and McFadden, 1994, section 4) would be

$$\tilde{\Sigma}_n = \left(J_n^{-1}(\hat{\theta})\tfrac{1}{n}\sum_i m_{ni}(\hat{\theta})m_{ni}(\hat{\theta})'J_n^{-1}(\hat{\theta})'\right)_{11}$$

where

$$m_{ni}(\theta) = \begin{pmatrix} z_i'\pi\phi_i(\beta) \\ \phi_i(\beta)\left(x_i - \psi_i(\beta)\gamma\right) \\ z_i\left(x_i - \psi_i(\beta)\gamma - z_i'\pi\right) \end{pmatrix}.$$

I demonstrate, in simulations, that $\tilde{\Sigma}_n$ overestimates Σ_n when there are many instruments.

4.3 Optimal and Optimal Robust Estimators

Corollary 4.4 below presents conditions under which the asymptotic variance of theorem 4.2 simplifies, and the efficiency results in lemma 4.5 and proposition 4.6 further below provides lower bounds for this simplified asymptotic variance.

Corollary 4.4 (Simplified Variance)**.** *Suppose $\hat{\theta}$ is indexed by ϕ and ψ that satisfy assumption 1. If one of the following is satisfied:*

(i) $\frac{1}{n}\sum_i (P_{ii} - \alpha)^2 = o_p(1)$*; or*

(ii) $u_i = \psi(\varepsilon_i)\gamma_0 + \eta_i$ *where* η_i *is independent of* ε_i*, and* $\mathbb{E}[\phi(\varepsilon_i)\psi'(\varepsilon_i)] = 0$*;*

then $\sqrt{n}\Sigma_n^*(\phi,\psi)^{-1/2}\left(\hat{\beta} - \beta_0\right) \xrightarrow{d} \mathcal{N}(0,1)$ *where*

$$\Sigma_n^*(\phi,\psi) = \frac{\mathbb{E}\left[\phi(\varepsilon_i)^2\right]}{\mathbb{E}\left[\phi(\varepsilon_i)s(\varepsilon_i)\right]^2\sigma_{xz}} + \frac{\alpha}{1-\alpha}\frac{\mathbb{E}\left[\phi(\varepsilon_i)^2\right]}{\mathbb{E}\left[\phi(\varepsilon_i)s(\varepsilon_i)\right]^2\sigma_{xz}}\frac{\mathbb{E}\left[(u_i - \psi(\varepsilon_i)\gamma_0)^2\right]}{\sigma_{xz}},$$

$s = f'/f$ *is the score function for* ε_i *and the notation makes it explicit that* Σ_n^* *depends on* ϕ *and* ψ*.*

There are a variety of primitive conditions on z_i that imply $\frac{1}{n}\sum_i (P_{ii} - \alpha)^2 = o_p(1)$ (see, e.g., Anatolyev and Yaskov, 2016), and one of the simplest is that z_i indicate group membership and that all groups have equal sizes (Bekker and van der Ploeg, 2005). An example of a primitive condition that leads to (ii) is that (ε_i, u_i) are jointly normal and ψ is linear. This corollary follows from theorem 4.2 upon observing that (i) or (ii) yields that the last two terms of Ω_n and the second term of D_n are $o_p(1)$. A continuity argument would imply that these terms are small whenever the conditions of corollary 4.4 are almost satisfied. Thus, the efficiency results below can be seen as focusing on the most relevant terms of the asymptotic variance.

For a fixed joint distribution of the errors, (ε_i, u_i), the following lemma characterizes the efficiency bound for estimators indexed by ϕ and ψ satisfying that

$$\mathbb{E}[\phi(\varepsilon_i)^2] < \infty,\ \mathbb{E}[\psi(\varepsilon_i)^2] < \infty,\ \text{and } \mathbb{E}[\phi(\varepsilon_i)\psi(\varepsilon_i)] \neq 0.$$

Furthermore, the result gives conditions under which the bound is attained by a specific estimator in the class. The result is a consequence of the following well-known inequalities,

$$\frac{\mathbb{E}\left[\phi(\varepsilon_i)^2\right]}{\mathbb{E}\left[\phi(\varepsilon_i)s(\varepsilon_i)\right]^2} \geq \frac{1}{\mathcal{I}_f} \quad \text{and} \quad \mathbb{E}\left[(u_i - \psi(\varepsilon_i)\gamma_0)^2\right] \geq \mathbb{E}\left[(u_i - \mathbb{E}[u_i \mid \varepsilon_i])^2\right].$$

Lemma 4.5 (Efficiency Bound)**.** *The largest lower bound on* $\Sigma_n^*(\phi,\psi)$ *is*

$$\inf_{\phi,\psi} \Sigma_n^*(\phi,\psi) = \frac{1}{\mathcal{I}_f\sigma_{xz}} + \frac{\alpha}{1-\alpha}\frac{1}{\mathcal{I}_f\sigma_{xz}}\frac{\mathbb{E}\left[(u_i - \mathbb{E}[u_i \mid \varepsilon_i])^2\right]}{\sigma_{xz}}.$$

If $\mathbb{E}[s(\varepsilon_i)u_i] \neq 0$*, then the bound is attained by* $(\phi(\varepsilon),\psi(\varepsilon)) = (s(\varepsilon), \mathbb{E}[u_i \mid \varepsilon_i = \varepsilon])$*.*

If $\mathbb{E}[s(\varepsilon_i)u_i] \neq 0$, then this result bounds any asymptotic variance, $\Sigma_n^*(\phi,\psi)$, from below by an asymptotic variance that lets $\phi(\varepsilon) = s(\varepsilon)$ and $\psi(\varepsilon) = \mathbb{E}[u_i \mid \varepsilon_i = \varepsilon]$. As $\Sigma_n^*(\phi,\psi)$ is

homogeneous of degree zero, it follows that any ϕ proportional to the score for ε_i, $s = f'/f$, and any ψ proportional to the conditional mean of u_i given ε_i, $\mathbb{E}[u_i \mid \varepsilon_i = \cdot]$, leads to a minimal $\Sigma_n^*(\phi, \psi)$. If $\mathbb{E}[s(\varepsilon_i)u_i] = 0$, then the bound may not be attainable, but the bound is attainable in the special case where $\mathbb{E}[u_i \mid \varepsilon_i] = 0$. This case can be thought of as no endogeneity and implies that any feasible choice (s, ψ) reaches the efficiency bound. An interpretation of this is that the choice of ψ is irrelevant, when there is no endogeneity, and a continuity argument would imply that the choice of ψ is less important than the choice of ϕ, when there is weak endogeneity. I demonstrate, in simulations, that the choice of ψ tends to affect the sampling variance less than the choice of ϕ.

If the joint distribution of (ε_i, u_i) is such that the optimal ϕ or ψ is unbounded, then it follows that there exists a small perturbation to the distribution of ε_i such that the previous optimal choice of ϕ and ψ leads to an infinite asymptotic variance. In the context of the regression model (which has a univariate error term), Huber (1964) characterized such an issue as nonrobustness, and introduced the idea of a contamination model in order to derive new robust estimators. The following extends Huber's contamination model to the setup of the linear IV model (which has a bivariate error term). Given the importance of the joint normal distribution (and LIML), I focus on the case of a contaminated normal distribution, but more generally one could consider any contaminated distribution (see, Huber, 1964, theorem 1).

Assume that the absolutely continuous density of ε_i is

$$f = (1 - \delta)\Phi' + \delta h$$

where $\delta \in [0, 1)$ is a fixed, small level of contamination and h is an unknown absolutely continuous (contamination) density. Restrict the possible contamination densities such that the Fischer information, $\mathcal{I}_f$, is bounded by one, i.e., $\mathcal{I}_f = \mathbb{E}_f[s(\varepsilon_i)^2] \leq 1$, where $\mathbb{E}_f$ denotes expectation when ε_i has density f. Furthermore, let u_i be generated from the model

$$u_i = s(\varepsilon_i) + \eta_i \sqrt{2 - \mathcal{I}_f}$$

where η_i has a standard normal distribution and is independent of ε_i. The contamination model for ε_i is the same as Huber's. The choice of contamination model for u_i is guided by two principles. First, the joint distribution of (ε_i, u_i) should be normal under no contamination ($\delta = 0$), and this is achieved since $\delta = 0$ implies that $s(\varepsilon_i) = -\varepsilon_i$, $\mathcal{I}_f = 1$, and

$u_i \mid \varepsilon_i = \varepsilon \sim \mathcal{N}(-\varepsilon, 1)$. Second, the variance of u_i should stay bounded under contamination ($\delta > 0$), as the estimators are not robust with respect to outliers in u_i.[4] The model for u_i achieves this, since $var(u_i) = 2$ and

$$u_i \mid \varepsilon_i = \varepsilon \sim \mathcal{N}(s(\varepsilon), 2 - \mathcal{I}_f).$$

For the above contamination model, the following proposition characterizes the minimax efficiency bound for estimators indexed by ϕ and ψ satisfying that

$$\sup_f \mathbb{E}_f[\phi(\varepsilon_i)^2] < \infty, \ \sup_f \mathbb{E}_f[\psi(\varepsilon_i)^2] < \infty, \text{ and } \inf_f \mathbb{E}_f[\phi(\varepsilon_i)\psi(\varepsilon_i)]^2 > 0.$$

Th optimal robust estimator is the estimator that achieves the bound.

Proposition 4.6 (Minimax Efficiency Bound)**.** *Index the asymptotic variance by f, i.e, write $\Sigma_n^*(\phi, \psi, f)$ instead of $\Sigma_n^*(\phi, \psi)$. The largest lower bound on $\sup_f \Sigma_n^*(\phi, \psi, f)$ is*

$$\min_{\psi,\phi} \sup_f \Sigma_n^*(\phi, \psi, f) = \Sigma_n^*(\phi_{\nu_0}, \phi_{\nu_0}, f_0)$$

where $\phi_{\nu_0}(\varepsilon) = \min\{\nu_0, \max\{\varepsilon, -\nu_0\}\}$, ν_0 solves $\frac{\delta}{1-\delta} = \frac{2\Phi'(\nu_0)}{\nu_0} - 2\Phi(-\nu_0)$, and

$$f_0(\varepsilon) = \begin{cases} \frac{1-\delta}{\sqrt{2\pi}} e^{-\varepsilon^2/2}, & \text{if } |\varepsilon| \leq \nu_0, \\ \frac{1-\delta}{\sqrt{2\pi}} e^{\nu_0^2/2 - |\varepsilon|\nu_0}, & \text{if } |\varepsilon| \geq \nu_0. \end{cases}$$

This result is a consequence of the following saddle point property

$$\min_{\phi} \frac{\mathbb{E}_{f_0}[\phi(\varepsilon_i)^2]}{\mathbb{E}_{f_0}[\phi(\varepsilon_i)s_0(\varepsilon_i)]^2} \geq \frac{\mathbb{E}_{f_0}[\phi_{\nu_0}(\varepsilon_i)^2]}{\mathbb{E}_{f_0}[\phi_{\nu_0}(\varepsilon_i)s_0(\varepsilon_i)]^2} \geq \sup_f \frac{\mathbb{E}_f[\phi_{\nu_0}(\varepsilon_i)^2]}{\mathbb{E}_f[\phi_{\nu_0}(\varepsilon_i)s(\varepsilon_i)]^2}$$

where $s_0 = f_0'/f_0$, which Huber (1964, 1973) used to characterize an optimal robust estimator in the regression model. As in the regression model, proposition 4.6 shows that the bound is achieved by the estimator that is indexed by ϕ and ψ equal to $\phi_{\nu_0}(\varepsilon)$.

[4] One could potentially take the estimators of this paper and generalize them to also consider robustness with respect to u_i, e.g., by introducing nonlinearities in the latter entries of the moment function m_n. Although an interesting extension, I do not consider it in this paper for the following two reasons. First, such an extension would (in general) have less influence on the asymptotic variance than robustness with respect to ε_i (for reasons analogous to ψ influencing the asymptotic variance less than ϕ, but also for reasons related to the number of instruments (see Bean, Bickel, El Karoui, and Yu, 2013)). Second, such an extension may require stronger assumptions on the instruments (see El Karoui, 2013, for further details on this in the context of high dimensional regression).

The function $\phi_{\nu_0}(\varepsilon)$ censors ε at $\pm\nu_0$, where the level of censoring depends on the amount of contamination δ. A common way to choose the level of censoring in robust estimation of the regression model is to pick ν_0 as the solution to

$$0.95 = \sigma_\varepsilon^2 \frac{\mathbb{E}\left[\phi'_{\nu_0}(\varepsilon_i)\right]^2}{\mathbb{E}\left[\phi_{\nu_0}(\varepsilon_i)^2\right]} \quad \text{when} \quad \varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2). \tag{7}$$

This can be thought of as losing 5% efficiency when there is no contamination, it leads to $\nu_0 = 1.345\sigma_\varepsilon$, and it is optimal for a contamination level of $\delta \approx 0.058$. The next section outlines a method that uses (7) to estimate ν_0 and the remaining parameters jointly.

5 Scale Invariance and Included Exogenous Covariates

This section presents two extensions of the estimation strategy covered in the previous sections. Both extensions involve additional nuisance parameters that are estimated jointly with the remaining parameters in a GMM framework similar to the one presented in section 2. The first extension involves a scale parameter which both serves as a way to estimate the level of censoring for the optimal robust estimator and as a way to make $\hat{\beta}$ asymptotically scale invariant. Scale invariance of an estimator is desirable since it makes the estimator independent of the unit of measurement. The second extension considers the model in (1) without the simplification that (δ_0, η_0) are known, i.e., with additional exogenous covariates that are included in both equations of the model. Under the assumptions of corollary 4.4, these extensions do not affect the asymptotic variance of $\hat{\beta}$ or the optimality results of section 4. However, this section also presents natural extensions of the asymptotic variance estimators that are consistent without the simplifying assumptions of corollary 4.4.

5.1 Scale Invariance of $\hat{\beta}$

The following estimation procedure adapts Huber's "proposal 2" to the current setup (Huber, 1964, p. 96). Each estimand is a vector $\theta_0 = (\beta_0, \nu_0, \gamma_0, \pi_0)$ where the additional nuisance parameter ν_0 solves

$$\mathbb{E}\left[\phi(\varepsilon_i/\nu_0)^2\right] = c_0$$

for some known $c_0 > 0$. When ϕ is the Gauss score, the value of c_0 is irrelevant, so a natural choice is to let $c_0 = 1$. Otherwise, a way to choose c_0, which is similar to (7), is to pick (c_0, ν_1) as the solution to

$$0.95 = \frac{\mathbb{E}\left[\phi'(\varepsilon_i/\nu_1)/\nu_1\right]^2}{\mathbb{E}\left[\phi(\varepsilon_i/\nu_1)^2\right]} \quad \text{and} \quad c_0 = \mathbb{E}\left[\phi(\varepsilon_i/\nu_1)^2\right] \quad \text{when} \quad \varepsilon_i \sim \mathcal{N}(0,1).$$

If ϕ is the Huber score or the Cauchy score, this yields $(c_0, \nu_1) = (0.393, 1.345)$ or $(c_0, \nu_1) = (0.09, 2.384)$, respectively. Each estimator $\hat{\theta} = (\hat{\beta}, \hat{\nu}, \hat{\gamma}, \hat{\pi}) \in \Theta_n \subset \mathbb{R}^{k+3}$ is an approximate minimizer of an objective function, i.e.,

$$\|m_n(\hat{\theta})\| \leq \inf_{\theta \in \Theta_n} \|m_n(\theta)\| + o_p\left(n^{-1/2}\right)$$

where $m_n(\theta) = \frac{1}{n}\sum_i m_{ni}(\theta)$,

$$m_{ni}(\theta) = \begin{pmatrix} z_i'\pi\phi_i(\theta) \\ \phi_i(\theta)^2 - c_0 \\ \phi_i(\theta)\left(x_i - \psi_i(\theta)\gamma\right) \\ z_i\left(x_i - \psi_i(\theta)\gamma - z_i'\pi\right) \end{pmatrix},$$

$\phi_i(\theta) = \phi(\varepsilon_i(\theta))$, $\psi_i(\theta) = \psi(\varepsilon_i(\theta))$, $\varepsilon_i(\theta) = (y_i - x_i\beta)/\nu$, and $\Theta_n = \left[\hat{\beta}_{\text{init}} \pm b_n\right] \times \mathcal{V} \times \mathbb{R}^{k+1}$. Furthermore, let $\hat{\Sigma}_n$ be defined as in (4) with

$$m_{ni}^s(\theta) = \begin{pmatrix} z_i'\pi\phi_i(\theta) \\ \phi_i(\theta)^2 - nc_0 \\ \phi_i(\theta)\left(x_i - \psi_i(\theta)\gamma\right) \\ 0_k \end{pmatrix}.$$

Assumption 2. (i) There exist a $\nu_1 \geq 0$ such that $\mathbb{E}[\phi(\varepsilon_i/\nu)^2]$ is strictly decreasing (as a function of ν) on (ν_1, ∞), and $\nu_0 \in \text{int}(\mathcal{V})$ where $\mathcal{V} \subset (\nu_1, \infty)$ is compact and $\text{int}(\mathcal{V})$ is the interior of $\mathcal{V}$.

(ii) There exist a $K > 0$ such that $|\varepsilon\phi'(\varepsilon)| \leq K|\phi(\varepsilon)|$ and $|\varepsilon\psi'(\varepsilon)| \leq K|\psi(\varepsilon)|$.

Condition (i) is a global identification condition on ν_0 and (ii) (together with assumption 1) implies existence of sufficiently many moments for a law of large numbers and a central limit theorem to apply. If ϕ and ψ are the Huber score, then assumption 2 is satisfied with $\nu_1 = 0$ provided that $f(0) > 0$.

One can show, using the arguments from the proofs of lemma 4.1, theorem 4.2, and lemma 4.3, that if $\hat{\theta}$ is indexed by ϕ and ψ that satisfy the conditions of theorem 4.2 and assumption 2, then $\sqrt{n}\hat{\Sigma}_n^{-1/2}(\hat{\beta} - \beta_0) \xrightarrow{d} \mathcal{N}(0,1)$. If, in addition, corollary 4.4(ii) is satisfied, then one can show that $\sqrt{n}\Sigma_n^*(\phi^*, \psi^*)^{-1/2}(\hat{\beta} - \beta_0) \xrightarrow{d} \mathcal{N}(0,1)$ where $(\phi^*(\varepsilon), \psi^*(\varepsilon)) = (\phi(\varepsilon/\nu_0), \psi(\varepsilon/\nu_0))$.

