Conditional Moment Models under Weak Identification

Bertille Antoine* and Pascal Lavergne†

September 20, 2011

Abstract

We consider models defined by a set of conditional moment restrictions where weak identification may arise. Weak identification is directly defined through the conditional moments that are allowed to flatten as the sample size increases. We propose a minimum distance estimator of the structural parameters that is robust to potential weak identification and that uses neither instrumental variables nor smoothing. Hence, its properties only depend upon identification weakness, and not on the interplay between some tuning parameter, as the growth rate of the number of instruments, and the unknown degree of weakness. Our estimator is consistent and asymptotically normal, and its rate of convergence is the same as competing estimators based on many weak instruments. Heteroskedasticity-robust inference is possible through Wald testing without prior knowledge of the identification pattern. In simulations, we find that our estimator is competitive with estimators based on many instruments.

Keywords: Weak identification, Heteroskedasticity, Conditional moments, Minimum distance estimation, Hypothesis testing.

JEL classification: Primary C130 ; Secondary C120.

The authors gratefully acknowledge financial support from the Social Sciences and Humanities Research Council of Canada. This work was presented at the $(EC)^2$ conference (Toulouse), the second French Econometrics Conference (Paris), the conference in Honor of H.L. White Jr (San Diego), ESEM (Oslo), CESG (Toronto), and in seminars at the University of Chicago (Department of Statistics) and Simon Fraser University. We thank participants and colleagues for helpful comments.

Correspondence to: Bertille Antoine, Department of Economics, 8888 University Drive, Burnaby, BC V5A1S6, Canada. Email: bertille_antoine@sfu.ca. Phone: (1)-778-782-4514.

*Simon Fraser University. Email: bertille_antoine@sfu.ca.

†Toulouse School of Economics. Email: pascal.lavergne@univ-tlse1.fr

1 Introduction

In many econometric models with endogenous variables Y and exogenous variables X, the parameter value is identified through restrictions of the form

$$\mathrm{E}\left[g(Z_i, \theta_0)|X_i\right] = \mathbf{0} \quad \text{a.s.} \tag{1.1}$$

where $g(Z_i, \theta)$ is a known function of the random vector of observations $Z_i = (Y_i', X_i')' \in \mathbb{R}^{d+q}$ and of the structural parameter of interest $\theta \in \Theta \subset \mathbb{R}^p$. We are interested in providing reliable inference on the parameter θ_0 in cases where its identification may be weak. Intuitively, identification is weak when information about some, or all, components of θ_0 accumulates at rates slower than the square-root of the sample size n. In the literature, it is customary to transform (1.1) into K unconditional moment restrictions

$$\mathrm{E}\left[A(X_i)g(Z_i, \theta_0)\right] = \mathbf{0}\,,$$

where the instruments $A(X_i) = (A_1(X_i), \ldots, A_K(X_i))$ are user-chosen functions of X_i. In such a setting, weak identification is modeled by assuming that these unconditional moments flatten around the true value as the sample size increases. Our first contribution is to model weak identification directly through the conditional moments $\mathrm{E}\left[g(Z_i, \theta)|X_i\right]$. We essentially allow the conditional moments themselves to flatten. From there, our second contribution is to propose a simple weighted minimum distance estimator, which directly relies on the conditional moment restrictions (1.1), and is robust to weak identification.

Staiger and Stock (1997) are first to study weak identification in the context of a linear IV regression. They assume that the moments $\mathrm{E}\left[A(X_i)g(Z_i, \theta)\right]$ tend to zero at rate $\sqrt{n}$ around θ_0, and they show that parameters cannot be estimated consistently. However, Hansen, Hausman and Newey's (2008) survey of the applied literature suggests that such a modeling may not be the most suitable for microeconometric applications. Hence, many authors rather consider a rate of decay to zero slower than $\sqrt{n}$. In this situation, one recovers consistency and asymptotic normality in estimation, though at a slower than parametric rate. Hahn and Kursteiner (2002) study IV estimators in a linear model, while Caner (2010) and Antoine and Renault (2009, 2010a) consider the nonlinear case. To gain efficiency, many authors consider a growing number of instruments K. Chao and Swanson (2005), Stock and Yogo (2005) and Hansen, Hausman, and Newey (2008) study k-class IV estimators in linear models with many weak instruments. Han and Phillips (2006) study consistency and asymptotic distribution of GMM estimators with many weak instruments in nonlinear models. Newey and Windmeijer (2009) study the Continuously Updated GMM Estimator (CUE) and other Generalized Empirical Likelihood estimators. A key finding of these papers is that the estimators' properties, including consistency, are contingent upon the relative standing of the number of instruments and the identification weakness. For instance, in a linear model with endogenous regressors and many weak instruments, 2SLS is inconsistent if the number of instruments is too large, see Chao and Swanson (2005). Han and Philips (2006) give a general thorough discussion of the conditions under which GMM estimators are consistent. Jun and Pinkse (2011) consider a semiparametric estimator that uses smoothing, and find

that tuning the smoothing parameter also affects the consistency and efficiency of their estimator. Therefore, in applications, the number of instruments, or the smoothing parameter, is likely to have strong influence on empirical results. Since, in practice, we never know the extent of identification weakness, selecting the number of instruments appears a delicate task.

In this work, we define weak identification directly through conditional moments. Specifically, we assume that they can be additively decomposed into different components, each local-to-zero around θ_0 at a specific unknown rate. As a result, various identification patterns appear for different components of the parameter. We then study a new Weighted Minimum Distance (WMD) estimator that directly exploits the conditional restrictions (1.1), thus avoiding the use of instruments. As a result, our estimator's properties only depend upon identification weakness. Our estimator is asymptotically normal with the same convergence rate as the one obtained in recent literature for estimators using many instruments. Heteroskedasticity-robust confidence intervals and tests can easily be built without a priori knowledge of identification weakness. The price to pay for our estimator's simplicity is a possibly greater dispersion compared to existing estimators. Our simulations results of Section 5 reveal that this price is often reasonable and that, overall, our estimator is competitive.

The paper is organized as follows. In Section 2, we focus on a linear model with endogenous regressors for expository purposes. In this context, we introduce our definition of weak identification based on conditional moments and our WMD estimator that is robust to such identification weakness. Our estimator resembles a k-class estimator, so that its computation is fast, and inference is straightforward. In Section 3, we expose our general framework and estimation method. In Section 4, we develop asymptotic theory for estimation and heteroskedasticity-robust inference. In Section 5, we study the small sample performance of our estimator through Monte-Carlo simulations and compare it to estimators recently proposed by Hausman, Newey, Woutersen, Chao and Swanson (2009). Section 6 concludes and Section 7 gathers our technical proofs.

2 Linear model with endogenous regressors

Let us consider the structural equation

$$y_i = \alpha_0 + Y_{1i}'\beta_0 + \varepsilon_i \qquad \mathrm{E}\left[\varepsilon_i | X_i\right] = 0\,,$$

for $i = 1, \dots n$ i.i.d. observations, where the exogenous X_i are continuous. This type of model is often encountered in microeconometrics, with possibly additional exogenous variables that are subsumed in the intercept for expository purposes. Formal identification of the structural parameter $\theta_0 = (\alpha_0 \ \beta_0')'$ depends on the conditional moment

$$\mathrm{E}\left[y_i - \alpha - Y_{1i}'\beta | X_i\right] = (\alpha_0 - \alpha) + \mathrm{E}\left[Y_{1i}' | X_i\right](\beta_0 - \beta)\,. \tag{2.1}$$

We assume global identifiability, i.e. that (2.1) is zero almost surely only for $\theta = \theta_0$. While the intercept is always well-identified, we consider that β_0 might be weakly identified. A formal definition of weak

identification does not require to transform the conditional moment restriction (2.1) into unconditional ones. Following the pioneer work of Staiger and Stock (1997), weak identification is modeled as a decreasing correlation between Y_{1i} and X_i, or equivalently between $\mathrm{E}[Y_{1i}|X_i]$ and X_i. Indeed, identification strength of β_0 directly depends on $\mathrm{E}[Y_{1i}|X_i]$, as seen from (2.1). We assume that $\mathrm{E}[Y_{1i}|X_i]$ flattens to zero, so that our reduced equation for Y_{1i} is

$$Y_{1i} = \frac{f(X_i)}{r_n} + U_i\,, \qquad \mathrm{E}[U_i|X_i] = 0\,,$$

with $r_n \to \infty$. Identification strength of β_0 thus depends on the unknown rate r_n.

For estimation purposes, since the vector of functions $f(\cdot)$ is unknown, it is customary to introduce approximating functions $A(X_i) = (A_1(X_i), \ldots, A_K(X_i))$, such as power series or splines, and use them as instruments. This yields K unconditional moments

$$\mathrm{E}\left[A(X_i)\left(y_i - \alpha - Y_{1i}'\beta\right)\right]\,.$$

The parameter θ_0 can then be estimated by 2SLS, GMM, LIML, or any variant of these methods. The number of instruments can be allowed to grow with the sample size to gain efficiency, as in Chao and Swanson (2005), Stock and Yogo (2005), Han and Phillips (2006), Hansen, Hausman, and Newey (2008), and Hausman et al. (2009). A feature of such procedures is that the properties of the associated estimators, including consistency, depend on the chosen number of instruments, and, specifically, on the interplay between the growth rate of the number of instruments and the unknown degree of weakness. Inference can be entertained through t-tests, provided one uses corrected standard errors that account for the presence of many instruments.

Our estimation procedure directly relies on the conditional moment restrictions, and thus avoids practical choices that can have key consequences on the estimators' properties. We motivate our estimator in Section 3.2. For now, we define it for the linear model and discuss its main properties. Let e be the $n \times 1$ vector of ones, $Y_1 = (Y_{11}, \ldots Y_{1n})'$, and $Y_1^* = [e\ Y_1]$. Our WMD estimator of θ_0 is

$$\tilde{\theta}_n = \arg\min_{\theta\in\Theta} \left[\frac{(y - Y_1^*\theta)'\tilde{K}(y - Y_1^*\theta)}{(y - Y_1^*\theta)'(y - Y_1^*\theta)}\right]\,, \tag{2.2}$$

where $\tilde{K}$ is the matrix of size n with diagonal elements equal to zero and off-diagonal elements $K_{ij} \equiv K(X_i - X_j)$ for some multivariate density $K(\cdot)$; see the next section for examples.[1] The above minimization problem can be easily solved and our WMD estimator explicitly writes

$$\tilde{\theta}_n = \left[Y_1^{*\prime}(\tilde{K} - \tilde{\lambda}\mathbf{I}_n)Y_1^*\right]^{-1}\left[Y_1^{*\prime}(\tilde{K} - \tilde{\lambda}\mathbf{I}_n)y\right]\,, \tag{2.3}$$

[1] It has been rightly suggested to us that the observations X_i may be scaled before being passed as arguments to $K(\cdot)$ to get scale invariance. Though we do not formally investigate this possibility, this can be done without affecting the main properties of our estimator.

where $\mathbf{I}_n$ is the identity matrix of size n. Here, $\tilde{\lambda}$ is the minimum value of the objective function, which can be explicitly computed as the smallest eigenvalue of $(Y^{*'}Y^*)^{-1}(Y^{*'}\tilde{K}Y^*)$ with $Y^* = [y\ Y_1^*]$. Details are provided in the Appendix.