5.2 Included Exogenous Variables

This subsection considers a class of estimators of $\theta_0 = (\beta_0, \delta_0, \gamma_0, \pi_0, \eta_0)$ where the additional nuisance parameters δ_0, η_0 were introduced in (1). Each estimator $\hat{\theta} = (\hat{\beta}, \hat{\delta}, \hat{\gamma}, \hat{\pi}, \hat{\eta}) \in \Theta_n \subset \mathbb{R}^{k+2G+2}$ is an approximate minimizer of an objective function, i.e.,

$$\|m_n(\hat{\theta})\| \leq \inf_{\theta \in \Theta_n} \|m_n(\theta)\| + o_p\left(n^{-1/2}\right)$$

where $m_n(\theta) = \frac{1}{n}\sum_i m_{ni}(\theta)$,

$$m_n(\theta) = \begin{pmatrix} z_i'\pi\phi_i(\theta) \\ w_i\phi_i(\theta) \\ \phi_i(\theta)\left(x_i - \psi_i(\theta)\gamma\right) \\ (z_i', w_i')'(x_i - \psi_i(\theta)\gamma - z_i'\pi - w_i'\eta) \end{pmatrix},$$

$\phi_i(\theta) = \phi(\varepsilon_i(\theta))$, $\psi_i(\theta) = \psi(\varepsilon_i(\theta))$, $\varepsilon_i(\theta) = y_i - x_i\beta - w_i'\delta$, and $\Theta_n = \left[\hat{\beta}_{\text{init}} \pm b_n\right] \times \mathcal{D} \times \mathbb{R}^{k+1+G}$. Furthermore, let $\hat{\Sigma}_n$ be defined as in (4) with

$$m_{ni}^s(\theta) = \begin{pmatrix} z_i'\pi\phi_i(\theta) \\ w_i\phi_i(\theta) \\ \phi_i(\theta)\left(x_i - \psi_i(\theta)\gamma\right) \\ 0_{k+G} \end{pmatrix}.$$

Assumption 3. (i) For any $\delta > 0$

$$\sup_{\substack{\delta \in \mathcal{D}, \\ \|\delta_0 - \delta\| > \delta}} \left\| \frac{1}{n}\sum_i w_i \mathbb{E}\left[\phi(\varepsilon_i + w_i'(\delta_0 - \delta)) \mid w_i\right] \right\|^{-1} = O_p(1),$$

where $\delta_0 \in \text{int}(\mathcal{D})$, $\mathcal{D} \subset \mathbb{R}^G$ is compact, and $\lambda_{\min}(\frac{1}{n}W'W)^{-1} = O_p(1)$.

(ii) $\{\|w_i\|^2\}_i$ is uniformly integrable

Condition (i) is a global identification condition on (δ_0, η_0). When ϕ is the Gauss score, the first part is satisfied provided that $\lambda_{\min}(\frac{1}{n}W'W)^{-1} = O_p(1)$, which is the identification condition for η_0.

One can show, using the argument from the proofs of lemma 4.1, theorem 4.2, and lemma 4.3, that if $\hat{\theta}$ is indexed by ϕ and ψ that satisfy the conditions of theorem 4.2 and assumption 3, then $\sqrt{n}\hat{\Sigma}_n^{-1/2}(\hat{\beta} - \beta_0) \xrightarrow{d} \mathcal{N}(0,1)$. If, in addition, corollary 4.4(i) is satisfied, then one can show that $\sqrt{n}\Sigma_n^*(\phi,\psi)^{-1/2}(\hat{\beta} - \beta_0) \xrightarrow{d} \mathcal{N}(0,1)$.

6 Simulations

This section presents the results of a simulation study which shows that the asymptotic results give good approximations to the finite sample behavior of the estimators considered in this paper. The simulations consider eight estimators which are the optimal robust estimator, LIML, 2SLS and five other combinations of ϕ and ψ as one of the Huber, Cauchy, or Gauss scores. The estimators incorporate both of the extensions described in section 5, and are implemented using a quasi-Newton method using LIML as the initial value.[5] 2SLS is mainly considered here in order to give an example of an estimator that leads to incorrect inference in the context of many instruments.

The simulations generate data from the model

$$\begin{aligned} y_i &= x_i\beta_0 + \delta_0 + \varepsilon_i \\ x_i &= z_i'\pi_0 + \eta_0 + u_i \qquad\qquad (i = 1, \cdots, 500) \end{aligned}$$

where $\delta_0 = \eta_0 = 0$, $\beta_0 = 1$, $z_i \sim \mathcal{N}(0, I_k)$, $\pi_0 = (\sqrt{\sigma_{xz}}, 0_{k-1}')'$, $k = 50$, and (ε_i, u_i) are *i.i.d.* with mean zero and covariance matrix

$$\Omega = \begin{bmatrix} 1 & \rho\sqrt{10} \\ \rho\sqrt{10} & 10 \end{bmatrix} \quad \text{where} \quad \rho \in [-1, 1]\,. \tag{8}$$

The simulations consider three levels for the strength of the instruments, σ_{xz}. The levels are $\sigma_{xz} = 1$ in Tables 1 and 2, $\sigma_{xz} = .5$ in Table 3 (left), and $\sigma_{xz} = .2$ in Table 3 (right). Additionally, the simulations consider two levels for the strength of endogeneity which is measured in terms of ρ (the correlation between ε_i and u_i). The levels are $\rho = -.7$ in Table

[5]The code, implemented in the statistical software *R*, is available on request.

Table 1: Simulation results, Normal errors

	$\rho = -.3$				$\rho = -.7$			
Estimator	Bias	Size, $\hat{\Sigma}_n$	Size, $\tilde{\Sigma}_n$	RE	Bias	Size, $\hat{\Sigma}_n$	Size, $\tilde{\Sigma}_n$	RE
LIML	−0.02	4.49	2.21	1.00	0.00	5.32	3.62	1.00
ϕ, ψ								
Gauss, Huber	−0.02	4.29	2.25	1.00	−0.01	5.22	3.63	0.96
Huber, Gauss	−0.01	4.49	2.25	0.94	0.00	5.20	3.57	0.93
Optimal Robust	−0.01	4.50	2.29	0.94	0.00	5.21	3.69	0.93
Gauss, Cauchy	−0.02	4.32	2.17	0.99	−0.01	5.85	4.11	0.89
Cauchy, Gauss	−0.01	4.33	2.20	0.96	0.00	5.18	3.64	0.93
Cauchy, Cauchy	−0.01	4.44	2.25	0.95	−0.05	5.39	3.74	0.94
2SLS	−1.54	33.55		4.23	−4.24	96.86		4.65

NOTE: 20,000 replications, 500 observations, 50 instruments. Bias is $med(\hat{\beta}-\beta_0)/(1.48 \cdot mad(\hat{\beta}))$, Size $\hat{\Sigma}_n$ uses $\hat{\Sigma}_n$ to estimate the asymptotic variance, Size $\tilde{\Sigma}_n$ uses the classical GMM variance estimator $\tilde{\Sigma}_n$, and RE is $mad(\hat{\beta}_{\text{LIML}})^2/mad(\hat{\beta})^2$.

1 (right) and $\rho = -.3$ elsewhere. Finally, the errors are generated such that $a\varepsilon_i$ has density f, $u_i = b(f'/f)(\varepsilon_i) + c\eta_i$, and η_i is standard normal and independent of ε_i. The density f is the density of a standard normal in Table 1 (in which case (ε_i, u_i) is joint normal), the density of a Huber distribution with parameter 1.345 in Table 2 (left), and the density of a $t(3)$ distribution in Table 2 (right) and Table 3. In each case the constants (a, b, c) are chosen such that (8) holds.

The strength of the instruments (together with the number of observations and the variance of u_i) implies that the value of the concentration parameter (Rothenberg, 1984) equals 50 in Tables 1 and 2, 25 in Table 3 (left), and 10 in Table 3 (right). It is well-known that the value of the concentration parameter tends to influence the quality of the asymptotic approximations based on many instruments asymptotics, so the simulations should reflect the values of the concentration parameter that tend to occur in empirical research. Hansen et al. (2008) conducted a survey ($n = 28$) of microeconomic studies published in AER, JPE, and QJE and found that 80% of the papers had a value of the concentration parameter between 8.95 and 588 with a median of 23.6. Thus, the range of the concentration parameter considered here is relevant for empirical research. The survey

also found a median value for $|\rho|$ of .279, so the value $\rho = -.3$ used here is relevant as well. Finally, the probability limit of the first stage F-statistic is $1 + \sigma_{xz}$, i.e., 2 in Tables 1 and 2, 1.5 in Table 3 (left), and 1.2 in Table 3 (right). Thus, the designs considered here involves instruments that are quite weak when measured by the F-statistic.

Table 2: Simulation results, Huber and $t(3)$ errors

	Huber errors				t(3) errors			
Estimator	Bias	Size, $\hat{\Sigma}_n$	Size, $\tilde{\Sigma}_n$	RE	Bias	Size, $\hat{\Sigma}_n$	Size, $\tilde{\Sigma}_n$	RE
LIML	0.00	4.42	1.91	1.00	−0.01	4.12	1.47	1.00
ϕ, ψ								
Gauss, Huber	0.00	4.37	1.99	1.00	0.00	4.17	1.70	1.01
Huber, Gauss	−0.01	4.32	2.15	1.14	0.00	3.67	1.72	1.69
Optimal Robust	−0.01	4.35	2.27	1.16	−0.01	3.93	1.77	1.74
Gauss, Cauchy	0.00	4.33	1.87	0.99	−0.01	4.21	1.70	1.00
Cauchy, Gauss	0.00	4.28	2.15	1.14	0.00	3.54	1.77	1.74
Cauchy, Cauchy	−0.01	4.32	2.14	1.16	−0.01	3.81	1.71	1.81
2SLS	−1.45	31.18		4.37	−1.22	24.52		4.24

NOTE: 20,000 replications, 500 observations, 50 instruments. Bias is $med(\hat{\beta} - \beta_0)/(1.48 \cdot mad(\hat{\beta}))$, Size $\hat{\Sigma}_n$ uses $\hat{\Sigma}_n$ to estimate the asymptotic variance, Size $\tilde{\Sigma}_n$ uses the naive GMM variance estimator $\tilde{\Sigma}_n$, and RE is $mad(\hat{\beta}_{\text{LIML}})^2/mad(\hat{\beta})^2$.

Tables 1-3 report four summary statistics from the study. The first statistic is the median bias of $\hat{\beta}$ standardized by 1.48 times the median absolute deviation (mad) of $\hat{\beta}$. $1.48 \times mad(\hat{\beta})$ is a robust estimate of the standard deviation of $\hat{\beta}$, so the bias is reported at the appropriate scale, and according to theorem 4.2, $\hat{\beta}$ has no asymptotic bias. The second statistic is a rejection percentage of the testing procedure that rejects the hypothesis that $\beta_0 = 1$ when $|\hat{\beta} - 1|/\sqrt{\hat{\Sigma}_n/n} > 1.96$, and according to lemma 4.3, this test has asymptotic size of 5%. The third statistic is the same as the second except that it uses the classical GMM variance estimator, $\tilde{\Sigma}_n$, instead of $\hat{\Sigma}_n$, and the disccusion following lemma 4.3 suggest that this testing procedure will have asymptotic size less than 5%. The fourth statistic is the square of $mad(\hat{\beta}_{\text{LIML}})/mad(\hat{\beta})$. This is a robust estimate of the relative efficiency (RE) of $\hat{\beta}$ and LIML. According to the discussion after proposition 4.6, the asymptotic relative efficiency of the optimal robust estimator and LIML under normal errors is (approximately)

Table 3: Simulation results, $t(3)$ errors

	$\sigma_{xz} = .5$				$\sigma_{xz} = .2$			
Estimator	Bias	Size, $\hat{\Sigma}_n$	Size, $\tilde{\Sigma}_n$	RE	Bias	Size, $\hat{\Sigma}_n$	Size, $\tilde{\Sigma}_n$	RE
LIML	−0.01	3.94	1.21	1.00	−0.07	4.31	1.20	1.00
ϕ, ψ								
Gauss, Huber	−0.01	4.09	1.44	1.02	−0.08	4.95	1.76	1.07
Huber, Gauss	−0.04	3.67	1.62	1.75	−0.12	3.98	1.69	1.73
Optimal Robust	−0.04	4.00	1.69	1.77	−0.13	4.70	1.97	1.76
Gauss, Cauchy	−0.03	4.27	1.61	1.05	−0.12	5.73	2.21	1.34
Cauchy, Gauss	−0.04	3.45	1.52	1.80	−0.12	3.62	1.59	1.75
Cauchy, Cauchy	−0.05	3.73	1.60	1.88	−0.15	4.40	1.87	2.12
2SLS	−1.41	30.96		9.65	−1.64	37.51		30.04

NOTE: 20,000 replications, 500 observations, 50 instruments. Bias is $med(\hat{\beta} - \beta_0)/(1.48 \cdot mad(\hat{\beta}))$, Size $\hat{\Sigma}_n$ uses $\hat{\Sigma}_n$ to estimate the asymptotic variance, Size $\tilde{\Sigma}_n$ uses the naive GMM variance estimator $\tilde{\Sigma}_n$, and RE is $mad(\hat{\beta}_{\text{LIML}})^2/mad(\hat{\beta})^2$.

0.95, i.e., an efficiency loss of 5%. Under the heavier tailed distributions in Tables 2 and 3, the asymptotic relative efficiency of the optimal robust estimator and LIML is greater than 1.

The "Bias" columns of Tables 1-3 show that all the estimators considered, except for 2SLS, are essentially median unbiased, and a comparison across Table 2 and Table 3 indicates that a small bias emerges when the strength of the instruments approaches zero. The "Size, $\hat{\Sigma}_n$" columns show that the testing procedure based on $\hat{\Sigma}_n$ has good size properties for all the parameter values considered, so the small bias that emerges in Table 3 does not seem to affect its size. A comparison within Table 1 reveals that the size properties are somewhat affected by the strength of the endogeneity, something which is not captured by the asymptotic analysis. The "Size, $\tilde{\Sigma}_n$" columns illustrate that the testing procedure based on $\tilde{\Sigma}_n$ is conservative (rejection percentages are too low) for the sample sizes considered here, and a comparison across Table 2 and Table 3 suggests that the rejection percentage diverges further towards zero when the strength of the instruments approaches zero.

The "RE" columns of Table 1 show that when the errors are jointly normal, there is about a 5% efficiency loss (relative to LIML) from letting ϕ be the Huber or Gauss scores.

On the other hand, there is a $0-11\%$ efficiency loss from letting ψ be the Huber or Gauss scores. This latter finding conforms with the observation made after lemma 4.5, that the efficiency loss from a suboptimal ψ depends on the strength of the endogeneity. The "RE" columns of Tables 2 and 3 show that the robust estimators are substantially more efficient than LIML under the thick-tailed distributions, e.g., the optimal robust estimator is roughly 75% more efficient than LIML in the case of $t(3)$ errors (Table 2 (right)). Furthermore, a comparison of the rows in Table 3 show that the efficiency gains relative to LIML from letting ϕ be the Huber or Gauss scores are generally larger than the efficiency gains from letting ψ be the Huber or Gauss scores.

7 Quarter of Birth and Returns to Schooling

This section considers the empirical example provided by the Angrist and Krueger (1991) study of the returns to schooling using quarter of birth as an instrument. The data comes from the 1980 U.S. Census and includes 329,509 males born 1930–1939. The structural equation includes a constant, year, and state dummies, and the reduced form equation includes 180 instruments which are quarter of birth times year or state of birth. This model corresponds to table 7 of Angrist and Krueger (1991). In this example, the estimated concentration parameter is 257 and the correlation between ε_i and u_i is estimated at -0.2. These observations and the simulations suggest that the asymptotic approximations should work well for this example.

Table 4 presents the OLS estimate and the estimates from the eight estimators considered in the simulation study. Additionally, Table 4 reports standard error estimates based on $\hat{\Sigma}_n$ for the estimators analyzed in this paper and classical standard error estimates for OLS and 2SLS. The latter two are only included for easy reference as they lead to confidence intervals with incorrect coverage. Finally, Table 4 includes the variance ratios $\hat{\Sigma}_n(\hat{\beta}_{\text{LIML}})/\hat{\Sigma}_n(\hat{\beta})$ for each estimator, which provides an estimate of the efficiency gains relative to LIML.

Table 4 shows that the robust estimators deliver point estimates that are similar to either LIML or 2SLS. Furthermore, the table indicates that the robust estimators can be substantially more efficient than LIML, e.g., the optimal robust estimator is estimated to be 83% more efficient than LIML. These efficiency gains are similar in size to the gains

Table 4: Returns to Schooling

Estimator	Estimate	Standard error	Variance ratio
LIML	0.1064	0.014 88	1.00
ϕ, ψ			
Gauss, Huber	0.1051	0.014 41	1.07
Huber, Gauss	0.0891	0.010 85	1.88
Optimal Robust	0.0894	0.010 99	1.83
Gauss, Cauchy	0.1043	0.014 01	1.13
Cauchy, Gauss	0.0869	0.010 40	2.05
Cauchy, Cauchy	0.0874	0.010 63	1.96
2SLS	0.0928	0.009 30	
OLS	0.0673	0.000 35	

NOTE: Males born 1930–1939, 1980 IPUMS, $n = 329,509$. Variance ratio is $\hat{\Sigma}_n(\hat{\beta}_{\text{LIML}})/\hat{\Sigma}_n(\hat{\beta})$.

in the simulation study with $t(3)$ errors in Table 2 (right). Furthermore, these gains are similar in size to the gains achieved by some of the nonlinear IV estimators proposed by Hansen et al. (2010) in a similar model using three instruments.

To further illustrate why the robust estimators are more efficient than LIML in this example, this section presents two figures that describe the distribution of the errors. Figure 1 presents a nonparametric estimate of the density of ε_i along with a normal and $t(3)$ density where the location and scale parameters are based on the median and *mad* of the LIML residuals. From this figure it is evident that the $t(3)$ density provides a better fit than the normal, although the normal provides a reasonable fit in the center of the distribution. Figure 2 depict nonparametric estimates of the optimal ϕ and ψ (f'/f and $\mathbb{E}[u_i \mid \varepsilon_i = \cdot]$) together with the appropriately scaled estimates of these functions implied by the estimators that sets both of ϕ and ψ equal to one of the Gauss (LIML), Huber (optimal robust), or Cauchy scores. From this figure it is clear that the Huber and Cauchy scores provide a better fit to the unknown optimal ϕ and ψ than the Gauss score does.

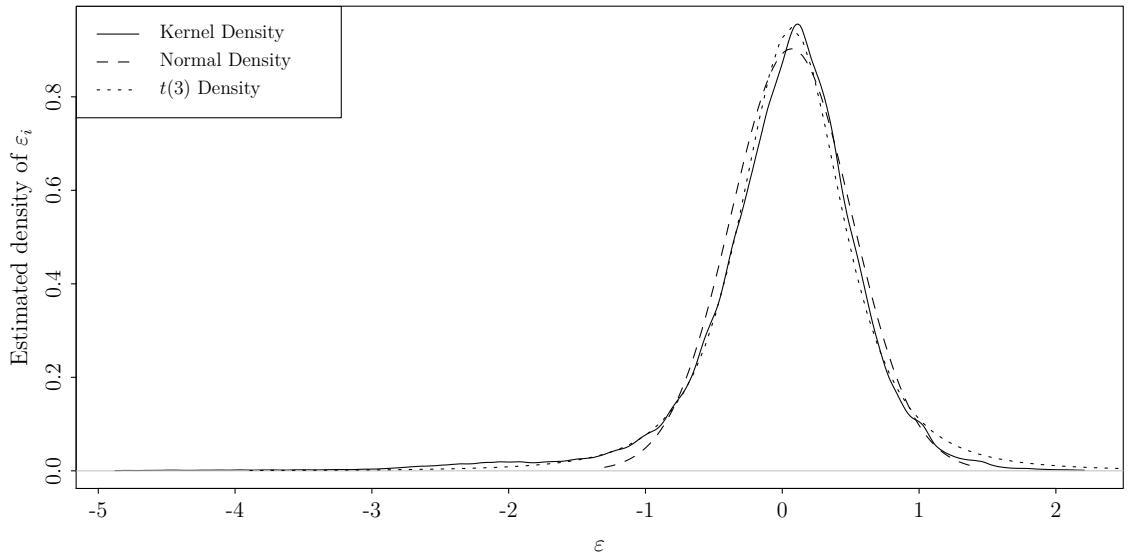

Figure 1: Estimates of $f_{\varepsilon_i}(\varepsilon)$ using LIML residuals

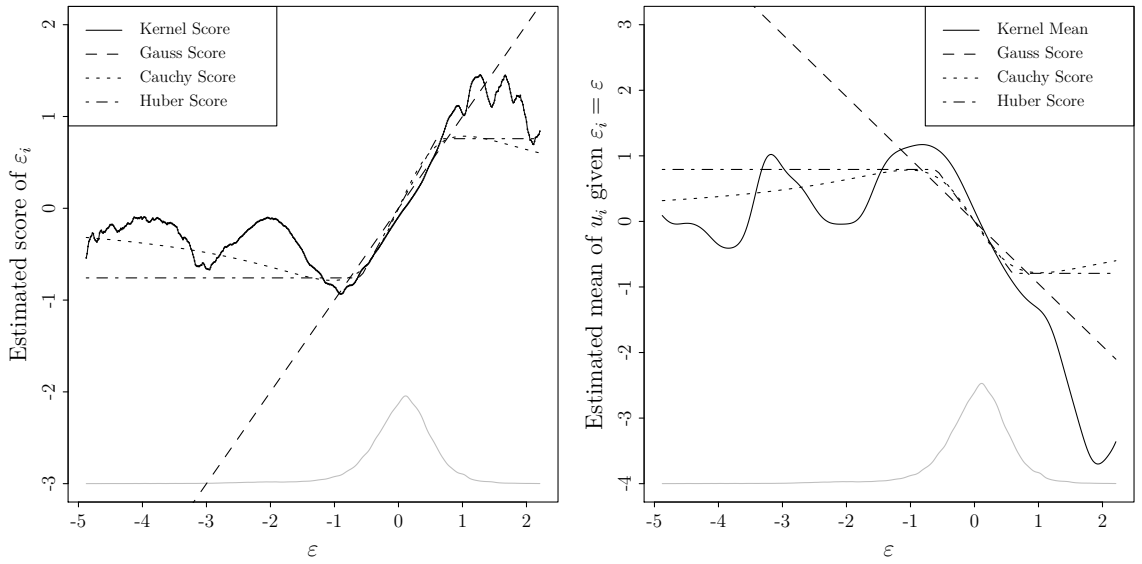

Figure 2: Estimates of f'/f and $\mathbb{E}[u_i \mid \varepsilon_i]$ using LIML residuals

8 Conclusion

This paper introduced a new class of robust estimators in a linear IV model with many instruments. Each estimator in the class was shown to be consistent and asymptotically normal under many instruments asymptotics, and the paper proposed consistent variance estimators that are of the "sandwich" type and can be used to conduct asymptotically correct inference. Furthermore, this paper characterized an optimal robust estimator among the members of the class. In the empirical example, the optimal robust estimator was approximately 80% more efficient than LIML.

Since the class of estimators introduced in this paper are generalizations of LIML, it is plausible that they can be generalized to accommodate conditional heteroskedasticity using leave-one-out ideas as in JIVE (Chao et al., 2012) and HLIM (Hausman et al., 2012). One approach that could achieve this and combines the estimators of this paper with the ideas behind HLIM is to let

$$(\hat{\beta}, \hat{\gamma}) = \arg\min_{\beta,\gamma} \|m_n(\beta, \gamma)\|,$$

where $m_n(\beta, \gamma) = \frac{1}{n} \sum_i m_{ni}(\beta, \gamma)$ and

$$m_{ni}(\beta, \gamma) = \begin{pmatrix} \sum_{j \neq i} P_{ij}(x_j - \psi_i(\beta)\gamma)\phi_i(\beta) \\ \phi_i(\beta)\left(x_i - \psi_i(\beta)\gamma\right) \end{pmatrix}.$$

With this formulation $(\hat{\beta}, \hat{\gamma})$ is essentially a zero of a U-process, and its asymptotic properties could potentially be analyzed using a suitable generalization of the theorems presented in Honoré and Powell (1994).

References

Ai, C. and X. Chen (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica 71*(6), 1795–1843.

Amemiya, T. (1982). Two stage least absolute deviations estimators. *Econometrica 50*(3), 689–711.

Anatolyev, S. and P. Yaskov (2016). Asymptotics of diagonal elements of projection matrices under many instruments/regressors. *Econometric Theory*, 1–22.

Anderson, T., N. Kunitomo, and Y. Matsushita (2010). On the asymptotic optimality of the liml estimator with possibly many instruments. *Journal of Econometrics 157*(2), 191–204.

Andrews, D. W. (1994). Empirical process methods in econometrics. *Handbook of Econometrics 4*, 2247–2294.

Angrist, J. D. and A. B. Krueger (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics 106*(4), 979–1014.

Bean, D., P. J. Bickel, N. El Karoui, and B. Yu (2013). Optimal m-estimation in high-dimensional regression. *Proceedings of the National Academy of Sciences 110*(36), 14563–14568.

Bekker, P. A. (1994). Alternative approximations to the distributions of instrumental variable estimators. *Econometrica 62*(3), 657–81.

Bekker, P. A. and J. van der Ploeg (2005). Instrumental variable estimation based on grouped data. *Statistica Neerlandica 59*(3), 239–267.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica 80*(6), 2369–2429.

Boucheron, S., G. Lugosi, and P. Massart (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press.

Cattaneo, M. D., R. K. Crump, and M. Jansson (2012). Optimal inference for instrumental variables regression with non-gaussian errors. *Journal of Econometrics 167*(1), 1–15.

Chamberlain, G. and G. Imbens (2004). Random effects estimators with many instrumental variables. *Econometrica 72*(1), 295–306.

Chao, J. C. and N. R. Swanson (2005). Consistent estimation with a large number of weak instruments. *Econometrica 73*(5), 1673–1692.

Chao, J. C., N. R. Swanson, J. A. Hausman, W. K. Newey, and T. Woutersen (2012). Asymptotic distribution of jive in a heteroskedastic iv regression with many instruments. *Econometric Theory 28*(01), 42–86.

Chatterjee, S. (2008). A new method of normal approximation. *The Annals of Probability 36*(4), 1584–1610.

Chen, L.-A. and S. Portnoy (1996). Two-stage regression quantiles and two-stage trimmed least squares estimators for structural equation models. *Communications in Statistics - Theory and Methods 25*(5), 1005–1032.

Chen, L. H. (1978). Two central limit problems for dependent random variables. *Probability Theory and Related Fields 43*(3), 223–243.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics 6*, 5549–5632.

Chen, X., O. Linton, and I. Van Keilegom (2003). Estimation of semiparametric models when the criterion function is not smooth. *Econometrica 71*(5), 1591–1608.

Chernozhukov, V. and C. Hansen (2006). Instrumental quantile regression inference for structural and treatment effect models. *Journal of Econometrics 132*(2), 491–525.

Chioda, L. and M. Jansson (2009). Optimal invariant inference when the number of instruments is large. *Econometric Theory 25*(03), 793–805.

Efron, B. and C. Stein (1981). The jackknife estimate of variance. *The Annals of Statistics*, 586–596.

El Karoui, N. (2013). Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results. *arXiv:1311.2445*.

El Karoui, N., D. Bean, P. J. Bickel, C. Lim, and B. Yu (2013). On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences 110*(36), 14557–14562.

Hahn, J. (2002). Optimal inference with many instruments. *Econometric Theory 18*(01), 140–168.

Hahn, J. and J. Hausman (2002). A new specification test for the validity of instrumental variables. *Econometrica 70*(1), 163–189.

Han, C. and P. C. B. Phillips (2006). Gmm with many moment conditions. *Econometrica 74*(1), 147–192.

Hansen, C., J. A. Hausman, and W. K. Newey (2008). Estimation with many instrumental variables. *Journal of Business & Economic Statistics 26*(4), 398–422.

Hansen, C. and D. Kozbur (2014). Instrumental variables estimation with many weak instruments using regularized jive. *Journal of Econometrics 182*(2), 290–308.

Hansen, C., J. B. McDonald, and W. K. Newey (2010). Instrumental variables estimation with flexible distributions. *Journal of Business & Economic Statistics 28*(1), 13–25.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica 50*(4), 1029–1054.

Hausman, J. A., W. K. Newey, T. Woutersen, J. C. Chao, and N. R. Swanson (2012). Instrumental variable estimation with heteroskedasticity and many instruments. *Quantitative Economics 3*(2), 211–255.

Honoré, B. E. and L. Hu (2004). On the performance of some robust instrumental variables estimators. *Journal of Business & Economic Statistics 22*(1), 30–39.

Honoré, B. E. and J. L. Powell (1994). Pairwise difference estimators of censored and truncated regression models. *Journal of Econometrics 64*(1-2), 241–278.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics 35*(1), 73–101.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, Berkeley, Calif., pp. 221–233. University of California Press.

Huber, P. J. (1973). Robust regression: asymptotics, conjectures and monte carlo. *The Annals of Statistics*, 799–821.

Huber, P. J. (1981). *Robust statistics*. Wiley.

Kleibergen, F. (2002). Pivotal statistics for testing structural parameters in instrumental variables regression. *Econometrica 70*(5), 1781–1803.

Kolesár, M. (2013). Estimation in an instrumental variables model with treatment effect heterogeneity. *Unpublished Working Paper*.

Kolesár, M. (2015). Minimum distance approach to inference with many instruments. *arXiv preprint arXiv:1504.02911*.

Krasker, W. S. and R. E. Welsch (1985). Resistant estimation for simultaneous-equations models using weighted instrumental variables. *Econometrica 53*(6), 1475–1488.

Kunitomo, N. (1980). Asymptotic expansions of the distributions of estimators in a linear functional relationship and simultaneous equations. *Journal of the American Statistical Association 75*(371), 693–700.

Morimune, K. (1983). Approximate distributions of k-class estimators when the degree of overidentifiability is large compared with the sample size. *Econometrica 51*(3), 821–841.

Newey, W. K. (1990). Efficient instrumental variables estimation of nonlinear models. *Econometrica 58*(4), 809–837.

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica 62*(6), 1349–1382.

Newey, W. K. and D. McFadden (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics 4*, 2111–2245.

Newey, W. K. and F. Windmeijer (2009). Generalized method of moments with many weak moment conditions. *Econometrica 77*(3), 687–719.

Pakes, A. and D. Pollard (1989). Simulation and the asymptotics of optimization estimators. *Econometrica 57*(5), 1027–1057.

Powell, J. L. (1983). The asymptotic normality of two-stage least absolute deviations estimators. *Econometrica 51*(5), 1569–1575.

Rothenberg, T. J. (1984). Approximating the distributions of econometric estimators and test statistics. *Handbook of Econometrics 2*, 881–935.

Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*, Berkeley, Calif., pp. 583–602. University of California Press.

Stein, C. (1986). Approximate computation of expectations. *Lecture Notes-Monograph Series 7*, i–164.

Wang, W. and M. Kaffo (2016). Bootstrap inference for instrumental variable models with many weak instruments. *Journal of Econometrics 192*(1), 231–268.

A Central Limit Theorem

This section presents a central limit theorem for random variables $W_n = W_n(V_n)$ where the elements of $V_n = (v_{n1}, \cdots, v_{nn})$ are mutually independent and W_n has mean zero and variance one. To simplify the presentation, I drop the subscript n on v_{ni}.

Let $\tilde{V}_n = (\tilde{v}_1, \cdots, \tilde{v}_n)$ be an independent copy of V_n. For each $i \in \{1, \cdots, n\}$, let $[i] = \{1, \cdots, i\}$ and define $V_n^{[i]} = (\tilde{v}_1, \cdots, \tilde{v}_i, v_{i+1}, \cdots, v_n)$ and $V_n^i = (v_1, \cdots, v_{i-1}, \tilde{v}_i, v_{i+1}, \cdots, v_n)$. For each $i \in [n]$ and measurable function g of V_n, define the following randomized derivatives of g along the ith coordinate as

$$\Delta_i g = g(V_n) - g(V_n^i) \quad \text{and} \quad \Delta_i g^{[i-1]} = g(V_n^{[i-1]}) - g(V_n^{[i]}).$$

Define $T_n = \frac{1}{2}\sum_i (\Delta_i W_n)(\Delta_i W_n^{[i-1]})$ and note that $\mathbb{E}[T_n] = \mathbb{E}[W_n^2] = 1$ (see, e.g., lemma $A.10$ further below).

Lemma A.1. *Let all terms be defined as above. If*

(i) $\mathbb{E}[T_n \mid V_n] \to 1$ *in* $\mathcal{L}^1$*;*

(ii) $\sum_i \mathbb{E}\left[(\Delta_i W_n)^2\right] = O(1)$*;*

(iii) $\sum_i \mathbb{E}\left[(\Delta_i W_n)^2 1_{\{|\Delta_i W_n| > \epsilon\}}\right] \to 0$ *for any* $\epsilon > 0$*;*

then $W_n \xrightarrow{d} \mathcal{N}(0,1)$.

Comments. 1. It is sufficient to verify (ii) and (iii) for $\Delta_i^0 W_n$, where

$$\Delta_i^0 W_n = W_n(V_n) - W_n(v_1, \cdots, v_{i-1}, 0, v_{i+1}, \cdots, v_n).$$

A sufficient condition for (iii) is $\sum_i \mathbb{E}\left[(\Delta_i W_n)^{2+\delta}\right] \to 0$ for some $\delta > 0$, and a sufficient condition for both (ii) and (iii) is uniform integrability of $\{n(\Delta_i W_n)^2\}_{i,n}$. Lemma A.2 provides a method to show that (i) is satisfied.

2. The proof of this lemma uses Stein's method (Stein, 1972, 1986) in the particular version given by Chen (1978), and combines it with the main ideas from Chatterjee (2008). (Chatterjee, 2008, theorem 2.2) imposes that $\mathbb{E}[T_n \mid V]$ converges in $\mathcal{L}^2$ and that $\sum_i \mathbb{E}\left[(\Delta_i W_n)^3\right] \to 0$. For the applications of lemma $A.1$ in this section, the "conditional variance" T_n, as defined here, is simpler to work with than the T_n in Chatterjee (2008).

Let $S_n = S_n(V_n)$ be random variables such that $\mathbb{E}|S_n| < \infty$ for all n.

Lemma A.2. *If one of the following is satisfied*

(i) $\sum_i \mathbb{E}\left[|\Delta_i S_n|\right] = O(1)$ *and* $\sum_i \mathbb{E}\left[|\Delta_i S_n| 1_{\{|\Delta_i S_n| > \epsilon\}}\right] \to 0$ *for any* $\epsilon > 0$*;*

(ii) $\sum_i \mathbb{E}\left[|\Delta_i S_n|^{1+\delta}\right] \to 0$ *for some* $\delta \in (0, 1]$*;*

then $S_n - \mathbb{E}\left[S_n\right] \to 0$ *in* $\mathcal{L}^1$.

Comments. 1. As in lemma $A.1$, it is sufficient to verify the conditions for $\Delta_i^0 S_n$.

2. The proof of this lemma combines the proof of the Efron-Stein inequality (see Efron and Stein, 1981; Boucheron, Lugosi, and Massart, 2013) with a truncation argument. Without the truncation argument, the sufficient condition becomes $\sum_i \mathbb{E}[(\Delta_i S_n)^2] \to 0$, which is the Efron-Stein inequality and implies convergence in $\mathcal{L}^2$. Corollary A.3 gives a simple example where lemma $A.2$ leads to weaker conditions than the Efron-Stein inequality.

A.1 Applications of lemmas $A.1$ and $A.2$

The following sequence of corollaries applies lemmas $A.1$ and $A.2$ to weighted averages, bi-linear, and tri-linear forms. Most (potentially all) of the results are known, but I present them here for easy reference. All of these corollaries are applied in the proofs of theorem 4.2 and lemma 4.3.

Suppose that $\{v_i\}_i$ is an *i.i.d.* sequence with $\mathbb{E}[v_1] = 0$, $\{w_{ni}\}_{i,n}$ are non-random weights, $M_n = (M_{nij})$ is a symmetric, non-random matrix with zeroes on the diagonal, and $\{M_{nijk}\}_{i,j,k,n}$ is a non-random array with $M_{nijk} = M_{nikj}$. For simplicity, I drop the subscript n on M_{ij} and M_{ijk}.

Corollary A.3. *Suppose that $S_n = \sum_i w_{ni} v_i$. If one of the following is satisfied:*

(i) $\mathbb{E}|v_1| < \infty$, $\max_i w_{ni} \to 0$, $\sum_i w_{ni} = 1$, *and* $w_{ni} \geq 0$; *or*

(ii) $\mathbb{E}[|v_1|^{1+\delta}] < \infty$ *and* $\sum_i |w_{ni}|^{1+\delta} \to 0$ *for some* $\delta \in (0, 1]$;

then $S_n \to 0$ in $\mathcal{L}^1$.

For this corollary, $\Delta_i^0 S_n = w_{ni} v_i$. It follows from (i) that $\sum_i \mathbb{E}\left[|\Delta_i^0 S_n|\right] = \mathbb{E}|v_i|$ and

$$\sum_i \mathbb{E}\left[|\tilde{\Delta}_i S_n| 1_{\{|\tilde{\Delta}_i S_n| > \epsilon\}}\right] \leq \sum_i w_{ni} \mathbb{E}\left[|v_i| 1_{\{w_{ni}|v_i| > \epsilon\}}\right] \leq \max_i \mathbb{E}\left[|v_i| 1_{\{|v_i| > C_n \epsilon\}}\right] \to 0$$

where $C_n = (\max_i w_{ni})^{-1} \to \infty$. From (ii) it follows that $\sum_i \mathbb{E}\left[|\Delta_i^0 S_n|^{1+\delta}\right] \to 0$.

Corollary A.4. *Suppose that $S_n = \sum_i \sum_{j \neq i} M_{ij} v_{i1} v_{j2}$ where $v_i = (v_{i1}, v_{i2})$. If one of the following is satisfied:*

(i) $\mathbb{E}[\|v_1\|^{1+\delta}] < \infty$ *for some* $\delta > 0$, $\max_i \sum_{j \neq i} M_{ij} \to 0$, $\sum_i \sum_{j \neq i} M_{ij} = 1$, *and* $M_{ij} \geq 0$; *or*

(ii) $\mathbb{E}[\|v_1\|^2] < \infty$, *and* $\sum_i \sum_{j \neq i} M_{ij}^2 \to 0$;

then $S_n \to 0$ in $\mathcal{L}^1$. Under (i) and without $\mathbb{E}[v_i] = 0$, it follows that $S_n - \mathbb{E}[S_n] \to 0$ in $\mathcal{L}^1$.