Our WMD estimator is consistent under weak identification as soon as $r_n = o(\sqrt{n})$. Its global rate of convergence is $r_n^{-1}\sqrt{n}$, the same as the one of previously proposed estimators under many weak instrument asymptotics. Moreover, as we will show in Section 4.1,

$$\begin{bmatrix} \sqrt{n}(\tilde{\alpha}_n - \alpha_0) \\ r_n^{-1}\sqrt{n}(\tilde{\beta}_n - \beta_0) \end{bmatrix}$$

is asymptotically normally distributed. Hence, as could be expected, the first component $\tilde{\alpha}_n$ is $\sqrt{n}$-asymptotically normal, while the remaining components $\tilde{\beta}_n$ are asymptotically normal with rate $r_n^{-1}\sqrt{n}$. Moreover, its asymptotic variance does not depend on higher moments.

Our estimator (2.3) resembles a k-class estimator, as the Fuller and LIML-like estimators recently proposed by Hausman et al. (2009), with the key difference that our WMD estimator does not use projection on instruments. Nonetheless, it retains the computational simplicity of this family of estimators. Its variance can be simply approximated in a heteroskedasticity-robust way, using an Eicker-White-type formula. Namely, one can use

$$\left[Y_1^{*'}(\tilde{K} - \tilde{\lambda}\mathbf{I}_n)Y_1^*\right]^{-1}\left[Y_1^{*'}(\tilde{K} - \tilde{\lambda}\mathbf{I}_n)\Omega_n(\tilde{K} - \tilde{\lambda}\mathbf{I}_n)Y_1^*\right]\left[Y_1^{*'}(\tilde{K} - \tilde{\lambda}\mathbf{I}_n)Y_1^*\right]^{-1},$$

where Ω_n is the diagonal matrix whose typical element is the squared residual $\left(y_i - Y_{1i}^{*'}\tilde{\theta}_n\right)^2$. We formally justify in Section 4.2 that inference can be conducted through a standard Wald test using the above formula. Our Monte-Carlo analysis shows that WMD is competitive with respect to the estimators proposed by Hausman et al. (2009). In particular, the dispersion of our estimator is slightly larger, but only for the weakly identified parameters β_0, while the coverage rates of confidence intervals are more accurate.

3 General framework

3.1 Weak identification

Let $g(Z_i, \theta)$ be a known r-vector valued function ($r \geq 1$) of the random vector of i.i.d. observations $Z_i = (Y_i', X_i')' \in \mathbb{R}^{d+q}$ and of the structural parameter $\theta \in \Theta \subset \mathbb{R}^p$. The parameter θ_0 is assumed to be identified through the conditional moment restrictions (CMR)

$$\mathrm{E}\left[g(Z_i, \theta_0)|X_i\right] = \mathbf{0} \quad \text{a.s.} \tag{3.1}$$

We assume that X are continuous, as discrete X would only yield a finite number of unconditional moment restrictions. We formally state our global identifiability assumption.

Assumption 1. *(Global Identifiability)*

(i) The parameter space Θ is compact.

(ii) θ_0 is the unique value in Θ satisfying (3.1), that is $\mathrm{E}[g(Z_i,\theta)|X_i] = \mathbf{0}$ a.s. $\Rightarrow \theta = \theta_0$.

Identification is weak when the above restrictions become less informative about some, or all, components of the parameter value as the sample size increases. We specifically assume that $\mathrm{E}[g(Z_i,\theta)|X_i]$ can be additively decomposed into components that are local-to-zero at different (unknown) rates. As a result, various identification patterns appear for different components of the structural parameter.

Assumption 2. *(Weak Identification)*

$$\tau(X_i,\theta) \equiv \mathrm{E}[g(Z_i,\theta)|X_i] = \sum_{l=1}^{s} r_{l,n}^{-1}\tau_l(X_i,\theta_1,\dots\theta_l)\,, \tag{3.2}$$

where $\theta_l,\ l = 1,\dots s$, are vectors of size p_l that form a partition of θ, and $r_{l,n}$ are real sequences such that (i) $r_{1,n} = 1$ or $\to\infty$, (ii) $r_{l,n} = o(r_{l+1,n}),\ l = 1,\dots s-1$, and (iii) $r_n \equiv \max_l[r_{l,n}] = o(\sqrt{n})$.

Our framework provides a natural extension of usual definitions of weak identification to conditional moments. To provide additional intuition, let us focus on the simple case where

$$\mathrm{E}[g(Z_i,\theta)|X] = \tau_\alpha(X_i,\alpha) + \frac{\tau_\beta(X_i,\alpha,\beta)}{r_n}\,. \tag{3.3}$$

Here, $\theta = (\alpha'\ \beta')'$ and $r_n \to \infty$ with $r_n = o(\sqrt{n})$. This corresponds to many models of interest where exogenous and endogenous variables enter in the estimating equations in an additive separable way. In the weak instruments literature, unconditional moments $m(\theta)$ are typically modeled as

$$m(\theta) \equiv m(\alpha,\beta) = m_1(\alpha) + \frac{m_2(\alpha,\beta)}{n^\tau}\,. \tag{3.4}$$

In their pioneer work, Stock and Wright (2000) consider $\tau = 1/2$ and show that consistent estimation of β_0 is not possible. Caner (2010) generalizes this formulation to consider additional components of θ that are near-weakly identified with $\tau < 1/2$. Clearly, weak identification as in (3.3) implies that any unconditional moment $m(\theta) = \mathrm{E}[A(X_i)g(Z_i,\theta)]$ writes as (3.4). The main difference is that, in our framework, identification weakness implicitly comes from the conditional distribution of Y given X, while in modelling unconditional moments as in (3.4), identification weakness may either come from this conditional distribution or from the marginal distribution of X. Indeed, it could well be the case that unconditional moments are local-to-zero because the distribution of X becomes degenerate as the sample size increases. However, in most of the literature, with the exception of some examples discussed in Han and Phillips (2006), this possibility is implicitly ruled out by regularity assumptions. Hence, our framework appears no less general than the one adopted up to now in the literature.

Our definition of weak identification assumes that the partition $\theta = (\theta_1',\cdots,\theta_s')'$ is known a priori, as is the case in the literature on weak instruments that originates in the work of Stock and Wright (2000). It is the case, for instance, in the linear model of Section 2. As we will see, we do not need in practice to know this partition or the different rates at which each subset is identified to estimate parameters or perform inference.

3.2 Estimation

Our above discussion of weak identification implies that the data generating process changes with n. As explained above, X are assumed to be continuous, since discrete X yield only a finite number of unconditional moment restrictions.

Assumption 3. *(Data Generating Process)*
The observations form a rowwise independent triangular array, where the marginal distribution of the continuously distributed X remains unchanged.

The assumption of a constant distribution of X could be weakened, but is made to formalize that weak identification comes from the conditional distribution of Y given X only. For the sake of simplicity, we will not use a double index for observations and will denote by $\{Z_1, \dots Z_n\}$ the independent copies from Z for a sample size n.
The restrictions (3.1) are equivalent to the continuum of unconditional restrictions

$$\mathrm{E}\left[g(Z_i, \theta_0) e^{it'X_i}\right] = 0 \quad \forall t \in \mathbb{R}^q \,,$$

see Bierens (1982) for a proof. The main idea is to built a theoretical criterion that combines the above continuum of restrictions in an integral. For a given strictly positive measure μ on $\mathbb{R}^p$, θ_0 minimizes the theoretical objective function

$$\sum_{k=1}^{r} \int_{\mathbb{R}^q} \left|\mathrm{E}\left[g^{(k)}(Z_i, \theta) e^{it'X_i}\right]\right|^2 d\mu(t) = \mathrm{E}\left[g'(Z_i, \theta) g(Z_j, \theta) \int_{\mathbb{R}^q} e^{it'(X_i - X_j)} d\mu(t)\right] ,$$

where $g^{(k)}(\cdot, \cdot)$, $k = 1, \dots, r$, are the components of $g(\cdot, \cdot)$ and $Z_j = (Y_j, X_j)$ is an independent copy of Z_i. If we denote by $K(X_i - X_j)$ the last integral, the objective function becomes

$$\mathrm{E}\left[g'(Z_i, \theta) g(Z_j, \theta) K(X_i - X_j)\right] .$$

A natural estimator of the latter is obtained after replacing the expectation by a double average. Therefore, an estimator could be chosen to minimize

$$\frac{1}{2n(n-1)} \sum_{i \neq j} g'(Z_i, \theta) g(Z_j, \theta) K(X_i - X_j) \,.$$

Such an estimator is a Smooth Minimum Distance (SMD) estimator, as introduced by Lavergne and Patilea (2010) under strong identification, where one chooses a fixed smoothing parameter equal to 1 for any n. One can show that the SMD estimator is consistent under weak identification. However, as the gradient of the objective function flattens under weak identification, the solution of the first-order conditions is very dispersed, as we have checked through unreported simulations. To avoid such a behavior, we consider instead

$$\tilde{\theta}_n = \arg\min_{\theta} \left[\frac{\sum_{i \neq j} g'(Z_i, \theta) g(Z_j, \theta) K(X_i - X_j)}{\sum_i g'(Z_i, \theta) g(Z_i, \theta)}\right] , \tag{3.5}$$

as our Weighted Minimum Distance (WMD) estimator. The first-order conditions imply

$$\sum_{i \neq j} \nabla_\theta g(Z_i, \theta) g(Z_j, \theta) K(X_i - X_j) - \lambda_n \sum_i \nabla_\theta g(Z_i, \theta) g(Z_i, \theta) = \mathbf{0}\,,$$

where λ_n is the minimum value of the objective function (3.5). This combines the gradient of the SMD estimator with the one of a least-squares criterion, assuming the functions in $g(\cdot, \theta_0)$ are homoskedastic and uncorrelated. This second gradient does not flatten, even under weak identification, and, thus, yields more stability in estimation. Clearly, least-squares estimation alone would yield a biased estimator. However, since λ_n is small, the former has no effect on the consistency of the WMD estimator. WMD is actually asymptotically equivalent to SMD, but we found that the former is much less variable and well-centered in small samples.[2]

The combination of the continuum of moments in our theoretical objective function, as well as in our estimator, is not optimal in general. Such an optimal combination is a difficult issue. Carrasco and Florens (2000) study this problem under strong identification. Generally, optimal combination of moments necessitates weighting depending on conditional variance of the moments, and estimation of this conditional variance can have adverse effects in practice. Hausman et al. (2009) found that this can degrade the finite sample performance of estimators such as CUE, which tend to have large dispersion under many weak instruments, suggesting a "moments problem". By contrast, our WMD estimator is well-behaved in simulations under either strong or weak identification.

4 Large Sample Theory

4.1 Asymptotic Normality

We now show that our WMD estimator is consistent and asymptotically normal, with different rates of convergence for the elements of the partition of θ_0 introduced in Assumption 2. We make the following regularity assumptions.

Assumption 4. *(Regularity of K)*
$K(\cdot)$ is a symmetric, bounded density on $\mathbb{R}^q$, with integral equal to one. Its Fourier transform is strictly positive on $\mathbb{R}^q$ and non-increasing on $(0, \infty)$.

Examples of suitable densities include products of triangular, normal, logistic (see Johnson, Kotz, and Balakrishnan, 1995, Section 23.3), Student (including Cauchy, see Hurst, 1995), or Laplace densities.
Let $\overline{\mathrm{E}}$ be the operator that maps any function $G(\cdot)$ of Z into $\limsup_n \mathrm{E}\, G(Z_n)$.

[2]Though we do not formally consider it, one could generalize our estimator to the case where some of the conditioning variables X are discrete, see e.g. Pacini (2011) for an application of SMD in this case.