For this corollary, $\Delta_i^0 S_n = v_{i1} \sum_{j \neq i} M_{ij} v_{j2} + v_{i2} \sum_{j \neq i} M_{ij} v_{j1}$ and I only treat the first of these. Condition (i), Jensen's inequality, and convexity of $|\cdot|^{1+\delta}$ implies that for $C_n = \max_i \sum_{j \neq i} M_{ij} \to 0$

$$\sum_i \mathbb{E}\left[\left|v_{i1} \sum_{j \neq i} M_{ij} v_{j2}\right|^{1+\delta}\right] \leq \mathbb{E}\left[|v_{11}|^{1+\delta}\right] \mathbb{E}\left[|v_{12}|^{1+\delta}\right] C_n^{\delta} \sum_i \sum_{j \neq i} M_{ij} \to 0.$$

Condition (ii) implies that

$$\sum_i \mathbb{E}\left[\left(v_{i1}\sum_{j\neq i} M_{ij}v_{j2}\right)^2\right] = \mathbb{E}[v_{11}^2]\mathbb{E}[v_{12}^2]\sum_i\sum_{j\neq i} M_{ij}^2 \to 0.$$

Corollary A.5. *Suppose that* $S_n = \sum_i \sum_{j\neq i}\sum_{k\neq i,j} M_{ijk}v_{i1}v_{j2}v_{k3}$ *where* $v_i = (v_{i1}, v_{i2}, v_{i3})$. *If*

$$\mathbb{E}[\|v_1\|^2] < \infty,\ \sum_i \sum_{j\neq i}\sum_{k\neq i,j} M_{ijk}^2 \to 0,\ \textit{and}\ \sum_i \sum_{j\neq i}\sum_{k\neq i,j} M_{jik}M_{kij} \to 0,$$

then $S_n \to 0$ *in* $\mathcal{L}^1$.

For this corollary,

$$\Delta_i^0 S_n = v_{i1}\sum_{j\neq i}\sum_{k\neq i,j} M_{ijk}v_{j2}v_{k3} + v_{i2}\sum_{j\neq i}\sum_{k\neq i,j} M_{jik}v_{j1}v_{k3} + v_{i3}\sum_{j\neq i}\sum_{k\neq i,j} M_{jik}v_{j1}v_{k2}.$$

It follows that

$$\sum_i \mathbb{E}\left[\left(v_{i1}\sum_{j\neq i}\sum_{k\neq i,j} M_{ijk}v_{j2}v_{k3}\right)^2\right] = \mathbb{E}[v_{11}^2]\mathbb{E}[v_{12}^2]\mathbb{E}[v_{13}^2]\sum_i\sum_{j\neq i}\sum_{k\neq i,j} 2M_{ijk}^2 \to 0,$$

$$\sum_i \mathbb{E}\left[\left(v_{i2}\sum_{j\neq i}\sum_{k\neq i,j} M_{jik}v_{j1}v_{k3}\right)^2\right] = \mathbb{E}[v_{11}^2]\mathbb{E}[v_{12}^2]\mathbb{E}[v_{13}^2]\sum_i\sum_{j\neq i}\sum_{k\neq i,j} (M_{jik}^2 + M_{jik}M_{kij}) \to 0$$

and the third term is similar to the second.

Corollary A.6. *Suppose that* $W_n = \sum_i w_{ni}'v_i$. *If*

$$\mathbb{E}[v_1v_1'] = \Omega,\ \sum_i w_{ni}'\Omega w_{ni} = 1,\ \textit{and}\ \max_i\|w_{ni}\|^2 \to 0,$$

then $W_n \xrightarrow{d} \mathcal{N}(0,1)$.

For this corollary, $\Delta_i W_n = w_{ni}'(v_i - \tilde{v}_i)$. Condition (ii) is satisfied since $\sum_i \mathbb{E}[(\Delta_i^0 W_n)^2] = 1$ and condition (iii) follows from the argument employed in corollary $A.3$(i) applied to $\|w_{ni}\|^2\|v_i\|^2$.

For condition (i), note that $\Delta_i W_n^{[i-1]} = w_{ni}'(v_i - \tilde{v}_i)$. Thus it follows that

$$T_n = \tfrac{1}{2}\sum_i \Delta_i W \Delta_i W^{[i-1]} = \tfrac{1}{2}\sum_i \left(w_{ni}'(v_i - \tilde{v}_i)\right)^2$$

and

$$\mathbb{E}\left[T_n \mid V_n\right] = \tfrac{1}{2}\sum_i w_{ni}'\left(v_iv_i' + \Omega\right)w_{ni} = 1 + \tfrac{1}{2}\sum_i w_{ni}'\left(v_iv_i' - \Omega\right)w_{ni}.$$

It follows from corollary $A.3$ that $\mathbb{E}\left[T_n \mid V_n\right] \to 1$ in $\mathcal{L}^1$.

Corollary A.7. *Suppose that* $W_n = \sum_i \sum_{j\neq i} M_{ij} v_{i1} v_{j2}$ *where* $v_i = (v_{i1}, v_{i2})$. *Let* $\bar{M} = M^2$. *If*

$$\mathbb{E}[\|v_1\|^4] < \infty,\ \sum_i \sum_{j\neq i} M_{ij}^2 = 1,\ \max_i \sum_{j\neq i} M_{ij}^2 \to 0,\ \textit{and}\ \sum_i \sum_{j\neq i} \bar{M}_{ij}^2 \to 0,$$

then $W_n/\sigma \xrightarrow{d} \mathcal{N}(0,1)$ *where* $\sigma^2 = \mathbb{E}[v_{11}^2]\mathbb{E}[v_{12}^2] + \mathbb{E}[v_{11}v_{12}]^2$.

I first treat the case where $v_{i1} = v_{i2} := v_i$. It follows that $\Delta_i W_n = 2(v_i - \tilde{v}_i)\sum_{j\neq i} M_{ij} v_j$. Condition (ii) follows from the argument employed in corollary $A.4$(ii), and (iii) follows from

$$\sum_i \mathbb{E}\left[\left|\Delta_i^0 W_n\right|^{2+2\delta}\right] \leq 2^{3+2\delta}\mathbb{E}[|v_1|^{2+2\delta}]^2 \max_i \left(\sum_{j\neq i} M_{ij}^2\right)^{\delta} \to 0 \text{ for any } \delta \in (0,1].$$

For condition (i), note that $\Delta_i W_n^{[i-1]} = 2(v_i - \tilde{v}_i)(\sum_{j<i} M_{ij}\tilde{v}_j + \sum_{j>i} M_{ij} v_j)$. Thus, it follows that

$$\mathbb{E}\left[\Delta_i W_n \Delta_i W_n^{[i-1]} \mid V_n\right] = 4(v_i^2 + \mathbb{E}[v_i^2])\left(\sum_{j>i} M_{ij} v_j\right)\left(\sum_{j\neq i} M_{ij} v_j\right),$$

and therefore that

$$\begin{aligned}\mathbb{E}\left[T_n \mid V_n\right] &= 2\sum_i (v_i^2 + \mathbb{E}[v_i^2])\left(\sum_{j>i} M_{ij} v_j\right)\left(\sum_{j\neq i} M_{ij} v_j\right) \\ &= \sum_i \sum_{j\neq i}\sum_{k\neq i} (v_i^2 + \mathbb{E}[v_i^2]) M_{ij} v_j M_{ik} v_k.\end{aligned}$$

Split $\mathbb{E}\left[T_n \mid V_n\right]$ into three different terms

$$\begin{aligned}a_n &= \sum_i \sum_{j\neq i} (v_i^2 + \mathbb{E}[v_i^2]) M_{ij}^2 v_j^2, \\ b_n &= 2\sum_i \sum_{j\neq i} \sum_{k\neq i,j} \mathbb{E}[v_i^2] M_{ij} v_j M_{ik} v_k = 2\mathbb{E}[v_1^2] \sum_j \sum_{k\neq j} \bar{M}_{jk} v_j v_k, \\ c_n &= \sum_i \sum_{j\neq i} \sum_{k\neq i,j} (v_i^2 - \mathbb{E}[v_i^2]) M_{ij} v_j M_{ik} v_k.\end{aligned}$$

Corollary A.3(i) and corollary $A.4$(i) implies that $a_n - 2\mathbb{E}[v_1^2]^2 \xrightarrow{\mathcal{L}^1} 0$. Corollary A.4(ii) together with $\sum_i \sum_{j\neq i} \bar{M}_{ij}^2 \to 0$ leads to $b_n \xrightarrow{\mathcal{L}^1} 0$. Corollary A.5 yields $c_n \xrightarrow{\mathcal{L}^1} 0$ since

$$\sum_i \sum_{j\neq i} \sum_{k\neq i,j} M_{ij}^2 M_{ik}^2 \leq \max_i \sum_{j\neq i} M_{ij}^2 \to 0$$

and

$$\sum_i \sum_{j\neq i} \sum_{k\neq i,j} M_{ij} M_{jk}^2 M_{ik} = \sum_j \sum_{k\neq j} \bar{M}_{jk} M_{jk}^2 \leq \max_{i\neq j} |\bar{M}_{ij}| \to 0$$

When $v_{i1} \neq v_{i2}$, then $\mathbb{E}\left[T_n \mid V_n\right]$ is composed of six terms. One converges to $\mathbb{E}[v_{11}^2]\mathbb{E}[v_{12}^2]$, one converges to $\mathbb{E}[v_{11}v_{12}]^2$, and the rest converges to zero.

Corollary A.8. *Suppose that* $W_n = \sum_i w_{ni}' v_{i1} + \sum_i \sum_{j\neq i} M_{ij} v_{i2} v_{j3}$ *where* $v_i = (v_{i1}', v_{i2}, v_{i3})$. *Let* $\bar{M} = M^2$. *If*

(i) $\sigma_n^2 = \sum_i w_{ni}' \Omega w_{ni} + \sum_i \sum_{j\neq i} M_{ij}^2 \left(\mathbb{E}[v_{12}^2]\mathbb{E}[v_{13}^2] + \mathbb{E}[v_{12}v_{13}]^2\right) > c > 0$;

(ii) $\mathbb{E}[v_{11}v_{11}'] = \Omega$, $\sum_i w_{ni}' \Omega w_{ni} \leq 1$, *and* $\max_i \|w_{ni}\|^2 \to 0$;

(iii) $\mathbb{E}[v_{12}^4 + v_{13}^4] < \infty$, $\sum_i \sum_{j\neq i} M_{ij}^2 \leq 1$, $\max_i \sum_{j\neq i} M_{ij}^2 \to 0$, *and* $\sum_i \sum_{j\neq i} \bar{M}_{ij}^2 \to 0$;

(iv) $\mathbb{E}[\|v_{11}\|^2 (v_{12}^2 + v_{13}^2)] < \infty$;

then $W_n/\sigma_n \xrightarrow{d} \mathcal{N}(0,1)$.

This corollary combines the two previous ones. The only thing to verify is that $\mathbb{E}[T_n \mid V_n]$ converges, and I do so for the case where $v_{i2} = v_{i3}$ and $v_{i1} \in \mathbb{R}$. The general case follows analogously. For this corollary,

$$\Delta_i W_n = w_{ni}(v_{i1} - \tilde{v}_{i1}) + 2(v_{i2} - \tilde{v}_{i2}) \sum_{j\neq i} M_{ij} v_{j2},$$

so it follows that

$$\begin{aligned}\mathbb{E}[T_n \mid V_n] = \tfrac{1}{2} \sum_i w_{ni}^2 \left(v_{i1}^2 + \mathbb{E}[v_{i1}^2]\right) + \sum_i \sum_{j\neq i} \sum_{k\neq i} (v_{i2}^2 + \mathbb{E}v_{i2}^2) M_{ij} v_{j2} M_{ik} v_{k2} \\ + 3 \sum_i \sum_{j\neq i} (v_{i1}v_{i2} + \mathbb{E}[v_{i1}v_{i2}]) w_{ni} M_{ij} v_{j2}.\end{aligned}$$

The first two terms were treated in the previous two corollaries. The third term can be split into two parts

$$\begin{aligned} a_n &= 6\mathbb{E}[v_{11}v_{12}] \sum_i \sum_{j\neq i} w_{nj} M_{ij} v_{i2} \\ b_n &= 3 \sum_i \sum_{j\neq i} (v_{i1}v_{i2} - \mathbb{E}[v_{i1}v_{i2}]) w_{ni} M_{ij} v_{j2}. \end{aligned}$$

Corollary A.3(ii) implies that $a_n \xrightarrow{\mathcal{L}^1} 0$, since

$$\sum_i (\sum_{j \neq i} w_{nj} M_{ij})^2 = \sum_i \sum_{j \neq i} \bar{M}_{ij} w_{ni} w_{nj} + \sum_i \sum_{j \neq i} M_{ij}^2 w_{nj}^2,$$

$\sum_i \sum_{j \neq i} M_{ij}^2 w_{nj}^2 \to 0$, and

$$\sum_i \sum_{j \neq i} \bar{M}_{ij} w_{ni} w_{nj} \leq \sum_i |w_{ni}| \sqrt{\sum_{j \neq i} \bar{M}_{ij}^2} \sqrt{\sum_{j \neq i} w_{nj}^2} \leq \sum_i w_{ni}^2 \sqrt{\sum_i \sum_{j \neq i} \bar{M}_{ij}^2} \to 0.$$

Furthermore, it follows from corollary $A.4$(ii) and $\sum_i \sum_{j \neq i} w_{ni}^2 M_{ij}^2 \to 0$, that $b_n \xrightarrow{\mathcal{L}^1} 0$.

A.2 Proofs of lemmas $A.1$ and $A.2$

The proof of lemma $A.1$ makes use of the following two lemmas. The first lemma reuses an idea of (Chen, 1978, lemma 2).

Lemma A.9. *Suppose that $\{W_n\}_n$ is a sequence of random variables with variance one. If there exist a sequence $\{T_n\}_n$ of random variables such that*

(i) $\mathbb{E}[T_n \mid W_n] \to 1$ *in* $\mathcal{L}^1$*;*

(ii) $\mathbb{E}\left[W_n e^{isW_n} - isT_n e^{isW_n}\right] \to 0$ *for any* $s \in \mathbb{R}$*;*

then $W_n \xrightarrow{d} \mathcal{N}(0,1)$.

Proof of lemma A.9. Fix an $s \in \mathbb{R}$ and note that (i) and (ii) implies that

$$\mathbb{E}\left[W_n e^{isW_n} - ise^{isW_n}\right] + is\mathbb{E}\left[(1 - T_n)e^{isW_n}\right] \to 0,$$

Uniform integrability of $\{W_n\}_n$ implies that every sub-sequence $\{k\}$, has a further sub-sequence $\{m\}$, along which $\{W_m\}_m$ converges in distribution. Let W be distributed according to the limit distribution and note that uniform integrability of $\{W_m\}_m$ implies that $\mathbb{E}|W| < \infty$, and since s was arbitrary that

$$\mathbb{E}\left[We^{isW} - ise^{isW}\right] = 0 \text{ for all } s \in \mathbb{R}. \tag{9}$$

It follows from (9) that $W \sim \mathcal{N}(0,1)$ (see, e.g., Chen, 1978, lemma 2), and therefore that $W_n \xrightarrow{d} \mathcal{N}(0,1)$. □

The next lemma reuses an idea of Chatterjee (2008, lemma 2.3).

Lemma A.10. *For any measurable functions f, g of V_n with $\mathbb{E}[f^2(V_n) + g^2(V_n)] < \infty$, we have*

$$cov(g(V_n), f(V_n)) = \tfrac{1}{2} \sum_i \mathbb{E}\left[\Delta_i g \Delta_i f^{[i-1]}\right]$$

Proof. Observe that $f(V_n) - f(V_n') = \sum_i \Delta_i f^{[i-1]}$ and that $g(V)\Delta_i f^{[i-1]} \sim -g(V_n^i)\Delta_i f^{[i-1]}$. This implies that

$$\begin{aligned} cov(g(V_n), f(V_n)) &= \mathbb{E}[g(V_n)(f(V_n) - f(V_n')] = \sum_i \mathbb{E}[g(V_n)\Delta_i f^{[i-1]}] \\ &= \tfrac{1}{2} \sum_i \mathbb{E}\left[\Delta_i g \Delta_i f^{[i-1]}\right]. \end{aligned}$$

□

*Proof of lemma A.*1. The conclusion of the lemma follows from lemma *A*.9, if

$$\mathbb{E}\left[W_n e^{isW_n} - isT_n e^{isW_n}\right] \to 0$$

for any $s \in \mathbb{R}$. Lemma A.10 with $W_n = f(V_n)$ and $g(V_n) = e^{isW_n}$ and the definition of T_n implies that

$$\begin{aligned} &\mathbb{E}\left[W_n e^{isW_n} - isT_n e^{isW_n}\right] \\ &= \tfrac{1}{2} \sum_j \mathbb{E}\left[\left(e^{isW_n} - e^{isW_n(V_n^j)}\right) \Delta_j W_n^{[j-1]} - is\left(\Delta_j W_n\right)\left(\Delta_j W_n^{[j-1]}\right) e^{isW_n}\right] \\ &= -\tfrac{1}{2} \sum_j \mathbb{E}\left[\left(e^{-is\Delta_j W_n} - 1 + is\Delta_j W_n\right) e^{isW_n} \Delta_j W_n^{[j-1]}\right]. \end{aligned}$$

A mean value expansion yields

$$\left|e^{-isx} - 1 + isx\right| \leq \min\left\{2|s||x|, \frac{|s|^2}{2}|x|^2\right\}.$$

Thus, it follows for any $\delta > 0$ that (ignoring $|s|$ and s^2)

$$\begin{aligned} &\left|\mathbb{E}\left[W_n e^{isW_n} - isT_n e^{isW_n}\right]\right| \\ &\leq \sum_j \mathbb{E}\left[\left|\Delta_j W_n^{[j-1]}\right| 1_{\left\{\left|\Delta_j W_n^{[j-1]}\right| > \delta\right\}} |\Delta_j W_n| + \delta(\Delta_j W_n)^2\right] \\ &\leq \left(\sum_j \mathbb{E}\left[(\Delta_j W_n)^2 1_{\{|\Delta_j W_n| > \delta\}}\right]\right)^{1/2} \left(\sum_j \mathbb{E}\left[(\Delta_j W_n)^2\right]\right)^{1/2} \\ &+ \delta \sum_j \mathbb{E}\left[(\Delta_j W_n)^2\right]. \end{aligned}$$

This converges to zero for δ converging slowly to zero by (ii) and (iii). □

Proof of lemma A.2. The expectation of S_n exist so all expectations in the theorem and the proof are well defined and the law of iterated expectations can be applied. Let $\mathcal{F}^0$ be the trivial σ-algebra and for each $i \in [n]$, let $\mathcal{F}^i = \sigma(\{v_j\}_{j \leq i})$.

A martingale decomposition yields

$$S_n - \mathbb{E}[S_n] = \sum_i \mathbb{E}\left[S_n \mid \mathcal{F}^i\right] - \mathbb{E}\left[S_n \mid \mathcal{F}^{i-1}\right] = \sum_i \mathbb{E}\left[\Delta_i S_n \mid \mathcal{F}^i\right].$$

Fix an $\epsilon > 0$, define $\bar{\Delta}_i = \Delta_i S_n 1_{\{|\Delta_i S_n| \leq \epsilon\}}$ and use $\mathbb{E}\left[\Delta_i S_n \mid \mathcal{F}^{i-1}\right] = 0$ to write

$$\begin{aligned}\sum_i \mathbb{E}\left[\Delta_i S_n \mid \mathcal{F}^i\right] = &\underbrace{\sum_i \mathbb{E}\left[\bar{\Delta}_i \mid \mathcal{F}^i\right] - \mathbb{E}\left[\bar{\Delta}_i \mid \mathcal{F}^{i-1}\right]}_{=a_n} \\ &+ \underbrace{\sum_i \mathbb{E}\left[\Delta_i S_n - \bar{\Delta}_i \mid \mathcal{F}^i\right]}_{=b_n} - \underbrace{\sum_i \mathbb{E}\left[\Delta_i S_n - \bar{\Delta}_i \mid \mathcal{F}^{i-1}\right]}_{=c_n}.\end{aligned}$$

The summands of a_n are mean zero and uncorrelated so

$$\mathbb{E}[a_n^2] = \sum_i \mathbb{E}\left(\mathbb{E}\left[\bar{\Delta}_i \mid \mathcal{F}^i\right] - \mathbb{E}\left[\bar{\Delta}_i \mid \mathcal{F}^{i-1}\right]\right)^2 \leq \sum_i \mathbb{E}[\bar{\Delta}_i^2] \leq \epsilon^{1-\delta} \sum_i \mathbb{E}\left[|\Delta_i S_n|^{1+\delta}\right],$$

where δ is given in the first condition of the lemma, or is zero under the second condition. For b_n and c_n, it follows that

$$\mathbb{E}[|b_n| + \mathbb{E}|c_n|] \leq 2 \sum_i \mathbb{E}|\Delta_i S_n - \bar{\Delta}_i| \leq 2 \sum_i \mathbb{E}\left[|\Delta_i S_n| 1_{\{|\Delta_i S_n| > \epsilon\}}\right].$$

The first condition of the lemma implies that $\mathbb{E}[a_n^2 + |b_n| + |c_n|] \to 0$ for any $\varepsilon > 0$, and the second condition implies that $\mathbb{E}[a_n^2 + |b_n| + |c_n|] \to 0$ for ε going slowly to zero. Thus, it follows that $S_n - \mathbb{E}[S_n] \to 0$ in $\mathcal{L}^1$.

□

B Proofs of Results in Section 3

This section presents the proofs of lemma 3.1, theorem 3.2, and proposition 3.3. The proofs are somewhat modified versions of the proofs of theorems 3.1 and 3.3 in Pakes and Pollard (1989).

Proof of lemma 3.1. Fix a $\delta > 0$. $\|\hat{\theta} - \bar{\theta}\| > \delta$ and (5) implies that

$$\inf_{\substack{\theta\in\Theta_n,\\ \|\theta-\bar{\theta}\|>\delta}} \|m_n(\theta)\| \leq \inf_{\substack{\theta\in\Theta_n,\\ \|\theta-\bar{\theta}\|\leq\delta}} \|m_n(\theta)\| + o_p\left(n^{-1/2}\right). \tag{10}$$

Let $1_n = 1\{\bar{\theta} \in \Theta_n\}$, and note that $\mathbb{P}(1_n = 1) \to 1$. When $1_n = 1$, (10) and (i) yields

$$\inf_{\substack{\theta\in\Theta_n,\\ \|\theta-\bar{\theta}\|>\delta}} \|m_n(\theta)\| \leq \|m_n(\bar{\theta})\| + o_p\left(n^{-1/2}\right) = o_p(1). \tag{11}$$

It follows from (ii) that there exist a $c > 0$ such that

$$\inf_{\substack{\theta\in\Theta_n,\\ \|\theta-\bar{\theta}\|>\delta}} \|m_n(\theta)\| > c + o_p(1),$$

so (11) implies that $c + o_p(1) < o_p(1)$, which happens with probability approaching zero. This line of reasoning can also be expressed as

$$\begin{aligned}
P\left(\|\hat{\theta} - \bar{\theta}\| > \delta\right) &\leq P\left(\inf_{\substack{\theta\in\Theta_n,\\ \|\theta-\bar{\theta}\|>\delta}} \|m_n(\theta)\| \leq \inf_{\substack{\theta\in\Theta_n,\\ \|\theta-\bar{\theta}\|\leq\delta}} \|m_n(\theta)\|\right) + o(1) \\
&\leq P\left(\inf_{\substack{\theta\in\Theta_n,\\ \|\theta-\bar{\theta}\|>\delta}} \|m_n(\theta)\| \leq \|m_n(\bar{\theta})\|, 1_n = 1\right) + \mathbb{P}(1_n = 0) + o(1) \\
&\leq P\left(\inf_{\substack{\theta\in\Theta_n,\\ \|\theta-\bar{\theta}\|>\delta}} \|m_n(\theta)\| \leq \delta_n\right) + \mathbb{P}(\|m_n(\bar{\theta})\| > \delta_n) + o(1) \\
&= o(1),
\end{aligned}$$

where $\delta_n \downarrow 0$ satisfies that $\mathbb{P}(\|m_n(\bar{\theta})\| > \delta_n) \to 0$. □

Proof of theorem 3.2. Condition (iv) implies that $\sqrt{n}\Sigma_n^{-1/2}(\beta^* - \beta) \xrightarrow{d} \mathcal{N}(0, 1)$, and since $\Sigma_n > c + o_p(1)$ for some $c > 0$ it follows that $\sqrt{n}(\hat{\beta} - \beta^*) = o_p(1)$ is sufficient for the conclusions that $\sqrt{n}\Sigma_n^{-1/2}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, 1)$ and that $\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, \Sigma)$ when $\Sigma_n \xrightarrow{p} \Sigma$. The latter follows from the Cramér-Wold device. The rest of this proof shows that $\sqrt{n}\|\hat{\theta} - \theta^*\| = o_p(1)$.

The definition of θ^* implies that

$$\|m_n(\bar{\theta})\| \geq \sigma_{\min}\left(J_n(\bar{\theta})\right)\|\theta^* - \bar{\theta}\|,$$

so (ii) and (iv) leads to $\|\theta^* - \bar{\theta}\| = O_p\left(n^{-1/2}\right)$.

Let $1_n = 1\{\bar{\theta} \in \Theta_n\}$, and note that $\mathbb{P}(1_n = 1) \to 1$ by assumption. When $1_n = 1$, it follows from (5) and (iv) that

$$\|m_n(\hat{\theta})\| \leq \inf_{\theta \in \Theta_n} \|m_n(\theta)\| + o_p\left(n^{-1/2}\right) \leq \|m_n(\bar{\theta})\| + o_p\left(n^{-1/2}\right) = O_p\left(n^{-1/2}\right). \quad (12)$$

The integral form of the mean value theorem and (i) yields

$$m_n(\theta) = m_n(\bar{\theta}) + \int_0^1 J_n\left(\bar{\theta} + t(\theta - \bar{\theta})\right)\, dt \times (\theta - \bar{\theta}), \quad (13)$$

which in turn implies that

$$\begin{aligned} &\|m_n(\hat{\theta})\| + \|m_n(\bar{\theta})\| + \sigma_{\max}\left(\int_0^1 J_n\left(\bar{\theta} + t(\hat{\theta} - \bar{\theta})\right) - J_n(\bar{\theta})\, dt\right)\|\hat{\theta} - \bar{\theta}\| \quad (14) \\ &\quad \geq \sigma_{\min}\left(J_n(\bar{\theta})\right)\|\hat{\theta} - \bar{\theta}\|. \end{aligned}$$

The assumption that $\|\hat{\theta} - \bar{\theta}\| = o_p(1)$ implies that there exist some sequence $\delta_n \downarrow 0$ such that $\mathbb{P}(\|\hat{\theta} - \bar{\theta}\| \geq \delta_n) \to 0$. Redefine 1_n to also include the event that $\|\hat{\theta} - \bar{\theta}\| < \delta_n$ and note that $\mathbb{P}(1_n = 1) \to 1$. When $1_n = 1$, it follows from Jensen's inequality and monotonicity of the integral that

$$\sigma_{\max}\left(\int_0^1 J_n\left(\bar{\theta} + t(\hat{\theta} - \bar{\theta})\right) - J_n(\bar{\theta})\, dt\right) \leq \sup_{\|\theta - \bar{\theta}\| < \delta_n} \sigma_{\max}\left(J_n(\theta) - J_n(\bar{\theta})\right), \quad (15)$$

which together with (12), (14), (iii), and (iv) leads to

$$O_p\left(n^{-1/2}\right) = \|m_n(\hat{\theta})\| + \|m_n(\bar{\theta})\| \geq \left(\sigma_{\min}\left(J_n(\bar{\theta})\right) - o_p(1)\right)\|\hat{\theta} - \bar{\theta}\|.$$

Therefore $\|\hat{\theta} - \bar{\theta}\| = O_p(n^{-1/2})$ by (ii).

The result that $\|\theta^* - \bar{\theta}\| = o_p(1)$ implies that the sequence $\delta_n \downarrow 0$ can be chosen such that $\mathbb{P}(\|\theta^* - \bar{\theta}\| \geq \delta_n) \to 0$. Redefine 1_n to also include the event that $\|\theta^* - \bar{\theta}\| < \delta_n$ and note that $\mathbb{P}(1_n = 1) \to 1$. Define $L_n(\theta) = m_n(\bar{\theta}) + J_n(\bar{\theta})(\theta - \bar{\theta})$ and observe that $L_n(\theta^*) = 0$. By (13), (15), and (iii) it follows that

$$\|m_n(\theta) - L_n(\theta)\| \leq \sup_{\|\theta - \bar{\theta}\| < \delta_n} \sigma_{\max}\left(J_n(\theta) - J_n(\bar{\theta})\right)\|\theta - \bar{\theta}\| = o_p\left(\|\theta - \bar{\theta}\|\right), \quad (16)$$

when θ equals $\hat{\theta}$ or θ^* and $1_n = 1$. Thus, it follows from $\|\theta^* - \bar{\theta}\| = O_p(n^{-1/2})$ and $\|\hat{\theta} - \bar{\theta}\| = O_p(n^{-1/2})$ that $\|m_n(\hat{\theta}) - L_n(\hat{\theta})\| = o_p(n^{-1/2})$ and $\|m_n(\theta^*)\| = o_p(n^{-1/2})$.

For the last time, redefine 1_n to also include the event that $\theta^* \in \Theta_n$. Since $\|\theta^* - \bar{\theta}\| = O_p(n^{-1/2})$, $\mathbb{P}\left(\bar{\theta} \in \Theta_n\right) \to 1$, and $\sqrt{n}\inf_{\theta \in \partial\Theta_n}\|\theta - \bar{\theta}\| \xrightarrow{p} \infty$, it follows that $\mathbb{P}(1_n = 1) \to 1$. When $1_n = 1$, it follows from $\|m_n(\hat{\theta}) - L_n(\hat{\theta})\| = o_p\left(n^{-1/2}\right)$, (5), $\theta^* \in \Theta_n$, and $\|m_n(\theta^*)\| = o_p\left(n^{-1/2}\right)$ that

$$\|L_n(\hat{\theta})\| - o_p\left(n^{-1/2}\right) \leq \|m_n(\hat{\theta})\| \leq \|m_n(\theta^*)\| + o_p\left(n^{-1/2}\right) = o_p\left(n^{-1/2}\right),$$

and since $L_n(\hat{\theta}) = J_n(\bar{\theta})(\hat{\theta} - \theta^*)$ that

$$o_p\left(n^{-1/2}\right) = \|L_n(\hat{\theta})\| \geq \sigma_{\min}\left(J_n(\bar{\theta})\right)\|\hat{\theta} - \theta^*\|.$$

Thus it follows from (ii) and $\mathbb{P}(1_n = 1) \to 1$ that $\sqrt{n}\|\hat{\theta} - \theta^*\| = o_p(1)$. □

Proof of proposition 3.3. The proof is essentially the same as the proof of theorem 3.2, but the integral form of the mean value theorem is applied to M_n rather than m_n which leads to two minor differences. The first difference is that (14) becomes

$$\begin{aligned}\|m_n(\hat{\theta})\| + \|m_n(\bar{\theta})\| + \sigma_{\max}\left(\int_0^1 J_n\left(\bar{\theta} + t(\hat{\theta} - \bar{\theta})\right) - J_n(\bar{\theta})\, dt\right)\|\hat{\theta} - \bar{\theta}\| \\ + \|m_n(\hat{\theta}) - m_n(\bar{\theta}) - (M_n(\hat{\theta}) - M_n(\bar{\theta}))\| \\ \geq \sigma_{\min}\left(J_n(\bar{\theta})\right)\|\hat{\theta} - \bar{\theta}\|,\end{aligned}$$

which by (i) and (ii) implies that

$$\|m_n(\hat{\theta})\| + \|m_n(\bar{\theta})\| - o_p\left(n^{-1/2}\right) \geq \left(\sigma_{\min}\left(J_n(\bar{\theta})\right) - o_p(1)\right)\|\hat{\theta} - \bar{\theta}\|$$

when $1_n = 1$ and therefore that $\|\hat{\theta} - \bar{\theta}\| = O_p(n^{-1/2})$.

The second difference is that (16) becomes

$$\begin{aligned}\|m_n(\theta) - L_n(\theta)\| &\leq \sup_{\|\theta - \bar{\theta}\| < \delta_n} \sigma_{\max}\left(J_n(\theta) - J_n(\bar{\theta})\right)\|\theta - \bar{\theta}\| \\ &\quad + \|m_n(\theta) - m_n(\bar{\theta}) - (M_n(\theta) - M_n(\bar{\theta}))\| \\ &= o_p\left(n^{-1/2} + \|\theta - \bar{\theta}\|\right)\end{aligned}$$

when θ equals $\hat{\theta}$ or θ^* and $1_n = 1$. Following the argument in the proof of theorem 3.2, it then follows from (i) and (ii) that $\|\hat{\theta} - \theta^*\| = o_p(n^{-1/2})$. □

C Proofs of Results in Sections 4

For simplicity, I introduce the following notation. Let Y, X, ε, u, $\phi(\varepsilon(\beta))$, and $\psi(\varepsilon(\beta))$ denote the stacked observations of y_i, x_i, ε_i, u_i, $\phi_i(\beta)$, and $\psi_i(\beta)$. Additionally, let $\bar{X}(\theta) = X - \psi(\varepsilon(\beta))\gamma$ and $\bar{u} = u - \psi(\varepsilon)\gamma_0$. For two matrices A and B of the same dimensions, let $A \circ B$ be the Hadamard (entrywise) product of A and B. Furthermore, let c be some positive and finite constant that varies for each use. I repeatedly use that $P = Z(Z'Z)^{-1}Z' = n^{-1}ZZ'$ as $n^{-1}Z'Z = I_k$.

The proof of lemma 4.1 verifies the conditions of lemma 3.1.

Proof of lemma 4.1. Assumption 1(ii) implies that $\mathbb{P}(\bar{\theta} \in \Theta_n) = \mathbb{P}(|\hat{\beta}_{\text{init}} - \beta_0| \leq b_n) \to 1$.

Note that

$$\bar{\pi} = (Z'Z)^{-1}Z'(X - \psi(\varepsilon)\gamma_0) = n^{-1}Z'(X - \psi(\varepsilon)\gamma_0),$$

which implies that the first two elements of $m_n(\bar{\theta})$ equals

$$\tfrac{1}{n}\sum_i \begin{pmatrix} \phi(\varepsilon_i)z_i'\pi_0 + P_{ii}\phi(\varepsilon_i)\bar{u}_i + \sum_{j\neq i} P_{ij}\phi(\varepsilon_i)\bar{u}_j \\ \phi(\varepsilon_i)z_i'\pi_0 + \phi(\varepsilon_i)\bar{u}_i \end{pmatrix}.$$

This has mean zero conditional on Z, and a variance calculation yields

$$\begin{aligned} var\left(\tfrac{1}{n}\sum_i \phi(\varepsilon_i)z_i'\pi_0 \mid Z\right) &= \tfrac{1}{n}\mathbb{E}[\phi(\varepsilon_i)^2]\sigma_{xz} \\ var\left(\tfrac{1}{n}\sum_i \phi(\varepsilon_i)\bar{u}_i \mid Z\right) &= \tfrac{1}{n}\mathbb{E}[\phi(\varepsilon_i)^2\bar{u}_i^2] \\ var\left(\tfrac{1}{n}\sum_i\sum_{j\neq i} P_{ij}\phi(\varepsilon_i)\bar{u}_j \mid Z\right) &= \tfrac{1}{n^2}\sum_i\sum_{j\neq i} P_{ij}^2\mathbb{E}[\phi(\varepsilon_i)^2]\mathbb{E}[\bar{u}_j^2] \leq \tfrac{1}{n}\mathbb{E}[\phi(\varepsilon_i)^2]\mathbb{E}[\bar{u}_i^2] \end{aligned}$$

As each of these terms is $O_p(n^{-1})$, it follows that $\|m_n(\bar{\theta})\| = O_p(n^{-1/2})$. Thus, lemma 3.1(i) (and the first part of theorem 3.2(iv)) is satisfied.

Let $\delta > 0$ and $\{\theta_n\}_n$ be a sequence of (random) vectors such that $\theta_n \in \Theta_n \subset \mathbb{R}^{k+2}$ and $\|\theta_n - \bar{\theta}\| > \delta$ for all n. From $b_n = o_p(1)$ and $\mathbb{P}(\bar{\theta} \in \Theta_n) \to 1$ it follows that $\beta_n - \beta_0 = o_p(1)$. Thus, I can choose $\delta_n \downarrow 0$ such that $\mathbb{P}(|\beta_n - \beta_0| < \delta_n) \to 1$, which implies that I can assume for the rest of the proof that $|\beta_n - \beta_0| < \delta_n$, as the probability of the complement goes to zero. Furthermore, I note that Lipschitz continuity of ϕ and ψ, Cauchy-Schwarz' inequality,

and existence of various second moments implies that

$$\left|\tfrac{1}{n}X'(\phi(\varepsilon(\beta_n))-\phi(\varepsilon))\right|^2+\left|\tfrac{1}{n}\phi(\varepsilon(\beta_n))'\psi(\varepsilon(\beta_n))-\tfrac{1}{n}\phi(\varepsilon)'\psi(\varepsilon)\right|^2+\left\|\tfrac{1}{n}Z'(\psi(\varepsilon(\beta_n))-\psi(\varepsilon))\right\|^2$$

$$\leq \delta_n^2\tfrac{c}{n}\|X\|^2\tfrac{1}{n}\left(\|X\|^2+\|\phi(\varepsilon)\|^2+\|\psi(\varepsilon)\|^2+n\right)=O_p(\delta_n^2). \tag{17}$$

Let $\bar{m}_n$ be the last $k+1$ entries of m_n and note that for any $\theta\in\Theta_n$

$$\|m_n(\theta)\|\geq\|m_n(\theta)-m_n(\bar{\theta})\|-\|m_n(\bar{\theta})\|\geq\|\bar{m}_n(\theta)-\bar{m}_n(\bar{\theta})\|-o_p(1),$$

where $\|m_n(\bar{\theta})\|=o_p(1)$ by lemma 3.1(i). From the definition of m_n, it follows that

$$\begin{aligned}\|\bar{m}_n(\theta)-\bar{m}_n(\bar{\theta})\|^2&=\left|\tfrac{1}{n}\phi(\varepsilon(\beta))'(X-\psi(\varepsilon(\beta))\gamma)-\tfrac{1}{n}\phi(\varepsilon)'(X-\psi(\varepsilon)\gamma_0)\right|^2\\&\quad+\left\|\tfrac{1}{n}Z'(X-\psi(\varepsilon(\beta))\gamma-Z\pi)-\tfrac{1}{n}Z'(X-\psi(\varepsilon)\gamma_0-Z\bar{\pi})\right\|^2\\&\geq(\gamma-\gamma_0)^2\left|\tfrac{1}{n}\phi(\varepsilon)'\psi(\varepsilon)\right|^2\\&\quad+\left\|\bar{\pi}-\pi+\tfrac{1}{n}Z'\psi(\varepsilon)(\gamma-\gamma_0)\right\|^2-r_n(\beta,\gamma),\end{aligned}$$

where

$$\begin{aligned}r_n(\beta,\gamma)&=\left|\tfrac{1}{n}X'\left(\phi(\varepsilon(\beta))-\phi(\varepsilon)\right)\right|^2\\&\quad+((\gamma-\gamma_0)^2+\gamma_0^2)\left(\left|\tfrac{1}{n}\phi(\varepsilon(\beta))'\psi(\varepsilon(\beta))-\tfrac{1}{n}\phi(\varepsilon)'\psi(\varepsilon)\right|^2+\left\|\tfrac{1}{n}Z'(\psi(\varepsilon(\beta))-\psi(\varepsilon))\right\|^2\right)\end{aligned}$$

and (17) yields $r_n(\beta_n,\gamma_n)=\left((\gamma_n-\gamma_0)^2+1\right)O_p(\delta_n^2)$.