Assumption 5. *(Regularity of g)*
(i) The families $\mathcal{G}_k = \{g^{(k)}(\cdot,\theta) : \theta \in \Theta\}$, $1 \leq k \leq r$, *are uniformly Euclidean for an envelope* G *with* $\overline{\mathrm{E}}G^4(Z) < \infty$.
(ii) Uniformly in n, $\mathrm{E}\, g(Z,\theta)g'(Z,\theta)$ *is continuous in* θ *and* $\mathrm{var}\,[g(Z,\theta_0)|X]$ *is almost surely positive definite and bounded away from infinity.*

Assumption 5 does not require the continuity of the functions $\theta \mapsto g(z,\theta)$, but guarantees that the family of functions

$$\{(z,\bar{z}) \mapsto g'(z,\theta)g(\bar{z},\theta) : \theta \in \Theta\}$$

is uniformly Euclidean for a squared integrable envelope, see Lemma 2.14-(ii) of Pakes and Pollard (1989). Here, *uniformly* means that the envelope and the constants in the definition of the Euclidean family are independent of n.

Assumption 6. *(Regularity of* τ*)*
The functions $\tau_l(x,\theta_1,\ldots\theta_l)$, $l = 1,\ldots s$, *satisfy Condition 1.*

Condition 1. *A function* $l(x,\theta)$ *satisfies Condition 1 if (i)* $\sup_\theta \| l(\cdot,\theta) \| f(\cdot)$ *is in* $L^1 \cap L^2$. *(ii) For all* x, *the map* $\theta \mapsto l(x,\theta)$ *is continuous. (iii) For any* x, *all second partial derivatives of* $l(x,\theta)$ *exist on a neighborhood* $\mathcal{N}$ *of* θ_0 *independent of* x. *Each component of* $\nabla_\theta l(\cdot,\theta_0)f(\cdot)$ *belongs to* $L^1 \cap L^2$ *and* $\mathrm{E}\,\|\nabla_\theta l(X,\theta_0)\nabla'_\theta l(X,\theta_0)\|^2 < \infty$. *On the neighborhood* $\mathcal{N}$ *of* θ_0, *each second-order partial derivative is uniformly Euclidean for a common envelope* H *with* $\overline{\mathrm{E}}H(X) < \infty$.

Assumption 6 implies that $\tau(x,\theta)$ itself fulfills Condition 1.
Let D_n be the $p \times p$ matrix

$$D_n = \begin{bmatrix} r_{1,n}\mathbf{I}_{p_1} & \mathbf{0} & \ldots & \ldots & \mathbf{0} \\ \mathbf{0} & r_{2,n}\mathbf{I}_{p_2} & \mathbf{0} & \ldots & \mathbf{0} \\ \ldots & & & & \ldots \\ \ldots & & & \mathbf{0} & r_{s,n}\mathbf{I}_{p_s} \end{bmatrix},$$

where $\mathbf{I}_k$ is the identity matrix of size k. The matrix D_n allows to rescale the different components of $\tau(\cdot,\cdot)$.

Assumption 7. *(Local Identifiability)*
$\forall n$, $a'D_n\nabla_\theta\tau(X,\theta_0) = \mathbf{0} \Rightarrow a = \mathbf{0}$.

To gain some insight on Assumption 7, consider the linear model

$$y_i = \alpha_0 + \beta_0 Y_{1i} + \gamma_0 Y_{2i} + \varepsilon_i\,, \qquad \mathrm{E}\,(\varepsilon_i|X_i) = 0\,,$$

so that

$$\begin{aligned} \tau(X_i,\theta) &= (\alpha-\alpha_0) + (\beta-\beta_0)\,\mathrm{E}\,[Y_{1i}|X_i] + (\gamma-\gamma_0)\,\mathrm{E}\,[Y_{2i}|X_i] \\ &= (\alpha-\alpha_0) + r_{2n}^{-1}\,(\beta-\beta_0)\,\tau_2\,(X_i) + r_{3n}^{-1}\,(\gamma-\gamma_0)\,\tau_3\,(X_i)\,. \end{aligned}$$

Our local identifiability assumption means that the functions $\tau_2(\cdot)$, $\tau_3(\cdot)$, and the constant function are not perfectly collinear. If they were, we would only be able to identify some linear combinations of the coefficients. In this linear-in-parameters model, local identifiability directly follows from global identifiability. In nonlinear setups however, this additional assumption is generally needed.
To state our main result, let us define the matrices

$$\begin{aligned} \Delta &= \mathrm{E}\left[\nabla_\theta \tau(X_1,\theta_0)\,\mathrm{var}\left[g(Z_2,\theta_0)|X_2\right]\nabla'_\theta \tau(X_3,\theta_0) K\left(X_1-X_2\right)K\left(X_3-X_2\right)\right] \\ \text{and } V &= \mathrm{H}_{\theta,\theta}\,\mathrm{E}\,M_n(\theta_0) = \mathrm{E}\left[\nabla_\theta \tau(X_1,\theta_0)\nabla'_\theta \tau(X_2,\theta_0) K\left(X_1-X_2\right)\right]. \end{aligned}$$

Lemma 7.2 in Section 7 shows that, under our assumptions, $D_n V D_n$ and $D_n \Delta D_n$ have strictly positive and finite eigenvalues.

Theorem 4.1. *(Asymptotic Normality)*
Under Assumptions 1–7, $\sqrt{n}\left(D_n \Delta D_n\right)^{-1/2}\left(D_n V D_n\right) D_n^{-1}\left(\tilde{\theta}_n - \theta_0\right) \xrightarrow{d} N\left(\mathbf{0}, \mathbf{I}_p\right)$.

Our result implies that each component θ_l, $l = 1, \ldots s$, is estimated at asymptotic rate $r_{l,n}^{-1}\sqrt{n}$. This matches results obtained for IV estimators with many weak instruments. However, asymptotic variance comparison between WMD and IV estimators is not possible. Whether one approach yields more efficient estimation generally depends on the true data generating process and on the number of instruments used in IV estimation, as the variance of IV estimators is inflated when the number of instruments grows with the sample size, see e.g. Bekker (1994) and Hansen, Hausman and Newey (2008).

4.2 Inference

We consider testing implicit parametric restrictions of the form

$$H_0 : h(\theta_0) = \mathbf{0}\,,$$

where $h(\cdot)$ is a function from $\Theta \subset \mathbb{R}^p$ on $\mathbb{R}^m$ with $m \leq p$. In that aim, our first task is to estimate the matrices involved in the estimator's variance. Let us assume that the functions $g(\cdot,\cdot)$ are differentiable around θ_0.[3] Let Ω_{nj} be the diagonal matrix with s-th diagonal element equal to the square of the s-th component of $g(Z_j, \tilde{\theta}_n)$ and consider

$$\begin{aligned} V_n &= \frac{1}{n(n-1)}\sum_{i\neq j} \nabla_\theta g(Z_i,\tilde{\theta}_n)\nabla'_\theta g(Z_j,\tilde{\theta}_n) K_{ij} \\ \text{and } \Delta_n &= \frac{1}{n(n-1)(n-2)}\sum_{i\neq j, j\neq k} \nabla_\theta g(Z_i,\tilde{\theta}_n)\Omega_{nj}\nabla'_\theta g(Z_k,\tilde{\theta}_n) K_{ij}K_{jk}\,, \end{aligned}$$

[3] If $g(\cdot,\cdot)$ was not differentiable, one could use numerical methods to approximate $\nabla_\theta \tau(\cdot,\theta_0)$ as done in Pakes and Pollard (1989).

where $K_{ij} = K(X_i - X_j)$. From Theorem 4.1, what we need is that $D_n V_n D_n$ and $D_n \Delta_n D_n$ consistently approximate $D_n V D_n$ and $D_n \Delta D_n$, respectively. This is shown in the proof of Theorem 4.2 below.
The Wald test statistic can then be constructed as

$$W_n = nh'(\tilde{\theta}_n) \left[\nabla'_\theta h(\tilde{\theta}_n) V_n^{-1} \Delta_n V_n^{-1} \nabla_\theta h(\tilde{\theta}_n) \right]^{-1} h(\tilde{\theta}_n) .$$

It is the very same statistic one would compute under strong identification, i.e. when $D_n = \mathbf{I}_p$. Hence its computation does not require to know whether there is identification weakness and its specific pattern. This is because V_n and Δ_n provides automatic inner corrections for identification weakness. Since the test statistic has the usual asymptotic chi-square distribution under the null hypothesis, in practice we can perform inference as if θ was strongly identified. Its power under local alternatives is, however, affected by weak identification. Intuitively, the test will have non-trivial power under local alternatives, but the rate of such local alternatives will depend on the rate at which parameters involved in each of the restrictions can be estimated. In what follows, we state consistency of the test, but for the sake of simplicity we do not study its local power.

Theorem 4.2. *(Wald Test)*
Assume that (i) for any x, $g(x, \cdot)$ is differentiable in a neighborhood $\mathcal{N}$ of θ_0 independent of x, with first derivative Euclidean on this neighborhhod for an envelope L with $\overline{\mathrm{E}} L^2(Z) < \infty$, and that (ii) $h(\cdot)$ is continuously differentiable with $\nabla_\theta h(\theta_0)$ of full rank m.
Then, under the assumptions of Theorem 4.1, W_n is asymptotically chi-square with m degrees of freedom under H_0, and $W_n \xrightarrow{p} +\infty$ whenever $h(\theta_0) \neq 0$.

5 Monte-Carlo Simulations

We investigate the small sample properties of our estimator in the linear structural model of Section 2, that is

$$\begin{aligned} y_i &= \alpha_0 + Y_{1i}\beta_0 + s(X_i)\varepsilon_i , \\ Y_{1i} &= \frac{\sqrt{c}}{r_n} f(X_i) + U_i , \end{aligned}$$

where Y_{1i} is univariate and X_i follows a standard univariate normal distribution. We set $\alpha_0 = \beta_0 = 0$ and $r_n = n^{0.45}$. We consider different specifications, depending on $f(\cdot)$, $s(\cdot)$, and the joint distribution of (ε_i, U_i). In all of them, $(s(X_i)\varepsilon_i, U_i)$ has mean $\mathbf{0}$, unit unconditional variances, and unconditional correlation 0.8. Also $f(X)$ has variance one. We compare the performance of our WMD estimator with standard normal $K(\cdot)$ to the estimators considered by Hausman et al. (2009), namely JIVE, HLIM, and HFUL. These estimators are robust to many weak instruments and unknown heteroskedasticity. They have the form

$$\hat{\theta} = \left[Y_1^{*'} (\hat{P} - \hat{\lambda} \mathbf{I}_n) Y_1^* \right]^{-1} \left[Y_1^{*'} (\hat{P} - \hat{\lambda} \mathbf{I}_n) y \right] ,$$

	$f(x)$	$s(x)$	(ε_i, U_i)
M_L	x	1	Normal
$M_{L,H}$	x	$\sqrt{(1+x^2)/2}$	Normal
$M_{NL,H}$	$\sqrt{2\pi\sqrt{27}}xe^{-x^2/2}$	$\sqrt{(1+x^2)/2}$	Normal
$M_{ST,H}$	x	$\sqrt{(1+x^2)/2}$	$(3/5)\times$Bivariate Student with 5 d.f.