Define $\epsilon=\delta^2/(4\max\{1,\mathbb{E}[\psi(\varepsilon_i)^2]\})$. When $(\gamma_n-\gamma_0)^2>\epsilon$, it follows from $(\tfrac{1}{n}\phi(\varepsilon)'\psi(\varepsilon))^2=\mathbb{E}[\phi(\varepsilon_i)\psi(\varepsilon_i)]^2+o_p(1)$, that

$$\begin{aligned}\|\bar{m}_n(\theta_n)-\bar{m}_n(\bar{\theta})\|^2&\geq(\gamma_n-\gamma_0)^2\left(\mathbb{E}[\phi(\varepsilon_i)\psi(\varepsilon_i)]^2-o_p(1)-O_p(\delta_n^2)\right)-O_p(\delta_n^2)\\&\geq\epsilon\mathbb{E}[\phi(\varepsilon_i)\psi(\varepsilon_i)]^2-o_p(1).\end{aligned}$$

When $(\gamma_n-\gamma_0)^2\leq\epsilon$, it follows from $\|\tfrac{1}{n}Z'\psi(\varepsilon)\|^2\leq\tfrac{1}{n}\|\psi(\varepsilon)\|^2=\mathbb{E}[\psi(\varepsilon_i)^2]+o_p(1)$ and

$$\|\pi_n-\bar{\pi}\|^2=\|\theta_n-\bar{\theta}\|^2-(\gamma_n-\gamma_0)^2-(\beta_n-\beta_0)^2>\delta^2-\delta^2/4-\delta_n^2,$$

that

$$\begin{aligned}\|\bar{m}_n(\theta_n)-\bar{m}_n(\bar{\theta})\|^2&\geq\|\pi_n-\bar{\pi}\|^2-(\gamma_n-\gamma_0)^2\left(\mathbb{E}[\psi(\varepsilon_i)^2]+o_p(1)\right)-o_p(1)\\&\geq\delta^2/2-o_p(1).\end{aligned}$$

As $\mathbb{E}[\phi(\varepsilon_i)\psi(\varepsilon_i)] \neq 0$, it follows from these bounds that

$$\|m_n(\theta_n)\|^2 \geq \min\{\epsilon \mathbb{E}[\phi(\varepsilon_i)\psi(\varepsilon_i)]^2, \delta^2/2\} - o_p(1) > c - o_p(1).$$

Lemma 3.1(ii) follows, since $\{\theta_n\}_n$ was arbitrary.

□

I split the proof of theorem 4.2 into two parts. The first part of the proof treats the case where ϕ and ψ are continuously differentiable (theorem 3.2), and the second part covers the complications introduced when ϕ or ψ are Lipschitz continuous (proposition 3.3).

Proof of theorem 4.2*, part 1.* First, note that

$$\sqrt{n} \inf_{\theta \in \partial\Theta_n} \|\theta - \bar{\theta}\| = \sqrt{n} \inf_{\beta \in \pm b_n} |\beta + \hat{\beta}_{\text{init}} - \beta_0| \geq \sqrt{n} b_n (1 - o_p(1)) \xrightarrow{p} \infty,$$

by assumption 1(ii).

Suppose that ϕ and ψ are continuously differentiable, and note that this makes m_n continuously differentiable. Differentiation yields

$$-J_n(\theta) = \begin{bmatrix} A_n(\theta) & B_n(\theta) \\ C_n'(\theta) & I_k \end{bmatrix},$$

where

$$\underset{2\times 2}{A_n(\theta)} = \tfrac{1}{n} \begin{bmatrix} \left(X \circ \phi'(\varepsilon(\beta))\right)' Z\pi & 0 \\ \left(X \circ \phi'(\varepsilon(\beta))\right)' \bar{X}(\theta) - \gamma\phi(\varepsilon(\beta))' \left(X \circ \psi'(\varepsilon(\beta))\right) & \phi(\varepsilon(\beta))'\psi(\varepsilon(\beta)) \end{bmatrix},$$

$$\underset{2\times k}{B_n(\theta)} = \tfrac{1}{n} \begin{bmatrix} -\phi(\varepsilon(\beta))'Z \\ 0 \end{bmatrix},$$

$$\underset{2\times k}{C_n(\theta)} = \tfrac{1}{n} \begin{bmatrix} -\gamma \left(X \circ \psi'(\varepsilon(\beta))\right)' Z \\ \psi(\varepsilon(\beta))'Z \end{bmatrix}.$$

Let $A_n = A_n(\bar{\theta})$, $B_n = B_n(\bar{\theta})$, $C_n = C_n(\bar{\theta})$, and note that the singular values of $-J_n(\bar{\theta})$ are unchanged under multiplication with the matrices

$$U_l = \begin{bmatrix} I_2 & -B_n \\ 0 & I_k \end{bmatrix} \quad \text{and} \quad U_r = \begin{bmatrix} I_2 & 0 \\ -C_n' & I_k \end{bmatrix},$$

as these matrices have all eigenvalues equal to one. Multiplication leads to

$$-U_l J_n(\bar{\theta}) U_r = \begin{bmatrix} A_n - B_n C_n' & 0 \\ 0 & I_k \end{bmatrix}.$$

Thus, it can be shown that $\sigma_{\min}(J_n(\bar{\theta})) > c + o_p(1)$, provided that

$$\text{trace}\left(\left(A_n - B_n C_n'\right)'\left(A_n - B_n C_n'\right)\right) = O_p(1) \quad \text{and} \quad |\det(A_n - B_n C_n')| > c + o_p(1).$$

Multiplication yields

$$A_n - B_n C_n' = \tfrac{1}{n}\begin{bmatrix} \left(X \circ \phi'(\varepsilon)\right)' Z\bar{\pi} - \gamma_0 \phi(\varepsilon)' P \left(X \circ \psi'(\varepsilon)\right) & \phi(\varepsilon)' P \psi(\varepsilon) \\ \left(X \circ \phi'(\varepsilon)\right)' \bar{X}(\bar{\theta}) - \gamma_0 \phi(\varepsilon)' \left(X \circ \psi'(\varepsilon)\right) & \phi(\varepsilon)'\psi(\varepsilon) \end{bmatrix}.$$

Cauchy-Schwarz' inequality, $\frac{1}{n}\|Z\bar{\pi}\|^2 \leq \frac{1}{n}\|\bar{X}(\bar{\theta})\|^2$, and

$$\tfrac{1}{n}\left(\|\phi(\varepsilon)\|^2 + \|\psi(\varepsilon)\|^2 + \tfrac{1}{n}\|X \circ \phi'(\varepsilon)\|^2 + \|X \circ \psi'(\varepsilon)\|^2 + \|\bar{X}(\bar{\theta})\|^2\right) = O_p(1), \qquad (18)$$

implies that each entry of $A_n - B_n C_n'$ is $O_p(1)$ and therefore that the trace condition is satisfied.

If $|\det(A - BC)/(\phi(\varepsilon)'\psi(\varepsilon)/n)| > c + o_p(1)$, then $|\det(A_n - B_n C_n')| > c + o_p(1)$, since $|\frac{1}{n}\phi(\varepsilon)'\psi(\varepsilon)| > c + o_p(1)$.

A calculation gives

$$\begin{aligned} \frac{\det(A - BC)}{\frac{1}{n}\phi(\varepsilon)'\psi(\varepsilon)} &= \tfrac{1}{n}\left(X \circ \phi'(\varepsilon)\right)' Z\bar{\pi} - \tfrac{\phi(\varepsilon)' P \psi(\varepsilon)}{\phi(\varepsilon)'\psi(\varepsilon)} \tfrac{1}{n}\left(X \circ \phi'(\varepsilon)\right)' \bar{X}(\bar{\theta}) \\ &\quad - \tfrac{\gamma_0}{n}\left(\phi(\varepsilon)' P \left(X \circ \psi'(\varepsilon)\right) - \tfrac{\phi(\varepsilon)' P \psi(\varepsilon)}{\phi(\varepsilon)'\psi(\varepsilon)} \phi(\varepsilon)' \left(X \circ \psi'(\varepsilon)\right)\right). \end{aligned}$$

It follows from corollary A.4 that $\frac{\phi(\varepsilon)' P \psi(\varepsilon)}{\phi(\varepsilon)'\psi(\varepsilon)} = \alpha + o_p(1)$, and therefore from (18) that

$$\begin{aligned} \frac{\det(A - BC)}{\frac{1}{n}\phi(\varepsilon)'\psi(\varepsilon)} &= (1 - \alpha)\tfrac{1}{n}\sum_i \phi'(\varepsilon_i) x_i z_i' \pi_0 \\ &\quad + \tfrac{1}{n}\sum_i (P_{ii} - \alpha)\left(\phi'(\varepsilon_i) x_i \bar{u}_i - \gamma_0 \phi(\varepsilon_i) x_i \psi'(\varepsilon_i)\right) \\ &\quad + \tfrac{1}{n}\sum_i \sum_{j \neq i} P_{ij}\left(\mathbb{E}[x_j \phi'(\varepsilon_j) \mid z_j]\bar{u}_i - \gamma_0 \phi(\varepsilon_i)\mathbb{E}[x_j \psi'(\varepsilon_j) \mid z_j]\right) \\ &\quad + \tfrac{1}{n}\sum_i \sum_{j \neq i} P_{ij}\left(x_j \phi'(\varepsilon_j) - \mathbb{E}[x_j \phi'(\varepsilon_j) \mid z_j]\right)\bar{u}_i \\ &\quad - \tfrac{\gamma_0}{n}\sum_i \sum_{j \neq i} P_{ij}\phi(\varepsilon_i)\left(x_j \psi'(\varepsilon_j) - \mathbb{E}[x_j \psi'(\varepsilon_j) \mid z_j]\right) + o_p(1). \end{aligned}$$

When the conditions of corollarys $A.3$ and $A.4$ are satisfied it follows that

$$\begin{aligned}\frac{\det(A-BC)}{\frac{1}{n}\phi(\varepsilon)'\psi(\varepsilon)} &= (1-\alpha)\tfrac{1}{n}\sum_i \mathbb{E}\left[\phi'(\varepsilon_i)x_i z_i'\pi_0 \mid z_i\right] \\ &\quad + \tfrac{1}{n}\sum_i (P_{ii}-\alpha) z_i'\pi_0 \mathbb{E}\left[\phi'(\varepsilon_i)\bar{u}_i - \gamma_0\phi(\varepsilon_i)\psi'(\varepsilon_i)\right] + o_p(1) \\ &= D_n + o_p(1),\end{aligned}$$

where the second equality follows from $\frac{1}{n}\sum_i z_i'\pi_0 = 0$. To see that the conditions of the corollaries are satisfied, note that

$$\begin{aligned}\max_i \tfrac{|z_i'\pi_0|}{\sqrt{n}} &\xrightarrow{p} 0, \\ \tfrac{1}{n^2}\sum_i \left(\sum_{j\neq i} P_{ij}\mathbb{E}[x_j\phi'(\varepsilon_j) \mid z_j]\right)^2 &\leq \tfrac{\mathbb{E}[u_i\phi'(\varepsilon_i)]^2}{n} + \tfrac{1}{n^2}\sum_i (z_i'\pi_0)^2\mathbb{E}[\phi'(\varepsilon_i)]^2 \xrightarrow{p} 0, \\ \tfrac{1}{n^2}\sum_i\sum_{j\neq i} P_{ij}^2\left(1+(z_j'\pi_0)^2\right) &\leq \max_i \tfrac{1+(z_i'\pi_0)^2}{n} \xrightarrow{p} 0.\end{aligned}$$

Thus, theorem 3.2(ii) follows from $|D_n| > c + o_p(1)$.

Let $\{\delta_n\}$ be a sequence of positive numbers converging to zero and let θ_n be such that $\|\theta_n - \bar{\theta}\| \leq \delta_n$. In order to show that $\|J_n(\bar{\theta}) - J_n(\theta_n)\| = o_p(1)$, it is enough to show that

$$\|A_n(\theta_n) - A_n\| + \|B_n(\theta_n) - B_n\| + \|C_n(\theta_n) - C_n\| = o_p(1),$$

which, in turn, follows from

$$\begin{aligned}\tfrac{1}{n}\left(\|\phi(\varepsilon)\|^2 + \|\psi(\varepsilon)\|^2 + \|X\circ\phi'(\varepsilon)\|^2 + \|X\circ\psi'(\varepsilon)\|^2 + \|\bar{X}(\bar{\theta})\|^2\right) &= O_p(1), \\ \tfrac{1}{n}\|\phi(\varepsilon(\beta_n)) - \phi(\varepsilon)\|^2 + \tfrac{1}{n}\|\psi(\varepsilon(\beta_n)) - \psi(\varepsilon)\|^2 &= o_p(1), \\ \tfrac{1}{n}\left\|X\circ\left(\phi'(\varepsilon(\beta_n)) - \phi'(\varepsilon)\right)\right\|^2 + \tfrac{1}{n}\left\|X\circ\left(\psi'(\varepsilon(\beta_n)) - \psi'(\varepsilon)\right)\right\|^2 &= o_p(1).\end{aligned}$$

The first line follows (18) and the second follows in the same way as (17). Boundedness and Lipschitz continuity of ψ' and ϕ' yields that each term in the third line can be bounded by

$$\tfrac{c}{n}\sum_i x_i^2 1_{\{x_i^2>K\}} + K^2\delta_n^2.$$

It follows from assumption 1(i), that $\frac{c}{n}\sum_i x_i^2 1_{\{x_i^2>K\}} = o_p(1)$ for any $K\to\infty$. Thus I can choose $K\to\infty$ such that $K\delta_n \to 0$.

When $\det(A - BC)$ is nonzero, it follows from the partitioned inverse formula and the argument leading to theorem 3.2(ii), that the first two elements of $-J_n^1(\bar{\theta})$ equals

$$D_n^{-1}(1, -\alpha) + o_p(1).$$

The first two elements of $\sqrt{n}m_n(\bar{\theta})$ times $(1, -\alpha)$ is

$$W_n = \tfrac{1}{\sqrt{n}} \sum_i \left((1-\alpha)\phi(\varepsilon_i) z_i' \pi_0 + (P_{ii} - \alpha)\phi(\varepsilon_i)\bar{u}_i \right) + \tfrac{1}{\sqrt{n}} \sum_i \sum_{j \neq i} P_{ij}\phi(\varepsilon_i)\bar{u}_j.$$

When the conditions of corollary $A.8$ are satisfied it follows that $W_n/\sqrt{\Omega_n} \xrightarrow{d} \mathcal{N}(0,1)$ where

$$\begin{aligned}
\Omega_n &= (1-\alpha)^2 \sigma_{xz} \mathbb{E}\left[\phi(\varepsilon_i)^2\right] + \tfrac{1}{n} \sum_i (P_{ii} - \alpha)^2 \mathbb{E}\left[\phi(\varepsilon_i)^2 \bar{u}_i^2\right] \\
&\quad + 2(1-\alpha)\tfrac{1}{n} \sum_i (P_{ii} - \alpha) z_i' \pi_0 \mathbb{E}\left[\phi(\varepsilon_i)^2 \bar{u}_i\right] \\
&\quad + \tfrac{1}{n} \sum_i \sum_{j \neq i} P_{ij}^2 \mathbb{E}[\phi(\varepsilon_i)^2] \mathbb{E}[\bar{u}_j^2] \\
&= (1-\alpha)^2 \sigma_{xz} \mathbb{E}\left[\phi(\varepsilon_i)^2\right] + \alpha(1-\alpha) \mathbb{E}\left[\phi(\varepsilon_i)^2\right] \mathbb{E}\left[\bar{u}_i^2\right] \\
&\quad + 2(1-\alpha)\tfrac{1}{n} \sum_i (P_{ii} - \alpha) z_i' \pi_0 \mathbb{E}\left[\phi(\varepsilon_i)^2 \bar{u}_i\right] \\
&\quad + \tfrac{1}{n} \sum_i (P_{ii} - \alpha)^2 cov\left(\phi(\varepsilon_i)^2, \bar{u}_i^2\right) > c + o_p(1).
\end{aligned}$$

To see that the conditions of the corollary are satisfied, note that for $\bar{P}_{ij} = \frac{1}{n} \sum_{k \neq i,j} P_{ik} P_{kj}$,

$$\begin{aligned}
\max_i \tfrac{(P_{ii} - \alpha)^2 + (z_i \pi_0)^2}{n} &\xrightarrow{p} 0, \text{ and} \\
\sum_i \sum_{j \neq i} \bar{P}_{ij}^2 \leq \tfrac{c}{n^2} \sum_i \sum_{j \neq i} P_{ij}^2 \leq \tfrac{c}{n} &\to 0.
\end{aligned}$$

It follows that $-J_n^1(\bar{\theta})\sqrt{n}m_n(\bar{\theta}) = D_n^{-1} W_n + o_p(1)$ and therefore that

$$-J_n^1(\bar{\theta})\sqrt{n}m_n(\bar{\theta}) / \sqrt{D_n^{-1} \Omega_n D_n^{-1}} \xrightarrow{d} \mathcal{N}(0,1).$$

theorem 3.2(iv) follows from this.

□

For the second part of the proof, I define the function M_n used in proposition 3.3.

Definition 1. For $\omega \in \{\phi, \psi\}$, let $\omega_{\sigma_n}(\varepsilon) = \int_{\mathbb{R}} \omega(\varepsilon + \sigma_n v)\Phi'(v)\, dv$ where σ_n is a sequence of positive numbers converging to zero. Let M_n be as m_n except with ϕ and ψ replaced by ϕ_{σ_n} and ψ_{σ_n}.

The proof of the following lemma is at the end of this section.

Lemma C.1. *Under assumption 1 it follows that*

(i) ϕ_{σ_n} *and* ψ_{σ_n} *are continuously differentiable with bounded derivatives.*

(ii) For $\omega \in \{\phi, \psi, \phi', \psi'\}$, $|\omega_{\sigma_n}(\varepsilon_i) - \omega(\varepsilon_i)| \xrightarrow{a.s.} 0$ *and* $\sup_{i,n}|\omega_{\sigma_n}(\varepsilon_i) - \omega(\varepsilon_i)| \leq c$.

(iii) For $\omega \in \{\phi, \psi\}$ *and any sequence* $\{\delta_n\}$ *of positive numbers converging to zero,*

$$\mathbb{E}\left[\sup_{|\tau|\leq\delta_n}\left(\omega'_{\sigma_n}(\varepsilon_i+\tau) - \omega'_{\sigma_n}(\varepsilon_i)\right)^2\right] \leq c(\delta_n + \sqrt{\sigma_n}).$$

(iv) For $\omega \in \{\phi, \psi\}$ *and any sequence* $\{\delta_n\}$ *of positive numbers converging to zero,*

$$\sup_{|\beta-\beta_0|\leq\delta_n} \tfrac{1}{n}\left\|\omega(\varepsilon(\beta)) - \omega_{\sigma_n}(\varepsilon(\beta)) - \left(\omega(\varepsilon) - \omega_{\sigma_n}(\varepsilon)\right)\right\|^2 = o_p(1) \sup_{|\beta-\beta_0|\leq\delta_n} |\beta-\beta_0|^2.$$

Proof of theorem 4.2, part 2. It follows from lemma C.1(i),(ii), dominated convergence, and the first part of the proof that M_n satisfies theorem 3.2(i),(ii), and that the first two elements of $-J_n^1(\bar{\theta})$ equals $D_n^{-1}(1, -\alpha) + o_p(1)$. Furthermore, theorem 3.2(iv)does not depend on smoothness of m_n.

In order to show that M_n satisfies theorem 3.2(iii), I only need to redo the part of the argument that depends on Lipschitz continuity of ϕ' and ψ'. Thus, I will show that

$$\tfrac{1}{n}\left\|X \circ \left(\phi'_{\sigma_n}(\varepsilon(\beta_n)) - \phi'_{\sigma_n}(\varepsilon)\right)\right\|^2 + \tfrac{1}{n}\left\|X \circ \left(\psi'_{\sigma_n}(\varepsilon(\beta_n)) - \psi'_{\sigma_n}(\varepsilon)\right)\right\|^2 = o_p(1)$$

where $|\beta_n - \beta_0| \leq \delta_n \downarrow 0$. The first of these terms is bounded by

$$\tfrac{c}{n}\sum_i x_i^2 1_{\{x_i^2>K\}} + K \sup_{|\tau|\leq K\delta_n} \tfrac{1}{n}\sum_i \left(\phi'_{\sigma_n}(\varepsilon_i+\tau) - \phi'_{\sigma_n}(\varepsilon_i)\right)^2,$$

where the last of these terms have an expectation that is bounded by $c(K^2\delta_n + K\sqrt{\sigma_n})$ (see lemma C.1(iii)). It follows from assumption 1(i), that $\frac{c}{n}\sum_i x_i^2 1_{\{x_i^2>K\}} = o_p(1)$ for any $K \to \infty$. Thus, I choose $K \to \infty$ such that $K^2\delta_n + K\sqrt{\sigma_n} \to 0$. The term involving ψ follows analogously.

Finally, I verify proposition 3.3(i). Let $\{\delta_n\}$ be a sequence of positive numbers converging to zero, and let θ_n be such that $\|\theta_n - \bar{\theta}\| \leq \delta_n$. I show that

$$\frac{\sqrt{n}\|m_n(\theta_n) - m_n(\bar{\theta}) - (M_n(\theta_n) - M_n(\bar{\theta}))\|}{1 + \sqrt{n}\left\|\theta_n - \bar{\theta}\right\|} \leq o_p(1)\frac{\sqrt{n}\left\|\theta_n - \bar{\theta}\right\|}{1 + \sqrt{n}\left\|\theta_n - \bar{\theta}\right\|}, \tag{19}$$

which leads to proposition 3.3(i). The definition of m_n and M_n yields

$$\begin{aligned}
&\|m_n(\theta) - m_n(\bar{\theta}) - (M_n(\theta) - M_n(\bar{\theta}))\|^2 \\
&\quad \leq \left| \tfrac{1}{n} \left(\phi(\varepsilon(\beta)) - \phi_{\sigma_n}(\varepsilon(\beta)) \right)' Z'\pi - \tfrac{1}{n} \left(\phi(\varepsilon) - \phi_{\sigma_n}(\varepsilon) \right)' Z'\bar{\pi} \right|^2 \\
&\quad + 2 \left| \tfrac{1}{n} \left(\phi(\varepsilon(\beta)) - \phi_{\sigma_n}(\varepsilon(\beta)) - \phi(\varepsilon) - \phi_{\sigma_n}(\varepsilon) \right)' X \right|^2 \\
&\quad + 2 \left| \tfrac{\gamma}{n} \left(\phi(\varepsilon(\beta))'\psi(\varepsilon(\beta)) - \phi_{\sigma_n}(\varepsilon(\beta))'\psi_{\sigma_n}(\varepsilon(\beta)) \right) - \tfrac{\gamma_0}{n} \left(\phi(\varepsilon)'\psi(\varepsilon) - \phi_{\sigma_n}(\varepsilon)'\psi_{\sigma_n}(\varepsilon) \right) \right|^2 \\
&\quad + \left\| \tfrac{\gamma}{n} Z' \left(\psi(\varepsilon(\beta))) - \psi_{\sigma_n}(\varepsilon(\beta)) \right) - \tfrac{\gamma_0}{n} Z' \left(\psi(\varepsilon)) - \psi_{\sigma_n}(\varepsilon) \right) \right\|^2.
\end{aligned}$$

And repeated applications of Cauchy-Schwarz inequality makes the following sufficient for (19). For $\omega \in \{\phi, \psi\}$,

$$\begin{aligned}
\tfrac{1}{n}\|Z\pi_n\|^2 + \tfrac{1}{n}\|X\|^2 + \tfrac{1}{n}\|\omega(\varepsilon(\beta_n))\|^2 &= O_p(1) \\
\tfrac{1}{n}\left\|\omega(\varepsilon) - \omega_{\sigma_n}(\varepsilon)\right\|^2 &= o_p(1), \\
\tfrac{1}{n}\left\|\omega_{\sigma_n}(\varepsilon(\beta_n)) - \omega_{\sigma_n}(\varepsilon)\right\|^2 &= O_p(1)|\beta_n - \beta_0|^2, \\
\tfrac{1}{n}\left\|\omega(\varepsilon(\beta_n)) - \omega_{\sigma_n}(\varepsilon(\beta_n)) - \left(\omega(\varepsilon) - \omega_{\sigma_n}(\varepsilon)\right)\right\|^2 &= o_p(1)|\beta_n - \beta_0|^2.
\end{aligned}$$

The first line follows from previously made arguments (see (18)) and $\|\theta_n - \bar{\theta}\| \leq \delta_n$, the second line from lemma C.1(ii) and monotone convergence, and the third line from lemma C.1(i) and $\frac{1}{n}\|X\|^2 = O_p(1)$. The fourth line follows from lemma C.1(iii). □

Proof of lemma 4.3. The variance estimator only depends on the first two elements of J_n^1 and m_n. Let J_n^{12} be the first two elements of J_n^1, and overload notation by letting m_{ni}^s be the first two elements of m_{ni}.[6] From theorem 3.2 (ii),(iii) it follows that $J_n^{12}(\hat{\theta}) = D_n^{-1}(1, -\alpha) + o_p(1)$. To see this note that

$$\begin{aligned}
\|J_n(\hat{\theta})^{-1} - J_n(\bar{\theta})^{-1}\| &= \|J_n(\hat{\theta})^{-1} \left(J_n(\bar{\theta}) - J_n(\hat{\theta}) \right) J_n(\bar{\theta})^{-1}\| \\
&\leq \sigma_{\min}\left(J_n(\hat{\theta}) \right)^{-1} \sigma_{\min}\left(J_n(\bar{\theta}) \right)^{-1} \|J_n(\bar{\theta}) - J_n(\hat{\theta})\|.
\end{aligned}$$

theorem 3.2 (ii),(iii) and $\|\hat{\theta} - \bar{\theta}\| = o_p(1)$ implies that this is $o_p(1)$, and therefore that

[6] In the main text $m_{ni}^s(\bar{\theta})$ has the same dimension as $m_{ni}(\bar{\theta})$, but there the last k entries of $m_{ni}^s(\bar{\theta})$ are all zero.

$J_n^{12}(\hat{\theta}) - J_n^{12}(\bar{\theta}) = o_p(1)$. Let $\tilde{\Omega}$ be the following infeasible estimator of Ω_n

$$
\begin{aligned}
\tilde{\Omega} &= \tfrac{1}{n}\sum_i \left((1,-\alpha)m_{ni}^s(\bar{\theta})\right)^2 \\
&= \tfrac{1}{n}\sum_i \left(\left((1-\alpha)\phi(\varepsilon_i)z_i'\pi_0 + (P_{ii}-\alpha)\phi(\varepsilon_i)\bar{u}_i\right) + \sum_{j\neq i} P_{ij}\phi(\varepsilon_i)\bar{u}_j\right)^2 \\
&= \tfrac{1}{n}\sum_i \left((1-\alpha)\phi(\varepsilon_i)z_i'\pi_0 + (P_{ii}-\alpha)\phi(\varepsilon_i)\bar{u}_i\right)^2 \\
&+ \tfrac{2}{n}\sum_j \left(\sum_{i\neq j}\left((1-\alpha)\mathbb{E}[\phi(\varepsilon_i)^2]z_i'\pi_0 + (P_{ii}-\alpha)\mathbb{E}[\phi(\varepsilon_i)^2\bar{u}_i]\right)P_{ij}\right)\bar{u}_j \\
&+ \tfrac{1}{n}\sum_i\sum_{j\neq i} P_{ij}^2\phi(\varepsilon_i)^2\bar{u}_j^2 \\
&+ \tfrac{1}{n}\mathbb{E}[\phi(\varepsilon_i)^2]\sum_i\sum_{j\neq i}\bar{P}_{ij}\bar{u}_i\bar{u}_j \\
&+ \tfrac{2}{n}\sum_i\sum_{j\neq i}\left((1-\alpha)\left(\phi(\varepsilon_i)^2 - \mathbb{E}[\phi(\varepsilon_i)^2]\right)z_i'\pi_0 + (P_{ii}-\alpha)\left(\phi(\varepsilon_i)^2\bar{u}_i - \mathbb{E}[\phi(\varepsilon_i)^2\bar{u}_i]\right)\right)P_{ij}\bar{u}_j \\
&+ \tfrac{1}{n}\sum_i\sum_{j\neq i}\sum_{k\neq i,j} P_{ij}P_{ik}\left(\phi(\varepsilon_i)^2 - \mathbb{E}[\phi(\varepsilon_i)^2]\right)\bar{u}_j\bar{u}_k
\end{aligned}
$$

where $\bar{P}_{ij} = \sum_{k\neq i,j} P_{ik}P_{kj}$. These terms are all in a form that corollarys $A.3$, $A.4$, and $A.5$ can be applied to. Thus $\tilde{\Omega} - \Omega_n \xrightarrow{p} 0$.

Finally,

$$
\begin{aligned}
&\left\|\tfrac{1}{n}\sum_i m_{ni}^s(\bar{\theta})m_{ni}^s(\bar{\theta})' - m_{ni}^s(\hat{\theta})m_{ni}^s(\hat{\theta})'\right\| \\
&\quad\leq \left|\tfrac{1}{n}\sum_i \phi^2(\varepsilon_i)(Z_i'\bar{\pi})^2 - \phi^2(\varepsilon_i(\hat{\beta}))(Z_i'\hat{\pi})^2\right| \\
&\quad + \left|\tfrac{1}{n}\sum_i \phi^2(\varepsilon_i)(X-\gamma_0\psi(\varepsilon_i))^2 - \phi^2(\varepsilon_i(\hat{\beta}))(X-\hat{\gamma}\psi(\varepsilon_i(\hat{\beta})))^2\right|.
\end{aligned}
$$

This will be $o_p(1)$, provided that

$$
\begin{aligned}
&\left(\|\bar{\pi}-\hat{\pi}\|^2 + \tfrac{1}{n}\|\gamma_0\phi(\varepsilon_i)\psi(\varepsilon_i) - \hat{\gamma}\phi(\varepsilon_i(\hat{\beta}))\psi(\varepsilon_i(\hat{\beta}))\|^2\right)\max_i \phi(\varepsilon_i(\hat{\beta}))^2 \\
&+ \left(\|\bar{\pi}\|^2 + \|\bar{X}(\bar{\theta}\|\right)\max_i\left(\phi(\varepsilon_i) - \phi(\varepsilon_i(\hat{\beta}))\right)^2
\end{aligned}
$$

is $o_p(1)$. For the first of these terms, observe that $\max_i \phi(\varepsilon_i(\hat{\beta}))^2 = o_p(n)$ and is multiplied by an $O_p(n^{-1})$-term (as $\|\hat{\theta}-\bar{\theta}\| = O_p(n^{-1})$). For the second of these terms, observe that $\max_i\left(\phi(\varepsilon_i) - \phi(\varepsilon_i(\hat{\beta}))\right)^2 = o_p(1)$ (as $\|\hat{\theta}-\bar{\theta}\| = O_p(n^{-1})$) and is multiplied by an $O_p(1)$-term.

Combining these observations leads to $\hat{\Sigma}_n^{-1}\Sigma_n \xrightarrow{p} 1$. From theorem 4.2 we thus have

$$\frac{\sqrt{n}\left(\hat{\beta}-\beta_0\right)}{\sqrt{\hat{\Sigma}_n}} \xrightarrow{d} \mathcal{N}(0,1),$$

by the continuous mapping theorem.

□

Proof of corollary 4.4. Under (i), it follows from Cauchy-Schwarz' inequality that

$$\Omega_n = (1-\alpha)^2\sigma_{xz}\mathbb{E}\left[\phi(\varepsilon_i)^2\right] + \alpha(1-\alpha)\mathbb{E}\left[\phi(\varepsilon_i)^2\right]\mathbb{E}\left[(u_i-\psi(\varepsilon_i)\gamma_0)^2\right] + o_p(1), \quad (20a)$$

$$D_n = (1-\alpha)\sigma_{xz}\mathbb{E}\left[\phi'(\varepsilon_i)\right] + o_p(1). \quad (20b)$$

Thus it follows from $\sigma_{xz} > c + o_p(1)$, $\alpha \in [0,1)$ and $\mathbb{E}[\phi'(\varepsilon_i)] \neq 0$, that $|D_n| > c + o_p(1)$. Similarly, it follows from $\mathbb{E}[\phi'(\varepsilon_i)] \neq 0$ and $\mathbb{E}[(f'/f)^2(\varepsilon_i)] < \infty$ that $\mathbb{E}[\phi(\varepsilon_i)^2] > 0$, and therefore that $\Omega_n > c + o_p(1)$. This leads to $\Sigma_n = \Sigma_n^* + o_p(1) > c + o_p(1)$, which together with theorem 4.2 yields the conclusion.

Under (ii), it follows from integration that

$$\begin{aligned}
\mathbb{E}\left[\phi(\varepsilon_i)^2(u_i-\psi(\varepsilon_i)\gamma_0)\right] &= 0 \\
cov\left(\phi(\varepsilon_i)^2, (u_i-\psi(\varepsilon_i)\gamma_0)^2\right) &= 0 \\
\mathbb{E}\left[\phi'(\varepsilon_i)(u_i-\psi(\varepsilon_i)\gamma_0) - \gamma_0\phi(\varepsilon_i)\psi'(\varepsilon_i)\right] &= 0,
\end{aligned}$$

which yields (20a) and (20b) and a repetition of the rest of the proof. □

Proof of lemma 4.5. The inequality that

$$\inf_{\phi,\psi}\Sigma_n^*(\phi,\psi) \geq \frac{1}{\mathcal{I}_f\sigma_{xz}} + \frac{\alpha}{1-\alpha}\frac{1}{\mathcal{I}_f\sigma_{xz}}\frac{\mathbb{E}\left[(u_i-\mathbb{E}[u_i \mid \varepsilon_i])^2\right]}{\sigma_{xz}}, \quad (21)$$

follows immediately from the inequalities mentioned in the main text. Thus, it remains to be shown that the inequality binds.

When $\mathbb{E}[s(\varepsilon_i)u_i] \neq 0$, it follows that $(\phi_0,\psi_0) = (s, \mathbb{E}[u_i \mid \varepsilon_i = \cdot])$ satisfies that

$$\mathbb{E}[\phi_0(\varepsilon_i)^2] = \mathcal{I}_f < \infty,\ \mathbb{E}[\psi_0(\varepsilon_i)^2] \leq \mathbb{E}[u_i^2] < \infty,\ \text{and } \mathbb{E}[\phi_0(\varepsilon_i)\psi_0(\varepsilon_i)] = \mathbb{E}[s(\varepsilon_i)u_i] \neq 0.$$

Furthermore,

$$\gamma_0 = \frac{\mathbb{E}[u_i\phi_0(\varepsilon_i)]}{\mathbb{E}[\psi_0(\varepsilon_i)\phi_0(\varepsilon_i)]} = \frac{\mathbb{E}[\psi_0(\varepsilon_i)\phi_0(\varepsilon_i)]}{\mathbb{E}[\psi_0(\varepsilon_i)\phi_0(\varepsilon_i)]} = 1,$$

which implies that $\Sigma_n^*(\phi_0, \psi_0)$ equals the right side of (21). When $\mathbb{E}[u_i \mid \varepsilon_i] = 0$ a.s., it similarly follows that $(\phi_0, \psi_0) = (s, s)$ satisfies that

$$\mathbb{E}[\phi_0(\varepsilon_i)^2] = \mathcal{I}_f < \infty \text{ and } \mathbb{E}[\phi_0(\varepsilon_i)\psi_0(\varepsilon_i)] = \mathcal{I}_f \neq 0.$$

Furthermore, $\mathbb{E}[u_i \mid \varepsilon_i] = 0$ implies that $\gamma_0 = 0$, so $\Sigma_n^*(\phi_0, \psi_0)$ equals the right side of (21).