Table 1: Specification of the simulation designs.

where $Y_1^* = [e\ Y_1]$, $\hat{P}$ is the projection matrix on the space spanned by instruments, but whose diagonal elements are set to zero, and $\hat{\lambda}$ differs depending on the method. JIVE is the well-known jackknife IV estimator where $\hat{\lambda}_{JIVE} = 0$, see Angrist, Imbens, and Krueger (1999). The HLIM estimator is jackknifed LIML, where $\hat{\lambda}_{HLIM}$ is the smallest eigenvalue of $(Y^{*'}Y^*)^{-1}(Y^{*'}\hat{P}Y^*)$ with $Y^* = [y\ Y_1^*]$. The HFUL estimator is the equivalent of Fuller's (1997) modification of LIML for HLIM, where

$$\hat{\lambda}_{HFUL} = [\hat{\lambda}_{HLIM} - (1 - \hat{\lambda}_{HLIM})/n)]/[1 - (1 - \hat{\lambda}_{HLIM})/n]\,.$$

To generate instruments, we consider piecewise linear approximating functions on a grid of 3, 6, and 12 intervals, which yield 6, 12, and 24 instruments respectively. The intervals are based on the quantiles of the normal distribution, the distribution of X_i. Formally, with $2K$ instruments, the i^{th} observation is

$$(1, X_i, D_{i1}, X_i D_{i1}, \cdots, D_{i,K-1}, X_i D_{i,K-1})\,,$$

where $D_{ik} = \mathbb{I}(X_i \in \Phi^{-1}(I_k))$, with $\mathbb{I}$ the indicator function, Φ the c.d.f. of a standard normal, and $I_k = \left[\frac{(k-1)}{K}, \frac{k}{K}\right]$, $k = 1, \cdots, K$. Note that for a linear $f(\cdot)$, any set of instruments allows to recover the conditional expectation of Y_{1i} on X_i.

Since HFUL was found to perform best by Hausman et al. (2009), we also consider WMDF, a Fuller-like modification of our WMD estimator, which writes as (2.3) with

$$\tilde{\lambda}_{WMDF} = [\tilde{\lambda}_{WMD} - (1 - \tilde{\lambda}_{WMD})/n)]/[1 - (1 - \tilde{\lambda}_{WMD})/n]\,.$$

We consider the four specifications in Table 1 for $n = 250$ and $c = 8$. The constant c tunes the strength of the relation between Y_1 and X. Our choice yields a theoretical R^2 of 5.27% for the linear regression of Y_1 on X, which corresponds to an expected first-stage F statistic of about 8. We also consider two variations of $M_{L,H}$, see below. We report in Tables 2 to 5 the performance of the different estimators, summarized by median bias (Med), standard deviation (Std), interdecile range (DecR), and nominal 5% rejection frequencies for univariate Wald tests on α_0 and β_0 (Rej). All simulations results are based on 10,000 replications.

Several common features emerge from the results. Among IV estimators, HLIM and JIVE are very dispersed, and especially so for a large number of instruments, while HFUL is the least dispersed. This is in line with

the findings of Hausman et al. (2009), who also observed that CUE was more dispersed than HLIM in their simulations. The estimator WMDF performs better than WMD, especially for the slope parameter, being usually less dispersed with similar median bias. Comparing WMDF and HFUL, the former is often a bit more dispersed in terms of standard deviation, but, for a large number of instruments, has a lower interdecile range in all specifications but the homoskedastic one M_L. Moreover, WMD and WMDF always have smaller median bias than HFUL. IV estimators have too high rejection percentages for β_0, and these increase with the number of instruments, e.g. for HFUL with 24 instruments, they are 10.2% and 12% in specification M_L and $M_{L,H}$, respectively. Corresponding figures for WMD and WMDF are always the closest to nominal level. As for rejection percentages for α_0, they are all well below nominal level, mostly between 2 and 3%, and for IV estimators decrease when the number of instruments grows.

We also consider two variations around model $M_{L,H}$. First, we evaluate the effect of doubling the sample size, see Table 6. This does not affect our main findings. Second, we let $c = 50$ in Table 7 to see how identification weakness affects our results. Both parameters are then well estimated by all methods, and rejection frequencies are close to 5%. A striking finding is that WMD and WMDF are less variable and better centered than IV estimators. Moreover, they exhibit rejection percentages closer to nominal level than competitors for the slope parameter.

To further investigate the discrepancy between rates of convergence of the WMD estimators $\tilde{\alpha}_n$ and $\tilde{\beta}_n$, we perform a regression of the logarithm of the ratio of their Monte-Carlo standard deviations on the logarithm of the sample size, that is

$$\log\left(\frac{\widehat{\text{var}}(\tilde{\beta}_n)}{\widehat{\text{var}}(\tilde{\alpha}_n)}\right)^{1/2} = \widehat{\gamma}_1 + \widehat{\gamma}_2 \log(n)\,. \tag{5.1}$$

From our results in Section 2.1, the theoretical slope in this regression should be $\log r_n / \log n = 0.45$. Results of the above regression are reported for model $M_{L,H}$ with $c = 50$ and sample sizes varying from 200 to 1000 with increment 50, see Figure 1. We have also checked through unreported simulations that these results were robust to different designs. The estimated $\widehat{\gamma}_2$ are 0.449 for WMD and 0.450 for WMDF. This simple experiment confirms the theoretical discrepancy between rates of convergence for the intercept and slope estimators.

6 Conclusion

We have considered models defined by a set of conditional moment restrictions that may be subject to weak identification. Weak identification has been modeled directly through these conditional restrictions by assuming that they can be additively decomposed into local-to-zero components. In this setup, we have proposed a new weighted minimum distance (WMD) estimator which does not rely on instrumental variables. We have proved that WMD is consistent and asymptotically normal. We have also shown how, in practice, estimation and heteroskedasticity-robust inference Wald testing can be entertained without prior knowledge of the weakness pattern.

We have extensively discussed the linear model with endogenous regressors. We have shown that WMD resembles a k-class estimator, and that its asymptotic variance can be estimated straightforwardly. In Monte-Carlo experiments, we have compared the small sample properties of WMD and its Fuller-modified version, WMDF, to three estimators recently studied by Hausman et al. (2009) that are robust to many weak instruments and heteroskedasticity. Overall, WMD and its variant WMDF are very competitive, outperforming HLIM and JIVE in terms of median bias and dispersion, while being pretty comparable to HFUL overall. Finally, the rejection frequencies for WMD and WMDF are closer to nominal level than all competitors throughout. Thus, we recommend that WMD, or its variant WMDF, be used in models defined by conditional moment restrictions when heteroskedasticity is present, which is common in microeconometrics.

7 Proofs

7.1 Preliminary Results

7.1.1 Convergence Rates for U-Processes

In our main proofs, we will often use results on convergence rates for U-statistics as derived by Sherman (1994), namely his Corollaries 4, 7, and 8. These results are derived for i.i.d. observations. However, it is easy to see that they extend to our setup of rowwise independent triangular array of data. Indeed, all these results directly derive from his Main Corollary, an inequality that holds for any finite n under the assumption that the envelope function $F(\cdot)$ satisfies $\mathrm{E}\, F(Z_n) < \infty$. It is then straightforward to extend Corollaries 4, 7 and 8 of Sherman (1994) to our setup. As an example, we state and prove the first result.

Corollary 7.1 (Sherman's (1994) Corollary 4)**.** *For a rowwise independent triangular array of observations* $\{Z_{1n}, \dots Z_{nn}\}$*, let* $\mathcal{F}$ *be a class of functions such that* $\forall f \in \mathcal{F}$, $\mathrm{E}\, f(s_1, \dots, s_{i-1}, \cdot, s_{i+1}, \dots s_k) \equiv 0$, $i = 1, \dots k$, *. Suppose* $\mathcal{F}$ *is Euclidean for an envelope* F *satisfying* $\limsup_n \mathrm{E}\, F^2(Z_n) < \infty$*. Then*

$$\sup_{\mathcal{F}} |n^{k/2} U_n^k f| = O_p(1)\,, \quad \textit{where} \quad U_n^k f \equiv (n)_k^{-1} \sum_{\mathbf{i}_k} f\left(Z_{i_1 n}, \dots Z_{i_k n}\right)$$

and $\mathbf{i}_k = (i_1, \dots i_k)$ *ranges over the* $(n)_k$ *ordered* k*-tuples of distinct integers from* $\{1, \dots n\}$*.*

Proof. Sherman's Main Corollary with $p = 1$ yields that for any $0 < \alpha < 1$ and any n

$$\mathrm{E} \sup_{\mathcal{F}} |n^{k/2} U_n^k f| \leq \Omega_n \left[\mathrm{E} \sup_{\mathcal{F}} \left(U_{2n}^k f^2\right)^\alpha\right]^{1/2},$$

where $\Omega_n = C_\alpha \left(\mathrm{E}\, F^2(Z_n)\right)^{\epsilon/2}$ for some $\epsilon \in (0,1)$. Now

$$\begin{aligned} \limsup_n \Omega_n^2 \left[\mathrm{E} \sup_{\mathcal{F}} \left(U_{2n}^k f^2\right)^\alpha\right] &\leq C_\alpha^2 \left(\overline{\mathrm{E}} F^2(Z)\right)^\epsilon \limsup_n \left[\mathrm{E}\left(U_{2n}^k F^2\right)^\alpha\right] \\ &\leq C_\alpha^2 \left(\overline{\mathrm{E}} F^2(Z)\right)^\epsilon \left(\overline{\mathrm{E}} F^2(Z)\right)^\alpha < \infty. \end{aligned}$$

Conclude from Chebyshev's inequality. □

7.1.2 Matrices

For a real-valued function $l(\cdot)$, denote by $\mathcal{F}[l](\cdot)$ its Fourier transform, and by $\overline{l}(\cdot)$ its conjugate.

Lemma 7.2. *Under Assumptions 4 to 7, D_nVD_n and $D_n\Delta D_n$ have eigenvalues uniformly bounded away from zero and infinity.*

Proof. Let $\delta_n(X) = D_n\nabla_\theta\tau(X,\theta_0)$. We have $D_nVD_n = \mathrm{E}[\delta_n(X_1)\delta_n'(X_2)K(X_1-X_2)]$. Denote convolution by $*$. From Assumptions 4, 6 (Condition 1-(i) and (iii)), and the properties of Fourier transforms,

$$\begin{aligned}
& a'\,\mathrm{E}[\delta_n(X_1)\delta_n'(X_2)K(X_1-X_2)]\,a \\
&= \int_{\mathbb{R}^q} a'\delta_n(X_1)f(X_1)(f\delta_n'a * K)(X_1)\,dX_1 = \int_{\mathbb{R}^q} \mathcal{F}[a'\delta_n f](t)\mathcal{F}[f\delta_n'a * K](t)\,dt \\
&= \int_{\mathbb{R}^q} \mathcal{F}[a'\delta_n f](t)\overline{\mathcal{F}[f\delta_n'a * K](-t)}\,dt = \int_{\mathbb{R}^q} \mathcal{F}[a'\delta_n f](t)\overline{\mathcal{F}[f\delta_n'a](-t)}\mathcal{F}[K](t)\,dt \\
&= \int_{\mathbb{R}^q} |\mathcal{F}[a'\delta_n f](t)|^2\mathcal{F}[K](t)\,dt\,.
\end{aligned}$$

From the strict positivity of $\mathcal{F}[K](t)$, all eigenvalues of D_nVD_n are non-negative. Since $\mathcal{F}[K](t) \le 1\ \forall t$, and from Assumption 6, the above quantity is bounded for any a of norm 1, so that eigenvalues of D_nVD_n are bounded. Moreover, the minimum eigenvalue is zero iff

$$\exists\, a \neq \mathbf{0}:\ a'\delta_n(X)f(X) = \mathbf{0}\ \text{a.s.} \qquad \Leftrightarrow \qquad \exists\, a \neq \mathbf{0}:\ a'\delta_n(X) = \mathbf{0}\,,$$

which would contradict Assumption 7.
The matrix $D_n\Delta D_n$ is the variance matrix of $\mathrm{E}[\delta_n(X_0)K(X-X_0)|X]\,g(Z,\theta_0)$, so that it is singular iff there exists $a \neq \mathbf{0}$ such that

$$a'\,\mathrm{E}[\delta_n(X_0)K(X-X_0)|X]\,g(Z,\theta_0) = \mathbf{0} \quad \text{a.s.}$$

Given that $g(Z,\theta_0)$ cannot be identically zero by Assumption 5-(ii), this is equivalent to

$$\begin{aligned}
& a'\,\mathrm{E}[\delta_n(X_0)K(X-X_0)|X] = a'\,(\delta_n f * K)(X) = \mathbf{0} \quad \text{a.s.} \\
& \Leftrightarrow \quad \mathcal{F}[a'\delta_n f](t)\mathcal{F}[K](t) = \mathbf{0} \quad \forall t \in \mathbb{R}^q \Leftrightarrow a'\delta_n(X) = \mathbf{0} \quad \text{a.s.}
\end{aligned}$$

which would contradict Assumption 7. □

7.2 Main Proofs

7.2.1 Proof of Theorem 4.1

For the sake of simplicity, we detail most of our arguments only for the simplest case (3.3), and explain how they easily adapt to the general case (3.2), that is in three directions: (1) there may be only one rate for the whole parameter, (2) there may be more than two rates, and (3) the slowest rate r_{1n} could diverge.