Now assume that $\mathbb{E}[s(\varepsilon_i)u_i] = 0$ and $\mathbb{P}(\mathbb{E}[u_i \mid \varepsilon_i] \neq 0) > 0$. For $t \in \mathbb{R}$, let

$$\phi_t = s + t\mathbb{E}[u_i \mid \varepsilon_i = \cdot] \quad \text{and} \quad \psi_0 = \mathbb{E}[u_i \mid \varepsilon_i = \cdot],$$

and observe that for any $t \neq 0$, (ϕ_t, ψ_0) satisfies that

$$\mathbb{E}[\phi_t(\varepsilon_i)^2] = \mathcal{I}_f + t^2\mathbb{E}[\mathbb{E}[u_i \mid \varepsilon_i]^2] < \infty,\ \mathbb{E}[\psi_0(\varepsilon_i)^2] \leq \mathbb{E}[u_i^2] < \infty, \text{ and}$$
$$\mathbb{E}[\phi_t(\varepsilon_i)\psi_0(\varepsilon_i)] = t\mathbb{E}[\mathbb{E}[u_i \mid \varepsilon_i]^2] \neq 0.$$

Furthermore,

$$\gamma_0 = \frac{\mathbb{E}[u_i\phi_t(\varepsilon_i)]}{\mathbb{E}[\psi_0(\varepsilon_i)\phi_t(\varepsilon_i)]} = \frac{\mathbb{E}[\psi_0(\varepsilon_i)\phi_t(\varepsilon_i)]}{\mathbb{E}[\psi_0(\varepsilon_i)\phi_t(\varepsilon_i)]} = 1$$

for any t. This implies that

$$\Sigma_n^*(\phi_t, \psi_0) = \frac{\mathbb{E}\left[\phi_t(\varepsilon_i)^2\right]}{\mathcal{I}_f^2\sigma_{xz}} + \frac{\alpha}{1-\alpha}\frac{\mathbb{E}\left[\phi_t(\varepsilon_i)^2\right]}{\mathcal{I}_f^2\sigma_{xz}}\frac{\mathbb{E}\left[(u_i - \mathbb{E}[u_i \mid \varepsilon_i])^2\right]}{\sigma_{xz}},$$

and $\lim_{t\to 0}\mathbb{E}\left[\phi_t(\varepsilon_i)^2\right] = \mathcal{I}_f$ implies that $\lim_{t\to 0}\Sigma_n^*(\phi_t, \psi_0)$ equals the right side of (21). □

Proof of proposition 4.6. Let $a = \frac{1}{\sigma_{xz}}$ and $b = \frac{\alpha}{(1-\alpha)\sigma_{xz}^2}$ be constants that do not depend on f, ϕ, or ψ. Define the feasible sets as

$$\mathcal{F} = \left\{f : f = (1-\delta)\Phi' + h,\ h \text{ is absolutely continuous },\ \mathbb{E}_f[s(\varepsilon_i)^2] \leq 1\right\}, \text{ and}$$
$$\mathcal{E} = \left\{(\phi, \psi) : \sup_{f\in\mathcal{F}}\mathbb{E}[\phi(\varepsilon_i)^2] < \infty,\ \sup_{f\in\mathcal{F}}\mathbb{E}[\psi(\varepsilon_i)^2] < \infty,\ \inf_{f\in\mathcal{F}}\mathbb{E}[\phi(\varepsilon_i)\psi(\varepsilon_i)]^2 > 0\right\}.$$

For any $(\phi, \psi) \in \mathcal{E}$ and $\omega \in \{\phi, \psi\}$, let

$$J(\omega, f) = \frac{\mathbb{E}_f[\omega(\varepsilon_i)^2]}{\mathbb{E}_f[\omega(\varepsilon_i)s(\varepsilon_i)]^2}, \text{ and}$$
$$\tilde{\Sigma}_n(\phi, \psi, f) = J(\phi, f)\left(a + b\left(2 - J(\psi, f)^{-1}\right)\right).$$

If the following four statements are correct, then the proposition follows immediately. (i) For any $(\phi, \psi) \in \mathcal{E}$ and $f \in \mathcal{F}$, $\Sigma_n^*(\phi, \psi, f) \geq \tilde{\Sigma}_n(\phi, \psi, f)$. (ii) For any $(\phi, \psi) \in \mathcal{E}$,

$$\sup_{f\in\mathcal{F}}\tilde{\Sigma}_n(\phi, \psi, f) \geq \tilde{\Sigma}_n(\phi, \psi, f_0) \geq \tilde{\Sigma}_n(\phi_{\nu_0}, \phi_{\nu_0}, f_0) = \sup_{f\in\mathcal{F}}\tilde{\Sigma}_n(\phi_{\nu_0}, \phi_{\nu_0}, f).$$

(iii) For any $f \in \mathcal{F}$, $\Sigma_n^*(\phi_{\nu_0}, \phi_{\nu_0}, f) = \tilde{\Sigma}_n(\phi_{\nu_0}, \phi_{\nu_0}, f)$. (iv) $(\phi_{\nu_0}, \phi_{\nu_0}) \in \mathcal{E}$.

First, note that (iv) is satisfied since $2\nu_0^2\Phi(-\nu_0) \leq \mathbb{E}_f[\phi_{\nu_0}(\varepsilon_i)^2] \leq 1 + \delta\nu_0^2$. For (i), note that a simple calculation and $\mathbb{E}_f[u_i \mid \varepsilon_i] = s(\varepsilon_i)$ yields

$$\begin{aligned}\mathbb{E}_f[(u_i - \psi(\varepsilon_i)\gamma_0)^2] &= \mathbb{E}_f[(u_i - \psi(\varepsilon_i)\gamma_1)^2] + (\gamma_0 - \gamma_1)^2\mathbb{E}_f[\psi(\varepsilon_i)^2] \\ &= 2 - J(\psi, f)^{-1} + (\gamma_0 - \gamma_1)^2\mathbb{E}_f[\psi(\varepsilon_i)^2],\end{aligned}$$

where

$$\gamma_1 = \frac{\mathbb{E}_f[u_i\psi(\varepsilon_i)]}{\mathbb{E}_f[\psi(\varepsilon_i)^2]} = \frac{\mathbb{E}_f[s(\varepsilon_i)\psi(\varepsilon_i)]}{\mathbb{E}_f[\psi(\varepsilon_i)^2]}.$$

(i) follows from positivity of $(\gamma_0 - \gamma_1)^2\mathbb{E}_f[\psi(\varepsilon_i)^2]$ and (iii) from the observation that $\gamma_0 = \gamma_1$ when $\phi = \psi$. The first inequality of (ii) follows if $f_0 \in \mathcal{F}$. To see that this is the case observe that

$$\begin{aligned}\int_{\mathbb{R}} f_0(\varepsilon)\, d\varepsilon &= (1-\delta)\left[1 - 2\Phi(-\nu_0) + \frac{2\Phi'(\nu_0)}{\nu_0}\right] = 1, \\ \mathbb{E}_{f_0}[s(\varepsilon_i)^2] &= \mathbb{E}_{f_0}[1\{|\varepsilon_i| \leq \nu_0\}] = (1-\delta)\left(\Phi(\nu_0) - \Phi(-\nu_0)\right) \leq 1,\end{aligned}$$

and $f_0(\varepsilon) \geq (1-\delta)\Phi'(\varepsilon)$. The last two parts of (ii) follows from (Huber, 1964, theorem 1) or at this stage from the observation that $\phi_{\nu_0} = f_0'/f_0$ and that for any $f \in \mathcal{F}$,

$$J(\phi_{\nu_0}, f) \leq \frac{(1-\delta)\mathbb{E}_{\Phi'}[\phi_{\nu_0}(\varepsilon_i)^2] + \delta\nu_0^2}{\mathbb{E}_{\Phi'}[\phi_{nu_0}(\varepsilon_i)s(\varepsilon_i)]^2} = J(\phi_{\nu_0}, f_0).$$

This implies that

$$\sup_{f\in\mathcal{F}} J(\phi_{nu_0}, f) = J(\phi_{nu_0}, f_0) = \inf_{(\phi,\psi)\in\mathcal{E}} J(\phi, f_0) = \inf_{(\phi,\psi)\in\mathcal{E}} J(\psi, f_0),$$

which is a saddle point result that implies the last two parts of (ii). □

Proof of lemma C.1. I only perform the proof for statements about ϕ as statements about ψ follow analogously. The set A refers to the finite number of points of non-differentiability of ϕ and ψ.

(i) A change of variables, differentiation of φ, and $\int_{\mathbb{R}} v\varphi(v)\, dv = 0$ leads to

$$\phi'_{\sigma_n}(\varepsilon) = \int_{\mathbb{R}} \frac{\phi(\varepsilon + \sigma_n v) - \phi(\varepsilon)}{\sigma_n v} v^2\varphi(v)\, dv, \tag{22}$$

and Lipschitz continuity of ϕ leads to boundedness and continuity of ϕ'_{σ_n}.

(ii) $\int_{\mathbb{R}} \varphi(v)\, dv = 1$ and $\int_{\mathbb{R}} v^2\varphi(v)\, dv = 1$ yields

$$|\phi_{\sigma_n}(\varepsilon) - \phi(\varepsilon)| = \left| \int_{\mathbb{R}} \left(\phi(\varepsilon + \sigma_n v) - \phi(\varepsilon)\right) \varphi(v)\, dv \right| \leq c\sigma_n,$$

and

$$|\phi'_{\sigma_n}(\varepsilon) - \phi'(\varepsilon)| = \left| \int_{\mathbb{R}} \left(\frac{\phi(\varepsilon + \sigma_n v) - \phi(\varepsilon)}{\sigma_n v} - \phi'(\varepsilon) \right) v^2 \varphi(v)\, dv \right|.$$

Lipschitz continuity of ϕ implies that $|\frac{\phi(\varepsilon+\sigma_n v)-\phi(\varepsilon)}{\sigma_n v} - \phi'(\varepsilon)| < c$ and for any $\varepsilon \notin A$ and $v \in \mathbb{R}$ it follows from differentiability of ϕ that $|\frac{\phi(\varepsilon+\sigma_n v)-\phi(\varepsilon)}{\sigma_n v} - \phi'(\varepsilon)| \to 0$. Dominated convergence leads to $|\phi'_{\sigma_n}(\varepsilon) - \phi'(\varepsilon)| \to 0$ for any $\varepsilon \notin A$.

(iii) Fix $\delta_n \downarrow 0$, let $A_{\delta_n} = \{x \in \mathbb{R} : \min_{a\in A} |x - a| \leq \delta_n\}$, $d_{\delta_n,\varepsilon} = \min_{a\in A_{\delta_n}} |\varepsilon - a|$, and consider ε with $d_{\delta_n,\varepsilon} \geq \delta_n + \sqrt{\sigma_n}$. From (22), $\int_{\mathbb{R}} v\varphi(v)\, dv = 0$, and Lipschitz continuity of ϕ it follows that

$$\phi'_{\sigma_n}(\varepsilon + \tau) - \phi'_{\sigma_n}(\varepsilon) = \int_{\mathbb{R}} \frac{\phi(\varepsilon + \sigma_n v + \tau) - \phi(\varepsilon + \sigma_n v)}{\sigma_n v} v^2\varphi(v)\, dv \tag{23a}$$

$$= \int_{\mathbb{R}} \int_0^\tau \frac{\phi'(\varepsilon + t + \sigma_n v) - \phi'(\varepsilon + t)}{\sigma_n v}\, dt\; v^2\varphi(v)\, dv. \tag{23b}$$

Use (23b) when $|\sigma_n v| \leq d_{\delta_n,\varepsilon}$ and (23a) when $|\sigma_n v| > d_{\delta_n,\varepsilon}$ to write

$$\left|\phi'_{\sigma_n}(\varepsilon + \tau) - \phi'_{\sigma_n}(\varepsilon)\right| \leq c\tau + c \int_{|v|>d_{\delta_n,\varepsilon}/\sigma_n} v^2\varphi(v)\, dv \leq c\delta_n + \tfrac{c}{\sigma_n} e^{-c/\sigma_n}.$$

As $\left|\phi'_{\sigma_n}(\varepsilon + \tau) - \phi'_{\sigma_n}(\varepsilon)\right|$ is bounded it follows that

$$\mathbb{E}\left[\sup_{|\tau|\leq\delta_n} \left(\phi'_{\sigma_n}(\varepsilon_i + \tau) - \phi'_{\sigma_n}(\varepsilon_i)\right)^2 \right] \leq c\delta_n^2 + \tfrac{c}{\sigma_n^2} e^{-c/\sigma_n} + cP\left(d_{\delta_n,\varepsilon_i} < \delta_n + \sqrt{\sigma_n}\right).$$

Furthermore,

$$\begin{aligned} P\left(d_{\delta_n,\varepsilon_i} < \delta_n + \sqrt{\sigma_n}\right) &\leq \sum_{a\in A} P\left(|\varepsilon_i - a| \leq 2\delta_n + \sqrt{\sigma_n}\right) \\ &\leq 2 \sup_{\varepsilon\in\mathbb{R}} f(\varepsilon) \sum_{a\in A} \left(2\delta_n + \sqrt{\sigma_n}\right) \\ &\leq c\left(\delta_n + \sqrt{\sigma_n}\right). \end{aligned}$$

This leads to a bound on the expectation of $c(\delta_n + \sqrt{\sigma_n})$ as δ_n and σ_n goes to zero.

(iv) Fix $\delta_n \downarrow 0$ and let $|\beta_n - \beta_0| \leq \delta_n$. Lipschitz continuity of ϕ and ϕ_{σ_n} leads to

$$\phi(\varepsilon_i(\beta)) - \phi_{\sigma_n}(\varepsilon_i(\beta)) - \Big(\phi(\varepsilon_i) - \phi_{\sigma_n}(\varepsilon_i)\Big) = x_i \int_0^{\beta-\beta_0} \phi'(\varepsilon_i - \tau x_i) - \phi'_{\sigma_n}(\varepsilon_i - \tau x_i)\, d\tau,$$

so boundedness of ϕ' and ϕ'_{σ_n} yields

$$\begin{aligned}&\tfrac{1}{n}\Big\|\phi(\varepsilon(\beta_n)) - \phi_{\sigma_n}(\varepsilon(\beta_n)) - \Big(\phi(\varepsilon) - \phi_{\sigma_n}(\varepsilon)\Big)\Big\|^2 \\ &\quad \leq |\beta_n - \beta_0| \left(\tfrac{c}{n} \sum_i x_i^2 1_{\{x_i^2 > K\}} + K^2 \sup_{|\tau| \leq K\delta_n} \tfrac{1}{n} \sum_i |\phi'(\varepsilon_i + \tau) - \phi'_{\sigma_n}(\varepsilon_i + \tau)|^2 \right).\end{aligned}$$

Assumption 1(i) implies that $\frac{c}{n}\sum_i x_i^2 1_{\{x_i^2 > K\}} = o_p(1)$ provided that $K \to \infty$. As in (iii), let $\tilde{\delta}_n = K\delta_n$, $A_{\tilde{\delta}_n} = \{x \in \mathbb{R} : \min_{a \in A} |x - a| \leq \tilde{\delta}_n\}$, $d_{\tilde{\delta}_n, \varepsilon} = \min_{a \in A_{\delta_n}} |\varepsilon - a|$, and consider ε with $d_{\delta_n,\varepsilon} \geq \sqrt{\sigma_n}$ and $\tau \leq \delta_n$.

From (22), $\int_{\mathbb{R}} v^2 \varphi(v)\, dv = 1$, and Lipschitz continuity of ϕ it follows that

$$\phi'(\varepsilon + \tau) - \phi'_{\sigma_n}(\varepsilon + \tau) = \int_{\mathbb{R}} \int_0^{\sigma_n v} \frac{\phi'(\varepsilon + \tau) - \phi'(\varepsilon + \tau + s)}{\sigma_n v} v^2 \varphi(v)\, dv.$$

Use Lipschitz continuity of ϕ' when $|\sigma_n v| \leq d_{\tilde{\delta}_n,\varepsilon}$ and boundedness of ϕ' when $|\sigma_n v| > d_{\delta_n,\varepsilon}$ to write

$$|\phi'(\varepsilon + \tau) - \phi'_{\sigma_n}(\varepsilon + \tau)| \leq c\sigma_n + c \int_{|v| > d_{\tilde{\delta}_n,\varepsilon}/\sigma_n} v^2 \varphi(v)\, dv \leq c\sigma_n + \tfrac{c}{\sigma_n} e^{-c/\sigma_n}.$$

As in (iii), it follows from boundedness of $\left|\phi'(\varepsilon + \tau) - \phi'_{\sigma_n}(\varepsilon + \tau)\right|$ that

$$\begin{aligned}\mathbb{E}\left[\sup_{|\tau| \leq \delta_n} \Big(\phi'(\varepsilon_i + \tau) - \phi'_{\sigma_n}(\varepsilon_i + \tau)\Big)^2\right] &\leq c\sigma_n^2 + \tfrac{c}{\sigma_n^2} e^{-c/\sigma_n} + cP\left(d_{\tilde{\delta}_n,\varepsilon_i} < \sqrt{\sigma_n}\right) \\ &\leq c(K\delta_n + \sqrt{\sigma_n}).\end{aligned}$$

Choosing $K \to \infty$ such that $K^2(K\delta_n + \sqrt{\sigma_n}) \to 0$ yields the desired conclusion.

□

D Interpretation of LIML as a GMM Estimator

This section verifies two claims made in section 2. The first claim is that the minimizer and maximizer of $Q_n(\beta) = \varepsilon(\beta)' P \varepsilon(\beta) / \varepsilon(\beta)'\varepsilon(\beta)$ are minimizers of $\|m_n(\theta)\|$ when $\phi(\varepsilon) =$

$\psi(\varepsilon) = \varepsilon$. The second claim is that under certain conditions, the maximizer of $Q_n(\beta)$ does not converge to β_0.

For the first claim let $\hat{\beta}$ be the minimizer or maximizer of $Q_n(\beta)$ and define

$$\hat{\gamma} = \frac{\varepsilon(\hat{\beta})'X}{\varepsilon(\hat{\beta})'\varepsilon(\hat{\beta})} \quad \text{and} \quad \hat{\pi} = (Z'Z)^{-1}Z'(X - \hat{\gamma}\varepsilon(\hat{\beta})),$$

then $\|m_n(\theta)\| = 0$ if $\varepsilon(\hat{\beta})'Z'\hat{\pi} = 0$. However, $\varepsilon(\hat{\beta})'Z'\hat{\pi} = \varepsilon(\hat{\beta})'PX - Q_n(\hat{\beta})\varepsilon(\hat{\beta})'X$, which is proportional to the derivative of $Q_n(\beta)$ and therefore zero at $\hat{\beta}$.

For the second claim suppose that the sample is *i.i.d.*, σ_{xz} is constant in n, $\mathbb{E}[\varepsilon_i^2]+\mathbb{E}[u_i^2] < \infty$, $\sigma_\varepsilon^2 \neq 0$, and $\sigma_{u\varepsilon} \neq 0$. One can show that under these conditions $\sup_\beta |Q_n(\beta) - Q(\beta)| = o_p(1)$ where

$$Q(\beta) = \frac{(\beta - \beta_0)^2\sigma_{xz} + \alpha\mathbb{E}\left[(\varepsilon_i - (\beta - \beta_0)u_i)^2\right]}{(\beta - \beta_0)^2\sigma_{xz} + \mathbb{E}\left[(\varepsilon_i - (\beta - \beta_0)u_i)^2\right]},$$

and some calculus shows that $Q(\beta)$ is maximized at $\beta_0 + \frac{\sigma_\varepsilon^2}{\sigma_{u\varepsilon}}$ (also $Q(\beta)$ is minimized at β_0). Finally,

$$\lim_{|\beta|\to\infty} Q(\beta) = \frac{\sigma_{xz} + \alpha\sigma_u^2}{\sigma_{xz} + \sigma_u^2} < \frac{\sigma_{xz} + \alpha\sigma_u^2(1-\rho^2)}{\sigma_{xz} + \sigma_u^2(1-\rho^2)} = Q\left(\beta_0 + \frac{\sigma_\varepsilon^2}{\sigma_{u\varepsilon}}\right)$$

where ρ is the correlation between ε and u. Thus, the maximizer of $Q_n(\beta)$ converges in probability to $\beta_0 + \frac{\sigma_\varepsilon^2}{\sigma_{u\varepsilon}}$.