(i) *Approximation of the criterion.* Let write our criterion $Q_n(\theta) = M_n(\theta)/\sigma_n^2(\theta)$ where

$$\begin{aligned} M_n(\theta) &= \frac{1}{2n(n-1)} \sum_{i \neq j} g'(Z_i, \theta) g(Z_j, \theta) K(X_i - X_j) \\ \text{and} \quad \sigma_n^2(\theta) &= \frac{1}{n} \sum_i g'(Z_i, \theta) g(Z_i, \theta) \,. \end{aligned}$$

From Assumption 5-(i) and Corollary 7.1, $\sup_\theta |\sigma_n^2(\theta) - \mathrm{E}\, g'(Z, \theta) g(Z, \theta)| = O_p(n^{-1/2})$. Moreover, from Assumptions 1 and 5, $\sigma^2(\theta) \equiv \mathrm{E}\, g'(Z, \theta) g(Z, \theta)$ is uniformly bounded away from zero and infinity. Hence $Q_n(\theta) = \frac{M_n(\theta)}{\sigma^2(\theta)} \left(1 + O_p(n^{-1/2})\right)$ uniformly in θ and

$$\begin{aligned} & Q_n(\theta) - Q_n(\theta_0) \\ &= [M_n(\theta) - M_n(\theta_0)] \frac{1}{\sigma^2(\theta)} \left(1 + O_p(n^{-1/2})\right) + M_n(\theta_0) \left[\frac{1}{\sigma^2(\theta)} - \frac{1}{\sigma^2(\theta_0)}\right] \left(1 + O_p(n^{-1/2})\right) \end{aligned} \tag{7.1}$$

uniformly in θ, since $M_n(\theta_0)$ is a degenerate second-order U-statistic.

(ii) *Consistency of* $\tilde{\alpha}_n$. The parameter value θ_0 is the unique minimizer of $\mathrm{E}\, M_n(\theta)$. Indeed, reason as in the proof of Lemma 7.2 to get that

$$\begin{aligned} \mathrm{E}\, M_n(\theta) &= \frac{1}{2} \mathrm{E}\left[\tau'(X_1, \theta) \tau(X_2, \theta) K\left(X_1 - X_2\right)\right] \\ &= \frac{1}{2} \sum_{k=1}^{r} \left\{ \int_{\mathbb{R}^q} \left| \mathcal{F}\left[\tau^{(k)}(\cdot, \theta) f(\cdot)\right](t) \right|^2 \mathcal{F}[K](t)\, dt \right\} \geq 0 \,, \end{aligned} \tag{7.2}$$

so that by Assumption 1

$$\begin{aligned} \mathrm{E}\, M_n(\theta) = 0 \quad &\Leftrightarrow \quad \mathcal{F}\left[\tau^{(k)}(\cdot, \theta) f(\cdot)\right](t) = 0 \quad \forall\, t \in \mathbb{R}^q,\ k = 1, \ldots r \\ &\Leftrightarrow \quad \mathrm{E}\left[g(Z, \theta) | X\right] = \mathbf{0} \quad \text{a.s.} \Leftrightarrow \theta = \theta_0 \,. \end{aligned}$$

By Assumption 6, $\mathrm{E}\, M_n(\theta)$ is continuous in θ, and then in α. Hence from Assumption 1, $\forall \varepsilon > 0$, $\exists \mu > 0$ such that $\inf_{\|\alpha - \alpha_0\| \geq \varepsilon} \mathrm{E}\, M_n(\theta) \geq \mu$. The family $\{g'(Z_1, \theta) g(Z_2, \theta) K(X_1 - X_2) : \theta \in \Theta\}$ is uniformly Euclidean for an envelope $F(\cdot)$ such that $\overline{\mathrm{E}} F(Z_1, Z_2)$ from Assumptions 4 and 5. By Corollary 7 of Sherman (1994), $\sup_{\theta \in \Theta} |M_n(\theta) - \mathrm{E}\, M_n(\theta)| = O_{\mathbb{P}}(n^{-1/2})$. Hence

$$\inf_{\|\alpha - \alpha_0\| \geq \varepsilon} M_n(\theta) - M_n(\theta_0) \geq \mu + O_p(n^{-1/2}) \,.$$

Using (7.1), $\inf_{\|\alpha - \alpha_0\| \geq \varepsilon} Q_n(\theta) - Q_n(\theta_0) \geq C + O_p(n^{-1/2})$, for some $C > 0$. Since $Q_n(\tilde{\theta}_n) \leq Q_n(\theta_0)$, it follows that $\forall \varepsilon > 0$ $\Pr\left[\|\tilde{\alpha}_n - \alpha_0\| < \varepsilon\right] \to 1$.

Extension: If the rate r_{1n} for α diverges, then use Hoeffding's decomposition of $M_n(\theta)$ and Corollary 7.1 to obtain $\sup_{\theta \in \Theta} |M_n(\theta) - \mathrm{E}\, M_n(\theta)| = O_p(n^{-1}) + O_p(n^{-1/2} r_{1n}^{-1})$ and note that $\forall \varepsilon > 0$, $\exists \mu > 0$ such that $\inf_{\|\alpha - \alpha_0\| \geq \varepsilon} \mathrm{E}\, M_n(\theta) \geq r_{1n}^{-2} \mu$. Then proceed as above.

(iii) *Consistency of* $\tilde{\beta}_n$. Apply Hoeffding's decomposition to $M_n(\theta) - M_n(\theta_0)$. The second order U-process in the decomposition of $M_n(\theta) - M_n(\theta_0)$ is $O_p(n^{-1})$ uniformly over θ by Assumption 5 and Corollary 7 of

Sherman (1994). Consider the first-order U-process $\mathbb{P}_n \tilde{l}_\theta$, where $\tilde{l}_\theta(Z_i) = \mathrm{E}[l_\theta(Z_i, Z_j) \mid Z_i] + \mathrm{E}[l_\theta(Z_i, Z_j) \mid Z_j] - 2\,\mathrm{E}[l_\theta(Z_i, Z_j)]$,

$$\begin{aligned} l_\theta(Z_i, Z_j) &= (1/2)\left(g'(Z_i,\theta)g(Z_j,\theta) - g'(Z_i,\theta_0)g(Z_j,\theta_0)\right) h^{-q} K\left((X_i - X_j)/h\right) \\ &= (1/2) g'(Z_i,\theta_0)\left(g(Z_j,\theta) - g(Z_j,\theta_0)\right) K\left(X_i - X_j\right) \\ &\quad + (1/2)\left(g(Z_i,\theta) - g(Z_i,\theta_0)\right)' g(Z_j,\theta_0) K\left(X_i - X_j\right) \\ &\quad + (1/2)\left(g(Z_i,\theta) - g(Z_i,\theta_0)\right)'\left(g(Z_j,\theta) - g(Z_j,\theta_0)\right) K\left(X_i - X_j\right) \\ &= l_{1\theta}(Z_i, Z_j) + l_{2\theta}(Z_i, Z_j) + l_{3\theta}(Z_i, Z_j)\,, \end{aligned}$$

and $l_{1\theta}(Z_i, Z_j) = l_{2\theta}(Z_j, Z_i)$ by the symmetry of $K(\cdot)$. Now

$$\begin{aligned} &2\,\mathrm{E}[l_{1\theta}(Z_i, Z_j) \mid Z_i] \\ &= g'(Z_i,\theta_0)\left\{\mathrm{E}\left[\tau_\alpha(X,\alpha) K\left(X_i - X\right) | X_i\right] + r_n^{-1}\,\mathrm{E}\left[\tau_\beta(X,\theta) K\left(X_i - X\right) | X_i\right]\right\}. \end{aligned} \tag{7.3}$$

The U-process based on the second part of (7.3) is $O_p(r_n^{-1} n^{-1/2})$ uniformly in θ. Using Assumption 6, the first term in (7.3) admits uniformly for α in a $o(1)$ neighborhood of α_0 the expansion

$$\begin{aligned} & g'(Z_i,\theta_0)\left[\int_{\mathbb{R}^q} \nabla'_\alpha \tau_\alpha(x,\alpha_0) f(x) K\left(X_i - x\right) dx\right](\alpha - \alpha_0) \\ + \;& \frac{1}{2} g'(Z_i,\theta_0) \sum_{k,l=1}^{p} \left(\alpha^{(k)} - \alpha_0^{(k)}\right)\left(\alpha^{(l)} - \alpha_0^{(l)}\right) \\ & \left[\int_{\mathbb{R}^q} \mathrm{H}_{\alpha^{(k)}\alpha^{(l)}} \tau_\alpha(x,\alpha_0) f(x) K\left(X_i - x\right) dx\right] + R_{n\alpha}(Z_i,\alpha)\,. \end{aligned} \tag{7.4}$$

The U-statistic based on the first element of (7.4) is an $\|\alpha - \alpha_0\| O_p(n^{-1/2})$. The one based on the second element of (7.4) is $\|\alpha - \alpha_0\|^2 O_p(n^{-1/2})$. The remaining term is a U-process of the form $(\alpha - \alpha_0)' C_n(\alpha)(\alpha - \alpha_0)$, where C_n has typical element

$$\frac{1}{2n}\sum_{i=1}^{n} g'(Z_i,\theta_0)\left[\int_{\mathbb{R}^q} \left(\mathrm{H}_{\alpha^{(k)}\alpha^{(l)}} \tau_\alpha(x,\bar{\alpha}) - \mathrm{H}_{\alpha^{(k)}\alpha^{(l)}} \tau_\alpha(x,\alpha_0)\right) f(x) K\left(X_i - x\right) dx\right],$$

where $\|\bar{\alpha} - \alpha_0\| \leq \|\alpha - \alpha_0\|$. The above function has a squared integrable envelope from Assumptions 5 and 6, and approaches zero when α tends to α_0. Use Corollary 8 of Sherman (1994) to obtain that the remaining term is an $\|\alpha - \alpha_0\|^2 o_p(n^{-1/2})$. Use similar arguments for $2\,\mathrm{E}[l_{3\theta}(Z_i, Z_j) \mid Z_i]$. We thus obtain that uniformly in β and uniformly over $o(1)$ neighborhoods of α_0

$$M_n(\theta) - M_n(\theta_0) = \mathrm{E}\, M_n(\theta) + \|\alpha - \alpha_0\| O_p(n^{-1/2}) + \|\alpha - \alpha_0\|^2 o_p(1) + O_p(n^{-1/2} r_n^{-1})\,. \tag{7.5}$$

From Assumption 6 and a Taylor expansion of $\tau_\alpha(X,\alpha)$, for α in a $o(1)$ neighborhood of α_0,

$$\begin{aligned} \mathrm{E}\, M_n(\theta) &\geq \mathrm{E}\left[\tau'_\alpha(X_1,\alpha)\tau_\alpha(X_2,\alpha) K\left(X_1 - X_2\right)\right] \\ &\geq (\alpha - \alpha_0)'\,\mathrm{E}\left[\nabla_\alpha \tau_\alpha(X_1,\alpha_0)\nabla'_\alpha \tau_\alpha(X_2,\alpha_0) K\left(X_1 - X_2\right)\right](\alpha - \alpha_0) + o\left(\|\alpha - \alpha_0\|^2\right). \end{aligned}$$

Since the above matrix is positive definite, see Lemma 7.2, then $\forall \varepsilon > 0$, $\exists \mu > 0$ such that $\inf \mathrm{E}\, M_n(\theta) \geq \mu \|\alpha - \alpha_0\|^2$. This and (7.5) imply that for some $C > 0$

$$\inf_{\|\alpha-\alpha_0\| \geq \varepsilon r_n^{-1}} M_n(\theta) - M_n(\theta_0) \geq \mu r_n^{-2} + o_p(r_n^2)\,.$$

Now (7.1) implies that for some $C > 0$

$$\inf_{\|\alpha-\alpha_0\| \geq \varepsilon r_n^{-1}} Q_n(\theta) - Q_n(\theta_0) \geq C r_n^{-2} + o_p(r_n^2)\,.$$

Since $Q_n(\tilde{\theta}) \leq Q_n(\theta_0)$, $\|\tilde{\alpha}_n - \alpha_0\| = o_p(r_n^{-1})$.
For $\|\alpha - \alpha_0\| = o(r_n^{-1})$, (7.5) yields $M_n(\theta) - M_n(\theta_0) = \mathrm{E}\, M_n(\theta) + o_p(r_n^{-2})$. As

$$\mathrm{E}\, M_n(\theta) \geq r_n^{-2}\, \mathrm{E}\left[\tau_\beta'(X_1, \theta)\tau_\beta(X_2, \theta) K\left(X_1 - X_2\right)\right]\,,$$

and the latter is continuous in β, we obtain that $\forall \varepsilon > 0$, $\exists \mu > 0$ such that $\inf_{\|\beta-\beta_0\| \geq \varepsilon} \mathrm{E}\, M_n(\theta) \geq \mu r_n^{-2} |\beta - \beta_0\|^2$, and then that for some $C > 0$

$$\inf_{\|\tilde{\alpha}_n-\alpha_0\|=o(r_n^{-1}), \|\beta-\beta_0\| \geq \varepsilon} Q_n(\theta) - Q_n(\theta_0) \geq C r_n^{-2} + o_p(r_n^{-2})\,.$$

Since $Q_n(\tilde{\theta}) \leq Q_n(\theta_0)$, this in turn yields $\|\tilde{\beta}_n - \beta_0\| = o_p(1)$.
Extension: If there are more than two rates, e.g. the case where $\theta = (\alpha, \beta, \lambda)$ with corresponding rates $(1, r_{2n}, r_{3n})$, proceed as in Part (iii) to show first that $\|\tilde{\beta}_n - \beta_0\| = o_p(r_{3n})$ and then that $\|\tilde{\lambda}_n - \lambda_0\| = o_p(1)$.
(iv) *Rate of convergence and asymptotic normality.* Apply once again Hoeffding's decomposition to $M_n(\theta) - M_n(\theta_0)$ as in the previous part. The second order U-process in this decomposition is $o_p(n)$ uniformly over $o(1)$ neighborhoods of θ_0 from Assumption 5 and Corollary 8 of Sherman (1994). To treat the first-order empirical process $\mathbb{P}_n \mathrm{E}\left[l_{1\theta}(Z_i, Z_j)|Z_i\right]$, use this time use a Taylor expansion in θ, that is,

$$\begin{aligned}
2\,\mathrm{E}[l_{1\theta}(Z_i, Z_j) \mid Z_i] &= g_n'(Z_i, \theta_0)\, \mathrm{E}\left[\left(g_n(Z, \theta) - g_n(Z, \theta_0)\right) K\left(X_i - X\right) | Z_i\right] \\
&= g_n'(Z_i, \theta_0) \left[\int_{\mathbb{R}^q} \nabla_\theta' \tau(x, \theta_0) f(x) K\left(X_i - x\right)\, dx\right] (\theta - \theta_0) \\
&\quad + \frac{1}{2} g_n'(Z_i, \theta_0) \sum_{k,l=1}^{p} \left(\theta^{(k)} - \theta_0^{(k)}\right)\left(\theta^{(l)} - \theta_0^{(l)}\right) \\
&\quad \left[\int_{\mathbb{R}^q} \mathrm{H}_{\theta^{(k)}\theta^{(l)}} \tau(x, \theta_0) f(x) K\left(X_i - x\right)\, dx\right] + R_{1n}(Z_i, \theta)\,. \qquad (7.6)
\end{aligned}$$

Use the structure of $\tau(\cdot, \cdot)$ to decompose the first element of (7.6) into

$$g_n'(Z_i, \theta_0) \left[\int_{\mathbb{R}^q} \left(\nabla_\alpha' \tau_\alpha(x, \alpha_0) + r_n^{-1} \nabla_\alpha' \tau_\beta(x, \theta_0)\right) f(x) K\left(X_i - x\right)\, dx\right] (\alpha - \alpha_0)$$

$$+ r_n^{-1} g_n'(Z_i, \theta_0) \left[\int_{\mathbb{R}^q} \nabla_\beta' \tau_\beta(x, \theta_0) f(x) K\left(X_i - x\right)\, dx\right] (\beta - \beta_0)\,.$$

Use the same reasoning as in Part (ii) to conclude that the corresponding U-statistic can be written as $n^{-1/2}A_n'D_n^{-1}\theta$, where $A_n = O_p(1)$. The U-statistic based on the second element of (7.6) can be decomposed as

$$(\alpha-\alpha_0)' B_{n\alpha\alpha}(\alpha-\alpha_0) + 2r_n^{-1}(\alpha-\alpha_0)' B_{n\alpha\beta}(\beta-\beta_0) + r_n^{-1}(\beta-\beta_0)' B_{n\beta\beta}(\beta-\beta_0)\,,$$

where $B_{n\alpha\alpha}$, $B_{n\alpha\beta}$, and $B_{n\beta\beta}$ are $O_p(n^{-1/2})$, so it is an $\|\alpha-\alpha_0\|^2 O_p(n^{-1/2}) + \|\beta-\beta_0\|^2 O_p(n^{-1/2}r_n^{-1}) + \|\alpha-\alpha_0\|\|\beta-\beta_0\| O_p(n^{-1/2})$. The empirical process in the remaining term can be shown to be of a smaller order similarly to what was done in Part (iii), using Assumption 6 and Corollary 8 of Sherman (1994). For the empirical process $\mathbb{P}_n \mathrm{E}\left[l_{3\theta}(Z_i,Z_j)|Z_i\right]$, use a similar expansion, the fact that

$$\mathrm{E}\left|(g(Z_i,\theta)-g(Z_i,\theta_0))'\left[\int_{\mathbb{R}^q} \nabla_\alpha'\tau(x,\theta_0)f(x)K(X_i-x)\,dx\right]\right| \to 0$$

and

$$\mathrm{E}\left|(g(Z_i,\theta)-g(Z_i,\theta_0))'\left[\int_{\mathbb{R}^q} \nabla_\beta'\tau_\beta(x,\theta_0)f(x)K(X_i-x)\,dx\right]\right| \to 0$$

as $\theta-\theta_0 \to 0$, and Corollary 8 of Sherman (1994) to conclude that it is of a smaller order than $\mathbb{P}_n \mathrm{E}\left[l_{1\theta}(Z_i,Z_j)|Z_i\right]$ uniformly in θ in a $o(1)$ neighborhood of θ_0.

Let us now gather the results. Adopting the reparametrization $\gamma = D_n^{-1}\theta$ yields

$$M_n(\theta) - M_n(\theta_0) = \mathrm{E}\,M_n(\theta) + \frac{1}{\sqrt{n}}A_n'(\gamma-\gamma_0) + \|\gamma-\gamma_0\|^2 o_p(1) + o_p(n^{-1})\,,$$

uniformly for γ in a $o(r_n^{-1})$ neighborhood of γ_0. For θ in a $o(1)$ neighborhood of θ_0, $\sigma^2(\theta) = \sigma^2(\theta_0) + o(1)$, which is bounded from below by Assumption 5. Equation (7.1) thus implies

$$Q_n(\theta) - Q_n(\theta_0) = \frac{\mathrm{E}\,M_n(\theta)}{\sigma^2(\theta_0)} + \frac{n^{-1/2}}{\sigma^2(\theta_0)}A_n'(\gamma-\gamma_0) + \|\gamma-\gamma_0\|^2 o_p(1) + o_p(n^{-1})\,. \tag{7.7}$$

Moreover, as $\nabla_\theta \mathrm{E}\,M_n(\theta_0) = 0$,

$$\begin{aligned}\frac{\mathrm{E}\,M_n(\theta) - \mathrm{E}\,M_n(\theta_0)}{\sigma^2(\theta_0)} &= \left[(\theta-\theta_0)'\nabla_\theta \mathrm{E}\,M_n(\theta_0) + \frac{1}{2}(\theta-\theta_0)' V (\theta-\theta_0) + R_1\right]\sigma^{-2}(\theta_0) \\ &= \frac{1}{2\sigma^2(\theta_0)}(\gamma-\gamma_0)' D_n V D_n (\gamma-\gamma_0) + o(\|\gamma-\gamma_0\|^2) \geq C\|\gamma-\gamma_0\|^2\,,\end{aligned}$$

for some $C>0$, by Assumption 6 and since D_nVD_n has eigenvalues bounded away from zero by Lemma 7.2. Use now (7.7) to deduce that $\|\tilde{\gamma}-\gamma_0\|^2 = O_p(n^{-1/2})$ by Theorem 1 of Sherman (1993), see also the extension of Lavergne and Patilea (2010) that allows to deal with a varying limit criterion. Therefore

$$Q_n(\theta) = Q_n(\theta_0) + \frac{1}{\sqrt{n}}A_n'(\gamma-\gamma_0) + \frac{1}{2}(\gamma-\gamma_0)' D_n V D_n (\gamma-\gamma_0) + o_p(n^{-1})\,,$$

uniformly over $O(n^{-1/2})$ neighborhoods of γ_0, and $\sqrt{n}(D_nVD_n)(\tilde{\gamma}_n-\gamma_0) + A_n = o_p(1)$ from Theorem 2 of Sherman (1993). By Lemma 7.2, the variance $D_n\Delta D_n$ of A_n is non-singular and bounded, and by a standard central limit theorem for triangular arrays, $(D_n\Delta D_n)^{-1/2}A_n$ is asymptotically normal with mean zero and variance identity. This concludes the proof.

7.2.2 Proof of Theorem 4.2

To simplify the exposition, most of the proof is performed with only two groups of parameters, i.e. $\theta = (\alpha', \beta')'$ and (3.3) holds, and we explain how this generalizes to more complex setups. Following Antoine and Renault (2010b), we proceed in two main steps. First, we define an equivalent formulation of H_0 as $\breve{h}(\theta_0) = 0$, where $\breve{h}$ is a linear transformation of h that separates the restrictions into (i) restrictions that involve components of α only and are therefore strongly identified, and (ii) restrictions that gathers the remaining restrictions. Second, we show that the Wald test statistic for testing $\breve{h}(\theta) = 0$ is numerically equal to the Wald statistic for testing H_0 and we derive its asymptotic behavior. The two extreme cases where all restrictions are identified at the same rate, whether strongly or weakly, do not require the first step.

The space $I_1 \equiv [\texttt{col}\,(\nabla_\theta h(\theta_0))] \cap [\texttt{col}\,(\nabla_\theta \alpha')]$ is the space of tested restrictions that are identified at the standard rate $\sqrt{n}$. Let its dimension be m_1 and ϵ_i, $i = 1, \ldots, m_1$, be vectors of $\mathbb{R}^m$ such that $[\nabla_\theta h(\theta_0)\epsilon_i]_{i=1,\cdots,m_1}$ is a basis of I_1. The remaining $(m - m_1)$ directions that are identified at the slower rate r_n belongs to the space $I_2 \equiv [\texttt{col}\,(\nabla_\theta h(\theta_0))] \cap [I_1]^\perp$. Let similarly ϵ_i, $i = m_1 + 1, \cdots, m$ be vectors of $\mathbb{R}^m$ such that $[\nabla_\theta h(\theta_0)\epsilon_i]_{i=m_1+1,\cdots,m}$ is a basis of I_R. Consider the invertible matrix $H' = [\epsilon_1\ \epsilon_2\ \cdots\ \epsilon_s]$. Then $H_0 :\ h(\theta_0) = \mathbf{0}$ is equivalent to $\breve{h}(\theta_0) = \mathbf{0}$ with $\breve{h}(\theta) = Hh(\theta)$.

Extension: In cases where they are more than two rates and $\theta = (\theta_1, \ldots \theta_s)$, one should define s spaces I_l, $l = 1, \ldots s$, such that I_l is the (possibly empty) space of tested directions that are identified at rate faster or equal than r_{ln}. The vectors ϵ_i, $i = 1, \ldots m$ and the matrix H are thus defined similarly as above.

The Wald statistics based on $h(\cdot)$ and $\breve{h}(\cdot)$ are numerically equal, since

$$\begin{aligned} W_n(h) &= nh'(\tilde{\theta}_n)\left[\nabla'_\theta h(\tilde{\theta}_n) V_n^{-1} \Delta_n V_n^{-1} \nabla_\theta h(\tilde{\theta}_n)\right]^{-1} h(\tilde{\theta}_n) \\ &= n(Hh)'(\tilde{\theta}_n)\left[H\nabla'_\theta h(\tilde{\theta}_n) V_n^{-1} \Delta_n V_n^{-1} \nabla_\theta h(\tilde{\theta}_n) H'\right]^{-1} Hh(\tilde{\theta}_n) \\ &= n\breve{h}'(\tilde{\theta}_n)\left[\nabla'_\theta \breve{h}(\tilde{\theta}_n) V_n^{-1} \Delta_n V_n^{-1} \nabla_\theta \breve{h}(\tilde{\theta}_n)\right]^{-1} \breve{h}(\tilde{\theta}_n) \equiv \breve{W}_n \,. \end{aligned}$$

Hence we can equivalently prove our theorem for $\breve{W}_n$. Now this statistic equals

$$\begin{aligned} &\left[\sqrt{n}\breve{D}_n^{-1}\breve{h}(\tilde{\theta}_n)\right]' \left[\left(\breve{D}_n^{-1}\nabla'_\theta \breve{h}(\tilde{\theta}_n) D_n\right)(D_n V_n D_n)^{-1}(D_n \Delta_n D_n)(D_n V_n D_n)^{-1}\left(D_n \nabla_\theta \breve{h}(\tilde{\theta}_n)\breve{D}_n^{-1}\right)\right]^{-1} \\ &\quad \times \left[\sqrt{n}\breve{D}_n^{-1}\breve{h}(\tilde{\theta}_n)\right], \end{aligned}$$

where

$$\breve{D}_n = \begin{pmatrix} \mathbf{I}_{s_1} & 0 \\ 0 & r_n \mathbf{I}_{s_2} \end{pmatrix}.$$

From the consistency of $\tilde{\theta}$, Assumption 5, the assumption on the derivative of $g(\cdot, \theta)$ on $\mathcal{N}$, and Hoeffding's decomposition,

$$D_n (V_n - V) D_n = O_p(r_n^2 n^{-1}) + O_p(r_n n^{-1/2}) = o_p(1)\,,$$

and similarly $D_n\left(\Delta_n - \Delta\right)D_n = o_p(1)$. From Lemma 7.2, D_nVD_n and $D_n\Delta D_n$ have finite and non-vanishing eigenvalues. From a mean-value expansion of $\breve{h}$ around θ_0,

$$\begin{aligned} \breve{h}(\tilde{\theta}_n) &= \breve{h}(\theta_0) + \nabla'_\theta \breve{h}(\overline{\theta}_n)(\tilde{\theta}_n - \theta_0) \\ \Leftrightarrow\ \sqrt{n}\breve{D}_n^{-1}(\breve{h}(\tilde{\theta}_n) - \breve{h}(\theta_0)) &= \left[\breve{D}_n^{-1}\nabla'_\theta \breve{h}(\overline{\theta}_n)D_n\right]\left[\sqrt{n}D_n^{-1}\left(\tilde{\theta}_n - \theta_0\right)\right], \end{aligned}$$

with $\|\overline{\theta}_n - \theta_0\| \leq \|\tilde{\theta}_n - \theta_0\|$. The desired result then follows from Theorem 4.1 if $D_n\nabla_\theta \breve{h}(\overline{\theta}_n)\breve{D}_n^{-1}$, and then $\breve{D}_n^{-1}\nabla'_\theta \breve{h}(\tilde{\theta}_n)D_n$, converges to a full rank matrix. Finally,

$$\begin{aligned} D_n\nabla_\theta \breve{h}(\tilde{\theta}_n)\breve{D}_n^{-1} &= D_n\nabla_\theta h(\tilde{\theta}_n)H'\breve{D}_n^{-1} \\ &= \begin{pmatrix} \mathbf{I}_{p_1} & \mathbf{0} \\ \mathbf{0} & r_n\mathbf{I}_{p_2} \end{pmatrix}\begin{pmatrix} \left[\nabla_\theta h(\tilde{\theta}_n)\epsilon_i\right]_{i=1,\cdots,s_1} \\ r_n^{-1}\left[\nabla_\theta h(\tilde{\theta}_n)\epsilon_i\right]_{i=s_1+1,\cdots,s} \end{pmatrix} \\ &= \begin{pmatrix} \left[\nabla_\alpha h(\tilde{\theta}_n)\epsilon_i\right]_{i=1,\cdots,s_1} & \mathbf{0} \\ r_n^{-1}\left[\nabla_\alpha h(\tilde{\theta}_n)\epsilon_i\right]_{i=s_1+1,\cdots,s} & \left[\nabla_\beta h(\tilde{\theta}_n)\epsilon_i\right]_{i=s_1+1,\cdots,s} \end{pmatrix} \\ &= \begin{pmatrix} \nabla_\alpha \breve{h}_\alpha(\tilde{\theta}_n) & \mathbf{0} \\ r_n^{-1}\nabla_\alpha \breve{h}_\beta(\tilde{\theta}_n) & \nabla_\beta \breve{h}_\beta(\tilde{\theta}_n) \end{pmatrix} \\ &\to \begin{pmatrix} \nabla_\alpha \breve{h}_\alpha(\theta_0) & \mathbf{0} \\ \mathbf{0} & \nabla_\beta \breve{h}_\beta(\theta_0) \end{pmatrix}, \end{aligned}$$

by the continuous mapping theorem, and this matrix is full rank by construction.

Appendix

We show how to derive the formula of our WMD estimator for the linear model of Section 2. We do not detail assumptions and we assume away suitable conditions on all matrices involved.
The first order conditions of (2.2) are equivalent to

$$\left[Y_1^{*'}(\tilde{K} - \tilde{\lambda}\mathbf{I}_n)Y_1^*\right]\tilde{\theta}_n - \left[Y_1^{*'}(\tilde{K} - \tilde{\lambda}\mathbf{I}_n)y\right] = \mathbf{0}\,.$$

This yields the estimators' formula Equation (2.3). To obtain $\tilde{\lambda}$, note that

$$\tilde{\lambda} = \min_{\gamma\in\mathbb{R}^{p+1}}\left[\frac{\gamma' Y^{*'}\tilde{K}Y^*\gamma}{\gamma' Y^{*'}Y^*\gamma}\right].$$

The first-order conditions yield $\left[Y^{*'}\tilde{K}Y^* - \tilde{\lambda}Y^{*'}Y^*\right]\tilde{\gamma} = \mathbf{0}$. Premultiply by $\left(Y^{*'}Y^*\right)^{-1}$ to obtain

$$\left[\left(Y^{*'}Y^*\right)^{-1}Y^{*'}\tilde{K}Y^* - \tilde{\lambda}\mathbf{I}_n\right]\left(Y^{*'}Y^*\right)^{-1}\tilde{\gamma} = \mathbf{0}\,.$$

Hence $\tilde{\lambda}$ should be the minimum eigenvalue of $\left(Y^{*'}Y^*\right)^{-1}\left(Y^{*'}\tilde{K}Y^*\right)$.

References

Angrist, J., G. Imbens, and A. Krueger (1999): "Jackknife Instrumental Variables Estimation," *Journal of Applied Econometrics*, 14, 57–67.

Antoine, B. and E. Renault (2009): "Efficient GMM with Nearly-weak Instruments," *Econometrics Journal*, 12, 135–171.

——— (2010a): *Efficient Inference with Poor Instruments: a General Framework*, Taylor & Francis, chap. 2 in Handbook of Empirical Economics and Finance.

——— (2010b): "Efficient Minimum Distance Estimation with Multiple Rates of Convergence," *forthcoming Journal of Econometrics*.

Bekker, P. (1994): "Alternative Approximations to the Distributions of Instrumental Variables estimators," *Econometrica*, 62, 657–681.

Bierens, H. (1982): "Consistent Model Specification Tests," *Journal of Econometrics*, 20, 105–134.

Caner, M. (2010): "Testing, Estimation in GMM and CUE with Nearly-Weak Instruments," *Econometrics Reviews*, 29, 330–363.

Chao, J. and N. Swanson (2005): "Consistent Estimation with a Large Number of Weak Instruments," *Econometrica*, 73, 1673–1692.

Fuller, W. (1977): "Some Properties of a Modification of the Limited Information Estimator," *Econometrica*, 45, 939–954.

Hahn, J. and G. Kuersteiner (2002): "Discontinuities of Weak Instruments Limiting Distributions," *Economics Letters*, 75, 325–331.

Han, C. and P. Phillips (2006): "GMM with Many Moment Conditions," *Econometrica*, 74, 147–192.

Hansen, C., J. Hausman, and W. Newey (2008): "Estimation With Many Instrumental Variables," *Journal of Business and Economic Statistics*, 26, 398422.

Hausman, J., W. Newey, T. Woutersen, J. Chao, and N. Swanson (2009): "Instrumental Variable Estimation with Heteroskedasticity and Many Instruments," Tech. rep., MIT.

Jun, S. and J. Pinkse (2011): "Testing under Weak Identification with Conditional Moment Restrictions," *forthcoming Econometric Theory*.

Lavergne, P. and V. Patilea (2010): "Smooth Minimum Distance Estimation and Testing with Conditional Estimating Equations: Uniform in Bandwidth Theory," Tech. rep., Toulouse School of Economics.

Newey, W. and F. Windmeijer (2009): "GMM with Many Weak Moment Conditions," *Econometrica*, 77, 687–719.

Pacini, D. (2011): "Identification and Estimation of a Semiparametric Binary Response Model from Repeated Cross Sections," Tech. rep., Toulouse School of Economics.

Pakes, A. and D. Pollard (1989): "Simulation and the Asymptotics of Optimization Estimators," *Econometrica*, 57, 1027–1057.

Sherman, R. (1993): "The Limiting Distribution of the Maximum Rank Correlation Estimator," *Econometrica*, 61, 123–137.

——— (1994): "Maximal Inequalities for Degenerate $U-$Processes with Applications to Optimization Estimators," *Annals of Statistics*, 22, 439–459.

Staiger, D. and J. Stock (1997): "Instrumental Variables Regression with Weak instruments," *Econometrica*, 65, 557–586.

Stock, J. and J. Wright (2000): "GMM with Weak Identification," *Econometrica*, 68, 1055–1096.

Stock, J. and M. Yogo (2005): *Asymptotic Distributions of Instrumental Variables Statistics With Many Instruments*, Cambridge, U.K.: Cambridge University Press, chap. 6 in Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg, 109–120.

				6 IV			12 IV			24 IV		
		WMDF	WMD	HFUL	HLIM	JIVE	HFUL	HLIM	JIVE	HFUL	HLIM	JIVE
α_0	**Med**	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.001	0.000	0.000	0.001
	Std	0.076	0.085	0.065	0.080	0.880	0.065	0.238	0.954	0.066	1.025	7.534
	DecR	0.177	0.179	0.162	0.173	0.215	0.162	0.176	0.227	0.162	0.181	0.243
	Rej	0.029	0.029	0.034	0.031	0.026	0.032	0.030	0.025	0.029	0.026	0.024
β_0	**Med**	0.002	-0.004	0.055	-0.004	-0.116	0.058	-0.004	-0.064	0.065	0.000	0.022
	Std	0.421	0.803	0.260	0.660	9.563	0.282	4.583	15.223	0.333	17.073	139.529
	DecR	0.833	0.852	0.642	0.813	1.623	0.686	0.900	1.954	0.764	1.040	2.689
	Rej	0.069	0.068	0.089	0.069	0.097	0.094	0.072	0.165	0.102	0.082	0.237

Table 2: Linear Homoskedastic Model M_L with $n = 250$ and $c = 8$.

				6 IV			12 IV			24 IV		
		WMDF	WMD	HFUL	HLIM	JIVE	HFUL	HLIM	JIVE	HFUL	HLIM	JIVE
α_0	**Med**	0.000	0.000	0.001	0.000	0.000	0.001	0.001	0.001	0.001	0.001	0.001
	Std	0.071	0.081	0.067	6.582	1.133	0.068	1.075	1.048	0.069	2.111	7.922
	DecR	0.159	0.160	0.160	0.171	0.222	0.161	0.175	0.232	0.162	0.182	0.248
	Rej	0.020	0.020	0.025	0.022	0.020	0.025	0.022	0.019	0.023	0.021	0.019
β_0	**Med**	-0.015	-0.020	0.083	0.031	-0.128	0.081	0.028	-0.074	0.090	0.032	0.024
	Std	0.493	0.906	0.376	86.415	12.363	0.404	12.265	16.443	0.447	22.772	148.869
	DecR	0.967	0.992	0.888	1.086	1.989	0.942	1.195	2.263	0.998	1.378	2.956
	Rej	0.062	0.060	0.098	0.082	0.096	0.106	0.091	0.149	0.120	0.102	0.204

Table 3: Linear Heteroskedastic Model $M_{L,H}$ with $n = 250$ and $c = 8$.

				6 IV			12 IV			24 IV		
		WMDF	WMD	HFUL	HLIM	JIVE	HFUL	HLIM	JIVE	HFUL	HLIM	JIVE
α_0	**Med**	0.001	0.001	0.000	0.001	0.000	0.001	0.001	0.002	0.001	0.001	0.001
	Std	0.069	0.074	0.066	0.485	4.601	0.067	0.237	1.028	0.068	0.573	9.966
	DecR	0.158	0.159	0.162	0.175	0.217	0.162	0.179	0.224	0.163	0.186	0.236
	Rej	0.024	0.024	0.027	0.024	0.025	0.026	0.024	0.024	0.026	0.023	0.022
β_0	**Med**	-0.014	-0.019	0.059	0.002	-0.107	0.062	0.003	-0.067	0.070	0.011	0.014
	Std	0.420	0.573	0.304	11.925	68.257	0.348	3.589	21.451	0.405	8.491	155.796
	DecR	0.847	0.864	0.691	0.907	1.693	0.757	1.020	1.980	0.861	1.237	2.668
	Rej	0.057	0.055	0.092	0.073	0.100	0.099	0.081	0.152	0.115	0.094	0.208

Table 4: Non-linear Heteroskedastic Model $M_{NL,H}$ with $n = 250$ and $c = 8$.

				6 IV			12 IV			24 IV		
		WMDF	WMD	HFUL	HLIM	JIVE	HFUL	HLIM	JIVE	HFUL	HLIM	JIVE
α_0	**Med**	0.001	0.001	-0.000	-0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.002
	Std	0.072	0.117	0.067	0.156	0.665	0.068	0.395	2.684	0.068	0.979	2.340
	DecR	0.160	0.161	0.160	0.172	0.218	0.161	0.176	0.226	0.162	0.182	0.232
	Rej	0.022	0.022	0.025	0.023	0.021	0.022	0.020	0.020	0.023	0.021	0.019
β_0	**Med**	-0.013	-0.019	0.076	0.023	-0.108	0.080	0.028	-0.057	0.090	0.030	0.028
	Std	0.510	1.139	0.378	2.042	10.749	0.402	26.078	43.056	0.441	17.148	35.398
	DecR	0.983	1.005	0.894	1.107	1.887	0.924	1.211	2.150	1.003	1.367	2.630
	Rej	0.063	0.062	0.099	0.082	0.101	0.108	0.090	0.143	0.116	0.098	0.197

Table 5: Linear Heteroskedastic Model $M_{ST,H}$ with Student errors, $n = 250$ and $c = 8$.

				6 IV			12 IV			24 IV		
		WMDF	WMD	HFUL	HLIM	JIVE	HFUL	HLIM	JIVE	HFUL	HLIM	JIVE
α_0	**Med**	0.001	0.001	-0.000	0.000	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000
	Std	0.050	0.063	0.048	0.376	0.825	0.048	0.150	0.387	0.049	0.159	4.546
	DecR	0.112	0.112	0.113	0.120	0.150	0.113	0.122	0.153	0.115	0.127	0.164
	Rej	0.019	0.019	0.020	0.019	0.019	0.021	0.019	0.018	0.020	0.019	0.019
β_0	**Med**	-0.006	-0.009	0.090	0.045	-0.107	0.093	0.047	-0.079	0.097	0.051	-0.004
	Std	0.534	1.211	0.386	5.192	15.276	0.419	4.440	11.064	0.465	11.028	76.746
	DecR	0.950	0.960	0.911	1.085	1.786	0.952	1.175	2.007	1.025	1.327	2.513
	Rej	0.061	0.060	0.102	0.085	0.085	0.108	0.092	0.125	0.118	0.104	0.185

Table 6: Linear Heteroskedastic Model $M_{L,H}$ with $n = 500$ and $c = 8$.

				6 IV			12 IV			24 IV		
		WMDF	WMD	HFUL	HLIM	JIVE	HFUL	HLIM	JIVE	HFUL	HLIM	JIVE
α_0	**Med**	0.000	0.000	0.001	0.001	0.001	0.001	0.001	0.001	0.000	0.000	0.001
	Std	0.058	0.058	0.064	0.065	0.066	0.064	0.065	0.066	0.064	0.065	0.067
	DecR	0.148	0.148	0.162	0.163	0.168	0.162	0.164	0.168	0.163	0.165	0.169
	Rej	0.048	0.048	0.050	0.050	0.047	0.050	0.049	0.048	0.051	0.049	0.048
β_0	**Med**	-0.004	-0.004	0.014	0.005	-0.026	0.014	0.004	-0.027	0.013	0.003	-0.026
	Std	0.136	0.136	0.150	0.154	0.167	0.151	0.155	0.174	0.155	0.160	0.189
	DecR	0.342	0.343	0.379	0.387	0.414	0.379	0.389	0.433	0.388	0.399	0.454
	Rej	0.046	0.046	0.061	0.057	0.055	0.061	0.056	0.062	0.064	0.058	0.075

Table 7: Linear Heteroskedastic Model $M_{L,H}$ with $n = 250$ and $c = 50$.

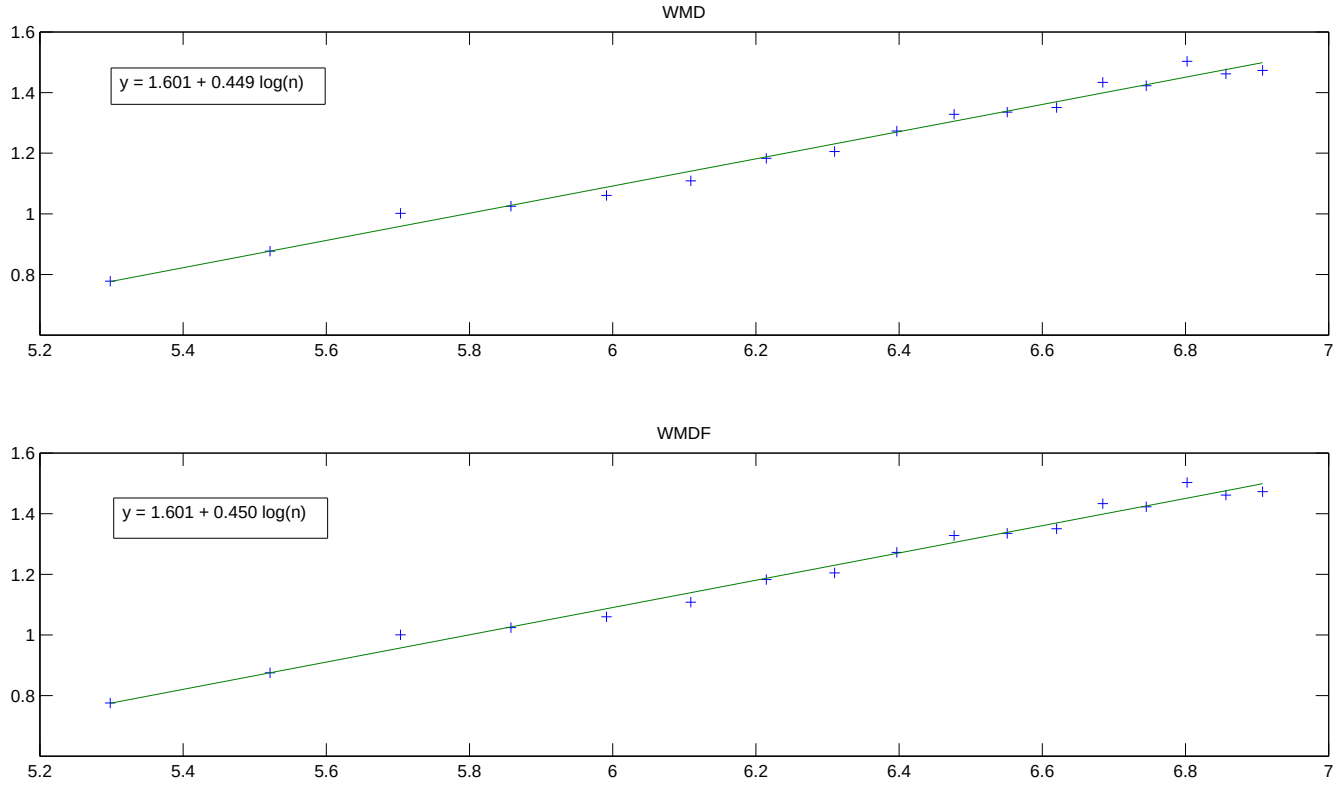

Figure 1: Linear regression (5.1) for Model $M_{L,H}$ with $c = 50$. Top panel: estimator WMD. Bottom panel: estimator WMDF.